SSPictR: A Biologically-plausible Image Representation

Anna Penzkofer[1,*], Karim Habashy[2], Chris Eliasmith[2] and Andreas Bulling[1]

[1]University of Stuttgart, Pfaffenwaldring 5A, 70569 Stuttgart, Germany
[2]University of Waterloo, 200 University Ave W, ON N2L 3G5 Waterloo, Canada

Abstract

Finding interpretable and generalisable representations of natural images has the potential to increase performance on various computer vision tasks, particularly those that require semantic information and spatial understanding, such as image segmentation or scene recognition. Drawing inspiration from a cognitive modelling framework, we propose SSPictR – a biologically plausible image representation based on spatial semantic pointers (SSPs). SSPictR encodes semantic labels of objects and their spatial locations extracted from segmentation maps. It only requires a single vector to capture a compressed but fully decodable neuro-symbolic representation of an image. We demonstrate the biological plausibility of SSPictR by performing representation similarity analysis, finding a significant correlation with fMRI data recorded from the early visual cortex. We further highlight the effectiveness and out-of-domain generalisability of SSPictR representations by training a compact model for scene recognition on standard benchmark datasets. Our simple neural network achieves performance on par with previous work, while having more than three times fewer trainable parameters. Taken together, SSPictR bridges the gap between biological plausibility and effective representations for tasks in computer vision and beyond.

Keywords

Neuro-symbolic Representation, Spatial Semantic Pointer, Vector Symbolic Algebra, Representation Similarity Analysis, Scene Recognition

1. Introduction

Machine learning models for computer vision have achieved human-level performance on many tasks, such as object detection [1], semantic segmentation [2] or depth estimation [3]. But these improvements come at the ever-increasing cost of inefficiency and lack of interpretability: Current models require training of millions of parameters from large amounts of data, and their learned internal representations are notoriously challenging to analyse and interpret. Furthermore, despite impressive performance on tasks that they were trained for, these models still lack generalisability to out-of-domain (OOD) data [4]. Aligning model and human representations has increased robustness in object detection [5], improved performance [6], and has advanced our understanding of human cognition through better computational models [7]. While there has been a lot of research in aligning deep neural networks (DNNs) in object classification tasks [7, 8, 9], human alignment for scene understanding remains under-explored [10].

At the same time, computational models of human perception have a long-standing history of research in cognitive science [11]. Most interesting are biologically plausible cognitive models, as they can be naturally integrated with DNNs while maintaining the interpretability of symbolic systems. A popular model is the semantic pointer architecture (SPA) [12] – a cognitive modelling framework based on vector symbolic algebras (VSAs). VSAs use hyperdimensional distributed vectors for efficient and robust representation of semantic concepts. There are many successful applications of cognitively-inspired VSAs, for example, in abstract reasoning [13], ego-motion prediction [14], path integration [15], or reinforcement learning [16]. Particularly interesting as representations for scene understanding are Komer et al. [17], Dumont and Eliasmith [18], and Penzkofer et al. [19], as they encode objects in a grid-like continuous vector space similar to cognitive maps found in the brain [20]. VSA-based encodings have also been shown to replicate the behaviour of visual working memory in humans [21].

AIC 2025: The 10th International Workshop on Artificial Intelligence and Cognition (held as part of ECAI 2025). October 25-26, 2025. Bologna, Italy
*Corresponding author.
anna.penzkofer@vis.uni-stuttgart.de (A. Penzkofer)

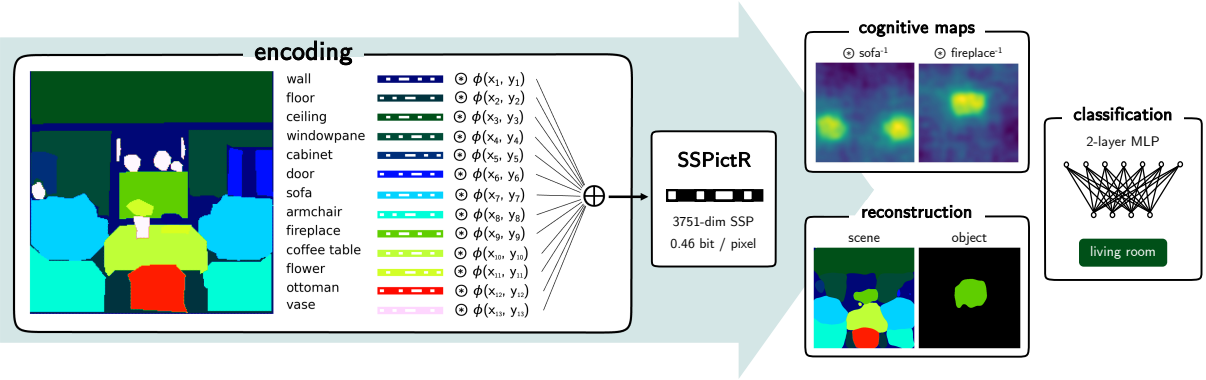


Figure 1: Overview of SSPictR: we construct a neuro-symbolic representation of a scene within a single 3,751-dimensional vector. SSPictR can then be used to decode cognitive maps of the scene, reconstruct the semantic segmentation map, or as input for downstream task models.

In this work, we bring together research in computer vision and cognitive science and propose SSPictR– Spatial Semantic Pointer Picture Representation – a cognitively inspired image representation that is efficient, interpretable, and generalisable (see Figure 1 for an overview). At its core, SSPictR encodes objects from segmentation maps into a continuous vector space that compresses the semantic and spatial information of the full visual scene. SSPictR offers a compression of 0.46 bits per pixel (bpp) compared to 8 bpp for segmentation maps [22], encoding an entire scene in a single 3,751-dimensional vector. This is a compression factor of more than 17.5, while we show that we can still reconstruct up to 57.3% of the original semantic segmentation on the COCOStuff dataset [23]. More importantly, we show that our neuro-symbolic representation can also be used as a feature embedding for downstream tasks, such as scene recognition, where a simple linear network achieves performance on par with more complicated models from previous works.

Taken together, this paper makes the following contributions: (1) We propose SSPictR– a novel neuro-symbolic representation of images with an encoding method from semantic segmentation maps, (2) we show the biological plausibility of SSPictR and its alignment with representations found in the early visual cortex of humans by performing Representation Similarity Analysis (RSA) [24], and (3) we show the efficacy of SSPictR for downstream tasks, such as scene recognition on popular subsets of Places365 [25], and demonstrate its generalisability on OOD data: SUN-RGBD [26] and ADE20K [27]. As such, SSPictR is an important step towards human-aligned image representations that are compact, interpretable, and generalisable, as well as applicable to computer vision tasks and beyond.

2. Related Work

The choice of data representation is a key factor for the performance of computer vision models [28]. Scene representation, in particular, is challenging due to complex configurations, because scenes are comprised of diverse objects in complex spatial layouts with substantial semantic ambiguity [29]. In contrast, humans are adept at encoding any scene efficiently [10]. Leveraging a cognitive modelling framework, we design a representation that is well aligned with neural response data in humans.

Representation Design. While current research usually employs representation learning to find intermediate representations, some work enforces a specific design on representations to enhance downstream task performance, similar to our work. For example, object-centric methods like slot attention [30], and adaptive slot attention (AdaSlot) [31] decompose scenes into discrete object representations, performing competitively on tasks such as object discovery and property prediction. Masked Autoencoders (MAE) [32] reconstruct image patches by predicting pixels, capturing primarily low-level details. Gaussian Masked Autoencoders (GMAE) [33] refine this approach: Their decoder parametrises Gaussians, adjusting centres, scale, colour, opacity, and rotation and employs a rendering-based loss to

focus attention on regions with higher information density, thereby capturing richer semantic cues. BEiT, on the other hand, [34] adopts a codebook-based approach, predicting high-level tokens instead of pixels, which abstracts object representations into a more semantic space. Vector-symbolic representations like the Hyperdimensional Inference Layer (HIL) [35] encode semantic information in hyper-dimensional vectors, demonstrating robust performance in tasks such as image classification.

Vector Symbolic Algebras. Vector symbolic algebras (VSAs) play an important role in cognitive architectures [36] and showed improved performance of machine learning methods, in ego-motion prediction [35], speech recognition [37], and object classification [38]. VSAs offer a unique way of encoding symbolic meaning in hyper-dimensional distributed representations, making them inherently interpretable [35] and robust to errors [39]. Furthermore, their compatibility with neuromorphic hardware makes them highly efficient, achieving speed gains of up to 100 times GPU performance [40]. The potential of VSAs in replicating cognitive maps for navigation has been shown via path integration [15] and reinforcement learning on navigation tasks [16]. Further, the potential in scene understanding has been shown via visual question answering [17, 41, 19]. For more application examples and a comprehensive literature review on VSAs, we refer to Kleyko et al. [36]. In this work, we build upon the semantic pointer architecture (SPA) [12], a cognitively-inspired VSA, which uses holographic reduced representations (HRRs) [42], i.e., a set of operations that can be applied for manipulation of hyper-dimensional vectors representing symbols. Semantic similarity of vectors is calculated by the dot product, while vector addition aggregates multiple concepts, and circular convolution – denoted by $\circledast$ – enables binding operations. Finally, Komer et al. [17] has introduced fractional binding: binding a vector with itself $k \in \mathbb{R}$ times, allowing for the encoding of continuous data such as spatial coordinates.

Human Alignment. Aligning representations of deep neural networks (DNNs) to humans is a promising avenue to increase the performance and generalisability of computer vision models [6]. While many works analysed object representations and their alignment to human similarity judgements [8, 9, 43], meaningful representations of full scenes remain under-explored. In contrast, Groen et al. [44] analysed scene recognition in humans and found that in addition to object co-occurrence statistics, other features across different levels of visual processing play an important role, such as spatial layouts, boundaries, and textures. Further, Bartnik and Groen [10] has shown that alignment of DNNs is focused on object representations instead of navigational affordances, which is the navigability of spatial layouts – a core skill for visual navigation. Following those works, we evaluate the human alignment of SSPictR via representation similarity analysis (RSA) [24]. RSA measures the correlation between two pairwise representational dissimilarity matrices (RDMs), which can be generated from any feature embeddings or human measurements like neural responses from fMRI data. For evaluating SSPictR, we selected the publicly available data set by Bonner and Epstein [45], as it features pre-computed RDMs for 50 images of indoor scenes. While their analysis focused on navigational affordances, we focus on the fMRI data obtained from the occipital place area (OPA), the parahippocampal place area (PPA), the retrosplenial complex (RSC), and the early visual cortex (EVC).

Scene Recognition. Scene recognition, the task of classifying scenes into categories, is considered a fundamentally important but challenging task in computer vision with a wide range of applications, from robot navigation [46] to disaster detection [47]. Current benchmark datasets include Places365 [25], ADE20K [27], and SUN-RGBD [26]. Methods trained on Places365 leverage the 10 million images, where large-scale DNNs significantly outperform previous approaches. However, to enable more focused and faster models, [48] and [49] proposed subsets of the Places365 dataset containing seven and 14 indoor classes, respectively. Building on this, Miao et al. [50] proposed a model that integrates knowledge from semantic segmentation maps: object-to-scene (OTS). OTS extracts object features through a pre-trained segmentation model and calculates object relations, outperforming both [48] and [49]. However, OTS also requires up to 255 million model parameters and, therefore, only achieves an inference speed of three fps. More recently, attentional graph convolutional network (AGCN) [51]

achieved higher performance than OTS on both datasets with only 85 million parameters. Song and Ma [52] proposed a semantic region relationship model (SRRM) and combined it with the PlacesCNN module by Zhao et al. [3], yielding CSRRM, which achieves the current state-of-the-art performance on Places365-7 and Places365-14. In follow-up work, Song et al. [53] focused on computational efficiency that is essential for the low-resource and high-speed requirements of edge devices in practical robotics applications, decreasing the number of trainable parameters to only 25 million. In contrast, we are the first to leverage a biologically plausible framework focusing on an interpretable representation design, which enables effective scene recognition with a simple linear network and requires four times fewer trainable parameters.

3. Method

3.1. Theoretical Background

Under the SPA, spatial semantic pointers (SSPs) [17] were proposed for spatial representations through the following encoding scheme:

$$\phi(x) = \mathscr{F}^{-1}\{e^{i\lambda^{-1}Ax}\}, \tag{1}$$

where $\phi : x \in \mathbb{R}^2 \mapsto \mathbb{R}^d$, λ defines the length scale of the encoded representation, $\mathscr{F}^{-1}$ denotes the inverse Fourier transform, and $A \in \mathbb{R}^{d\times 2}$ is a phase matrix whose columns consist of phasors representing different frequencies [54]. For real-valued spatial representations, the phase matrix is conjugate symmetric. Additional biological constraints ensure grid-cell-like firing patterns [18], aligning the representation with the hyper-toroidal manifold [55]. The dimensionality d is given by $d = n_{\text{scales}} \cdot n_{\text{rotates}} \cdot 3 \cdot 2 + 1$, where n_{scales} denotes the scale of the firing pattern activity, n_{rotates} denotes the orientation of the grid cells, 3 denotes triplets, 2 for conjugate symmetry, and +1 for the 0-frequency term.

Using the SSP representation and the set of operations given by HRR [56], we can construct cognitive maps. For example, a map M encoding a table, chair, and person is given by:

$$\mathbf{M} = \mathbf{TABLE} \circledast \phi(x_1, y_1) + \mathbf{CHAIR} \circledast \phi(x_2, y_2) + \mathbf{PERSON} \circledast \phi(x_3, y_3)$$

Then, to query an object's location, unbinding (binding with a vector's pseudo-inverse) can be used :

$$\mathbf{M} \circledast \mathbf{PERSON}^{-1} = \phi(x_3, y_3) + noise$$

This noise is due to unbinding's distributive property over the map representation, yielding the terms $\mathbf{TABLE} \circledast \phi(x_1, y_1) \circledast \mathbf{PERSON}^{-1}$ and $\mathbf{CHAIR} \circledast \phi(x_2, y_2) \circledast \mathbf{PERSON}^{-1}$.

3.2. SSPictR Encoding

In previous works, objects have been encoded as point sources [15] or a set of bounding box coordinates [19]. However, for a more accurate representation, we encoded segmentation maps of a scene as follows:

$$\mathbf{M} = \sum_i \left[\text{obj}_i \circledast \int_{A_i} \phi(x, y) dx dy \right] \tag{2}$$

In this hyper-dimensional representation, $\text{obj}_i \in \mathbb{R}^d$ is a semantic pointer (SP), representing the class of a given object i. This object is bound with a bundle of SSPs representing the area occupied by this object in the representation. In our preliminary analysis, we found that encoding all pixels of a given object degrades representation accuracy due to overloading the SSP bundle. Therefore, we sampled the points for each object, either via uniform sampling within the mask or by sampling from a Gaussian distribution with its mean at each mask's centre of mass and covariance matrix given by the segmentation mask. Points outside the mask are not encoded.

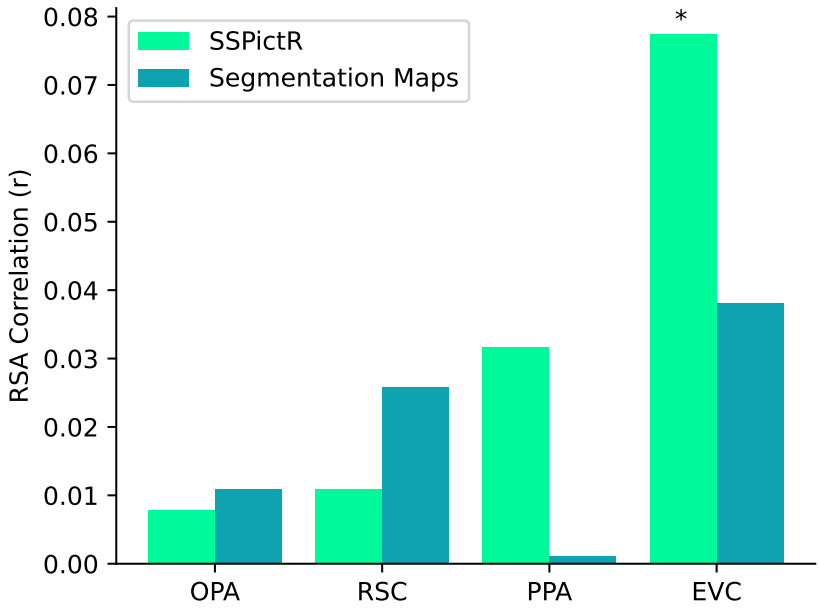


Figure 2: Representation similarity analysis (RSA) on fMRI data in different areas of the brain – comparing SSPictR to the underlying segmentation maps, where we find ($\star$) a significant correlation of SSPictR to the early visual cortex (EVC) with $p < 0.01$.

To evaluate the quality of the SSP representation, we decoded the masks used to generate the scene representation $\mathbf{M}$. We calculated the similarity map ψ_j of an object j in the scene by taking the dot product between a grid of SSPs and the encoded scene bound with the inverse of that object's SP:

$$\psi_j = \left\langle M \circledast \text{obj}_j^{-1}, \int_{A_{\text{grid}}} \phi(x, y) dx dy \right\rangle. \tag{3}$$

This yields an approximate similarity value between the SSP grid A_{grid} that represents each possible location and the scene SSP $\mathbf{M}$ that is queried for the object of interest. The set of similarity values larger than some threshold $\psi_j > \tau$ was used as the decoded mask for object j. Finally, we calculated the intersection-over-union (IoU) using the ground truth and decoded masks. This approach can be used to probe for every object in the scene, to determine what is at a specific location, or to verify that an object exists in the scene, making this representation inherently interpretable. We then optimised the hyperparameters most important for this encoding, i.e., the length scale parameter λ, threshold τ for reconstructing object masks, the percentage/number of encoded points, and the SSP dimensionality. The results of this hyperparameter search and the evaluation of the two encoding schemes are in Appendix B.

4. Experiments

First, we evaluated the human alignment of our proposed neuro-symbolic representation SSPictR. Then, we computed the segmentation reconstruction accuracy to evaluate the quality of our representation in comparison to its compression factor. Further, we applied SSPictR to a downstream task, namely indoor scene recognition on popular subsets of Places365 [25]. To evaluate the generalisation ability of our representation, we tested the simple linear model trained on Places365 on two out-of-domain (OOD) data sets: SUN-RGBD [26] and ADE20K [27].

4.1. Human Alignment

To showcase the biological plausibility of our method, we evaluated the representation alignment of our manually generated SSPs with human fMRI data. Here we have selected the navigational affordances data set provided by Bonner and Epstein [45], as it provides fMRI data of three scene-selective areas of the brain, the occipital place area (OPA), the parahippocampal place area (PPA), and the retrosplenial complex (RSC), as well as the early visual cortex (EVC). While they evaluated the representation similarity of the brain areas to navigational affordances, subsequent works (*e.g.* [57]) used the dataset to evaluate the representational alignment of deep learning models. We followed their approach and performed a RSA [24] with the open source Python framework `rsatoolbox`[1].

For the RSA, we computed RDMs, which consist of pair-wise correlation of our SSP representation for each of the 50 images in the data set, yielding a correlation matrix across all images. The RDMs of the neural data from different brain areas are already provided in the dataset. RSA then computes the correlation of the RDMs as Spearman rank correlation r. We visualised the correlation between SSPictR and the different brain areas in Figure 2, as well as the same analysis for the unprocessed segmentation maps. We tested the significance of all correlations with the Fisher transformation of r and found a statistically significant correlation ($p = 0.00695$) for the EVC and SSPictR. In contrast, the same analysis performed with segmentation maps does not yield a significant correlation ($p > 0.175$). In short, our representation SSPictR is well aligned with neuron activations in the EVC. This is unsurprising as the EVC, more specifically V2, is known for representing objects and their segmentation surfaces [58].

While we anticipated SSPictR to be more similar to the scene selective regions in the brain, i.e., OPA, RSC, and PPA, we find that in contrast to the segmentation maps, our representation is more correlated with the PPA than with the RSC and OPA. Dilks et al. [59] found that the PPA is most likely involved in scene categorisation in contrast to visual navigation (OPA) and map-based navigation (RSC). This aligns well with our findings, as we will see that a simple model on SSPictR performs on par with previous works on three benchmark scene recognition datasets for predicting scene categories.

4.2. Segmentation Reconstruction

In early experiments, we evaluated different encoding schemes and fine-tuned our hyperparameters on object-level samples of the COCOStuff training set [23]. Full details on the grid search for the best hyperparameters can be found in Appendix B. Taking those fine-tuned parameters and the uniform encoding scheme, we then evaluated the segmentation reconstruction on the COCOStuff validation split. To find all objects in the scene, we created a vocabulary of all objects SSPs and performed unbinding with our representation as described in Equation 3. This yields a similarity map $\psi_j(x, y)$ for each object j, indicating the probability of the label assignment to the specific pixel (x, y). By selecting all pixels above a similarity threshold τ to belong to the object mask, we fully reconstructed the scene. We achieved 45.4 mean IoU (mIoU) on the COCOStuff validation split with the similarity threshold set to $\tau = 0.7$. Additionally, following He et al. [32], we evaluated SSPictR on the ADE20K [27] segmentation data set, where we achieved 37.8 mIoU with $\tau = 0.3$.

However, for both data sets, we further increased the reconstruction accuracy by training a UNet model [60] to refine the object masks based on the SSP similarity maps $\psi(x, y)$ instead of using a threshold τ. Our U-Net model consists of four encoder and decoder layers with a total of 465K trainable parameters. We trained the U-Net model on object-level data, i.e. we extracted our similarity map predictions from the training sets for all individual objects, yielding 30,435 object samples for ADE20K and 10,000 for a subset of COCOStuff. When extracting the similarity maps, we already enforced the minimum similarity threshold $\tau_{\text{min}} = 0.004$. This excludes objects that achieve a low similarity score; however, please note that this threshold is enforced on the unnormalised similarity maps, therefore, not comparable to other thresholds. The U-Net model then received the similarity map of one object as input and learned to predict a refined segmentation map (for qualitative examples see Appendix C).

[1] https://pypi.org/project/rsatoolbox

Table 1
Scene recognition results and comparison to previous work. Our simple 2-layer MLP model uses SSPictR as input and is on par with previous work, while also generalising to out-of-domain (OOD) data sets.

Method	Parameters	Places365-7	Places365-14	SUN-RGBD^{OOD}	ADE20K^{OOD}
OTS [50]	255 M	90.1	85.9	70.6	-
AGCN [51]	85 M	<u>91.7</u>	86.0	-	-
CSRRM [52]	50 M	**93.4**	**88.7**	**75.4**	-
GLS + BCL [53]	25 M	90.6	<u>86.6</u>	-	-
SSPictR (Ours)	**7 M**	90.1	82.2	<u>71.8</u>	**94.5**

With this fine-tuning approach, we achieved a performance increase of 8.3 percentage points, up to 46.1 mIoU for ADE20K and an increase of 13.2 percentage points to 57.3 mIoU on COCOStuff. The segmentation accuracy for ADE20K is on par to BEiT [34] and MAE [32], which are self-supervised vision models. It is important to note, however, that our current method relies on pre-trained semantic segmentation networks, which makes a quantitative comparison to other methods unfair. In the future, we plan to provide an end-to-end method and are therefore taking these results as an indication of the feasibility of semantic segmentation reconstruction from SSPictR. Further, they provide a measure of information loss in comparison to the 17.5 factor compression.

4.3. Scene Recognition

SSPictR provides a compressed, interpretable neuro-symbolic scene representation. We evaluate the quality of the representation by predicting scene classes solely from SSPictR, a method known as linear probing, which is generally used to determine intermediate representation quality in self-supervised models [61]. More specifically, we trained a small classification model on the SSP representations on subsets of the Places365 dataset [25] for indoor scene recognition. Here, we followed the setup by Miao et al. [50] and consecutive works [51, 52, 53]. The Places365-7 training set consists of 35,000 images, and the Places365-14 training set consists of 55,000 images. As Places365 does not offer segmentation maps, we ran the pre-trained semantic segmentation model VPD [3], which was trained on ADE20K [27]. Therefore, we used the same object classes for both datasets. After generating the 3,751-dimensional SSPs for all samples, we trained a two-layer linear neural network (NN) model with approx. seven million parameters. The linear NN takes the SSPs as input and has a hidden dimension of 1,875, ultimately reducing the features to the output dimension of seven or 14 classes. We used batch normalisation and dropout layers and set the batch size to 1,024 and $p_{\text{dropout}} = 0.4$. We used AdamW [62] as an optimiser with a learning rate of 0.00195. We trained the model for 25 epochs and evaluated its performance on the held-out official validation set with 700 images and 1,100 images for the seven and 14 classes, respectively.

Table 1 summarises our scene recognition results compared to previous methods. As can be seen from the table, our method achieves competitive performance in terms of accuracy but only requires the 3,751-dim SSP vectors as input. Furthermore, it allows for the use of a simple linear model with only two layers and significantly fewer parameters to train. This reduces the experimentation and training time, as well as allows for generalisation to other datasets. We evaluated the generalisation performance of the trained Places365-7 model on SUN-RGBD [26] and ADE20K as OOD data and achieved a performance of 71.8% and 94.5% classification accuracy, respectively. This is the second best for SUN-RGBD, while ADE20K was not evaluated by previous methods. The generally high performance on the ADE20K dataset is likely due to the availability of ground truth segmentation maps, since the quality of the segmentation maps is a limiting factor of current scene recognition methods [53].

5. Discussion

Our overall goal was to develop a compact image representation that is biologically plausible and can be deployed end-to-end on edge devices for efficient scene recognition and other downstream tasks. While the current method requires segmentation maps or uses a pre-trained segmentation model like VPD [3], SSPictR is an important first step towards this goal. In fact, SSPictR achieves a 17.5 factor compression compared to standard segmentation maps and only requires 0.46 bits per pixel (bpp). For comparison, JPEG typically achieves a compression around 1 bpp [22], and a compression of < 0.1 bpp is considered extreme [63], where distortion of the image content is expected.

Another advantage of SSPictR is the inherent interpretability of the representation. As showcased by the scene reconstruction, we can always probe the representation to decode the information of objects and their spatial location. This allows SSPictR to potentially be used for complex probes and efficiently shifting objects in its high-dimensional embedding space [17] or being integrated with visual question answering methods based on VSAs [19]. We showed the alignment of SSPictR to neural data from different brain regions with a significant correlation to the early visual cortex. The basis of SSPictR is a biologically plausible cognitive modelling framework, which should facilitate the conversion into a fully neural implementation, allowing for efficient deployment on neuromorphic hardware in the future.

Additionally, in future work, we plan to train an end-to-end image-to-SSPictR model that allows the direct integration of SSPictR into cognitive architectures, as well as end-to-end trainable downstream tasks in computer vision and beyond. In general, we believe SSPictR is uniquely suitable for robotics applications, such as embodied visual navigation [46] or for applications in cognitive science, e.g. visual spatial reasoning [64] or symbolic reasoning [13]. Further, we would like to analyse the alignment of SSPictR to human scene representations on a larger data set, e.g. BOLD5000 [65].

6. Conclusion

We proposed SSPictR: a biologically-plausible, neuro-symbolic image representation that is inherently interpretable and enables the deployment of simple models for downstream tasks like scene recognition with robust generalisation to out-of-domain data. Our representation is compact with a compression factor of 17.5, while maintaining semantic and spatial information comparable to trained vision models for semantic image reconstruction. SSPictR was motivated by the goal of aligning machine perception more closely with human perception; therefore, we followed the methods of a biologically plausible cognitive modelling framework. By performing a representation similarity analysis, we found that our representation is indeed significantly ($p < 0.01$) correlated with fMRI data found in the early visual cortex of human subjects. In short, SSPictR provides a strong, interpretable foundation for future research into human-aligned, generalisable vision models.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

Acknowledgments

Anna Penzkofer was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016. Chris Eliasmith was funded in part by the Canada Research Chairs program, CFI and OIT infrastructure funding, NSERC Discovery grant 261453, NUCC NRC File A-0028850, and AFOSR grant FA9550-17-1-0026.

References

[1] Z. Zong, G. Song, Y. Liu, DETRs with Collaborative Hybrid Assignments Training, in: ICCV, 2023.

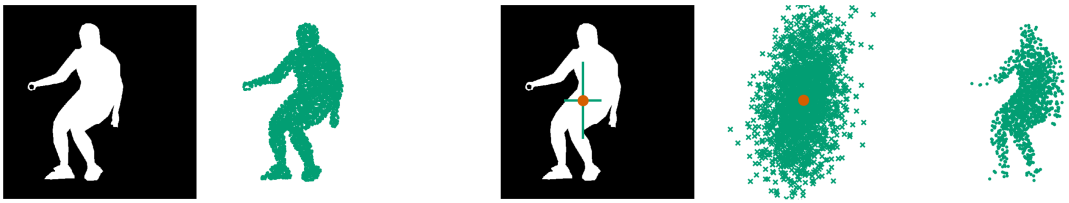
[2] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, Y. Qiao, Vision Transformer Adapter for Dense Predictions, in: ICLR, 2022.
[3] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, J. Lu, Unleashing Text-to-Image Diffusion Models for Visual Perception, in: ICCV, 2023. doi:10.1109/ICCV51070.2023.00527.
[4] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, F. A. Wichmann, Generalisation in humans and deep neural networks, in: NeurIPS, 2018.
[5] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, W. Brendel, ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, in: ICLR, 2019.
[6] I. Sucholutsky, L. Muttenthaler, A. Weller, A. Peng, A. Bobu, B. Kim, B. C. Love, E. Grant, I. Groen, J. Achterberg, J. B. Tenenbaum, K. M. Collins, K. L. Hermann, K. Oktar, K. Greff, M. N. Hebart, N. Jacoby, Q. Zhang, R. Marjieh, R. Geirhos, S. Chen, S. Kornblith, S. Rane, T. Konkle, T. P. O'Connell, T. Unterthiner, A. K. Lampinen, K.-R. Müller, M. Toneva, T. L. Griffiths, Getting aligned on representational alignment, 2023. doi:10.48550/arXiv.2310.13018.
[7] F. P. Mahner, L. Muttenthaler, U. Güçlü, M. N. Hebart, Dimensions underlying the representational alignment of deep neural networks with humans, 2024.
[8] L. Muttenthaler, J. Dippel, L. Linhardt, R. A. Vandermeulen, S. Kornblith, Human alignment of neural network representations, in: ICLR, 2022.
[9] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, W. Brendel, Partial success in closing the gap between human and machine vision, in: NeurIPS, 2021.
[10] C. G. Bartnik, I. Groen, Human and Deep Neural Network Alignment in Navigational Affordance Perception, in: Re-Align workshop, 2024.
[11] I. Kotseruba, J. K. Tsotsos, 40 years of cognitive architectures: core cognitive abilities and practical applications, Artif. Intell. Rev. (2020). doi:10.1007/s10462-018-9646-y.
[12] C. Eliasmith, How to Build a Brain: A Neural Architecture for Biological Cognition, OUP, 2013. doi:10.1093/acprof:oso/9780199794546.001.0001.
[13] M. Hersche, M. Zeqiri, L. Benini, A. Sebastian, A. Rahimi, A Neuro-vector-symbolic Architecture for Solving Raven's Progressive Matrices, Nat. Mach. Intell. (2023). doi:https://doi.org/10.1038/s42256-023-00630-8.
[14] A. Mitrokhin, P. Sutor, C. Fermüller, Y. Aloimonos, Learning sensorimotor control with neuromorphic sensors: Toward hyperdimensional active perception, Sci. Robot. (2019). doi:10.1126/scirobotics.aaw6736.
[15] N. S.-Y. Dumont, J. Orchard, C. Eliasmith, A model of path integration that connects neural and symbolic representation, CogSci (2022).
[16] M. Bartlett, T. C. Stewart, J. Orchard, Fast Online Reinforcement Learning with Biologically-Based State Representations, in: ICCM, 2022.
[17] B. Komer, T. Stewart, A. Voelker, C. Eliasmith, A neural representation of continuous space using fractional binding, in: CogSci, 2019.
[18] N. S.-Y. Dumont, C. Eliasmith, Accurate representation for spatial cognition using grid cells, Proceedings of the Annual Meeting of the Cognitive Science Society 42 (2020).
[19] A. Penzkofer, L. Shi, A. Bulling, VSA4VQA: Scaling A Vector Symbolic Architecture To Visual Question Answering on Natural Images, in: CogSci, 2024.
[20] E. Bermudez-Contreras, B. J. Clark, A. Wilber, The Neuroscience of Spatial Navigation and the Relationship to Artificial Intelligence, Front. Comput. Neurosci. 14 (2020). doi:10.3389/fncom.2020.00063.
[21] A. Penzkofer, M. Furlong, C. Eliasmith, A. Bulling, A Cognitively Plausible Visual Working Memory Model, in: Proc. Annual Meeting of the Cognitive Science Society (CogSci), 2025, pp. 1–6.
[22] J. Dotzel, B. Kotb, J. Dotzel, M. S. Abdelfattah, Z. Zhang, Exploring the Limits of Semantic Image Compression at Micro-bits per Pixel, in: ICLR Tiny Papers, 2024. doi:10.48550/arXiv.2402.13536.
[23] H. Caesar, J. Uijlings, V. Ferrari, COCO-Stuff: Thing and Stuff Classes in Context, in: CVPR, 2018. doi:10.1109/CVPR.2018.00132.

[24] N. Kriegeskorte, M. Mur, P. Bandettini, Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience, Front. Syst. Neurosci (2008). doi:10.3389/neuro.06.004.2008.
[25] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 Million Image Database for Scene Recognition, TPAMI (2018). doi:10.1109/TPAMI.2017.2723009.
[26] S. Song, S. P. Lichtenberg, J. Xiao, SUN RGB-D: A RGB-D scene understanding benchmark suite, in: CVPR, 2015. doi:10.1109/CVPR.2015.7298655.
[27] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene Parsing Through ADE20K Dataset, in: CVPR, 2017.
[28] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, TPAMI (2013). doi:10.1109/TPAMI.2013.50.
[29] L. Xie, F. Lee, L. Liu, K. Kotani, Q. Chen, Scene recognition: A comprehensive survey, Pat. Cog. (2020). doi:10.1016/j.patcog.2020.107205.
[30] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, T. Kipf, Object-Centric Learning with Slot Attention, in: NeurIPS, volume 33, 2020.
[31] K. Fan, Z. Bai, T. Xiao, T. He, M. Horn, Y. Fu, F. Locatello, Z. Zhang, Adaptive Slot Attention: Object Discovery with Dynamic Slot Number, in: CVPR, 2024. doi:10.1109/CVPR52733.2024.02176.
[32] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, R. Girshick, Masked Autoencoders Are Scalable Vision Learners, in: CVPR, 2022. doi:10.1109/CVPR52688.2022.01553.
[33] J. Rajasegaran, X. Chen, R. Li, C. Feichtenhofer, J. Malik, S. Ginosar, Gaussian Masked Autoencoders, 2025. doi:10.48550/arXiv.2501.03229.
[34] H. Bao, L. Dong, S. Piao, F. Wei, BEiT: BERT Pre-Training of Image Transformers, in: ICLR, 2021.
[35] A. Mitrokhin, P. Sutor, D. Summers-Stay, C. Fermüller, Y. Aloimonos, Symbolic Representation and Learning With Hyperdimensional Computing, Front. Robot. AI (2020). doi:10.3389/frobt.2020.00063.
[36] D. Kleyko, D. Rachkovskij, E. Osipov, A. Rahimi, A Survey on Hyperdimensional Computing aka Vector Symbolic Architectures, Part II: Applications, Cognitive Models, and Challenges, ACM Comput. Surv. (2023). doi:10.1145/3558000.
[37] M. Imani, C. Huang, D. Kong, T. Rosing, Hierarchical hyperdimensional computing for energy efficient classification, in: DAC, 2018. doi:10.1145/3195970.3196060.
[38] S. I. Gallant, P. Culliton, Positional binding with distributed representations, in: ICIVC, 2016. doi:10.1109/ICIVC.2016.7571282.
[39] A. Rahimi, P. Kanerva, J. M. Rabaey, A Robust and Energy-Efficient Classifier Using Brain-Inspired Hyperdimensional Computing, in: ISLPED, 2016. doi:10.1145/2934583.2934624.
[40] P. Blouw, X. Choo, E. Hunsberger, C. Eliasmith, Benchmarking Keyword Spotting Efficiency on Neuromorphic Hardware, in: NICE workshop, ACM, 2019. doi:10.1145/3320288.3320304.
[41] T. Lu, A. Voelker, B. Komer, C. Eliasmith, Representing spatial relations with fractional binding, in: CogSci, 2019.
[42] T. A. Plate, Holographic reduced representations, IEEE Trans. Neural Netw. 6 3 (1995).
[43] M. N. Hebart, C. Y. Zheng, F. Pereira, C. I. Baker, Revealing the multidimensional mental representations of natural objects underlying human similarity judgements, Nat. Hum. Behav. (2020). doi:10.1038/s41562-020-00951-3.
[44] I. I. A. Groen, E. H. Silson, C. I. Baker, Contributions of low- and high-level properties to neural processing of visual scenes in the human brain, Philos. Trans. R. Soc. B (2017). doi:10.1098/rstb.2016.0102.
[45] M. F. Bonner, R. A. Epstein, Coding of navigational affordances in the human visual system, PNAS (2017). doi:10.1073/pnas.1618228114.
[46] K. Yadav, R. Ramrakhya, A. Majumdar, V.-P. Berges, S. Kuhar, D. Batra, A. Baevski, O. Maksymets, Offline Visual Representation Learning for Embodied Navigation, in: Reincarnating RL Workshop, 2023.
[47] K. Muhammad, J. Ahmad, S. W. Baik, Early fire detection using convolutional neural networks during surveillance for effective disaster management, Neurocom. (2018). doi:10.1016/j.neucom.

2017.04.083.
[48] A. Pal, C. Nieto-Granda, H. I. Christensen, DEDUCE: Diverse scEne Detection methods in Unseen Challenging Environments, in: IROS, 2019. doi:10.1109/IROS40897.2019.8968108.
[49] B. X. Chen, R. Sahdev, D. Wu, X. Zhao, M. Papagelis, J. K. Tsotsos, Scene Classification in Indoor Environments for Robots using Context Based Word Embeddings, in: ICRA workshop, 2019. doi:10.48550/arXiv.1908.06422.
[50] B. Miao, L. Zhou, A. S. Mian, T. L. Lam, Y. Xu, Object-to-Scene: Learning to Transfer Object Knowledge to Indoor Scene Recognition, in: IROS, 2021. doi:10.1109/IROS51168.2021.9636700.
[51] L. Zhou, Y. Zhou, X. Qi, J. Hu, T. L. Lam, Y. Xu, Attentional Graph Convolutional Network for Structure-Aware Audiovisual Scene Classification, IEEE TIM (2023). doi:10.1109/TIM.2023.3260282.
[52] C. Song, X. Ma, SRRM: Semantic Region Relation Model for Indoor Scene Recognition, in: IJCNN, 2023. doi:10.1109/IJCNN54540.2023.10191605.
[53] C. Song, H. Wu, X. Ma, Y. Li, Semantic-embedded similarity prototype for scene recognition, Pattern Recognit. (2024). doi:10.1016/j.patcog.2024.110725.
[54] N. S.-Y. Dumont, A. Stöckel, P. M. Furlong, M. Bartlett, C. Eliasmith, T. C. Stewart, Biologically-Based Computation: How Neural Details and Dynamics Are Suited for Implementing a Variety of Algorithms, Brain Sciences (2023). doi:10.3390/brainsci13020245.
[55] R. J. Gardner, E. Hermansen, M. Pachitariu, Y. Burak, N. A. Baas, B. A. Dunn, M.-B. Moser, E. I. Moser, Toroidal topology of population activity in grid cells, Nature (2021).
[56] T. A. Plate, Holographic Reduced Representation: Distributed Representation for Cognitive Structures, CSLI, 2003.
[57] K. Dwivedi, R. M. Cichy, G. Roig, Unraveling Representations in Scene-selective Brain Regions Using Scene Parsing Deep Neural Networks, J. Cogn. Neurosci. 33 (2021). doi:10.1162/jocn_a_01624.
[58] J. S. Bakin, K. Nakayama, C. D. Gilbert, Visual Responses in Monkey Areas V1 and V2 to Three-Dimensional Surface Configurations 20 (????). doi:10.1523/JNEUROSCI.20-21-08188.2000.
[59] D. D. Dilks, F. S. Kamps, A. S. Persichetti, Three cortical scene systems and their development, Trends in cognitive sciences 26 (2022) 117–127. doi:10.1016/j.tics.2021.11.002.
[60] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: MICCAI, 2015. doi:10.1007/978-3-319-24574-4_28.
[61] N. Mu, A. Kirillov, D. Wagner, S. Xie, SLIP: Self-supervision Meets Language-Image Pre-training, in: ECCV, 2022. doi:10.1007/978-3-031-19809-0_30.
[62] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: ICLR, 2018.
[63] E. Lei, Y. B. Uslu, H. Hassani, S. S. Bidokhti, Text + Sketch: Image Compression at Ultra Low Rates, in: ICML Neural Compression Workshop, 2023. doi:10.48550/arXiv.2307.01944.
[64] F. Liu, G. Emerson, N. Collier, Visual Spatial Reasoning, Transactions of the Association for Computational Linguistics 11 (2023) 635–651. doi:10.1162/tacl_a_00566.
[65] N. Chang, J. A. Pyles, A. Marcus, A. Gupta, M. J. Tarr, E. M. Aminoff, BOLD5000, a public fMRI dataset while viewing 5000 visual images, Scientific Data (2019). doi:10.1038/s41597-019-0052-3.

A. Sampling Methods

In our preliminary analysis (see Table 2), we found that only a percentage of the pixels of a given object can be encoded without a significant loss in representation accuracy. Therefore, we tested two sampling techniques: uniform sampling and Gaussian sampling. Uniform sampling randomly picks points within the mask, while Gaussian sampling draws from a Gaussian distribution around the mask's centre of mass and covariance matrix given by the segmentation mask and points outside the mask are discarded. Both sampling methods are visualised in Figure 3, where we see that the Gaussian sampling technique focuses on the centre of the object.



(a) Uniform sampling.

(b) Gaussian sampling.

Figure 3: Comparison of sampling methods: **(a)** uniform sampling and **(b)** Gaussian sampling.

B. Hyper-parameter Selection

Table 2
Hyper-parameter analysis on 50 samples of the COCO-Stuff dataset. Uncertainty corresponds to one std.

Encoding	Dimensions	λ	τ	# Points ↓	Time [s] ↓	IoU ↑
Full scene	1,015	27.5	0.65	179,950	8.8 ±1.1	36.1 ±12.9%
	1,945	30.0	0.7	179,950	37.7 ±4.8	43.4 ±13.1%
	3,751	22	0.7	179,950	40.1 ±5.1	47.6 ±13.5%
Uniform	1,015	27.5	0.65	3,706	1.5 ±0.3	37.5 ±13.1%
	1,945	30	0.7	3,706	6.0 ±1.1	41.8 ±12.5%
	3,751	22	0.7	3,706	6.3 ±1.1	47.1 ±13.6%
Gaussian	1,015	20.0	0.6	16,202	**2.2 ±0.4**	40.5 ±9.4%
	1,945	15.0	0.6	16,202	4.7 ±1.4	45.3 ±9.3%
	3,751	17	0.55	16,202	5.1 ±1.5	**47.9 ±9.5%**

For fine-tuning our hyperparameters: lengthscale λ and threshold τ, we performed a grid search on 50 samples from the COCOStuff dataset [23]. We encoded all objects with the full, uniform, or Gaussian encoding scheme and evaluated reconstruction performance in terms of IoU between ground truth object segmentation and predicted masks, i.e. taking all points above the threshold τ in the similarity map after unbinding with the object's inverse SSP. Table 2 aggregates the results with the average number of encoded points, the average encoding time, and the average object IoU for each best representation configuration with tuned λ and τ. Overall, the IoU accuracy increases with an increase in SSP dimensions, from 1,015 (top) to 3,751 dimensions (bottom), due to the increased capacity of higher-dimensional vectors. Similarly, encoding time also increases with SSP dimensionality, which is due to the larger dimensionality of all vectors x in Equation 2, while the number of encoded points does not. This is expected as the point selection only depends on the encoding scheme and the size of the object masks. We find that uniform sampling is on par with the full encoding scheme, while being significantly faster. Gaussian sampling achieves the best results in terms of IoU and encoding time as it uses a fixed number of samples to draw, which also increases decoding accuracy for smaller objects.

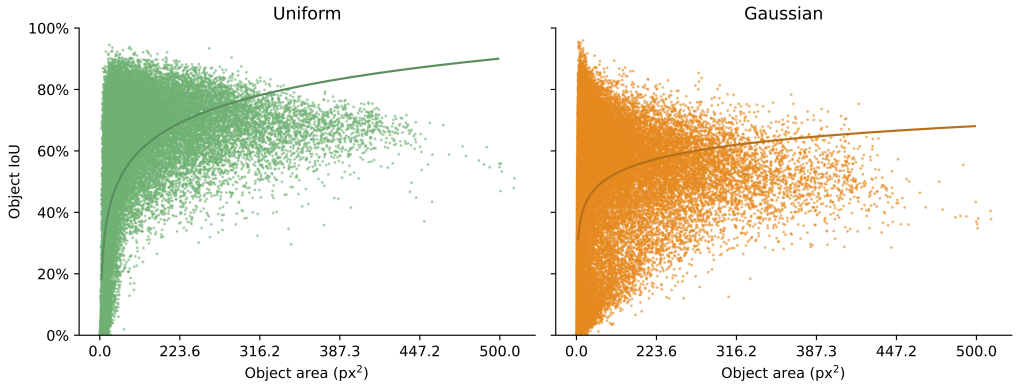


Figure 4: Object-level IoU dependent on object area. Uniform encoding (left) and Gaussian encoding (right).

Based on this preliminary analysis, we selected 3,751 dimensions for the SSPictR representation. We strived to keep the representation as compact as possible, i.e., at the lowest SSP dimension that achieves reasonable results. We further evaluated the effect of object areas on the IoU accuracy, where a weak correlation was reported in previous works [41, 19]. Figure 4 shows that the uniform encoding scheme struggles with small objects, i.e., object areas $< 200px^2$, but achieves higher object IoU in general compared to the Gaussian encoding. The fitted logarithmic curves with $-0.72 + 0.13 \cdot \log(x)$ (Uniform) and $-0.15 + 0.07 \cdot \log(x)$ (Gaussian) emphasise this point. Hence, we selected uniform encoding for all analyses that follow.

C. Scene Reconstruction Refinement

Qualitative examples of the scene reconstruction refinement are shown in Figure 5. In the first example (top) the similarity map (b) captures the holes in the segmentation mask, however, when thresholding with a fixed τ this information is lost (c). The UNet model is able to recapture this information (d) from the similarity map it receives as input by allowing a learned threshold per image and area. However, the second example (bottom) highlights that both approaches struggle with small objects, indicating that the representation does not capture small objects well.

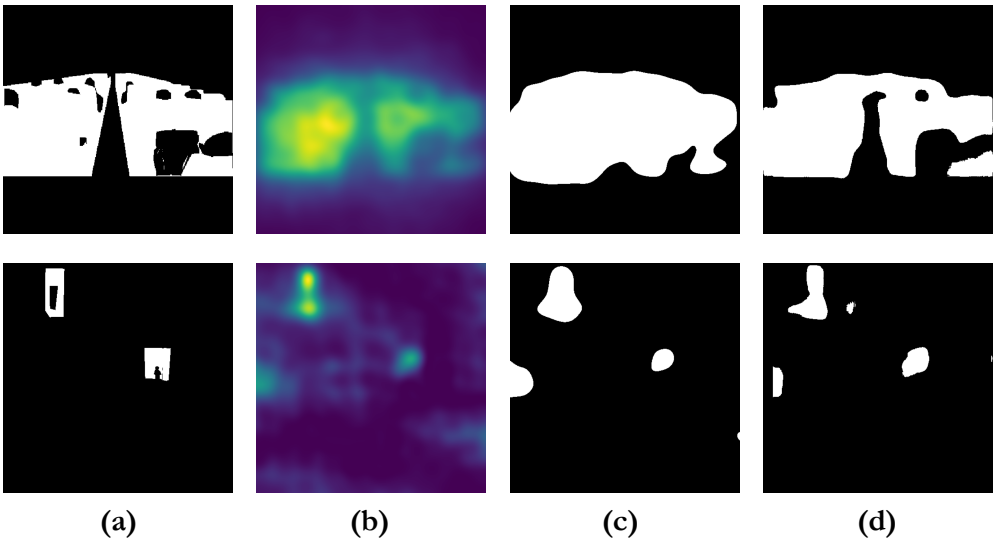


Figure 5: Example **(a)** ground truth segmentation masks from ADE20K with **(b)** SSPictR similarity maps, **(c)** threshold reconstruction, and **(d)** refined UNet model reconstruction.