# SSPICTR: SPATIAL SEMANTIC POINTER PICTURE REPRESENTATION

Anonymous authors

Paper under double-blind review

## Abstract

The development of image representations that capture semantic and spatial information efficiently, which are also interpretable and generalisable, remains unsolved. Drawing from a cognitive modelling framework, we propose SSPictR - a biologically plausible image representation based on spatial semantic pointers (SSPs). SSPictR encodes semantic labels and their spatial locations extracted from segmentation maps and only requires a single vector to capture a fully decodable neuro-symbolic representation of a natural scene. It is inherently interpretable, offers a high compression factor and significantly faster inference speed on downstream tasks, such as scene recognition. We evaluate the efficiency and generalisability of SSPictR on the popular Places365, and ADE20K datasets for scene recognition, on COCOStuff for segmentation reconstruction, and on VISC and Savoias for prediction of visual complexity. We show that the scene representations provided by SSPictR are more generalisable within and across these tasks while only requiring a fraction of model parameters and, therefore, offer 25 times higher inference speed, with comparable accuracy. As such, SSPictR opens up a new direction for future research on cognitively-inspired image representations that are not only significantly smaller but also more interpretable and generalisable.

027 028 029

030

003 004

010 011

012

013

014

015

016

017

018

019

021

023

025

026

## 1 INTRODUCTION

031 Machine learning models for computer vision have achieved human-level performance on a number 032 of tasks, such as object detection (Zong et al., 2023), semantic segmentation (Chen et al., 2022), 033 or depth estimation (Zhao et al., 2023). But these improvements come at an ever-increasing cost: 034 These models are highly inefficient and lack interpretability, i.e. they require training of millions of parameters from large amounts of data and their learned internal representations are notoriously challenging to analyse and understand. Furthermore, despite impressive performance on tasks that they were trained for, they still lack generalisability to out-of-domain (OOD) data (Geirhos et al., 037 2018; Mahner et al., 2024). Aligning model and human representations has increased robustness in object detection (Geirhos et al., 2019), improved performance (Sucholutsky et al., 2023), and has advanced our understanding of human cognition through better computational models (Mahner 040 et al., 2024). While there has been a lot of research in aligning deep neural networks (DNNs) in 041 object classification tasks (Mahner et al., 2024; Muttenthaler et al., 2022; Geirhos et al., 2021), 042 human alignment for scene understanding remains under-explored (Bartnik & Groen, 2024). 043

Computational models of human perception have a long-standing history of research in cognitive 044 science (Kotseruba & Tsotsos, 2020). Most interesting are biologically-plausible cognitive models, as they can be naturally integrated with DNNs while maintaining the interpretability of symbolic 046 systems. A popular model is the semantic pointer architecture (SPA; Eliasmith 2013) – a cognitive 047 modelling framework based on vector symbolic algebras (VSAs). VSAs use hyperdimensional dis-048 tributed vectors for efficient and robust representation of concepts (Kleyko et al., 2022). There are many successful applications of cognitively-inspired VSAs, for example, in abstract reasoning (Hersche et al., 2023), ego-motion prediction (Mitrokhin et al., 2019), path integration (Dumont et al., 051 2022), and reinforcement learning (Bartlett et al., 2022). Particularly interesting as representations for scene understanding are the works by Komer et al. (2019) and Penzkofer et al. (2024), as they 052 encode objects in a grid-like continuous vector space similar to cognitive maps found in the brain (Bermudez-Contreras et al., 2020).



Figure 1: Overview of SSPictR – a cognitively-inspired image representation that only requires a single vector, which can be used to predict scene categories, visual complexity ratings, or even decode a full cognitive map of the scene.

In this work, we bring together this research and propose SSPictR- a neurally inspired image repre-067 sentation that is more efficient, interpretable, and generalisable than existing image representations. 068 At its core, SSPictR encodes objects from segmentation maps into a continuous vector space that 069 compresses the semantic and spatial information of a scene. Figure 1 shows an overview. In contrast to existing representations, SSPictR offers a significantly better compression factor of only 0.46 bits 071 per pixel (vs 8 bits for segmentation maps) and an inference speed of 2,857 fps for scene recog-072 nition. The compressed vector representation only has 3,751 dimensions and can be queried, used 073 to fully reconstruct the segmentation map, or used as a feature embedding for further downstream 074 tasks, such as visual complexity prediction. We evaluate different encoding methods for this novel 075 representation based on reconstruction error of the segmentation map. We also show that SSPictR 076 is highly generalisable across tasks (scene recognition, visual complexity) and datasets, providing 077 competitive performance for OOD scene recognition on the ADE20K dataset (Zhou et al., 2017).

In summary, we present a novel scene representation: *SSPictR*. SSPictR is highly efficient, interpretable, and generalisable. We provide a construction method from segmentation maps and evaluate different encoding schemes. Further, we evaluate the representation quality via linear probing on scene recognition, its generalisability on OOD data, and highlight potential applications in computer vision, cognitive science, and robotics.

083 084

062

063

064 065 066

## 2 RELATED WORK

085 086 087

## 2.1 Scene Representations

The choice of data representation is a key factor for the performance of machine learning models (Bengio et al., 2013). Scene representation, in particular, is challenging due to complex configurations, i.e., scenes are comprised of diverse objects in complex spatial layouts with substantial semantic ambiguity (Xie et al., 2020). Furthermore, scenes can also be characterised as environments for embodied agents to navigate in (Malcolm et al., 2016).

Scene Recognition. Scene recognition, the task of classifying scenes into categories, is considered 094 a fundamentally important but challenging task in computer vision (Zeng et al., 2021), with a wide 095 range of applications, from robot navigation (Yadav et al., 2023) to disaster detection (Muhammad 096 et al., 2018). Current benchmark datasets include Places365 (Zhou et al., 2018), ADE20K (Zhou 097 et al., 2017), SUN397 (Xiao et al., 2010), and MIT67 (Quattoni & Torralba, 2009). While MIT67 098 only contains 15 thousand images classified into 67 indoor scenes, the most recent Places365, con-099 sists of 10 million images with annotations for 434 scene classes in three macro-classes: Indoor, Nature, and Urban. The advance of the available datasets for scene recognition has significantly im-100 pacted scene recognition methods. Previously, methods focused on enhancing specific features, e.g. 101 semantic features (López-Cifuentes et al., 2020), multi-layer features (Liu et al., 2019), or multi-102 view features (Seong et al., 2020), achieving state-of-the-art results on MIT67 and SUN397 (Zeng 103 et al., 2021). Methods trained on Places365, however, leverage the large amount of data so that 104 standard CNNs significantly outperform previous approaches (Zhou et al., 2018). 105

Since SSPictR is primarily designed for robotics applications, such as visual navigation, our fo cus is particularly on indoor scene recognition. Pal et al. (2019) and Chen et al. (2019) proposed
 subsets of the Places365 dataset containing seven and 14 classes, respectively. Building on this,

108 Miao et al. (2021) proposed a novel model that resembles our approach by integrating knowledge 109 from semantic segmentation maps: object-to-scene (OTS). OTS extracts object features through a 110 pre-trained segmentation model and calculates object relations, outperforming both Pal et al. (2019) 111 and Chen et al. (2019). However, OTS also requires up to 255 million model parameters and, there-112 fore, only achieves an inference speed of three fps. More recently, attentional graph convolutional network (AGCN; Zhou et al. 2023) achieved higher performance than OTS on both datasets while 113 also increasing inference time to 27 fps. Song & Ma (2023) proposed a semantic region relation-114 ship model (SRRM) and combined it with the PlacesCNN module by Zhao et al. (2023), yielding 115 CSRRM, which achieves the current state-of-the-art performance on Places365-7 and Places365-116 14. In follow-up work (Song et al., 2024), the authors focused on computational efficiency that is 117 essential for the low-resource and high-speed requirements of edge devices in practical robotics ap-118 plications. However, their method is specific for enhancing scene recognition models and does not 119 transfer to other tasks. 120

Human Alignment. Aligning representations of deep neural networks (DNNs) to humans is a 121 promising avenue to increase performance and generalisability of computer vision models (Sucholut-122 sky et al., 2023; Chang et al., 2019). While many works analysed object representations and their 123 alignment to human similarity judgements (Muttenthaler et al., 2022; Geirhos et al., 2021; Hebart 124 et al., 2020; King et al., 2019), meaningful representations of full scenes remain under-explored. 125 Groen et al. (2017) analysed scene recognition in humans and found that in addition to object co-126 occurrence statistics as found by Stansbury et al. (2013), other features across different levels of 127 visual processing play an important role, such as spatial layouts, boundaries, and textures. This en-128 courages building scene representations based on object-level statistics, such as the presence and fea-129 tures of certain objects, but also highlights the importance of additional spatial information. Hence, we construct SSPictR from segmentation maps of images, which provide both object information as 130 well as their spatial layout in the scene. 131

132 Visual Complexity. Segmentation maps and object-level statistics were also successfully used to 133 predict the visual complexity of images (Nath et al., 2024). Visual complexity ratings of images are 134 important for cognitive science studies, as they impact attention, engagement, and memorability of 135 stimuli. Furthermore, visual complexity is relevant for practical applications, such as user experience on webpages and brand logo design (Kyle-Davidson et al., 2023). To understand and effectively 136 model visual complexity, most methods rely on hand-crafted features to explain human complexity 137 ratings, i.e., number of regions, frequency factor, and colours (Corchs et al., 2016), clutter and patch 138 symmetry (Kyle-Davidson et al., 2023), or number of classes and segments (Nath et al., 2024). Su-139 pervised methods achieve the highest correlation with human ratings, but they require large datasets 140 for training, where currently only IC9600 (Feng et al., 2023) is publicly available, offering 9,600 141 images across eight categories. Smaller datasets include Savoias (Saraee et al., 2020) consisting of 142 1,400 images with seven categories and VISC (Kyle-Davidson et al., 2023) containing 800 images 143 across 12 sub-categories.

144 145

## 146 2.2 VECTOR SYMBOLIC ALGEBRAS

147 Vector symbolic algebras (VSAs) play an important role in cognitive architectures (Kleyko et al., 148 2023; Stewart et al., 2012) and showed improved performance of machine learning methods, e.g., 149 ego-motion prediction (Mitrokhin et al., 2020), speech recognition (Imani et al., 2018), or object 150 classification (Gallant & Culliton, 2016). VSAs offer a unique way of encoding symbolic meaning 151 in hyper-dimensional distributed representations, making them inherently interpretable (Mitrokhin 152 et al., 2020) and robust to errors (Rahimi et al., 2016). Furthermore, their potential application in neuromorphic hardware make them highly efficient, achieving speed gains of up to 100 times 153 GPU performance (Blouw et al., 2019). The potential of VSAs in replicating cognitive maps for 154 navigation has been shown via path integration (Dumont et al., 2022) and reinforcement learning on 155 navigation tasks (Bartlett et al., 2022), while the potential in scene understanding has been shown 156 via visual question answering (Komer et al., 2019; Penzkofer et al., 2024). For more application 157 examples and a comprehensive literature review on VSAs, we refer to Kleyko et al. (2022; 2023). 158

In this work, we build upon the semantic pointer architecture (SPA) (Eliasmith, 2013), a cognitively inspired VSA, which uses holographic reduced representations (HRRs) (Plate, 1995), 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. Vectors can be bundled to represent

170 171

172

173

174

175 176

177

178

179

180

181 182

183

185

202

205

209

215



(a) Uniform sampling.

(b) Gaussian sampling.

Figure 2: Comparison of sampling methods on the same segmentation mask. For uniform sampling (a), a fixed percentage of points is drawn from the given mask. For Gaussian sampling (b), the centre of mass (orange) is computed. Then, a fixed number of samples from Gaussian distribution with mean as centre of mass and covariance of segmentation mask are drawn.

multiple concepts, computed by vector addition, and they can be bound together to represent concepts that belong together, e.g. a red apple. Binding – denoted by  $\circledast$  – is circular convolution in HRR and the inverse operation (unbinding) is binding with a vector's pseudo-inverse. Komer et al. (2019) have introduced fractional binding, i.e. binding the vector with itself  $k \in \mathbb{R}$  times, to encode continuous data, which enables the encoding of spatial locations.

## 3 Method

## 3.1 THEORETICAL BACKGROUND

<sup>186</sup> Under the SPA, spatial semantic pointers (SSPs) were proposed for spatial representations through the following encoding scheme: <sup>188</sup> (1)  $\tau^{-1}(x)$  (1)

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

189 where  $\phi : x \in \mathbb{R}^2 \mapsto \mathbb{R}^d$ ,  $\lambda$  defines the length scale of the encoded representation,  $\mathcal{F}^{-1}$  denotes the 190 inverse Fourier transform, and  $A \in \mathbb{R}^{d \times 2}$  is a phase matrix whose columns consist of phasors repre-191 senting different frequencies. For real-valued spatial representations, the phase matrix is conjugate 192 symmetric. For biological realism of the SSP representation, further constraints are applied to the 193 phase matrix that enable the replication of grid cell firing patterns as seen in Dumont & Eliasmith (2020). This is achieved by setting triplets of rows in the phase matrix  $120^{\circ}$  apart, resulting in grid-194 ded interference patterns. As given by biological experimental findings, this data lies on the hyper-195 toroidal manifold (Gardner et al., 2021). The dimensionality d is given by  $d = n_{\text{scales}} \cdot n_{\text{rotates}} \cdot 3 \cdot 2 + 1$ , 196 where  $n_{\text{scales}}$  denotes the scale of the firing pattern activity,  $n_{\text{rotates}}$  denotes the orientation of the grid 197 cells, 3 denotes triplets, 2 for conjugate symmetry, and +1 for the 0-frequency term. 198

Using the SSP representation and the set of operations given by HRR, we can construct cognitive maps, for example, we can construct the map M of a cat at location  $(x_1, y_1)$ , a mouse at location  $(x_2, y_2)$ , and cheese at location  $(x_3, y_3)$ :

$$\mathbf{M} = \mathbf{CAT} \circledast \phi(x_1, y_1) + \mathbf{MOUSE} \circledast \phi(x_2, y_2) + \mathbf{CHEESE} \circledast \phi(x_3, y_3),$$
(2)

then, to query an object's location, unbinding can be used :

$$\mathbf{M} \circledast \mathbf{CHEESE}^{-1} = \phi(x_3, y_3) + noise \tag{3}$$

This noise is due to the nature of the unbinding operation: since the unbinding of CHEESE distributes over the map representation, CAT  $\circledast \phi(x_1, y_1) \circledast$  CHEESE<sup>-1</sup> and MOUSE  $\circledast \phi(x_2, y_2) \circledast$ CHEESE<sup>-1</sup> produces random noise.

## 210 3.2 ENCODING SCHEME

In previous works, objects have been encoded as point sources (Dumont et al., 2022), or a set of bounding box coordinates (Penzkofer et al., 2024). However, for a more accurate representation, we instead encode segmentation maps of a scene as follows:

$$\mathbf{S} = \sum_{i} \left[ \operatorname{obj}_{i} \circledast \int_{A_{i}} \phi(x) dx \right]$$
(4)

231 232 233



Figure 3: Object-level IoU dependent on object area – comparison between encoding schemes: uniform encoding (left) and Gaussian encoding (right).

In this hyper-dimensional representation,  $obj_i \in \mathbb{R}^d$  is a semantic pointer (SP), representing the 234 class of a given object i. This object is bound with a bundle of SSPs representing the area occupied 235 by this object in the representation. In our preliminary analysis (see Table 1) we found that only a 236 percentage of the pixels of a given object can be encoded without a significant loss in representation 237 accuracy. To this end, we sample the representation either via uniform sampling within the mask 238 or by sampling from a Gaussian distribution with the mean set as the mask's centre of mass and 239 covariance matrix given by the segmentation mask. Points outside the mask are not encoded. Both 240 sampling methods are visualised in Figure 2, where we see that the Gaussian sampling technique 241 focuses on the center of the object.

242 To evaluate the quality of the SSP representation, we decode the masks used to generate the scene 243 representation S. We calculate the similarity map of an object j in the scene by taking the dot 244 product between a grid of SSPs and the encoded scene bound with the inverse of that object's SP: 245  $\left\langle S \circledast \operatorname{obj}_{j}^{-1}, \int_{A_{\operatorname{grid}}} \phi(x) dx \right\rangle$ . This yields an approximate similarity value between the SSP grid  $A_{\operatorname{grid}}$ 246 that represents each possible location and the scene SSP S that is queried for the object of interest. 247 The set of similarity values larger than some threshold  $\tau$  is used as the decoded mask. Finally, the 248 intersection-over-union (IoU) is calculated using the ground truth and the decoded mask. This ap-249 proach can be used to probe for every object in the scene, to determine what is at a specific location, 250 or to verify that an object exists in the scene, making this representation inherently interpretable. We 251 then optimise the hyperparameters most important for this encoding, i.e., the length scale, threshold, 252 percentage of encoded points, and SSP dimensionality. 253

Table 1 shows the average number of encoded points, the average encoding time, and the average 254 object IoU for each best representation configuration with tuned lengthscale  $\lambda$  and threshold  $\tau$  pa-255 rameters. For details on finetuning the parameters to find the best configuration, see Figure 5 in 256 the Appendix. We perform a grid search on 50 samples from the COCOStuff (Caesar et al., 2018) 257 dataset, which encodes 8.34 objects on average. Increasing the SSP dimensionality also increases 258 the average decoding accuracy for all encoding schemes. Similarly, encoding time also increases 259 with SSP dimensionality, but the number of encoded points does not. This is expected as the point 260 selection only depends on the encoding scheme and the size of the object masks. The increase in 261 encoding time is due to the larger dimensionality of all vectors x in equation 4. Most interestingly, we find that uniform sampling is on par with the full encoding scheme while being significantly 262 faster. Gaussian sampling achieves the best results in terms of IoU and encoding time as it uses a 263 fixed number of samples to draw, which increases decoding accuracy for smaller objects. 264

Based on this preliminary analysis, we select 3,751 dimensions for the SSPictR representation.
We strive to keep the representation as compact as possible, i.e., at the lowest SSP dimension that
achieves reasonable results. This allows for a compression rate of only 0.46 bits per pixel for a
512x512 image, similar to the highest possible compression rate of JPEG (Dotzel et al., 2024).
However, as both sampled encoding schemes achieve similar results in terms of IoU, we further
evaluate them. To this end, we analyse the effect of object areas on the IoU accuracy, where a weak

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

Table 1: Hyper-parameter analysis on 50 samples of the COCO-Stuff dataset. Uncertainty terms correspond to one standard deviation.

correlation was reported in (Lu et al., 2019; Penzkofer et al., 2024). Figure 3 shows that the uniform encoding scheme struggles with small objects, i.e.,  $< 200px^2$ , but achieves higher object IoU in general compared to the Gaussian encoding. The fitted logarithmic curves with -0.72+0.13\*log(x) (Uniform) and -0.15+0.07\*log(x) (Gaussian) emphasise this point. Hence, we selected uniform encoding for all analyses that follow.

290 291 292

293

283 284 285

286

287

288

289

## 4 EXPERIMENTS

We first evaluate the representation quality by calculating the reconstruction accuracy for semantic
segmentation on the COCOStuff 2017 dataset (Caesar et al., 2018) with 118K images. Furthermore,
we perform linear probing by training a small scene recognition model on a reduced Places365
dataset (Zhou et al., 2018) and test generalisation for out-of-domain (OOD) scene recognition on
ADE20K (Zhou et al., 2017). As an additional downstream application, we predict the visual complexity of images with SSPictR as features for a regression model.

300 301

## 4.1 REPRESENTATION QUALITY

302 For evaluating the reconstruction accuracy, we compute all SSPs with 3751 dimensions and the 303 uniform encoding scheme for the COCOStuff train and validation set (Caesar et al., 2018). Then, 304 we reconstruct the segmentation maps by unbinding the scene representation with each object vector 305 separately, yielding the similarity map for the queried object. We select all points above  $\tau = 0.7$ 306 similarity to belong to the queried object. We compare this mask with the ground truth mask of the 307 object and calculate the pixel-wise IoU. With the uniform encoding scheme, we achieve  $45.36\% \pm$ 308 13.05% average IoU on the COCOStuff validation split. Here, we encode 3,770 points on average 309 with an encoding time of  $6.13 \pm 1.18$ s per image. We further increase the reconstruction accuracy by training a UNet model (Ronneberger et al., 2015) to refine the object masks based on the SSP 310 similarity maps instead of using the threshold  $\tau$ . 311

312 Our UNet model consists of four encoder and decoder layers with a total of 465K trainable param-313 eters. We trained the model on a subset of the COCOStuff 2017 train dataset Caesar et al. (2018), 314 i.e. we use 10K samples of similarity maps, which amounts to a total of 1,235 training images, split 315 into train and validation. We evaluate the model on the official validation split (not used for validation during training) of 5,000 images and achieve an average IoU of  $57.3 \pm 7.1\%$ , a considerable 316 13.2% increase. As shown in Figure 4, the model learns to refine the given similarity maps, remov-317 ing excess noise. Based on further examples, we find that the model learns to adapt the threshold 318 of the similarity map based on object-specific features, mainly size, as the SSPictR representation 319 struggles with small objects (see Figure 3). To conclude, while we only trained on a small subset 320 of the available dataset, we achieved a significant reconstruction performance increase. We believe 321 this could be further increased by more training or a more elaborate network structure. 322

We then evaluate the quality of our representation via linear probing, a standard method for determining intermediate representation quality in self-supervised models (Mu et al., 2022). More specif-



Figure 4: Example (a) similarity maps with (b) ground truth masks, (c) threshold prediction, and (d) refined Unet model prediction.

ically, we train a small classification model on the SSP representations of the popular Places365-7 (Pal et al., 2019) and Places365-14 (Chen et al., 2019) subsets of the Places365 dataset Zhou et al. (2018) for indoor scene recognition. We follow the setup by Miao et al. (2021). The Places365-7 training set consists of 35,000 images, and the Places365-14 training set of 55,000 images. As Places365 does not offer segmentation maps, we run the pre-trained VPD model (Zhao et al., 2023) for semantic segmentation, which was trained on the ADE20K dataset. Therefore, we use the same object classes for both datasets. After generating the 3,751-dimensional SSPs for all samples, we train a two-layer linear neural network (NN) model with approx. seven million parameters. The lin-ear 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 use both batch normalisation and dropout layers, where the batch size is set to 1,024 and  $p_{dropout} = 0.4$ . We use AdamW (Loshchilov & Hutter, 2018) as an optimiser with a learning rate of 0.00195. We train the model for 25 epochs and evaluate its performance on the held-out official validation set with 700 images and 1,100 images for the seven and 14 classes, respectively. 

Table 2 summarises our scene recognition results, also in comparison with previous methods. As can be seen from the table, our method achieves comparable performance in terms of accuracy but only takes the 3,751-dim SSP vectors as inputs and achieves faster inference speed of 2,875 frames-per-second (fps), 25 times more than the next best method. This showcases the potential of SSPictR to be used as a low-memory, high-efficiency image representation for edge devices. We further evaluate the generalisation performance of our trained Places365-7 model on the ADE20K dataset as OOD data and achieve a remarkable performance of 94.5% classification accuracy. This even outperforms our SVM baseline (RBF kernel with c = 5) trained on ADE20K directly, achieving 94.2% test accuracy on a hold-out set of 817 images. The higher performance on the ADE20K dataset might be explained by the availability of ground truth segmentation maps. The quality of segmentation maps is a limiting factor of current scene recognition methods (Song et al., 2024). 

Table 2: Scene recognition results and comparison to previous state-of-the-art.

Method	Parameters	Inference [fps]	Places365-7	Places365-14
OTS Miao et al. (2021)	255 M	3	90.1	85.9
AGCN Zhou et al. (2023)	85 M	27	91.7	86.0
CSRRM Song & Ma (2023)	50 M	-	93.4	88.7
GLS + BCL Song et al. (2024)	25 M	115	90.6	86.6
SSPictR (Ours)	7 M	2,857	90.1	82.2

Table 3: Visual complexity prediction and comparison to other handcrafted methods, as well as supervised methods trained on larger datasets. We report Spearman rank correlation r with human complexity ratings.

Method	<b>Savoias</b> Art	<b>Savoias</b> Scenes	<b>Savoias</b> Int. Design	VISC
handcrafted features				
clutter + symmetry Kyle-Davidson et al. (2023)	0.55	0.54	0.74	0.60
#seg + #classes Nath et al. (2024)	0.73	0.78	0.61	0.56
#seg + #classes + symmetry Nath et al. (2024)	-	-	0.80	0.68
supervised models				
ComplexityNet Kyle-Davidson et al. (2023)	0.30	0.36	0.56	-
ICNet Feng et al. (2023)	0.81	0.79	0.89	0.72
SSPs	0.42 (KR)	0.52 (RF)	0.45 (KR)	0.53 (RF)
SSPs + # classes	0.45 (KR)	0.60 (KR)	0.51 (RF)	0.54 (RF)
SSPs + symmetry	0.45 (RF)	0.54 (RF)	0.49 (RF)	0.57 (RF)

For a further visual analysis of the SSP representations from the ADE20K dataset in comparison to Places365, see Appendix A.2.

## 4.2 VISUAL COMPLEXITY

381 382

396 397

399

400 401

402

403 To evaluate the generalisability of SSPictR to other downstream tasks, we predict human visual 404 complexity ratings of images. Following previous work (Nath et al., 2024) we use handcrafted 405 features, i.e., our SSP representations, and train simple regressors to predict the complexity scores 406 that go from zero to one hundred. We evaluate the following models: support vector regression 407 (SVR), kernel ridge regression (KR), gradient boosting (GB), and random forest regression (RF). 408 KR with a cosine kernel, i.e., cosine similarity as a comparative measure between different vectors, 409 yields the best results. This is intuitive as the cosine similarity has been shown to be the same as the 410 distance between two unitary SSPs in the Fourier domain (Voelker, 2020). We evaluate our models 411 on 3 classes of the Savoias dataset (Saraee et al., 2020) and on the full VISC dataset with a 7-fold 412 cross-validation. Unfortunately, we did not get access to the larger scale IC9600 (Feng et al., 2023) dataset and are therefore unable to test a supervised method. We perform the same preprocessing 413 pipeline as for the scene recognition model on Places365; first, we generate segmentation maps 414 for each image with the pre-trained VPD model (Zhao et al., 2023), then, we compute our SSP 415 representations with the uniform encoding scheme. For a visualisation of samples from the different 416 datasets, see Appendix A.3. The art category (Savoias) is the most difficult for the VPD model 417 to segment, as it is comprised of paintings or simple drawings, which was not in the training data 418 (ADE20K; Zhou et al. 2017) of VPD. Further, we found that the interior design category, while close 419 to the interior scene images in ADE20K, has significantly more objects encoded per image: 22.26 420 on average compared to 10.87 (VISC, Savoias SCENES) and 8.23 (Savoias ART). Additionally, the 421 average object size is also significantly smaller.

422 Table 3 summarises relevant prior work with Spearman rank correlation coefficient r between pre-423 dictions and the human complexity ratings (values taken over from Nath et al. (2024)). Our results 424 show that we achieve reasonable performance on the Savoias SCENES dataset, i.e., on par with 425 Kyle-Davidson et al. (2023) and similarly on VISC. However, SSPs as features significantly strug-426 gle with the ART and INTERIOR DESIGN category; we believe this is due to limited segmentation 427 accuracy (ART) and the high number of objects (INTERIOR DESIGN). Further, we test whether 428 additional features can increase the correlation coefficient. Inspired by Nath et al. (2024), we use the number of classes extracted from the number of unique labels in the segmentation map. Addi-429 tionally, we also calculate the patch symmetry, as proposed by Kyle-Davidson et al. (2023). Both 430 features improve the predictions, however, since the added feature is only one of 3752 the impact is 431 not significant. Overall, we show that SSPictR as features for visual complexity prediction perform similar to previous handcrafted features, but there is still room to improve. We believe a supervised
method is better suited to work with the large feature space of 3751 dimensions, and we will test
this hypothesis when a suitable dataset becomes available.

## 5 DISCUSSION AND CONCLUSION

438 We have presented SSPictR – a cognitively-inspired image representation that is inherently inter-439 pretable and efficient. We evaluated different encoding schemes and found the best hyperparameters 440 for encoding a full scene based on segmentation maps. Further, we evaluated the representation qual-441 ity by calculating the reconstruction accuracy on COCOStuff (Caesar et al., 2018), where we found 442 that a simple mask refinement model can significantly enhance reconstruction IoU to  $57.3 \pm 7.1\%$ . 443 The key advantage of SSPictR is that the representations are compact, only requiring 0.46 bits per 444 pixel and that they can directly be used for downstream tasks, such as scene recognition and visual 445 complexity prediction.

446 For scene recognition, a key task for robotic navigation (Xie et al., 2020; Miao et al., 2021), we 447 trained a small neural network that achieved comparable performance on two indoor scene subsets 448 of the popular Places365 dataset (Zhou et al., 2018). Similar to previous work (Song et al., 2024), 449 we found that scene recognition accuracy is limited by segmentation quality, i.e. where ground 450 truth segmentation maps are available, we achieved a remarkable scene recognition accuracy of 451 94.2% with a SVM on the ADE20K dataset (Zhou et al., 2017). Further, the model trained on 452 Places365 achieves an even higher performance on ADE20K as out-of-distribution data with 94.5%, highlighting the generalisation ability of our representation. Additionally, SSPictR achieves 25 times 453 higher processing speed than previous methods. 454

- 455 Moreover, SSPictR is able to generalise across different tasks, as we have shown by performing vi-456 sual complexity prediction. Visual complexity prediction is an important task in cognitive science, as 457 perceived complexity is linked to engagement and attention, influencing subjects' reaction to given 458 stimuli (Nath et al., 2024). While we achieve comparable performance to other handcrafted feature methods on datasets that represent scenes, i.e., VISC (Kyle-Davidson et al., 2023) and Savoias 459 SCENES (Saraee et al., 2020), the SSP representations struggle with abstract images (Savoias ART) 460 and images with a significantly higher amount of objects (Savoias INTERIOR DESIGN). The latter 461 could potentially be addressed by increasing the capacity of the SSPs by allowing a higher dimen-462 sionality. However, the more abstract art images do not encompass real-world scenes with spatial 463 layouts, which is the intended application area of SSPictR. 464
- Our overall goal is to develop a compact image representation that can be deployed end-to-end on 465 edge devices for efficient visual navigation and other downstream tasks. SSPictR is an important first 466 step towards this goal. In future work, we would like to address some of the discussed limitations 467 by integrating a larger object vocabulary and more fine-grained segmentation, e.g., by using the 468 SAM model (Kirillov et al., 2023), which detects segments at different scales. Further, we plan to 469 train an end-to-end image to SSP model, potentially with a fully neural implementation to achieve 470 the highest efficiency. In general, we believe SSPictR is uniquely suitable for robotics applications, 471 such as embodied visual navigation (Yadav et al., 2023; Ramakrishnan et al., 2021), as well as for 472 further cognitive applications. More specifically, we would like to analyse the alignment of SSPictR 473 to human scene representations, e.g., by performing similarity analysis on fMRI data (Chang et al., 474 2019), or integrate it into symbolic reasoning systems (Hersche et al., 2023).

As we have shown, SSPictR opens up a new direction of research, leveraging the advantages of cognitively-inspired image representations.

478

436

437

## 479 REFERENCES

- Madeleine Bartlett, Terrence C. Stewart, and Jeff Orchard. Fast Online Reinforcement Learning with Biologically-Based State Representations. In 20th International Conference on Cognitive Modeling, 2022. URL https://iccm-conference.neocities.org/2022/papers/860.pdf.
- 484
- 485 Clemens Georg Bartnik and Iris Groen. Human and Deep Neural Network Alignment in Navigational Affordance Perception. In *ICLR 2024 Workshop on Representational Alignment (Re-Align)*,

512

486 March 2024. URL https://openreview.net/forum?id=FS5Lq9Flep&noteId= iSzjRKFXHJ.

- Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013. doi: 10.1109/TPAMI.2013.50. URL http://ieeexplore.ieee.org/document/6472238/.
- Edgar Bermudez-Contreras, Benjamin J. Clark, and Aaron Wilber. The Neuroscience of Spatial Navigation and the Relationship to Artificial Intelligence. Frontiers in Computational Neuroscience, 14, July 2020. doi: 10.3389/fncom.2020.00063. URL https://www.frontiersin.org/journals/computational-neuroscience/ articles/10.3389/fncom.2020.00063/full.
- Peter Blouw, Xuan Choo, Eric Hunsberger, and Chris Eliasmith. Benchmarking Keyword Spotting Efficiency on Neuromorphic Hardware. In *Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop*, NICE '19, pp. 1–8, New York, NY, USA, 2019. Association for Computing Machinery. doi: 10.1145/3320288.3320304. URL https://dl.acm.org/doi/10.1145/3320288.3320304.
- Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and Stuff Classes in Context. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1209–1218, Salt Lake City, UT, USA, June 2018. IEEE. doi: 10.1109/CVPR.2018.00132. URL https://ieeexplore.ieee.org/document/8578230/.
- Nadine Chang, John A. Pyles, Austin Marcus, Abhinav Gupta, Michael J. Tarr, and Elissa M. Aminoff. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, 6(1):49, May 2019. doi: 10.1038/s41597-019-0052-3. URL https://www.nature.com/articles/s41597-019-0052-3.
- Bao Xin Chen, Raghavender Sahdev, Dekun Wu, Xing Zhao, Manos Papagelis, and John K. Tsotsos.
  Scene Classification in Indoor Environments for Robots using Context Based Word Embeddings. In 2018 IEEE International Conference of Robotics and Automation (ICRA) Workshop. arXiv, August 2019. doi: 10.48550/arXiv.1908.06422. URL http://arxiv.org/abs/1908. 06422.
- Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision Transformer Adapter for Dense Predictions. In *The Eleventh International Conference on Learning Representations*, September 2022. URL https://openreview.net/forum? id=plKu2GByCNW.
- Silvia Elena Corchs, Gianluigi Ciocca, Emanuela Bricolo, and Francesca Gasparini. Predicting Complexity Perception of Real World Images. *PLOS ONE*, 11(6):e0157986, June 2016.
   doi: 10.1371/journal.pone.0157986. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0157986.
- Jordan Dotzel, Bahaa Kotb, James Dotzel, Mohamed S. Abdelfattah, and Zhiru Zhang. Exploring the Limits of Semantic Image Compression at Micro-bits per Pixel. In *The Second Tiny Papers Track at ICLR 2024*, March 2024. URL https://openreview.net/forum?id=sfwtoH5GdD.
- Nicole Sandra-Yaffa Dumont and Chris Eliasmith. Accurate representation for spatial cognition
   using grid cells. In Annual Meeting of the Cognitive Science Society, 2020. URL https:
   //api.semanticscholar.org/CorpusID:221134847.
- Nicole Sandra-Yaffa Dumont, Jeff Orchard, and Chris Eliasmith. A model of path integration that connects neural and symbolic representation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44(44), 2022. URL https://escholarship.org/uc/item/ 3pf7f9b4.
- 538 Chris Eliasmith. How to Build a Brain: A Neural Architecture for Biological Cognition. Oxford
   539 University Press, June 2013. doi: 10.1093/acprof:oso/9780199794546.001.0001. URL https: //academic.oup.com/book/6263.

- Tinglei Feng, Yingjie Zhai, Jufeng Yang, Jie Liang, Deng-Ping Fan, Jing Zhang, Ling Shao, and Dacheng Tao. IC9600: A Benchmark Dataset for Automatic Image Complexity Assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8577–8593, July 2023. doi: 10.1109/TPAMI.2022.3232328. URL https://ieeexplore.ieee.org/ document/9999482.
- Stephen I. Gallant and Phil Culliton. Positional binding with distributed representations. In 2016 International Conference on Image, Vision and Computing (ICIVC), pp. 108–113, August 2016. doi: 10.1109/ICIVC.2016.7571282. URL https://ieeexplore.ieee.org/abstract/document/7571282.
- Richard J. Gardner, Erik Hermansen, Marius Pachitariu, Yoram Burak, Nils A. Baas, Benjamin A.
   Dunn, May-Britt Moser, and Edvard I. Moser. Toroidal topology of population activity in grid cells. *Nature*, 602:123 128, 2021. URL https://api.semanticscholar.org/
   CorpusID:232081647.
- Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge,
   and Felix A. Wichmann. Generalisation in humans and deep neural networks. In Ad *vances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.,
   URL https://proceedings.neurips.cc/paper\_files/paper/2018/
   hash/0937fb5864ed06ffb59ae5f9b5ed67a9-Abstract.html.
- Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and
   Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias
   improves accuracy and robustness. In *International Conference on Learning Representations*,
   2019. URL https://openreview.net/forum?id=Bygh9j09KX.
- Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge,
   Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and
   machine vision. In Advances in Neural Information Processing Systems, volume 34, pp. 23885–
   23899. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/
   paper/2021/hash/c8877cff22082a16395a57e97232bb6f-Abstract.html.
- Iris I. A. Groen, Edward H. Silson, and Chris I. Baker. Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160102, February 2017. doi: 10.1098/rstb.2016.0102. URL https://royalsocietypublishing.org/doi/full/10.1098/rstb.2016.0102.
- Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multidi mensional mental representations of natural objects underlying human similarity judgements. *Na- ture Human Behaviour*, 4(11):1173–1185, November 2020. doi: 10.1038/s41562-020-00951-3.
   URL https://www.nature.com/articles/s41562-020-00951-3.
- 578 Michael Hersche, Mustafa Zeqiri, Luca Benini, Abu Sebastian, and Abbas Rahimi. A Neuro-vector 579 symbolic Architecture for Solving Raven's Progressive Matrices, March 2023. URL http:
   580 //arxiv.org/abs/2203.04571.
- Mohsen Imani, Chenyu Huang, Deqian Kong, and Tajana Rosing. Hierarchical hyperdimensional computing for energy efficient classification. In *Proceedings of the 55th Annual Design Automation Conference*, DAC '18, pp. 1–6, New York, NY, USA, June 2018. Association for Computing Machinery. doi: 10.1145/3195970.3196060. URL https://dl.acm.org/doi/10.1145/3195970.3196060.
- Marcie L. King, Iris I. A. Groen, Adam Steel, Dwight J. Kravitz, and Chris I. Baker. Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, 197:368–382, August 2019. doi: 10.1016/j.neuroimage. 2019.04.079. URL https://www.sciencedirect.com/science/article/pii/s1053811919303702.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete
   Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick.
   Segment Anything, April 2023. URL http://arxiv.org/abs/2304.02643.

594 Denis Kleyko, Dmitri A. Rachkovskij, Evgeny Osipov, and Abbas Rahimi. A Survey on Hyper-595 dimensional Computing aka Vector Symbolic Architectures, Part I: Models and Data Transfor-596 mations. ACM Computing Surveys, 55(6):130:1–130:40, 2022. doi: 10.1145/3538531. URL 597 https://dl.acm.org/doi/10.1145/3538531. 598 Denis Kleyko, Dmitri Rachkovskij, Evgeny Osipov, and Abbas Rahimi. A Survey on Hyperdimensional Computing aka Vector Symbolic Architectures, Part II: Applications, Cognitive 600 Models, and Challenges. ACM Computing Surveys, 55(9):175:1–175:52, January 2023. doi: 601 10.1145/3558000. URL https://dl.acm.org/doi/10.1145/3558000. 602 603 Brent Komer, Terrence Stewart, Aaron Voelker, and Chris Eliasmith. A neural representation of con-604 tinuous space using fractional binding. In Annual Meeting of the Cognitive Science Society, July 605 2019. URL https://compneuro.uwaterloo.ca/files/publications/komer. 606 2019.pdf. 607 Iuliia Kotseruba and John K. Tsotsos. 40 years of cognitive architectures: core cognitive abilities 608 and practical applications. Artificial Intelligence Review, 53(1):17-94, January 2020. doi: 10. 609 1007/s10462-018-9646-y. URL https://doi.org/10.1007/s10462-018-9646-y. 610 611 Cameron Kyle-Davidson, Elizabeth Yue Zhou, Dirk B. Walther, Adrian G. Bors, and Karla K. 612 Evans. Characterising and dissecting human perception of scene complexity. Cognition, 613 231:105319, February 2023. doi: 10.1016/j.cognition.2022.105319. URL https://www. 614 sciencedirect.com/science/article/pii/S0010027722003080. 615 616 Shaopeng Liu, Guohui Tian, and Yuan Xu. A novel scene classification model combining ResNet 617 based transfer learning and data augmentation with a filter. *Neurocomputing*, 338:191–206, April 2019. doi: 10.1016/j.neucom.2019.01.090. URL https://www.sciencedirect.com/ 618 science/article/pii/S0925231219301833. 619 620 Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In International Con-621 ference on Learning Representations, September 2018. URL https://openreview.net/ 622 forum?id=Bkg6RiCqY7. 623 624 Thomas Lu, Aaron Voelker, Brent Komer, and Chris Eliasmith. Representing spatial relations with 625 fractional binding. In Annual Meeting of the Cognitive Science Society, July 2019. URL https: //compneuro.uwaterloo.ca/files/publications/lu.2019.pdf. 626 627 Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, and Álvaro García-Martín. 628 Semantic-aware scene recognition. Pattern Recognition, 102:107256, June 2020. doi: 10.1016/j. 629 patcog.2020.107256. URL https://www.sciencedirect.com/science/article/ 630 pii/S0031320320300613. 631 632 Florian P. Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N. Hebart. Dimensions underlying 633 the representational alignment of deep neural networks with humans, June 2024. URL http: 634 //arxiv.org/abs/2406.19087. 635 George L. Malcolm, Iris I. A. Groen, and Chris I. Baker. Making Sense of Real-World Scenes. 636 Trends in Cognitive Sciences, 20(11):843–856, November 2016. doi: 10.1016/j.tics.2016.09. 637 003. URL https://www.cell.com/trends/cognitive-sciences/abstract/ 638 S1364-6613(16)30146-2. 639 640 Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation 641 and Projection for Dimension Reduction, September 2020. URL http://arxiv.org/abs/ 642 1802.03426. 643 644 Bo Miao, Liguang Zhou, Ajmal Saeed Mian, Tin Lun Lam, and Yangsheng Xu. Object-to-Scene: Learning to Transfer Object Knowledge to Indoor Scene Recognition. In 2021 IEEE/RSJ In-645 ternational Conference on Intelligent Robots and Systems (IROS), pp. 2069–2075, September 646 2021. doi: 10.1109/IROS51168.2021.9636700. URL https://ieeexplore.ieee.org/ 647 abstract/document/9636700.

- 648 A. Mitrokhin, P. Sutor, C. Fermüller, and Y. Aloimonos. Learning sensorimotor control with neuro-649 morphic sensors: Toward hyperdimensional active perception. Science Robotics, 4(30):eaaw6736, 650 May 2019. doi: 10.1126/scirobotics.aaw6736. URL https://www.science.org/doi/ 651 abs/10.1126/scirobotics.aaw6736. 652 Anton Mitrokhin, Peter Sutor, Douglas Summers-Stay, Cornelia Fermüller, and Yiannis Aloi-653 monos. Symbolic Representation and Learning With Hyperdimensional Computing. Fron-654 tiers in Robotics and AI, 7:63, June 2020. doi: 10.3389/frobt.2020.00063. URL https: 655 //www.frontiersin.org/article/10.3389/frobt.2020.00063/full. 656 Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. SLIP: Self-supervision Meets 657 Language-Image Pre-training. In Computer Vision – ECCV 2022: 17th European Conference, 658
- Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI, pp. 529–544, Berlin, Heidelberg, 659 2022. Springer-Verlag. doi: 10.1007/978-3-031-19809-0\_30. URL https://doi.org/10. 660 1007/978-3-031-19809-0\_30. 661
- Khan Muhammad, Jamil Ahmad, and Sung Wook Baik. Early fire detection using convolutional 662 neural networks during surveillance for effective disaster management. Neurocomputing, 288:30-663 42, May 2018. doi: 10.1016/j.neucom.2017.04.083. URL https://www.sciencedirect. 664 com/science/article/pii/S0925231217319203. 665
- 666 Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. 667 Human alignment of neural network representations. In The Eleventh International Conference on Learning Representations, September 2022. URL https://openreview.net/forum? 668 id=ReDQ10UQR0X. 669
- 670 Surabhi S. Nath, Kevin Shen, Aenne Annelie Brielmann, and Peter Dayan. Simplicity in Com-671 plexity. March 2024. URL https://openreview.net/forum?id=DHvVdakpqO& 672 referrer=%5Bthe%20profile%20of%20Peter%20Dayan%5D(%2Fprofile% 673 3Fid%3D~Peter\_Dayan1).
- 674 Anwesan Pal, Carlos Nieto-Granda, and Henrik I. Christensen. DEDUCE: Diverse scEne De-675 tection methods in Unseen Challenging Environments. In 2019 IEEE/RSJ International Con-676 ference on Intelligent Robots and Systems (IROS), pp. 4198-4204, November 2019. doi: 10. 677 1109/IROS40897.2019.8968108. URL https://ieeexplore.ieee.org/document/ 678 8968108. 679
- Anna Penzkofer, Lei Shi, and Andreas Bulling. VSA4VQA: Scaling A Vector Symbolic Architec-680 ture To Visual Question Answering on Natural Images. In Proceedings of the Annual Meeting of 681 the Cognitive Science Society, volume 46, 2024. URL https://escholarship.org/uc/ 682 item/26j7v1nf. 683
  - Tony A. Plate. Holographic reduced representations. *IEEE transactions on neural networks*, 63: 623-41, 1995. URL https://api.semanticscholar.org/CorpusID:2352281.

685 686

687

688

691

692

- Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 413–420, June 2009. doi: 10.1109/CVPR.2009. 5206537. URL https://ieeexplore.ieee.org/abstract/document/5206537.
- 689 Abbas Rahimi, Pentti Kanerva, and Jan M. Rabaey. A Robust and Energy-Efficient Classifier Using 690 Brain-Inspired Hyperdimensional Computing. In Proceedings of the 2016 International Symposium on Low Power Electronics and Design, ISLPED '16, pp. 64-69, New York, NY, USA, August 2016. Association for Computing Machinery. doi: 10.1145/2934583.2934624. URL https://dl.acm.org/doi/10.1145/2934583.2934624. 694
- Santhosh Kumar Ramakrishnan, Tushar Nagarajan, Ziad Al-Halah, and Kristen Grauman. Environment Predictive Coding for Visual Navigation. In International Conference on Learning Repre-696 sentations, October 2021. URL https://openreview.net/forum?id=DBiQQYWykyy. 697
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, 699 and Alejandro F. Frangi (eds.), Medical Image Computing and Computer-Assisted Interven-700 tion – MICCAI 2015, pp. 234–241, Cham, 2015. Springer International Publishing. doi: 10.1007/978-3-319-24574-4\_28.

702	Elham Saraee, Mona Jalal, and Margrit Betke. Visual complexity analysis using deep intermediate-
703	layer features. <i>Computer Vision and Image Understanding</i> , 195:102949, June 2020. doi: 10.1016/
704	j.cviu.2020.102949. URL https://www.sciencedirect.com/science/article/
705	pii/S1077314220300333.
706 707 708 709	Hongje Seong, Junhyuk Hyun, and Euntai Kim. FOSNet: An End-to-End Trainable Deep Neural Network for Scene Recognition. <i>IEEE Access</i> , 8:82066–82077, 2020. doi: 10.1109/ACCESS. 2020.2989863. URL https://ieeexplore.ieee.org/document/9076601.
710	Chuanxin Song and Xin Ma. SRRM: Semantic Region Relation Model for Indoor Scene Recog-
711	nition. In 2023 International Joint Conference on Neural Networks (IJCNN), pp. 01–08, June
712	2023. doi: 10.1109/IJCNN54540.2023.10191605. URL https://ieeexplore.ieee.
713	org/abstract/document/10191605.
714	Chuanxin Song, Hanbo Wu, Xin Ma, and Yibin Li. Semantic-embedded similarity prototype for
715	scene recognition. <i>Pattern Recognition</i> , 155:110725, November 2024. doi: 10.1016/j.patcog.
716	2024.110725. URL https://www.sciencedirect.com/science/article/pii/
717	S003132032400476X.
718 719 720 721	Dustin E. Stansbury, Thomas Naselaris, and Jack L. Gallant. Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. <i>Neuron</i> , 79(5):1025–1034, September 2013. doi: 10.1016/j.neuron.2013.06.034. URL https://www.sciencedirect.com/science/article/pii/S0896627313005503.
722 723 724 725	Terrence Stewart, Feng-Xuan Choo, and Chris Eliasmith. Spaun: A Perception-Cognition-Action Model Using Spiking Neurons. <i>Proceedings of the Annual Meeting of the Cognitive Science</i> <i>Society</i> , 34(34), 2012. URL https://escholarship.org/uc/item/168466tf.
726	Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim,
727	Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M.
728	Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qi-
729	uyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia
730	Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller,
731	Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment, Novem-
732	ber 2023. URL http://arxiv.org/abs/2310.13018.
733 734	Aaron R. Voelker. A short letter on the dot product between rotated Fourier transforms, July 2020. URL http://arxiv.org/abs/2007.13462.
735	Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database:
736	Large-scale scene recognition from abbey to zoo. In 2010 IEEE Computer Society Conference on
737	Computer Vision and Pattern Recognition, pp. 3485–3492, June 2010. doi: 10.1109/CVPR.2010.
738	5539970. URL https://ieeexplore.ieee.org/abstract/document/5539970.
739	Lin Xie, Feifei Lee, Li Liu, Koji Kotani, and Qiu Chen. Scene recognition: A comprehensive survey.
740	<i>Pattern Recognition</i> , 102:107205, June 2020. doi: 10.1016/j.patcog.2020.107205. URL https:
741	//www.sciencedirect.com/science/article/pii/S003132032030011X.
742 743 744 745 746	Karmesh Yadav, Ram Ramrakhya, Arjun Majumdar, Vincent-Pierre Berges, Sachit Kuhar, Dhruv Batra, Alexei Baevski, and Oleksandr Maksymets. Offline Visual Representation Learning for Embodied Navigation. In <i>Workshop on Reincarnating Reinforcement Learning at ICLR 2023</i> , March 2023. URL https://openreview.net/forum?id=Spfbts_vNY.
747	Delu Zeng, Minyu Liao, M. Tavakolian, Yulan Guo, Bolei Zhou, D. Hu, M. Pietikäinen,
748	and Li Liu. Deep Learning for Scene Classification: A Survey. ArXiv,
749	January 2021. URL https://www.semanticscholar.org/paper/
750	Deep-Learning-for-Scene-Classification%3A-A-Survey-Zeng-Liao/
751	3c3840c188518b80e53ce3f2f5cddf26b0f66a28.
752	Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing
753	Text-to-Image Diffusion Models for Visual Perception. In 2023 IEEE/CVF International Con-
754	ference on Computer Vision (ICCV), pp. 5706–5716, Paris, France, October 2023. IEEE. doi:
755	10.1109/ICCV51070.2023.00527. URL https://ieeexplore.ieee.org/document/

10377753/.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Tor-Scene Parsing Through ADE20K Dataset. In Proceedings of the IEEE Conralba. ference on Computer Vision and Pattern Recognition, pp. 633-641, 2017. URL https://openaccess.thecvf.com/content\_cvpr\_2017/html/Zhou\_Scene\_ Parsing\_Through\_CVPR\_2017\_paper.html. Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6):1452–1464, June 2018. doi: 10.1109/TPAMI.2017.2723009. URL https://ieeexplore.ieee.org/abstract/document/7968387. Liguang Zhou, Yuhongze Zhou, Xiaonan Qi, Junjie Hu, Tin Lun Lam, and Yangsheng Xu. Attentional Graph Convolutional Network for Structure-Aware Audiovisual Scene Classifica-tion. IEEE Transactions on Instrumentation and Measurement, 72:1-15, 2023. doi: 10.1109/ TIM.2023.3260282. URL https://ieeexplore.ieee.org/abstract/document/ 10078844. Zhuofan Zong, Guanglu Song, and Yu Liu. DETRs with Collaborative Hybrid Assignments Training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6748-6758, 2023. URL https://openaccess.thecvf.com//content/ICCV2023/ html/Zong\_DETRs\_with\_Collaborative\_Hybrid\_Assignments\_Training\_ ICCV\_2023\_paper.html. 

#### 810 A APPENDIX



Figure 5: Lengthscale  $\lambda$  and threshold  $\tau$  fine-tuning on 50 samples of COCOStuff (Caesar et al., 2018) dataset.

For fine-tuning the lengthscale  $\lambda$  parameter, we performed a grid search on 50 samples from the COCOStuff dataset (Caesar et al., 2018). Results are presented in Figure 5. 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 our predicted mask, i.e. taking all points above the threshold  $\tau$  in the similarity map after unbinding with the object's inverse SSP. We also optimised  $\tau$  (bottom of y-axis) and marked the best configuration with a star. Overall, the IoU accuracy increases with an increase in SSP dimensions, from 1945 (top) to 3751 dimensions (bottom). This is due to the increased capacity of higher dimensional vectors. 





Figure 7: Comparison of images and generated segmentation maps between VISC (Kyle-Davidson et al., 2023) and Savoias (Saraee et al., 2020) three classes: art, scenes, and interior design.