Human and Deep Neural Network Alignment in Navigational Affordance Perception

Clemens G. Bartnik
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
c.g.bartnik@uva.nl

Iris I.A. Groen
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
i.i.a.groen@uva.nl

Abstract

Moving through the world requires extracting navigational affordances from the immediate visual environment. How do humans compute this information from visual inputs? Over the last decade, Deep Neural Networks (DNNs) trained on visual recognition tasks have been shown to predict human perception remarkably well in the domain of object recognition, but their alignment with humans in other visual task contexts, such as spatial navigation, remain less understood. Here, we investigated the alignment of DNNs with human-perceived navigational affordances in a broad variety of visual environments by using explainable AI and different model training objectives. We curated a diverse set of naturalistic real-world indoor, outdoor man-made, and outdoor natural scenes. For each scene, we gathered human annotations identifying the objects present and collected drawings of path trajectories that participants would take through each scene. Quantitative analysis of the path annotations highlights that participants perceive and choose similar paths in each environment and thus diagnostic features for navigational affordances are present in the images. Using representational similarity analysis, we discovered that DNN features exhibit low correlations with information relevant to navigational affordance, such as mean pathways and floor segmentation. They demonstrate slightly better correlations with estimated depth information. However, these correlations are substantially lower than with the representational space of the contained objects. These findings illustrate that DNNs represent object information rather than representations of navigational affordances. This highlights that our path annotations are a rich and challenging benchmark to study human-DNN alignment and that current commonly used DNNs are yet not capturing navigational affordance representations well.

1 Introduction

Humans effortlessly perceive their environment as a coherent whole at a remarkable speed (Potter, 1975). Our perception extends beyond mere object recognition, encompassing a rich understanding of the surrounding scene which facilitates a variety of behavioral goals. Human psychophysical research over the years, including studies by Biederman et al. (1982) and Bar & Ullman (1996), has shown that scene perception involves both local properties, such as objects and their relationships, and global properties, including fixed immovable elements such as walls and sky. Perception of global properties is evident from humans' ability to infer scene meaning from unspecified blobs in specific spatial arrangements (Schyns & Oliva, 1994). Such perception of the 'gist' of a scene seems to be driven by visual properties like openness, depth, and navigability that are perceived before or in parallel with object information (Greene & Oliva, 2009). These findings inspired influential early computational models of object and scene perception such as HMAX (Poggio & Serre, 2013) and the GIST model (Oliva & Torralba, 2001).

While the representations of these models are fully understandable as they are constructed from 'handcrafted' features, inspired by hallmarks of the visual cortex, DNNs greatly outperform them in computer vision benchmarks (Krizhevsky et al., 2012; Kietzmann et al., 2019). Using supervised or unsupervised training regimes on large data sets DNNs learn rich feature representations optimized for the task at hand. While deep learning models now achieve human-level performance at many visual tasks, it is not a necessity that their learned representations should resemble those of humans. Therefore, a new field of research has evolved studying potential determinants of representational alignment between humans and machine learning systems (Sucholutsky et al., 2023). So far, most studies focused on alignment of human behavior in the context of object or scene categorization tasks (Muttenthaler et al., 2023; Peterson et al., 2017; King et al., 2019; Groen et al., 2018).

Here, we delve into the alignment of humans and DNNs in a different domain: human navigational affordance perception, aiming to identify visual features that are important for spatial navigation. To achieve this, we collected human path drawings on a large set of 231 diverse scene pictures as well as ratings of contained objects. By comparing these representations with those of DNNs trained with different task objectives and with explainable AI (XAI) extracted representations, we show that commonly used networks align well with the annotations of contained objects but fail to capture navigational affordance perception. Specifically, our contributions are as follows:

- We collected path drawings for a diverse set of scenes containing indoor, outdoor natural and outdoor man-made environments. Quantitative analysis using Fréchet Distance highlights that humans consistently annotate possible pathways and thus that diagnostic features for navigational affordances are detectable.
- We confirm findings from previous research indicating that DNNs are well aligned with object ratings. However, we also demonstrate that the representational space of average path trajectories is not well captured by commonly used DNNs including various architectures, training sets, and task objectives.
- We test other DNN-derived feature importance visualization maps and show that explicit spatial representation of potentially navigational relevant information, such as floor segmentations, are not well captured, but that estimated depth information is better captured by the set of DNNs we tested here.

Overall, our data provide a new, rich, and interesting benchmark that captures human navigational affordances perception, but it is not yet well aligned with commonly used DNN representations.

2 Related Work

About a decade ago, DNNs started to reach human-level performances in object recognition (Krizhevsky et al., 2012) whilst also outperforming shallow computer vision models in predicting neural responses in intermediate and higher areas of the ventral stream of the visual cortex (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014) across various modalities including electrophysiology (Cadieu et al., 2014), fMRI (Guclu & van Gerven, 2015), MEG (Cichy et al., 2016), and EEG (Greene & Hansen, 2018) making them promising computational models to understand the human visual system (see e.g., Kietzmann et al., 2019; Storrs & Kriegeskorte, 2019; Serre, 2019, for reviews). Internal representations of DNNs also have been aligned with human perceptual assessments of objects, revealing that they mirror the perception of object similarity observed in both monkeys (Rajalingham et al., 2018) and humans (Kubilius et al., 2016). While most prior research focused on object recognition, DNNs also exhibit accurate predictions of neural responses during scene (Cichy et al., 2017; Greene & Hansen, 2018; King et al., 2019) and action recognition (Güçlü & Gerven, 2017). However, these studies primarily focused on basic object or scene categorization tasks.

How can we incorporate more challenging behavior, specifically spatial navigation behavior, in the study of alignment of DNNs and human representations, in order to better understand the visual system and improve computational models? Bonner & Epstein (2017) introduced path drawings of indoor environments as a suitable representation to uncover where in the brain navigational affordances are processed. Follow-up work by Bonner & Epstein (2018) showed that mid-level layer activations of a scene-trained AlexNet show the highest alignment with navigational affordance representations, and determined that high spatial frequencies and cardinal orientations in the lower visual field are mainly driving these activations. Dwivedi et al. (2021) furthermore showed that representations of scene parsing networks better explain the same set of brain activity recordings in scene-selective regions over scene classification-trained models.

3 METHODS

3.1 STIMULI

We collected a set of 231 high-resolution color photographs (resolution: 1024×1024 pixels) from Flickr, to obtain a novel set of images that are not part of any commonly used large-scale image database (e.g. ImageNet or Places). Images depicted typical everyday environments devoid of prominent objects, humans, or animals and were captured from a human-scale, eye-level perspective (see Supplementary Figure 5 for examples). Diverging from previous research that predominantly utilized indoor imagery (e.g., Bonner & Epstein, 2017; Zamir et al., 2018), our study uses scenes from three distinct types of environments (indoor, outdoor natural, and outdoor man-made) a common taxonomy used in scene classification datasets (Zhou et al., 2018).

3.2 COLLECTING BEHAVIORAL RATINGS

3.2.1 PROCEDURE

We conducted an online experiment designed using the Gorilla Experiment Builder (Anwyl-Irvine et al., 2020) to quantify navigational affordances of our stimuli set gathering possible path trajectories similar to the paradigm used by (Bonner & Epstein, 2017). Initially, 167 participants were recruited from prolific.ac (Palan & Schitter, 2018), of whom 81 were female. 15 participants did not complete the study, resulting in a final count of 152 participants. The participants had a median age of 25 years old (range 18 - 74 years, *SD* = 14). We used stringent selection criteria provided by Prolific to screen for reliable participants. All participants had normal or corrected-to-normal vision. Participants were compensated with 7.5 £ . The experiment began with an introductory screen that informed participants about the experiment, and a screen where they gave informed consent to participate in our study. The study was approved by the ethical committee of our institution.

3.2.2 OBJECT RATINGS

To gather annotations of the presence of certain objects in the scene we had the participants view each of our 231 images for 1 second followed by a response screen with six button options (Building/Wall, Tree/Plant, Road/Street, Furniture, Body of water, Rocks/Stones). This list of response options was inspired by previous research Fei-Fei et al. (2007) characterizing objects in scenes.

3.2.3 TASK DESCRIPTION PATH DRAWING

During the task, each trial consisted of one of our 231 images with a yellow overlay (25 x 1024 pixel) at the bottom indicating where participants should start drawing their path. Participants used their computer mouse to draw a red line indicating a path they would use to move through the depicted scene (see 1 **A** for an example drawing). Each participant indicated one possible pathway. As Gorilla only provides JPEG images of the path annotations, paths were manually redrawn in a self-written tool in Python. The paths were saved as a CSV file containing sequences of x, and y coordinates in the image space. For each image, all path sequences were resampled to match the same length of points as the longest trajectory. These were resampled to have the same length.

With (*M* = 21.56, *SD* = 0.97) path annotations per image, we obtained a set of trajectories $\{T_1, T_2, \ldots, T_m\}$. Each trajectory is represented as a sequence of points $\{(x(t_i), y(t_i))\}$ in the 2D image space, capturing the x and y coordinates at discrete time intervals t_i. These trajectories enabled us to create heatmaps by averaging the 2D path annotations for each image 1 **A**. As each participant annotated only one path, this setup provided an opportunity to examine the variability and similarity between the individual paths.

3.2.4 QUANTIFYING NAVIGATIONAL AFFORDANCES

To perform a quantitative comparison of these trajectories, we utilized the discrete Fréchet Distance (Eiter & Mannila, 1994). For two trajectories T_a and T_b, the Fréchet Distance $F(T_a, T_b)$ is expressed as:

$$F(T_a, T_b) = \inf_{\alpha,\beta} \max_{t\in[0,1]} \{dt(T_a(\alpha(t)), T_b(\beta(t))) \quad (1)$$

where α and β are continuous non-decreasing functions mapping the interval $[0, 1]$ to the domains of T_a and T_b, respectively, and d represents the Euclidean distance. This metric is intuitively understood as the minimum leash length necessary for a person to walk a dog along two separate paths (the trajectories) from start to end without backtracking.

Subsequently, we computed the average Fréchet Distance across all possible pairs of trajectories for each image, providing insights into the degree of agreement regarding the most apparent and natural pathway through each scene. To put the mean Fréchet Distances into perspective we also created a distribution of mean Fréchet Distance representing randomly assigned path trajectories. This involved randomly allocating 20 pathways to each image from the entire collection of path annotations gathered during our online experiment. Subsequently, we calculated the pairwise Fréchet Distances using the previously described method. To test if the random pathways exhibit a higher average Fréchet Distance, we conducted a t-test to determine if the two distributions of average average Fréchet Distances are significantly different from each other.

There are various ways to transform path annotations into a navigational affordance feature space. Bonner & Epstein (2017) adopted an approach where they classified their path annotations based on whether they appeared on the left, center, or right side of the image. Here we utilized a different approach dividing each image into segments of 20x20 pixels and averaging the path information within each segment. This method offers the advantage of a more detailed representation where fine grained spatial information of navigational affordances are preserved. Additionally, with this approach alignment can be computed with any form of heatmap (e.g. XAI depictions or other visual hypotheses) by simply vectorizing the image.

3.3 Vision Models

We considered a diverse set of commonly used pre-trained neural networks with different architectures, trained on different datasets. We used ResNet50 (He et al., 2015), AlexNet (Krizhevsky et al., 2012), VGG16 (Simonyan & Zisserman, 2014) as vanilla CNN architectures trained on ImageNet-1K (Deng et al., 2009). To examine a possible influence of the dataset used for training the network, we also included a ResNet50 trained for scene classification on Places365 (Zhou et al., 2018) and a R50-FPN (Wu et al., 2019) trained for Scene parsing on ADE20k (Zhou et al., 2017), as well as two action recognition models from PySlowFast (Fan et al., 2020) with X3D m and SlowFast using a ResNet101 backbone. Lastly, we used three CLIP models from OpenCLIP (Radford et al., 2021) with ResNet101, VIT-B-16, and VIT-B-32 backbones. Feature activations of our 231 images were extracted from the pre-trained networks using the Net2Brain toolbox (Bersch et al., 2022). For each network, we extracted activations from the Net2Brain predefined layers. Extracted features were standardized by removing the mean and scaling to unit variance.

3.4 Image derived visual feature visualizations

To better understand which visual information is important for navigational affordances we explored various DNN-derived feature spaces that we hypothesize to contain navigational affordance related information.

3.4.1 Floor Segmentation

An intuitive approach to identifying navigational affordances involves focusing on the ground plane of a scene, as it naturally serves as the foundation for pathways. Unlike the method proposed by Bonner & Epstein (2017), in which participants indicate all possible pathways that they perceive which estimates the full navigable area, our single path annotations focus on the most likely path therefore being only a subregion in the floor segmentation. To estimate the ground plane in our images, we used a Cascade Segmentation Module trained on the ADE20k dataset (Zhou et al., 2018), which segments various elements of the scene. Subsequently, we isolated and generated binary segmentation maps for labels indicative of the ground plane. These maps were then used to calculate pairwise correlations, serving as the basis for our RDMs in this representational space.

3.4.2 Monocular depth estimation

To navigate through an environment it's important to perceive open unobstructed areas. Therefore depth information might be an important driver for navigational affordance perception. To test this we utilized a state-of-the-art monocular depth estimation model (Miangoleh et al., 2021). This model generates high-resolution depth maps from single RGB images by first producing depth estimations at multiple resolutions. Afterwards merging these into a single high-resolution depth map considering content-specific details and discrepancies across different resolutions. As described above we computed the pairwise correlation between those depth maps across our images resulting in an RDM representing which scenes are dissimilar to each other in regards to depth information.

3.4.3 Layer-wise Relevance propagation

Bonner & Epstein (2018) showed that a scene-trained AlexNet represents navigational affordance representations likely by encoding the presence of high spatial frequencies, rectilinear features, and cardinal orientations. A popular explainable AI used to visualize neural network decisions is Layer-wise Relevance Propagation (LRP; Bach et al., 2015). This method backpropagates the prediction output through the layers in reverse order, attributing relevance scores to each input pixel. This process results in a fine-grained heat map highlighting the most influential pixels in the input image. As displayed in 4 these maps highlight edges and high spatial frequencies, making this a promising feature representation to test the importance of these features further. We passed our images through a VGG16 trained on Places 365 (Zhou et al., 2018) and created LRP heatmaps for each image using the investigate toolbox (Alber et al., 2019) and created RDMs using pairwise correlations of LRP heatmaps.

3.4.4 Grad-CAM

As salient objects in a scene could inform navigational affordances we utilized Gradient-weighted Class Activation Mapping (Grad-CAM; Selvaraju et al., 2020) to visualize the areas in the input image that are most important for the predictions for the respective class. We used a VGG16 trained on Places365 for scene classification (Zhou et al., 2018) and created Grad-CAM heat maps for the third layer or the convolution block five, the last feature extraction block. The pairwise correlations of the resulting heatmaps again formed the RDM for this feature space.

3.5 Representational Similarity Analysis

We utilized representational similarity analysis (RSA; Kriegeskorte, 2008) to measure the correlation between human path annotations and feature activations of neural networks and DNN-derived visual feature visualizations. For each representational space, we created representational dissimilarity matrices (RDMs) based on pairwise 1-Pearson correlation. For both the path annotations and the visual features derived from images, we transformed the averages of the 20x20 image segments into vector form to serve as feature vectors. Regarding the vision models, we utilized the standardized activations from each layer as feature vectors. For the object representational space we computed RDMs using pairwise Euclidean correlation distances from the proportion of participants that annotated the presence or possibility of a given label (e.g., proportion of participants that annotated the scene as containing a road). The Euclidean distance metric was chosen due to the sparsity of the annotation options. Alignment was measured through Spearman's rank correlation coefficient between the RDMs.

4 Results

We used path drawings to study human and machine learning systems alignment in the domain of navigational affordance perception. We aim to understand what visual features underlie human visual processing for spatial navigation and to what extent DNNs capture these features. First, we present our path annotation data and determine their reliability and consistency. Then, we show how well a variety of different DNNs with varying architecture and training datasets capture navigational affordances in the form of path drawings and object representations via annotations of contained objects. Subsequently, we explore the connection between different 2D spatial DNN-derived visual features, such as segmentation network outputs or estimated depth maps, and explainable AI feature importance maps (LRP and Grad-CAM), examining their alignment with navigational affordances. Finally, we delve deeper into how the DNN feature activations relate to these individual visual feature representations.

4.1 Path annotations in complex environments are consistent across participants

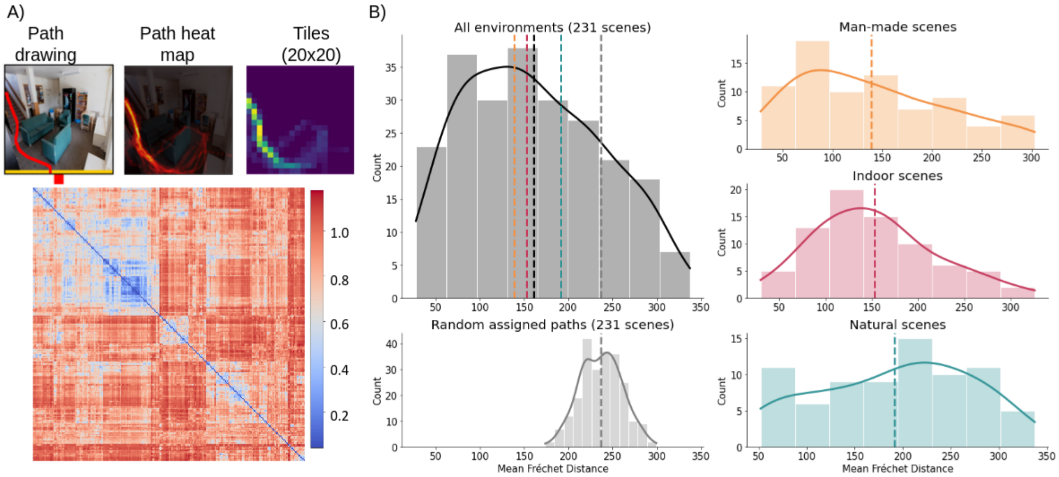

Figure 1: Capturing human navigational affordances: (A) illustrates the process of deriving an RDM from path annotations made by participants across various scenes. Path drawings were first averaged into a single heat map per scene, then split into 20x20 tiles representing the average path information in each tile (top). Subsequently, all tiles were flattened and then pairwise correlated using a 1-Pearson correlation, yielding a pairwise distance matrix reflecting the degree of overlap in navigational affordance locations between different images (bottom). (B) depicts the distribution of mean Fréchet Distances across all 231 images (left) or split by their environment label (right). Below is a distribution of mean Fréchet Distances for randomly assigned pathways. Overall we can see that the path annotations are consistent as the mean value for randomly assigned paths is substantially higher.

By using the Fréchet Distance (Eiter & Mannila, 1994) we measured how well participants agreed on the most likely path to traverse a given scene. The left panel of 1 **B** depicts a histogram of mean Fréchet Distances across our full 231 image dataset. Participants tended to agree on where to move through the scene which is reflected by the overall mean Fréchet Distances score (*M* = 167.73, *SD* = 77.63) which is significantly lower [$t(460) = 14.38$, $p < 0.001$] compared to the mean Fréchet Distances when randomizing paths across images (*M* = 236.88, *SD* = 23.43) (see 1 **B**). We also computed the path consistency for each environment type separately and found the lowest mean Fréchet Distances in outdoor man-made environments (*M* = 150.67, *SD* = 74.1), closely followed by indoor scenes (*M* = 158.23, *SD* = 71.65). These environments show no significant difference [$t(150) = 1.19$, $p = 0.23$] to each other suggesting that these types of environments have well-defined pathways that participants would choose to navigate. In natural scenes (*M* = 194.27, *SD* = 79.81) we found the highest mean Fréchet Distance which was significantly different from man-made environments [$t(150) = 4.21$, $p < 0.001$] but not significantly different from indoor environments [$t(150) = 3.28$, $p = 0.0013$], and still significantly lower than randomized paths [$t(150) = 7.64$, $p < 0.001$]. This highlights that in natural scenes on average more diverse possible pathways are perceived. However, paths are still consistently drawn in the ground plane even if somewhat more dispersed than in man-made environments.

Overall, this quantification of path annotations shows that across multiple types of environments, participants are highly consistent in annotating possible pathways, suggesting that diagnostic visual features for spatial navigation are present in the images. This makes these annotations an intriguing space to study human and machine alignment in the context of navigational affordance perception.

4.2 Alignment of DNNs with mean path and contained object annotations

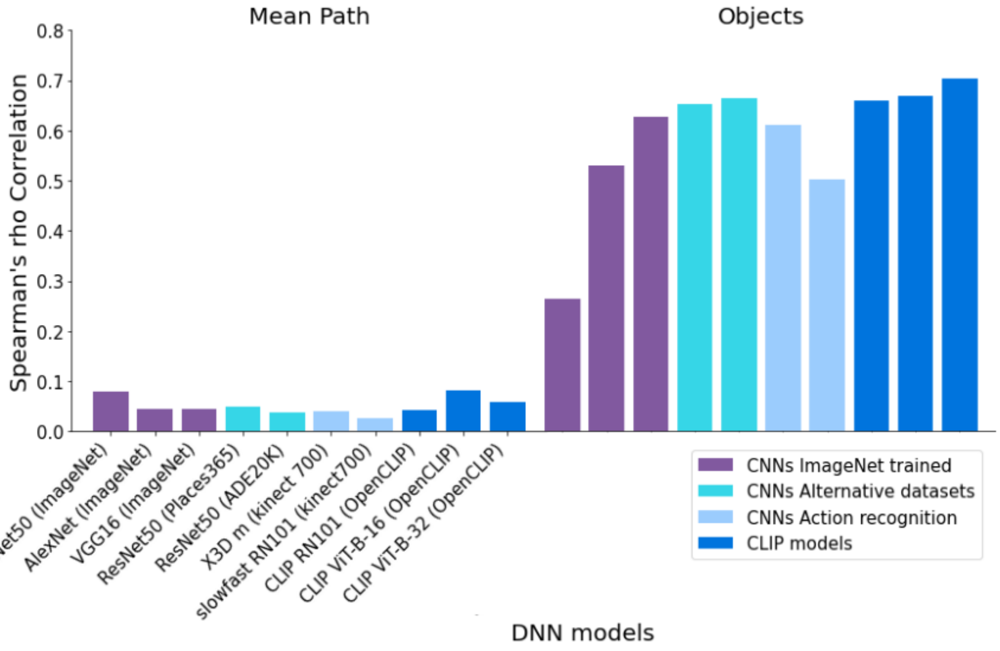

Figure 2: Comparison of DNN models best-correlating layer with mean path RDM, and object ratings in the online experiment using Spearman's rho correlation coefficient.

Next, we evaluated how well these navigational affordances in the form of path annotations are captured by feature activations of common DNN models. Hence, we used a variety of pre-trained DNNs covering different architectures, training datasets, and training regimes (see Methods). The correlations between path annotation-derived RDMs and these networks are depicted in 4.2. While the DNNs capture some information about navigational affordances overall correlations are very low. We observe the highest correction of the mean paths with a ResNet50 trained on ImageNet (*rho* = 0.08, p <0.001) and CLIP ViT-B-16 (*rho* = 0.08, p <0.001) trained on the OpenCLIP dataset. Different training sets like Places365, ADE20k, or Kinect700 seem to not improve the DNN alignment with path annotations.

In contrast, we observe that the feature activations of the DNNs correlate highly with annotations of the objects present in the scenes. Here we find the highest correlation with the CLIP ViT-B-32 (*rho* = 0.7, p <0.001) model. Interestingly the ResNet50 model trained on ImageNet exhibits the lowest correlations (*rho* = 0.26, p <0.001). Additionally, we can observe that ResNet50s are better aligned with the object representation when trained on alternative datasets like Places365 (*rho* = 0.65, p <0.001) or for scene parsing on ADE20k (*rho* = 0.66, p <0.001).

These results confirm prior research that the feature activations of DNNs are well aligned with human behavior regarding objects contained in scenes, but show that they fail to capture mean path annotations.

4.3 Comparative Analysis of Feature Spaces in Scene Representation

As we showed in the previous section feature activations of DNNs do not capture navigational affordance well. But what kind of information might be important for navigational affordances and are these better represented by DNNs? To explore this question we compute the interrelations between different DNN-derived visual feature importance maps for all our 231 images 4.3 **B**. We find the highest correlation between the mean path annotations and floor segmentation representations (*rho* = 0.24, p <0.001), which is intuitive as pathways need to be on the ground plane. Still, our path annotations rather highlight a preferred pathway, and not an annotation of the full navigable surface

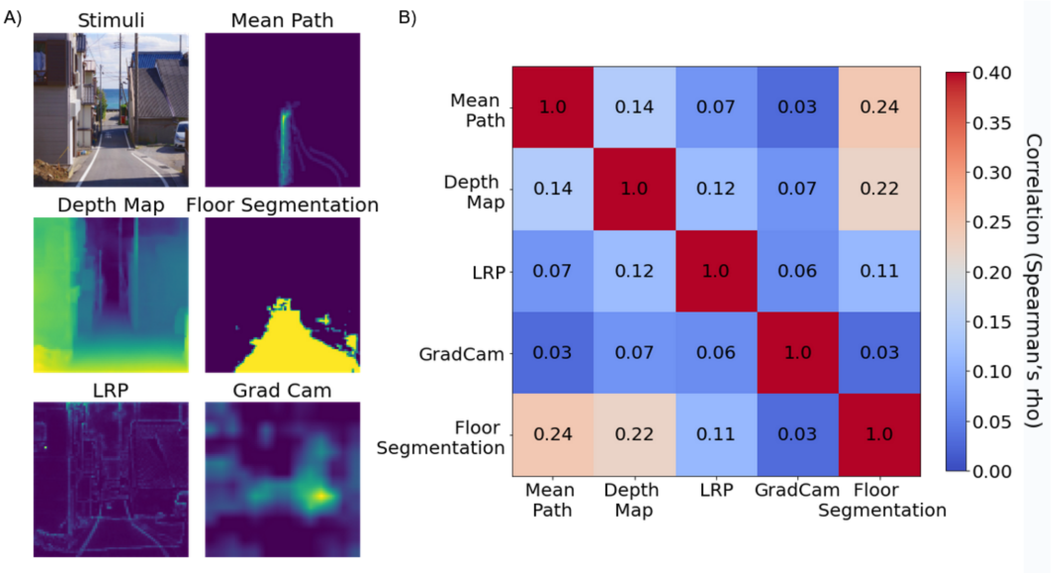

Figure 3: (A) Examples of DNN-derived spatial importance maps. The initial row presents stimuli and the average path annotation. The subsequent row showcases monocular depth estimation alongside floor segmentation maps extracted using ADE20k. The final row depicts spatial importance maps from explainable AI techniques, LRP and Grad-CAM, applied to VGG16 trained for scene classification on Places365. (B) Correlation matrix comparing DNN-derived feature importance maps with mean path annotations and with one another using Spearman's rho. The matrix displays the inter-relatedness of the various spaces (blue for lower, red for higher correlations).

area, hence the correlation is still relatively modest. The mean path also shows a notable correlation with the depth map ($\mathit{rho} = 0.14$, $p < 0.001$), suggesting that the perceived path trajectory is somewhat aligned with depth cues in the scene. However, we find a comparatively higher correlation between floor segmentation and depth estimation representations ($\mathit{rho} = 0.22$, $p < 0.001$). Besides using the ’classic’ way of aligning DNNs with human behavior through feature activations we can utilize other visual hypotheses to test DNN alignment. Here we use LRP to capture high spatial frequency information ($\mathit{rho} = 0.07$, $p < 0.001$) and Grad-CAM ($\mathit{rho} = 0.03$, $p < 0.001$). The low Grad-CAM correlation shows that the areas in the image that are most decisive for classification, in this case, the scene class, are not part of the most preferred pathways.

Overall, these correlations reveal which feature spaces are associated with navigational affordances, especially floor segmentation and depth estimation, and enable us to test how well DNNs capture these visual feature hypotheses.

Figure 4 illustrates how well DNNs capture these other visual feature spaces in comparison to the mean path annotations. DNNs demonstrate marginally higher correlations with the floor segmentation representation compared to the mean path representation. Here we find that CLIP R101 ($\mathit{rho} = 0.11$, $p < 0.001$) correlates highest. This showcases once more that DNNs don’t capture navigational affordance related representations well. While higher correlations are observed between DNN layer activations and the estimated depth map representation, they remain substantially lower than those for object information. The VGG16 model trained on ImageNet exhibits the highest correlation ($\mathit{rho} = 0.24$, $p < 0.001$) among them. This suggests that the representations captured by the DNNs are more closely associated with depth information, as opposed to navigational affordance features.

Together, these results showcase that common DNNs are not effective at capturing navigational affordances, such as mean pathways and related features like floor segmentation, and instead represent object information. Nonetheless, our approach enables a deeper analysis of how well certain features align with DNN representations, distinguishing between those that are captured and those that are not

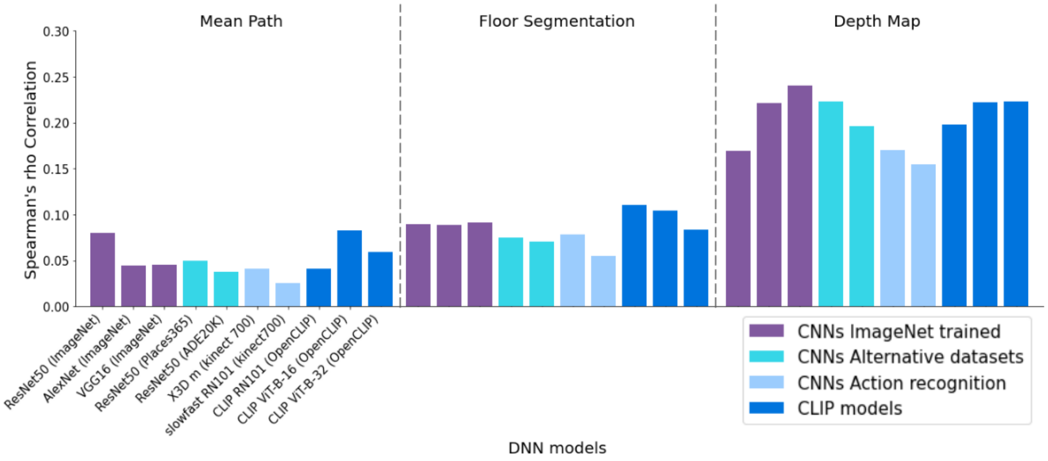

Figure 4: Comparison of DNN models best-correlating layer with Comparison mean path, floor segmentation, and estimated depth map RDMs.

5 Discussion

In this study, we investigate how well DNNs align with human spatial navigation, specifically in perceiving navigational affordances through path drawings. Our findings reveal a striking discrepancy: although participants consistently annotated pathways, these navigational representations are poorly captured by feature activations in DNNs. Instead, DNNs primarily represent object representations. Further analysis using 2D spatial visual features derived from DNNs further demonstrates their inadequacy in representing other visual features related to navigation, such as floor segmentation.

Using the Fréchet Distance as a quantitative measure, we show that humans annotate scene images in our dataset with high consistency. However, while the Fréchet Distance provides a clear method for assessing path coherence, it sometimes underestimates how coherent the paths actually are. This is particularly evident in scenarios where a scene presents two distinct yet viable paths, leading to a higher average Fréchet Distance—surprisingly, sometimes even higher than that of randomized pathways (e.g., Supplementary Figure 6). Given these limitations, future research should explore alternative measures for evaluating path consistency. This will further enhance the human navigational affordance benchmark with which alignment can be computed.

Our tiling approach involves segmenting images into tiles to analyze feature representations, providing a more rigorous assessment of alignment by specifically looking for spatially specific alignment. This approach reveals that DNNs are better at recognizing depth information rather than navigational cues. To gain a comprehensive understanding of the features DNNs represent, future research should explore a wider range of visual hypotheses

To improve the alignment of DNNs with human navigational affordance perception, future studies could consider training models using mean path heatmaps or segmentation masks. However, such an approach would necessitate a significantly larger dataset of human-annotated paths than what we have collected. An alternative, less resource-demanding strategy might involve using mean path heatmaps to refine the focus of DNNs, as suggested by Fel et al. (2022). This method would adjust the networks' attention mechanisms to prioritize areas highlighted by human annotations. Ultimately, this could enable DNNs to develop representations that more accurately reflect the characteristics humans use to perceive navigational affordances within a scene.

REFERENCES

Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. iNNvestigate Neural Networks! *Journal of Machine Learning Research*, 20(93):1–8, 2019. URL http://jmlr.org/papers/v20/18-540.html.

Alexander L. Anwyl-Irvine, Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K. Evershed. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52(1):388–407, February 2020. ISSN 1554-3528. doi: 10.3758/s13428-019-01237-x. URL http://link.springer.com/10.3758/s13428-019-01237-x.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, July 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130140. URL https://dx.plos.org/10.1371/journal.pone.0130140.

Moshe Bar and Shimon Ullman. Spatial context in recognition. *Perception*, 25(3):343–352, March 1996. ISSN 0301-0066, 1468-4233. doi: 10.1068/p250343. URL http://journals.sagepub.com/doi/10.1068/p250343.

Domenic Bersch, Kshitij Dwivedi, Martina Vilas, Radoslaw M. Cichy, and Gemma Roig. Net2Brain: A Toolbox to compare artificial vision models with human brain responses, 2022. URL https://arxiv.org/abs/2208.09677.

Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177, April 1982. ISSN 00100285. doi: 10.1016/0010-0285(82)90007-X. URL https://linkinghub.elsevier.com/retrieve/pii/001002858290007X.

Michael F. Bonner and Russell A. Epstein. Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, 114(18):4793–4798, May 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1618228114. URL https://pnas.org/doi/full/10.1073/pnas.1618228114.

Michael F. Bonner and Russell A. Epstein. Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLOS Computational Biology*, 14(4):e1006111, April 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006111. URL https://dx.plos.org/10.1371/journal.pcbi.1006111.

Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, 10(12):e1003963, December 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003963. URL https://dx.plos.org/10.1371/journal.pcbi.1003963.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):27755, September 2016. ISSN 2045-2322. doi: 10.1038/srep27755. URL http://www.nature.com/articles/srep27755.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, and Aude Oliva. Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153:346–358, June 2017. ISSN 10538119. doi: 10.1016/j.neuroimage.2016.03.063. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811916300076.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Kshitij Dwivedi, Radoslaw Martin Cichy, and Gemma Roig. Unraveling representations in scene-selective brain regions using scene-parsing deep neural networks. *Journal of Cognitive Neuroscience*, 33(10):2032–2043, September 2021. ISSN 0898-929X, 1530-8898. doi: 10.1162/jocn_a_01624. URL https://direct.mit.edu/jocn/article/33/10/2032/97376/Unraveling-Representations-in-Scene-selective.

Thomas Eiter and Heikki Mannila. Computing discrete Fréchet distance. 1994. Publisher: Technical Report CD-TR 94/64, Christian Doppler Laboratory for Expert ....

Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. PySlowFast, 2020. URL https://github.com/facebookresearch/slowfast.

Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1):10, January 2007. ISSN 1534-7362. doi: 10.1167/7.1.10. URL http://jov.arvojournals.org/article.aspx?doi=10.1167/7.1.10.

Thomas Fel, Ivan Felipe, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. 2022. doi: 10.48550/ARXIV.2211.04533. URL https://arxiv.org/abs/2211.04533. Publisher: arXiv Version Number: 2.

Michelle R. Greene and Bruce C. Hansen. Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLOS Computational Biology*, 14(7):e1006327, July 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006327. URL https://dx.plos.org/10.1371/journal.pcbi.1006327.

Michelle R. Greene and Aude Oliva. The briefest of glances: The time course of natural scene understanding. *Psychological Science*, 20(4):464–472, April 2009. ISSN 0956-7976, 1467-9280. doi: 10.1111/j.1467-9280.2009.02316.x. URL http://journals.sagepub.com/doi/10.1111/j.1467-9280.2009.02316.x.

Iris I. A. Groen, Michelle R. Greene, Christopher Baldassano, Li Fei-Fei, Diane M. Beck, and Chris I. Baker. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife*, 7:e32962, March 2018. ISSN 2050-084X. doi: 10.7554/eLife.32962.

U. Guclu and M. A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, July 2015. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.5023-14.2015. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.5023-14.2015.

Umut Güçlü and Marcel A. J. van Gerven. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*, 145:329–336, 2017. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2015.12.036. URL https://www.sciencedirect.com/science/article/pii/S1053811915011490.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. 2015. doi: 10.48550/ARXIV.1512.03385. URL https://arxiv.org/abs/1512.03385. Publisher: arXiv Version Number: 1.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, 10(11):e1003915, November 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003915. URL https://dx.plos.org/10.1371/journal.pcbi.1003915.

Tim C. Kietzmann, Patrick McClure, and Nikolaus Kriegeskorte. Deep Neural Networks in Computational Neuroscience. In *Oxford Research Encyclopedia of Neuroscience*. Oxford University Press, January 2019. ISBN 978-0-19-026408-6. doi: 10.1093/acrefore/9780190264086.013.46. URL https://oxfordre.com/neuroscience/view/10.1093/acrefore/9780190264086.001.0001/acrefore-9780190264086-e-46.

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. ISSN 10538119.

doi: 10.1016/j.neuroimage.2019.04.079. URL https://linkinghub.elsevier.com/retrieve/pii/S1053811919303702.

Nikolaus Kriegeskorte. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2008. ISSN 16625137. doi: 10.3389/neuro.06.004.2008. URL http://journal.frontiersin.org/article/10.3389/neuro.06.004.2008/abstract.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Jonas Kubilius, Stefania Bracci, and Hans P. Op De Beeck. Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLOS Computational Biology*, 12(4):e1004896, April 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004896. URL https://dx.plos.org/10.1371/journal.pcbi.1004896.

S. Mahdi H. Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yağız Aksoy. Boosting Monocular Depth Estimation Models to High-Resolution via Content-Adaptive Multi-Resolution Merging. In *Proc. CVPR*, 2021.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith. Human alignment of neural network representations, April 2023. URL http://arxiv.org/abs/2211.01201. arXiv:2211.01201 [cs, q-bio].

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001. ISSN 09205691. doi: 10.1023/A:1011139631724. URL http://link.springer.com/10.1023/A:1011139631724.

Stefan Palan and Christian Schitter. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, March 2018. ISSN 22146350. doi: 10.1016/j.jbef.2017.12.004. URL https://linkinghub.elsevier.com/retrieve/pii/S2214635017300989.

Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. 2017. doi: 10.48550/ARXIV.1706.02417. URL https://arxiv.org/abs/1706.02417. Publisher: arXiv Version Number: 3.

T. Poggio and T. Serre. Models of visual cortex. *Scholarpedia*, 8(4):3516, 2013. doi: 10.4249/scholarpedia.3516.

Mary C. Potter. Meaning in visual search. *Science*, 187(4180):965–966, March 1975. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1145183. URL https://www.sciencemag.org/lookup/doi/10.1126/science.1145183.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. 2021. doi: 10.48550/ARXIV.2103.00020. URL https://arxiv.org/abs/2103.00020. Publisher: arXiv Version Number: 1.

Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *The Journal of Neuroscience*, 38(33):7255–7269, August 2018. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0388-18.2018. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0388-18.2018.

Philippe G. Schyns and Aude Oliva. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, 5(4):195–200, July 1994. ISSN 0956-7976, 1467-9280. doi: 10.1111/j.1467-9280.1994.tb00500.x. URL http://journals.sagepub.com/doi/10.1111/j.1467-9280.1994.tb00500.x.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, February 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://link.springer.com/10.1007/s11263-019-01228-7.

Thomas Serre. Deep Learning: The Good, the Bad, and the Ugly. *Annual Review of Vision Science*, 5(1):399–426, September 2019. ISSN 2374-4642, 2374-4650. doi: 10.1146/annurev-vision-091718-014951. URL https://www.annualreviews.org/doi/10.1146/annurev-vision-091718-014951.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014. doi: 10.48550/ARXIV.1409.1556. URL https://arxiv.org/abs/1409.1556. Publisher: arXiv Version Number: 6.

Katherine R. Storrs and Nikolaus Kriegeskorte. Deep Learning for Cognitive Neuroscience. 2019. doi: 10.48550/ARXIV.1903.01458. URL https://arxiv.org/abs/1903.01458. Publisher: arXiv Version Number: 1.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. 2023. doi: 10.48550/ARXIV.2310.13018. URL https://arxiv.org/abs/2310.13018. Publisher: arXiv Version Number: 2.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2, 2019. URL https://github.com/facebookresearch/detectron2.

Amir R. Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. *CoRR*, abs/1804.08328, 2018. URL http://arxiv.org/abs/1804.08328. arXiv: 1804.08328.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5122–5130, 2017. doi: 10.1109/CVPR.2017.544.

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, 2018. doi: 10.1109/TPAMI.2017.2723009.

A Appendix

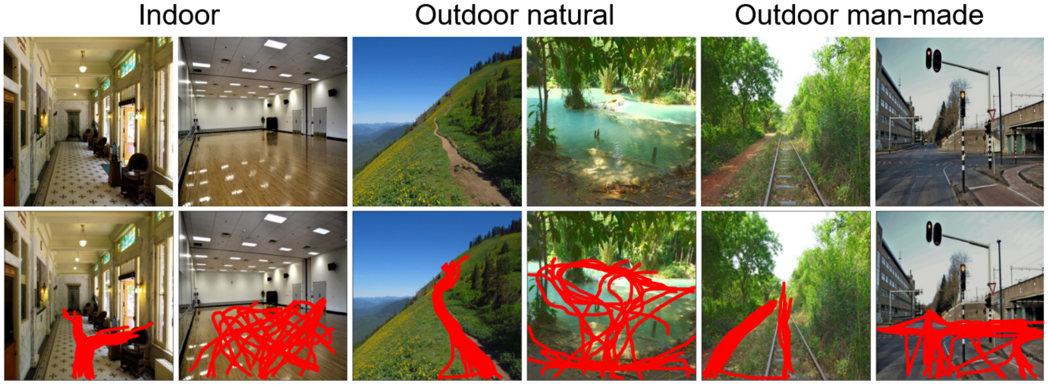

Figure 5: Example Stimuli for our three different types of environments and the respective path annotations from the online experiment. Each image was annotated by over 20 participants marking a single pathway they would use to navigate the depicted scene.

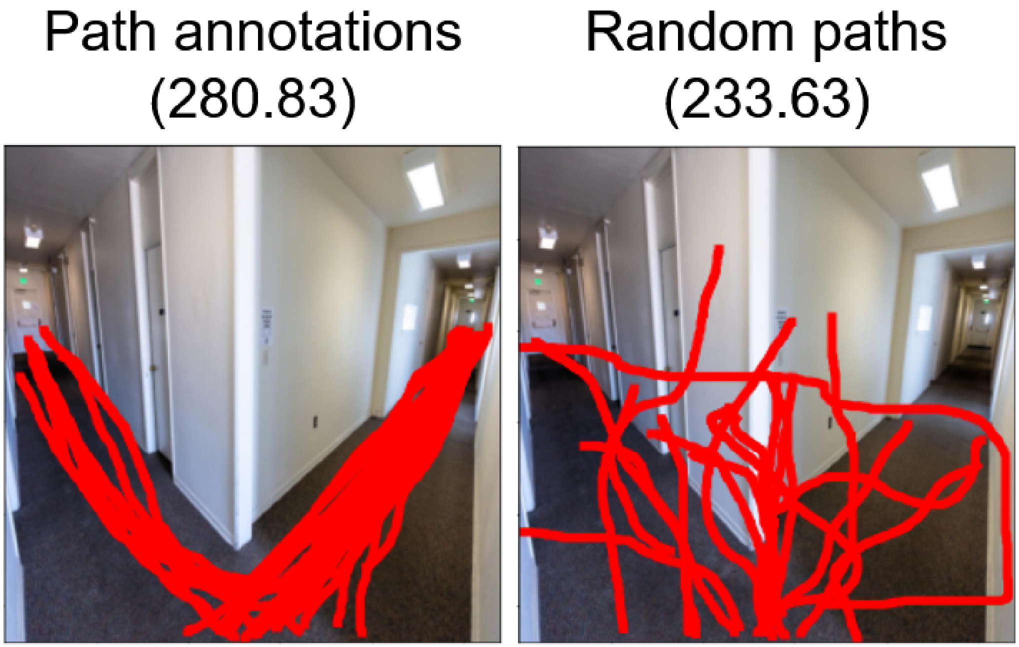

Figure 6: Fréchet Distance as a quantitative measure for path similarity underestimates the consistency of our path annotations. As there are two diverging path options the overall mean pairwise Fréchet distance between all paths is quite large (280.83) larger than for randomly placed path trajectories (233.63), yet we visually can observe that the trajectories are very similar between the participants.