Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-021-02395-x

SHORT REPORT

Meaning maps detect the removal of local semantic scene content but deep saliency models do not

Taylor R. Hayes[1] · John M. Henderson[1,2]

Accepted: 12 October 2021

Abstract

Meaning mapping uses human raters to estimate different semantic features in scenes, and has been a useful tool in demonstrating the important role semantics play in guiding attention. However, recent work has argued that meaning maps do not capture semantic content, but like deep learning models of scene attention, represent only semantically-neutral image features. In the present study, we directly tested this hypothesis using a diffeomorphic image transformation that is designed to remove the meaning of an image region while preserving its image features. Specifically, we tested whether meaning maps and three state-of-the-art deep learning models were sensitive to the loss of semantic content in this critical diffeomorphed scene region. The results were clear: meaning maps generated by human raters showed a large decrease in the diffeomorphed scene regions, while all three deep saliency models showed a moderate increase in the diffeomorphed scene regions. These results demonstrate that meaning maps reflect local semantic content in scenes while deep saliency models do something else. We conclude the meaning mapping approach is an effective tool for estimating semantic content in scenes.

Keywords Scene perception · Semantics · Image saliency · Meaning maps · Deep learning

Stored semantic knowledge gained from experience is thought to play an important role in how we guide our attention in real-world scenes (Henderson, 2007; Henderson et al., 2009). Unfortunately, estimating the role of semantic content in scenes is difficult, which has limited the study of scene semantics. To address this limitation, we recently proposed a meaning mapping approach that harnesses human raters' semantic knowledge to estimate the distribution of local semantic content across an entire scene (Henderson & Hayes, 2017; 2018). However, recent work has argued that meaning maps do not estimate local semantic content, but instead like computational deep saliency models, reflect semantically-neutral high-level image features (Pedziwiatr et al., 2021). While we believe the test used by Pedziwiatr et al. (2021) was fundamentally flawed (see Henderson et al. 2021), it is worth directly validating the assumption that meaning maps are sensitive to changes in local semantic content and are not reducible to non-semantic image features. In the present paper, we directly tested whether meaning maps and deep saliency models are sensitive to changes in local semantic content.

Cognitive guidance theory anchors our work, proposing that when visually perceiving the real world, visual-spatial attention is driven in large part by our understanding and interpretation of what we are seeing, along with what we are trying to accomplish (Henderson, 2003; 2007; 2011). That is, attention is driven by semantic representations. The evidence supporting this general idea has a long history in visual cognition (Buswell, 1935; Yarbus, 1967) and is supported by a growing body of behavioral and neural evidence that attention in scenes is strongly influenced by semantic content, which often overrides physical properties in the control of attention (Einhäuser et al., 2008; Tatler et al., 2011; Torralba et al., 2006; Henderson et al., 2020; Kiat et al., 2022; Hayes & Henderson, 2021b). This evidence has been observed in traditional attention paradigms (Malcolm et al., 2016; Shomstein et al., 2019), in simplified object displays (Nuthmann et al., 2019), and in scene perception (Võ et al., 2019; Williams & Castelhano, 2019; Wu et al., 2014; Hwang et al., 2011; Malcolm et al., 2016; Haas et al., 2019; Hayes & Henderson, 2021b).

✉ Taylor R. Hayes
taylor.r.hayes@gmail.com

[1] Center for Mind and Brain, University of California, Davis, CA, USA

[2] Department of Psychology, University of California, Davis, CA, USA

Published online: 09 February 2022

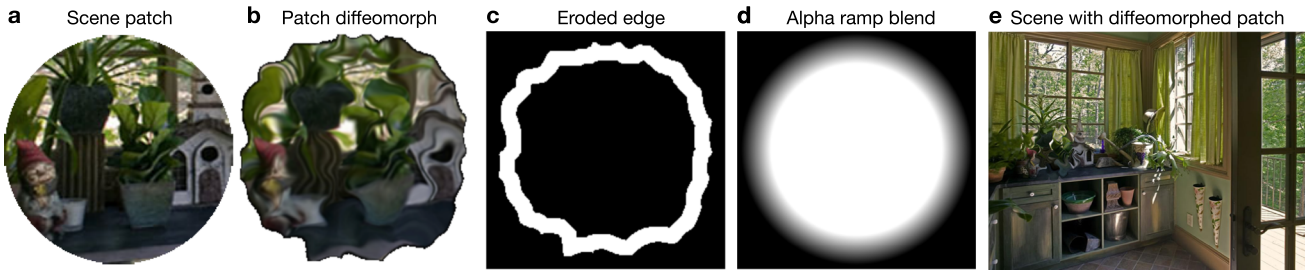

Fig. 1 *Illustration of the patch diffeomorph procedure.* Each scene had the meaning of a single patch (**a**) removed using a diffeomorphic transformation (**b**). After the patch was diffeomorphed, the patch diffeomorph's edge was eroded (**c**) and an alpha ramp (**d**) was used to smoothly blend the patch into the broader scene (**e**)

Meaning maps have served as an important tool to investigate cognitive guidance theory by examining how the spatial distribution of different types of semantic features in a scene are associated with visual attention (Henderson & Hayes, 2017; 2018; Henderson et al., 2019; Rehrig et al., 2020; Henderson et al., 2021). It is worth emphasizing that the meaning mapping approach is just a methodological tool for generating the scene distribution of different semantic features. Meaning maps are not a global model of scene attention and no meaning map is meant to simultaneously capture all aspects of scene semantics. In fact, quite the opposite. We have generally used the meaning mapping approach to target specific theoretical questions and/or to isolate specific types of semantic features. For example, the meaning map approach was initially introduced to estimate local semantic density in scenes based on how recognizable and informative scene regions are to directly compare the predictions of image guidance and cognitive guidance theories of scene attention (Henderson & Hayes, 2017; 2018). And in subsequent work, we have used the meaning mapping approach to evaluate other types of semantic features in scenes that are thought to guide attention such as graspability and reachability (Rehrig et al., 2020; Henderson et al., 2019). This subsequent work highlights both the flexibility of the meaning mapping method to study different isolated semantic features, and the principle that the original local semantic density meaning maps are only one semantic lens among many others. Therefore, the semantic 'meaning' represented by a particular type of meaning map is always explicitly defined by the rating instructions in each study (i.e., informative, recognizable, graspable, reachable, etc.). What these high-level semantic features have in common is that they primarily rely on stored semantic representations of objects and scene categories.

Given this theoretical and empirical context, it was recently proposed that the meaning maps from Henderson and Hayes (2017) based on informative and recognizable rating instructions do not estimate local semantic content, but rather 'semantically neutral high-level features' (Pedziwiatr et al., 2021). The basic argument offered by Pedziwiatr et al. (2021) is that since a deep learning model without semantics produces similar or better eye movement prediction than the original meaning maps both generally, and specifically for object-scene semantic inconsistency, meaning maps must not represent semantic content. There are a number of logical problems with this argument (see Henderson et al. 2021) and it is not entirely clear what is meant by 'semantically neutral high-level features'[1]. What is clear from Pedziwiatr et al. (2021), is that in their view 'semantically neutral high-level features' are distinct from semantic content and that neither meaning maps nor deep learning models directly reflect semantic content.

Therefore, the current experiment focused on this clear distinction centered around semantic content by experimentally testing whether meaning maps and/or deep saliency models reflect local semantic content. To test this idea, we created scene images in which local semantic content was eliminated in a circumscribed, local scene region (Fig. 1). Semantic information was locally eliminated using a diffeomorphic transformation. The diffeomorphic transformation was designed to preserve "the basic perceptual properties of the image while removing meaning" providing an ideal test for our question of interest by serving as an adversarial image (Stojanoski & Cusack, 2014). If humans rate the patches that are used to create meaning maps on the basis of visual rather than semantic features, as has been proposed (Pedziwiatr et al., 2021), then the regions without semantic content should not be rated lower than those with semantic content. That is, if meaning maps represent only semantically neutral image features, then loss of local meaning but preservation of local image features should have little effect on them. Furthermore, if meaning maps and deep saliency models both reflect the

[1] Our best guess is they are referring to the features captured in the late layers of the object recognition models like VGG-16 and VGG-19 that feed into deep learning models.

same type of non-semantic scene content, then elimination of semantics should affect them both in the same way. On the other hand, if meaning maps represent the semantic content that crowd-sourced workers were asked to rate, then eliminating that content should lead to a large decrease in rated meaning and a resulting drop in represented meaning at that location in the generated meaning maps, whereas deep saliency models based on high-level visual features should not show a reduction in salience value at those same map locations.

Method

Participants

University of California, Davis undergraduate students (N=164) with normal or corrected-to-normal vision participated in the meaning rating study in exchange for course credit. All participants were naive concerning the purposes of the experiment and provided verbal or written informed consent as approved by the University of California, Davis Institutional Review Board.

Stimuli

The stimuli were 40 real-world scene images from Henderson and Hayes (2017). These 40 scene stimuli (1024 x 768 pixels) were altered by transforming one circular region (205 pixel diameter) to remove its meaning while preserving its image features (see Diffeomorphic Transformation section below for details). Therefore, the full scene set contained 80 scene images total: the 40 original scenes from Henderson and Hayes (2017) and the same 40 scenes with a single diffeomorphed region.

Diffeomorphic transformation procedure

To remove the semantic content of a scene region while preserving its image properties, we applied a diffeomorphic transformation. The diffeomorphic transformation has been shown to remove image meaning while preserving the image properties better than phase scrambling, box scrambling, or texture scrambling (Stojanoski & Cusack, 2014). We applied the diffeomorphic transformation from Stojanoski and Cusack (2014) using the default parameters (i.e., max distortion=15, step number=10) to one circular region (205 pixel diameter, 4.2% of scene) in each of the 40 scenes. We then blended the diffeomorphed patch into the scene using a radially symmetric linear alpha ramp (15 pixel diameter) centered on the eroded edge of the patch. This allowed us to seamlessly blend each diffeomorphed patch into its broader scene (See Fig. 1).

Patch selection

The original scene meaning maps (Henderson & Hayes, 2017) and GBVS image saliency maps (Harel et al., 2006) were used to identify the circular region (205 pixel diameter) in each scene that was both high in meaning and low in image salience, since our focus was manipulating meaning while minimizing the changes to the underlying image features. We identified patches that were high in meaning and low in image salience using a simple weighted sorting algorithm. The sorting algorithm computed the average meaning value and GBVS image salience value for each of the 108 candidate coarse patches in each scene. The 108 patches for given scene were then sorted with a weight of 0.8 for their meaning value and a weight of 0.2 for their image salience value to ensure a high-meaning region was selected. The sixth item in the sorted list was selected as it had the best tradeoff of high meaning values and lower image salience values across the 40 scenes. This procedure was repeated for each scene to select the diffeomorphed region in each scene.

Meaning maps

Meaning maps were then generated for each diffeomorphed scene using the same meaning mapping procedure and same instructions as Henderson and Hayes (2017) and Henderson and Hayes (2018) (see https://osf.io/654uh/ for the code and rating instructions and https://osf.io/ptsvm/ for the 40 scene meaning maps). Specifically, a meaning map was created for each diffeomorphed scene by cutting the entire scene into a dense array of overlapping circular patches at a fine spatial scale (300 patches, diameter=87 pixels) and coarse spatial scale (108 patches, diameter=205 pixels). This procedure resulted in 12000 fine patches and 4320 coarse patches. Raters (N=164) then provided ratings of 300 random coarse or fine scene patches based on how informative or recognizable they thought they were on a 6-point Likert scale (Henderson & Hayes, 2017; Mackworth & Morandi, 1967). Patches were presented in random order and without scene context, so ratings were based on context-independent judgments. Each unique patch was rated by three unique raters, but due to the patch overlap and multiple scales, each patch contained between 6 and 30 unique ratings (mean=18.9, median=18).

A meaning map (Fig. 2) was generated for each diffeomorphed scene by averaging the patch rating data at each spatial scale separately, then averaging the spatial scale maps together, and then smoothing the grand average rating map with a Gaussian filter (i.e., Matlab 'imgaussfilt' with $\sigma = 10$, FWHM=23 pixels). These new diffeomorphed meaning maps were then compared to the original meaning maps that did not contain a diffeomorphed region (Henderson & Hayes 2017, https://osf.io/ptsvm/).

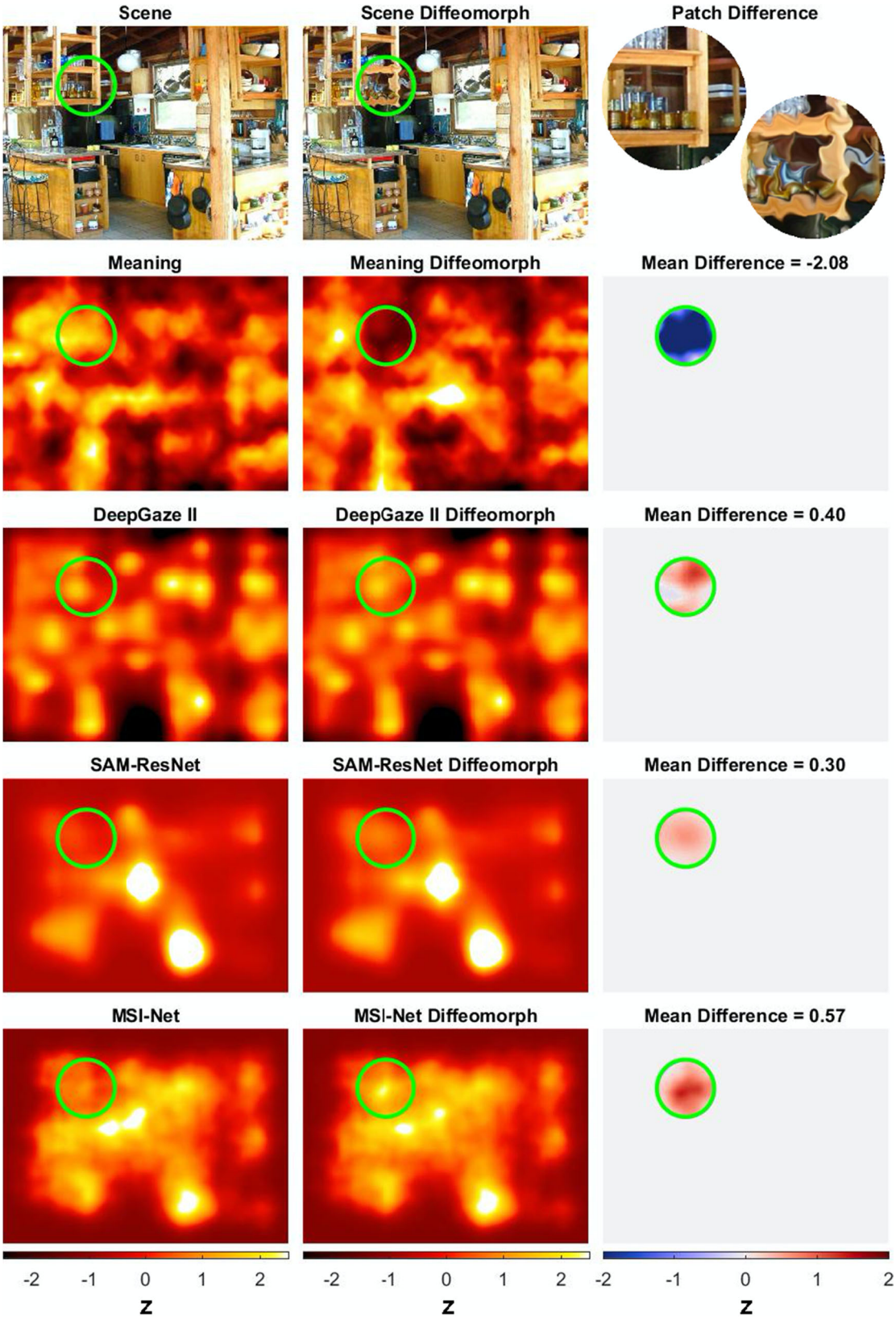

Fig. 2 *Example scene, scene diffeomorph, patch difference, and meaning and deep saliency model map critical patch comparisons.* Each scene had a region that had its meaning removed using a diffeomorphic transformation. The diffeomorphed scene region was then used to compare how meaning maps, DeepGaze II, SAM-ResNet, and MSI-Net responded to the original patch and its diffeomorph. The right column shows the mean difference in the diffeomorphed region activity for each map type

Deep saliency models

We compared three state-of-the-art deep saliency models (Bylinskii et al., 2012): DeepGaze II (Kümmerer et al., 2016), the multi-scale information network (MSI-Net, Kroner et al. 2020), and the saliency attentive model (SAM-ResNet, Cornia et al. 2018) on the scenes with and without the diffeomorphed patches. Each deep saliency model takes a scene image as input and produces a predicted saliency map as output that reflects where the model predicts human observers will look in that scene. All of the deep saliency models contain pre-trained object recognition weights that are fixed and then subsequent network layers that are trained on human data in the form of fixation and/or mouse-contingent density maps that reflect where humans focus their attention in scenes (Deng et al., 2009; Simonyan & Zisserman, 2015; Kroner et al., 2020; Kümmerer et al., 2016; Cornia et al., 2018). The deep saliency model weights are fixed following training, and then the models are evaluated on new scenes and fixation data. DeepGaze II, MSI-Net, and SAM-ResNet each have distinct network architectures, training regimens, center bias priors, and loss functions and represent a good cross section of deep convolutional neural network models of scene attention.

Map standardization

In order to allow for comparisons across the different map types (i.e., Meaning, DeepGaze II, SAM-ResNet and MSI-Net) and conditions (original and diffeomorphed) every map was standardized to a common unit of standard deviations (i.e., mean=0 and std=1) prior to the statistical analyses.

Results

The main result comparing the diffeomorphed scene patches to the original scene patches within each model are clear (Fig. 3). The meaning maps showed a significant decrease in meaning ratings for the diffeomorphed patch relative to the original scene patch ($t(39) = -22.81$, $p_{adj} < .001$, 95% CI $[-1.60, -1.34]$). This finding is consistent with the idea that meaning maps represent local semantic content, not simply semantically neutral image features. In contrast, all the deep saliency models showed a significant increase in values for the diffeomorphed patch relative to the original scene patch (DeepGaze II, $t(39) = 5.39$, $p_{adj} < .001$, 95% CI [0.14, 0.30]; SAM-ResNet, $t(39) = 4.75$, $p_{adj} < .001$, 95% CI [0.18, 0.45]; MSI-Net, $t(39) = 5.08$, $p_{adj} < .001$, 95% CI [0.20, 0.48]; p_{adj}=Bonferroni correction). That is, the deep saliency models did not just fail to detect that the semantic content had been removed, they actually showed increased activity in the diffeomorphed patch relative to the original patch. Together these findings starkly demonstrate that meaning maps generated by human raters directly reflect local semantic content whereas deep saliency models do not.

In addition, we compared the diffeomorph effect between models using a one-way Analysis of Variance (ANOVA) on the map type difference values (Fig. 3c). The results indicated a significant difference between the map types ($F(3, 156)$=211.79, $p < .001$). A post-hoc comparison using Tukey's HSD test for multiple comparisons (FWER=0.05) indicated that meaning maps were significantly different from each deep saliency model (DeepGaze II: $p_{adj} < 0.001$, 95% CI [1.47, 1.91]; SAM-ResNet: $p_{adj} < 0.001$, 95% CI [1.56, 2.01]; MSI-Net: $p_{adj} < 0.001$, 95% CI [1.59, 2.03]). The deep saliency models were not significantly different from one another (DeepGaze II and MSI-Net: $p_{adj} = 0.48$, 95% CI $[-0.10, 0.35]$; DeepGaze II and SAM-ResNet: $p_{adj} = 0.64$, 95% CI $[-0.12, 0.32]$; MSI-Net and SAM-ResNet: $p_{adj} = 0.90$, CI $[-0.25, 0.20]$). These results indicate that the meaning map diffeomorph effect was significantly different from all the deep saliency models, while the deep saliency models exhibited comparable diffeomorph effects among themselves.

Finally, we estimated the mean difference between the original scenes and the diffeomorphed scenes for all the *non-diffeomorphed* scene regions for each map type. The results indicated the non-diffeomorphed regions were highly consistent for each map type (Meaning: mean=-0.0025, std=0.0178; DeepGaze II: mean=-0.0006, std=0.0056; SAM-ResNet: mean=0.0005, std=0.006; MSI-Net: mean=0.0001 std=0.004) and were not significantly different from zero (Meaning: $t(39) = -0.89$, $p = 0.38$, 95% CI $[-0.008, 0.003]$; DeepGaze II: $t(39) = -0.71$, $p = 0.48$, 95% CI $[-0.002, 0.001]$; SAM-ResNet: $t(39) = 0.49$, $p = 0.63$, 95% CI $[-0.002, 0.002]$; MSI-Net: $t(39) = 0.22$, $p = 0.83$, 95% CI $[-0.001, 0.001]$). These findings strongly suggest our main results are not due to variability in the different rater groups used in the two meaning map studies or any kind of anomalous behavior in the deep saliency models we tested.

Discussion

In our work, we argue that attention in real-world scenes is primarily driven by semantic representations of what we are seeing and our current goals. Under this theoretical view, image features play a role in defining potential targets for attention, but it is semantic content that determines attentional priority. Here we provided a simple demonstration that one of the tools we have been using to test this theory (i.e., meaning maps) do reflect local semantic content, while deep saliency models do not.

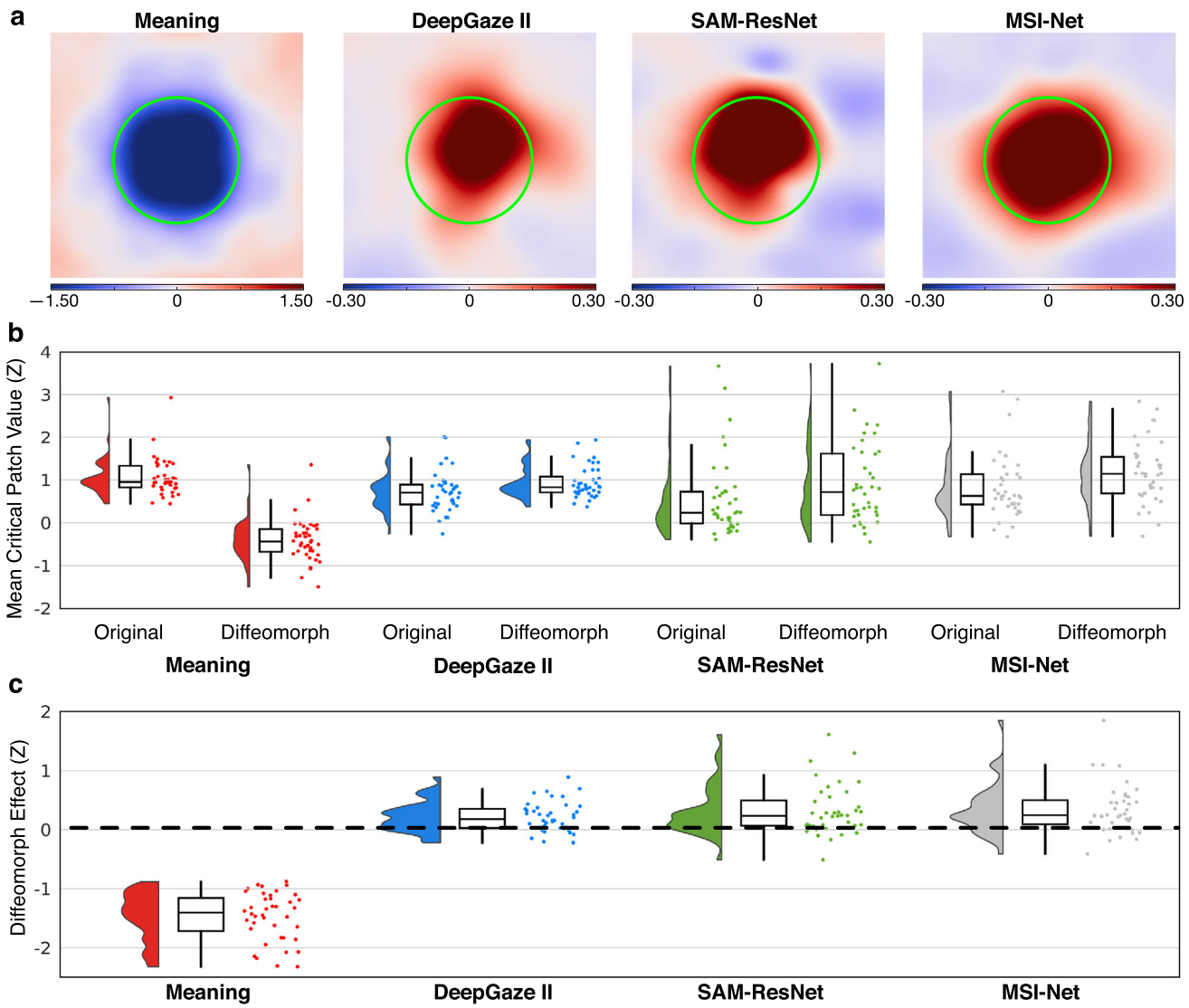

Fig. 3 *Critical patch meaning map and deep saliency model comparisons: original and diffeomorph spatial difference, distributions, and differences by map type.* Each sub-panel in panel (**a**) shows the mean difference spatially across the critical diffeomorphed scene region (indicated by neon green circle in each panel) averaged across all 40 scenes for each map type (Meaning, DeepGaze II, SAM-ResNet, and MSI-Net). Panel (**b**) shows the mean critical patch value for each scene (original vs. diffeomorph) for each map type. Panel (**c**) shows the difference in the diffeomorphed condition relative to the original scene value for each map type. Only meaning maps consistently detected the removal of semantic content from the diffeomorphed scene regions

That is, we found that meaning maps showed a large localized decrease when semantic content was removed from a scene region using a diffeomorphic transformation. In comparison, all of the deep saliency models we tested showed increased localized activity when the semantic content was removed. These findings demonstrate both that human generated meaning maps estimate local semantic content in scenes, and that meaning maps are not reducible to the features captured by deep saliency models as suggested by Pedziwiatr et al. (2021).

The present work supplements our previous work in a couple of important ways. First, while our previous work provides indirect evidence that meaning maps capture semantic content (e.g., when the rating instructions to the participants change, the meaning maps change), the current study provides direct evidence by actively manipulating local meaning while keeping the rating instructions constant. Second, the current study through its direct comparison of meaning maps and deep saliency models on the same active local meaning manipulation, is the only study that demonstrates that meaning maps are not reducible to deep saliency models. The indirect evidence we previously provided (Henderson et al., 2021) in conjunction with the direct evidence provided in current study, strongly suggests that meaning maps generated from human raters reflect semantic content.

Moving forward, it is important to keep in mind that the meaning mapping approach is not a theory, it is a tool to test theory. For example, when we first introduced meaning maps, they were used as a targeted method to generate a semantic analogue to a low-level image saliency map. This allowed us to test cognitive guidance theory against image guidance theory. That is, we used meaning maps to determine whether contrasts in local pre-semantic image features, or local semantic density, were a stronger predictor of attention across entire scenes (Henderson & Hayes, 2017). Since our original paper the meaning mapping approach has been used to test cognitive theory in a wide range of areas including scene memory (Bainbridge et al., 2019; Ramey et al., 2020), language production (Ferreira & Rehrig, 2019; Henderson et al., 2018; Rehrig et al., 2020), mind wandering (Krasich et al., 2020; Zhang et al., 2021), active vision in VR environments (Haskins et al., 2020), infant development (Klotz et al., 2021), and in the brain (Henderson et al., 2020; Kiat et al., 2022). The approach has also been extended to investigate other types of semantic features in scenes, such as reachability and graspability (Rehrig et al., 2020). These examples highlight the productive role the meaning mapping approach can play as a tool for testing cognitive theories of attention in scenes.

It is also worth repeating that meaning maps are not meant to be a global model of scene attention. That is, meaning maps are not trying to predict the maximum amount of total variance possible in looking behavior. The purpose of meaning maps is actually the opposite of a global model; meaning maps target and isolate specific semantic components that are theorized to be relevant to attention in scenes. In contrast, deep saliency models are global models of scene attention because they are directly trained on human fixation data over scenes and are optimized to account for the maximum variance possible in those data. As a result, deep saliency models learn to leverage a very broad set of different image features and regularities in human fixation data to predict attention (Hayes & Henderson, 2021a; Kümmerer et al., 2019). For this reason, to directly compare the overall prediction of deep saliency models to the isolated features of the meaning mapping approach (Pedziwiatr et al., 2021) is to miss the purpose of the meaning mapping approach entirely (Henderson et al., 2021). Succinctly, meaning maps are a flexible tool useful for testing the role of isolated components of scene semantics, which can then be used to guide explanatory theory building. They are not a global predictive model of scene attention to be benchmarked.

In summary, we used a local diffeomorphic transform to test whether meaning maps and visually-based deep saliency models would reflect the loss of local meaning. The results were clear: meaning maps generated by human raters showed a large decrease in the diffeomorphed scene regions, while all three deep saliency models showed a moderate increase in the diffeomorphed scene regions. Therefore, we conclude that meaning maps (Henderson & Hayes, 2017) estimate local semantic content and can continue to serve as a flexible tool for studying semantics in real-world scenes.

Acknowledgements This research was supported by the National Eye Institute of the National Institutes of Health, under award number R01EY027792. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors declare no competing financial interests.

Open Practices Statement Data are available from the authors upon reasonable request. The study was not preregistered.

References

Bainbridge, W. A., Hall, E. A. H., & Baker, C. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature Communications*, *10*, 1–13.

Buswell, G. T. (1935). *How people look at pictures*. Chicago: University of Chicago Press.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & et al. (2012). MIT Saliency Benchmark. http://saliency.mit.edu/.

Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). Predicting human eye fixations via an LSTM-based saliency attentive model. *IEEE Transactions on Image Processing*, *27*(10), 5142–5154.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). ImageNet: a large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, (pp. 248–255).

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2), 1–19.

Ferreira, F., & Rehrig, G. (2019). Linearisation during language production: evidence from scene meaning and saliency maps. *Language, Cognition and Neuroscience*, *34*, 1129–1139.

Haas, B. d. e., Iakovidis, A. L., Schwarzkopf, D., & Gegenfurtner, K. (2019). Individual differences in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 11687–11692.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Proceedings of the 19th international conference on neural information processing systems*, (pp. 545–552). MIT Press: Cambridge.

Haskins, A. J., Mentch, J., Botch, T. L., & Robertson, C. E. (2020). Active vision in immersive, 360° real-world environments. *Scientific Reports*, *10*(1), 14304.

Hayes, T. R., & Henderson, J. M. (2021a). Deep saliency models learn low-, mid-, and high-level features to predict scene attention. *Scientific Reports*, *11*, 1–13.

Hayes, T. R., & Henderson, J. M. (2021b). Looking for semantic similarity: what a vector space model of semantics can tell us about attention in real-world scenes. *Psychological Science*, *32*, 1262–1270.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504.

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science*, *16*, 219–222.

Henderson, J. M. (2011). Eye movements and scene perception. In Liversedge, I. S. P., Gilchrist, D., Everling, S., & Henderson, J. M. (Eds.), (pp. 593–606): Oxford University Press.

Henderson, J. M., Goold, J. E., Choi, W., & Hayes, T. R. (2020). Neural correlates of fixated low- and high-level scene properties during active scene viewing. *Journal of Cognitive Neuroscience*, *32*(10), 2013–2023.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes rereveal by meaning maps. *Nature Human Behaviour*, *1*, 743–747.

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6:10), 1–18.

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2019). Meaning and attentional guidance in scenes: a review of the meaning map approach. *Vision*, *2*(19), 1–10.

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2021). Meaning maps capture the density of local semantic features in scenes: a reply to Pedziwiatr, Kummerer, Wallis, Bethge and Teufel (2021). *Cognition*, 104742.

Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning guides attention during real-world scene description. *Scientific Reports*, *8*, 1–9.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin and Review*, *16*, 850–856.

Hwang, A. D., Wang, H. C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research*, *51*(10), 1192–1205.

Kiat, J. E., Hayes, T. R., Henderson, J. M., & Luck, S. J. (2022). Rapid extraction of the spatial distribution of physical saliency and semantic informativeness from natural scenes in the human brain. *Journal of Neuroscience*, *42*(1), 97–108.

Klotz, S., Hayes, T. R., Pomaranski, K., Henderson, J. M., & Oakes, L. (2021). Experience and age guide infants' attention to meaning in scenes. *Society for Research in Child Development*.

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). Deepgaze II: reading fixations from deep features trained on object recognition. Retrieved from arXiv:1610.01563.

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2019). Deepgaze III: using deep learning to probe interactions between scene content and scanpath history in fixation selection. *Proceedings of Cognitive Computational Neurocience* (542).

Krasich, K., Huffman, G., Faber, M., & Brockmole, J. (2020). Where the eyes wander: the relationship between mind wandering and fixation allocation to visually salient and semantically informative static scene content. *Journal of Vision*, *20*(9), 10.

Kroner, A., Senden, M., Driessens, K., & Goebel, R. (2020). Contextual encoder-decoder network for visual saliency prediction. *Neural Networks : the Official Journal of the International Neural Network Society*, *129*, 261–270.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception and Psychophysics*, *2*(11), 547–552.

Malcolm, G. L., Groen, I., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends in Cognitive Sciences*, *20*(11), 843–856.

Malcolm, G. L., Rattinger, M., & Shomstein, S. (2016). Intrusive effects of semantic information on visual selective attention. *Attention, Perception, and Psychophysics*, *78*, 2066–2078.

Nuthmann, A., Groot, F. d. e., Huettig, F., & Olivers, C. (2019). Extrafoveal attentional capture by object semantics. *PLOS ONE*, *14*, 1–19.

Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S. A., Bethge, M., & Teufel, C. (2021). Meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations. *Cognition, 206*(104465).

Ramey, M. M., Yonelinas, A., & Henderson, J. (2020). Why do we retrace our visual steps? Semantic and episodic memory in gaze reinstatement. *Learning and Memory*, *27*(7), 275–283.

Rehrig, G., Peacock, C. E., Hayes, T. R., Henderson, J., & Ferreira, F. (2020). Where the action could be: speakers look at graspable objects and meaningful scene regions when describing potential actions. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*(9), 1659–1681.

Shomstein, S., Malcolm, G., & Nah, J. (2019). Intrusive effects of task-irrelevant information on visual selective attention: semantics and size. *Current Opinion in Psychology*, *29*, 153–159.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. CoRR, arXiv:1409.1556.

Stojanoski, B., & Cusack, R. (2014). Time to wave good-bye to phase scrambling: creating controlled scrambled images using diffeomorphic transformations. *Journal of Vision*, *14*(12), 1–16.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: reinterpreting salience. *Journal of Vision*, *11*(5), 1–23.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, *113*, 766–786.

Võ, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: how scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, *29*, 205–210.

Williams, C. C., & Castelhano, M. S. (2019). The changing landscape: high-level influence on eye movement guidance in scenes. *Vision*, *3*(3), 33.

Wu, C. C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*, *5*, 1–13.

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum.

Zhang, H., Anderson, N. C., & Miller, K. F. (2021). Scene meaningfulness guides eye movements even during mind-wandering. *Attention, Perception, and Psychophysics*.