Vision Transformers are Inherently Saliency Learners (Supplementary materials)

Yasser Abdelaziz Dahou Djilali[1,2]
yasser.dahoudjilali2@mail.dcu.ie
Kevin McGuinness[1]
kevin.mcguinness@dcu.ie
Noel O'Connor[1]
Noel.OConnor@dcu.ie

[1] Dublin City University (DCU),
Dublin, Ireland.
[2] Technology Innovation Insitute,
Abu Dhabi, UAE.

A Other ViT variants

In addition to the ViT-base/16 model that was experimentally evaluated in the main paper, we also explore several other variants of the Vision Transformer (ViT) architecture. These variants include ViT-base/8, ViT-large/16, and ViT-large/32, which offer different representational capacities. These models have been pre-trained on large-scale datasets such as ImageNet [1] in a self-supervised learning paradigm, similar to the ViT-base/16 model used in the main experiments. By investigating a range of ViT variants, we aim to gain insights into the performance characteristics and generalization abilities of different ViT configurations when integrated into the proposed saliency framework. We experiment with the Raw Attention and Salieny Attention, and report the results on the Salicon validation set in Table A.1. It appears that the patch size is more important than the model size. Indeed, the ViT-S/16 surpasses the ViT-B/8 both on the Raw attention and Salient Attention. This might be explained by the fact that smaller patches result in an over estimated saliency maps because the number of total batches is higher in this case.

B Supervised vs Unsupervised self-attention maps

As shown in Figure A.1, Self-attention heads from the last layer are utilized to analyze the raw attention in both self-supervised and supervised DINO. In our investigation, we specifically examine the attention map when employing the [CLS] token. Comparing the attention maps between the two approaches, notable differences emerge. Through self-supervised learning, the attention mechanism demonstrates a superior ability to capture objects within the images. The attention maps generated by the self-supervised DINO model exhibit a remarkable level of detail and clarity. Objects of interest are highlighted with precision, enabling the model to focus on relevant regions effectively.

On the other hand, in the supervised DINO model, the attention maps appear to be relatively less refined. The supervised approach, relying on labeled data, seems to exhibit a relatively coarser representation of object attention. The attention maps produced by the

Table A.1: Comparative performance study on: Salicon. Unsupervised stands for using DINO self-supervised weights.

Models		Salicon					MIT300				
		SIM	s-AUC	CC	NSS	KLD	SIM	s-AUC	CC	NSS	KLD
ViT-S/16	Raw attention	0.72	0.72	0.80	2.20	0.60	0.64	0.71	0.70	2.08	0.69
	Salient Attention	0.78	0.73	0.84	2.42	0.41	0.66	0.75	0.76	2.39	0.52
ViT-S/8	Raw attention	0.66	0.64	0.71	1.86	0.68	0.68	0.69	0.65	1.78	0.84
	Salient Attention	0.74	0.71	0.80	2.13	0.49	0.60	0.72	0.72	2.02	0.53
ViT-B/16	Raw attention	0.73	0.71	0.81	2.24	0.61	0.63	0.72	0.71	2.06	0.71
	Salient Attention	0.78	0.74	0.86	2.41	0.42	0.65	0.74	0.76	2.38	0.51
ViT-B/8	Raw attention	0.70	0.68	0.77	2.01	0.58	0.60	0.71	0.67	2.01	0.70
	Salient Attention	0.77	0.72	0.82	2.23	0.45	0.62	0.74	0.75	2.31	0.53

supervised DINO model lack the same level of sharpness and fine-grained focus as those generated through self-supervised learning. This might be explained by the fact the the model explores more spurious correlations to solve the downstream classification task.

This disparity in attention map quality suggests that self-supervised DINO models excel in capturing the inherent structure and salient features of objects within the images. By learning representations from the data itself, the self-supervised approach uncovers meaningful patterns and relationships without explicit human annotations. Consequently, the resulting attention maps reflect a stronger alignment with the underlying objects and their contextual significance. This justifies the results of this work, where saliency is easily captured from self-sipervised DINO.

C Saliency for low-level features

As shown in Figure A.2 and Figure A.3, the attention maps produced by self-supervised transformers not only capture high-level semantic information but also provide insights into low-level visual features. The model is able to capture the intricate interplay of these low-level features, resulting in attention maps that provide a comprehensive representation of the visual content. Highlighting their remarkable capacity to capture low-level features such as color, intensity, and shape. These features, akin to the elemental constituents of the human visual system, form the foundational building blocks of human visual system. It selectively attends to relevant information, allowing for efficient processing and analysis of visual input. The captured color information aids in discerning and categorizing objects based on their distinctive hues, while intensity features contribute to edge detection and contour perception. Moreover, shape features play a crucial role in recognizing and understanding complex object structures. By attending to these low-level features, Vision Transformers gain a comprehensive understanding of visual input, reflecting the hierarchical processing observed in the human visual system. This capacity to capture and integrate low-level features aligns with the Feature Integration Theory, shedding light on the model's remarkable capabilities.

Figure A.1: Self-attention heads from the last layer for the raw attention. We look at the attention map when using the [CLS] token for self-supervised vs supervised DINO. It can be seen that the self-supervised capture the objects better and produce sharper attention maps.

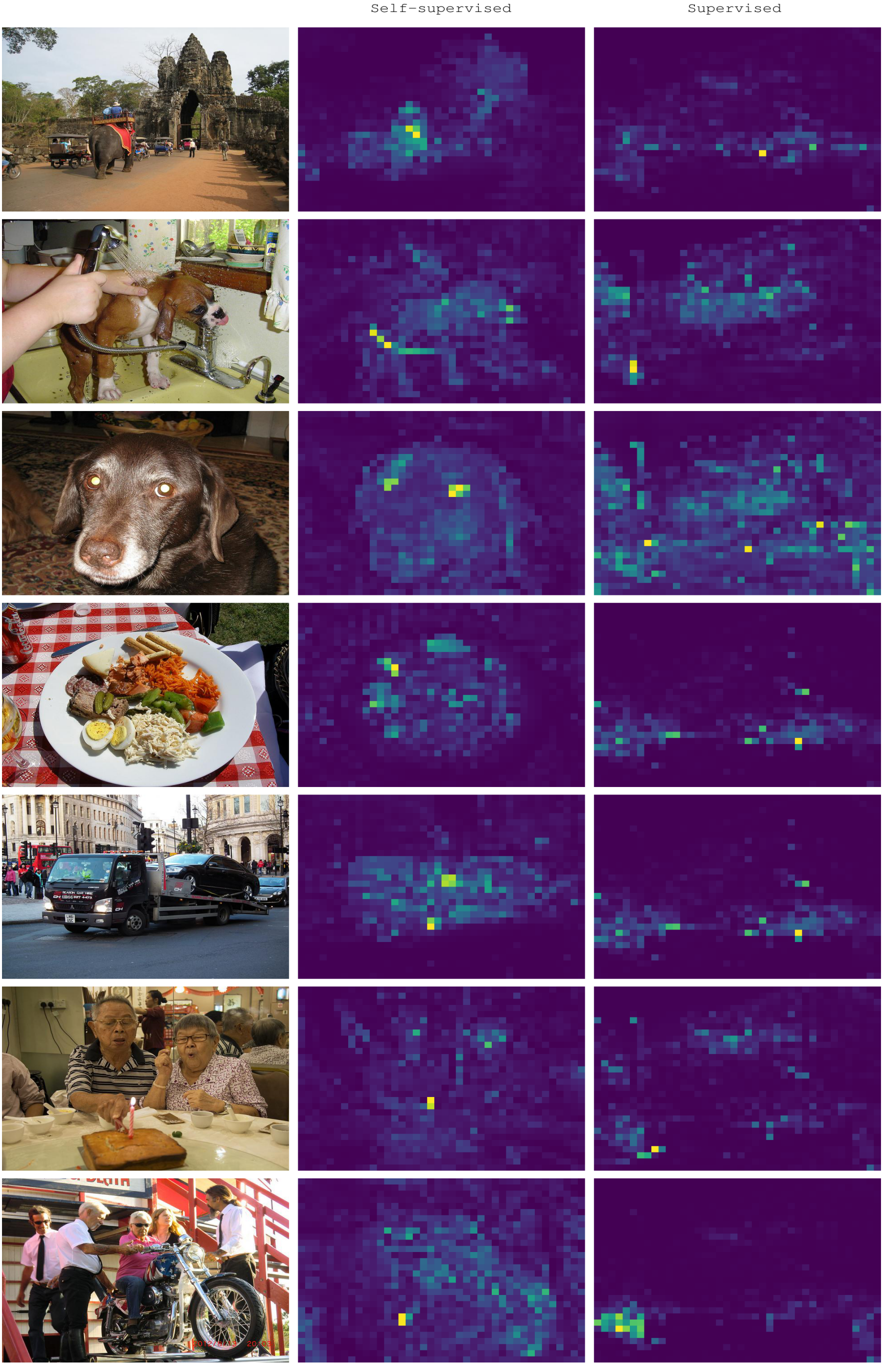


Figure A.2: Self-attention heads from the last layer for the raw attention. We look at the attention map when using the [CLS] token for self-supervised on the O3 dataset.

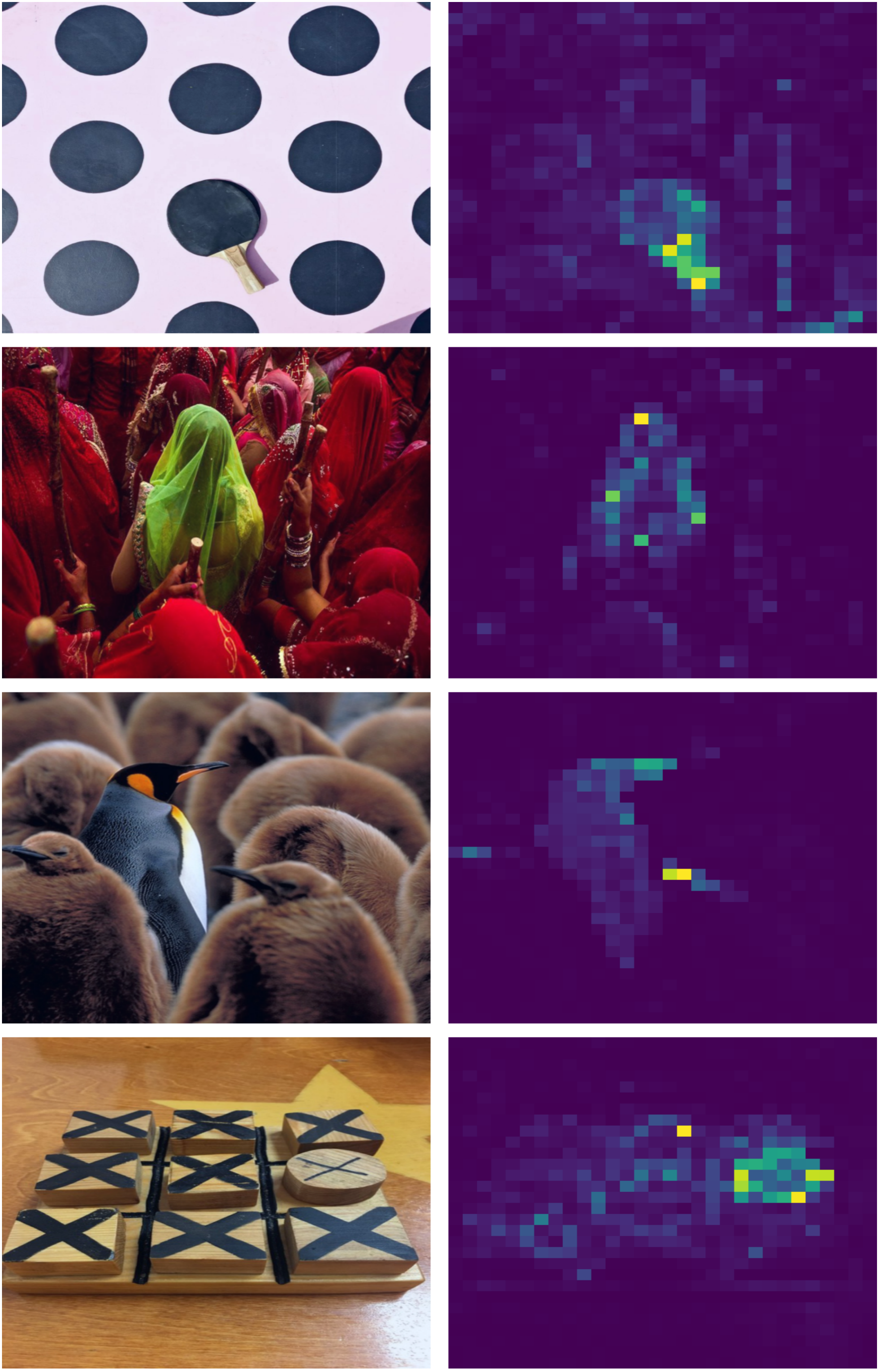
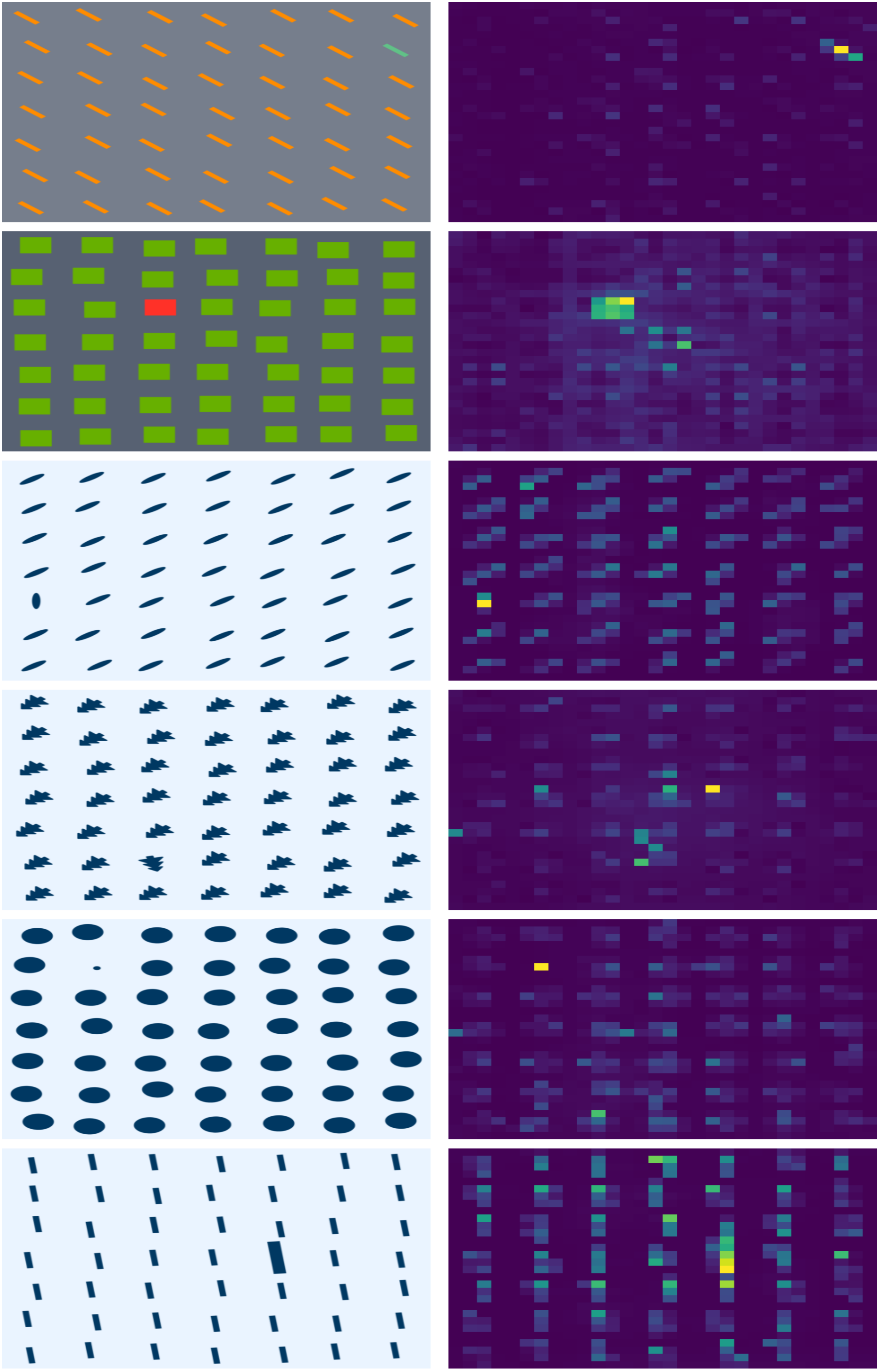


Figure A.3: Self-attention heads from the last layer for the raw attention. We look at the attention map when using the [CLS] token for self-supervised on the P3 dataset.



References

[1] 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*, pages 248–255. Ieee, 2009.