

# Supplementary Material for GTA: Guided Transfer of Spatial Attention from Self-supervised Models

Anonymous ICCV submission

Paper ID XXXXXX

## A. Release of code

Our code has been submitted for reproducibility. In order to ensure strict anonymization, we have removed all identifiable information such as URLs, copyright, and licenses. Upon acceptance, we will upload an un-anonymized version of the code on a public GitHub repository.

## B. Comparison of self-attention maps

In this section, we show additional visual comparisons of the self-attention maps obtained from pre-trained, fine-tuned, and GTA-trained models on multiple datasets (see Figure S1) [1, 2, 3, 4, 5]. The self-attention maps allow us to understand where the model attends to different parts of the input image.

For each dataset, we randomly select a sample image and visualize the self-attention maps. We observe that the self-attention maps of the fine-tuned model are much scattered over non-meaningful areas, in contrast to the pre-trained model which demonstrates focused attention on important regions. Such behavior could lead to the loss of well-trained spatial information, eventually resulting in lower performance. However, by introducing GTA, we show that it is possible to avoid this issue by explicitly regularizing the attention logits between target and source models. We present a visual comparison of the self-attention maps from these models to illustrate the effectiveness of the proposed method in guiding attention towards important regions during training. The visualization results demonstrate that GTA-trained models outperform fine-tuned models on multiple datasets.



Figure S1. **Comparison of self-attention maps from pre-trained, naïvely fine-tuned, and GTA-trained models across multiple datasets.** We consider CUB, Cars, Aircraft, Dogs, and Pets datasets. The self-attention maps of the multiple heads are aggregated with maximum values, and visualized in red color. Each column shows the attention maps from the models that are pre-trained using SSL, fine-tuned, and fine-tuned with GTA on 15% and 100% of training data, respectively. GTA shows that it is capable of fully leveraging object-centric representations learned by the SSL model.

## References

- [1] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer, 2011.
- [2] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.
- [3] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- [4] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.
- [5] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.