Robust Manipulation with Spatial Features

Anonymous Author(s)
Affiliation
Address
email

Abstract: Our goal is to develop visual pre-training strategies that enable more robust and efficient manipulation policy learning. We find that a Vision Transformer trained with a distillation loss that biases representations towards shape exhibits strong zero-shot transfer performance on the kitchen shift suite, even when compared to baselines trained on larger and more task-relevant datasets. When finetuned, the attention heads of a transformer trained with a shape bias can be visualized as a spatial feature map, which emergently segments manipulation-relevant objects in an image. By leveraging each of these insights, we are able to improve the average zero-shot performance of policies trained on the sliding door task within the FrankaKitchen environment by nearly 2x compared to the next best method. Additionally, we are able to improve maximum success in distribution by 13% by masking out attention heads that attend to distractors.

Keywords: Manipulation, visual pre-training, self-supervision

1 Introduction

Our goal is to learn robotic manipulation policies from images. For many computer vision tasks, models can be applied off-the-shelf in new environments with little to no task-specific tuning. In spite of this success, robotic policies learned from pixels remain surprisingly brittle. One common approach to learning from pixels follows a formula that is familiar to many computer vision practitioners: pre-train a self-supervised network on a broad and diverse image dataset before fine-tuning on task specific data and labels. We expect this strategy to yield a model that's capable of generalizing the downstream task across visually diverse environments, but when roboticists try this same formula, learned policies break in the presence of a distractor, under subtle lighting changes, and after a slight change in the camera position.

Recent work posits that the missing piece is a large dataset of object interactions across diverse environments—the ImageNet or CommonCrawl of manipulation. Indeed, training on large datasets of first person human interaction data increases policy performance downstream. However, these policies remain brittle to even small distribution shifts that commonly occur during deployment. Why are robotic policies learned from visual features so sensitive to distribution shift compared to other tasks that rely on visual information?

Successful manipulation requires spatial reasoning. To that end, past work introduced structured representations (e.g., keypoints) that capture the spatial aspects of the visual observation space at the cost of expressivity. Instead of introducing explicit structure into the representation, we leverage an encoder pre-trained with a self-supervised distillation loss (DiNo) [1] that biases the representation towards shape. We show that policies learned on top of these representations are more robust in the presence of visual distribution shift even when compared to representations learned from larger and more task-relevant datasets. Unlike more structured approaches, pre-training with DiNo can be applied to any architecture or dataset and doesn't require explicit supervision.

Another benefit of this encoder choice is that the transformer attention heads can be visualized as a spatial heatmap. We find that the visualized attention heads can be interacted with in predictable

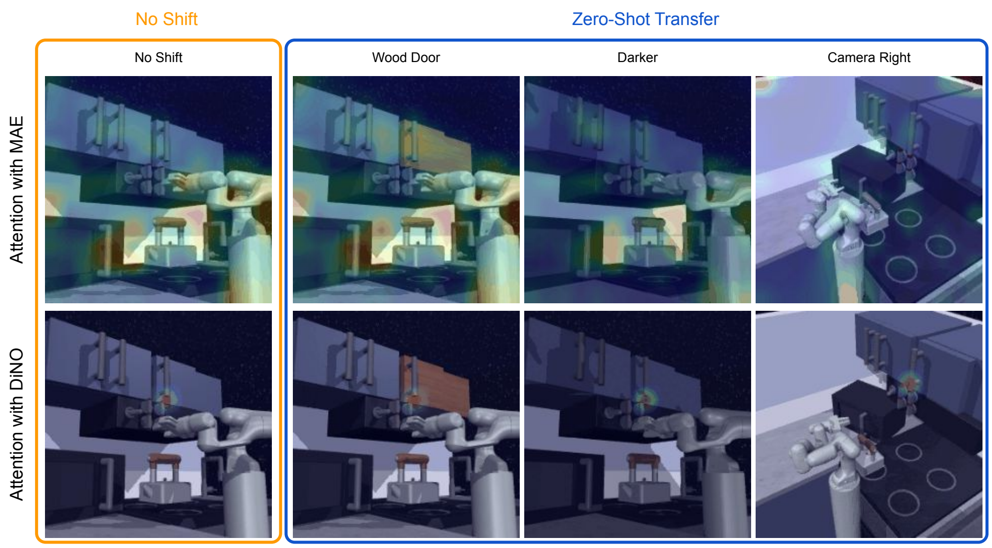


Figure 1: Encoders trained with losses that induce spatial heatmaps transfer to new viewpoints and textures zero-shot.

ways. Concretely, by masking out attention heads that segment task irrelevant parts of the image, we can improve the average performance of a policy trained on top of this architecture. This hints at the possibility of using the segmentation performance of the attention heads as a heuristic to identify good models to serve as representations for manipulation.

This work demonstrates that pre-training a visual encoder with a loss that is biased towards shape can dramatically improve policy performance under distribution shifts. We demonstrate this across 27 different training settings of the door sliding task within the FrankaKitchen [2] environment and evaluate across 7 different texture and lighting changes adapted from the Kitchen Shift benchmark [3]. We find that a vision transformer trained with a shape-biased distillation loss (DiNo) strongly outperforms both Transformer [4] and ResNet [5] architectures trained on much larger and task relevant datasets. The attention heads from a shape-biased Vision Transformer can be visualized as spatial heatmaps, which can be visually inspected to idenitfy task relevant and irrelevant heads. We explore this idea by masking attention heads that attend to irrelevant background pixels and observe a 13% boost to policy performance. After fine-tuning on a robotics task, these attention maps also emergently highlight task-relevant keypoints in the scene that are robust across visual distribution shifts. This is true for even dramatic visual shifts such as camera viewpoint. We believe that shape-biased losses could be a new standard for pre-training visiual encoders for manipulation and that leveraging the attention heads of Vision Transformers (e.g., by masking) could lead to further policy improvements.

2 Related Work

Policy adaptation. Policies learned from pixels are known to be sensitive to distractors. Policy adaptation approaches aim to resolve this instability by continuing to train self-supervised visual representations between during deployment [6], improving the transferability of encoders through augmentations [7, 8], or collecting exploration data in the target environment to align source and target represnetations [9]. Unlike this work, our method doesn't require any target domain data or hand-designed augmentations.

Representation learning for manipulation. The correct approach to visual representation learning for robotics is still an open question. Some work has analysed the transfer quality of a variety different supervised vision tasks to robotics tasks [10]. Unlike this work, training with DiNo

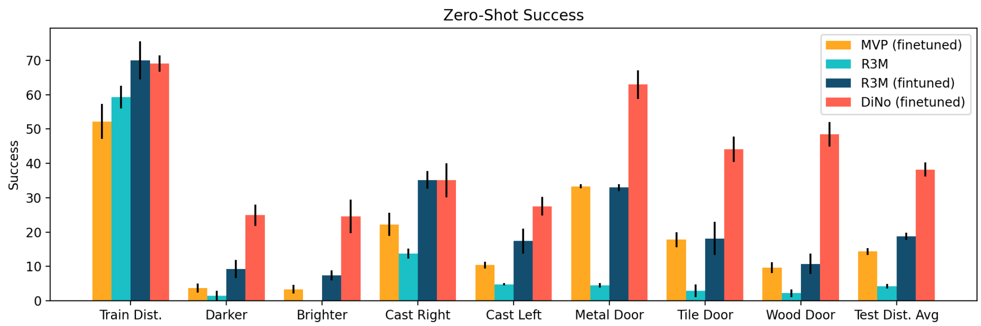


Figure 2: Zero shot performance on the kitchen shift suite.

does not require any labels and so it can be readily adapted to more robotics-relevant datasets. We compare directly against works that have developed self-supervised losses for manipulation [11] on manipulation-relevant datasets [12] or directly evaluated existing self-supervision approaches on such datasets [13].

3 Robust Manipulation with Spatial Features

We study two questions: (1) Can encoders trained with a shape-biased loss perform better under visual distribution shifts than other self-supervised losses? (2) Can the intuitive interpretations of attention map visualizations be leveraged to improve policy performance during training?

Shape bias improves zero-shot transfer. We follow the same evaluation protocol as R3M on the sliding door task. On top of each encoder, we train a two-layer MLP with imitation learning to perform the sliding door task. We compare across 3 seeds, 3 levels of demonstrations, and 3 camera angles. We then evaluate the performance of the policy and encoder across a subset of visual distribution shifts in the Kitchen Shift benchmark. This includes changing lighting—making the lighting darker, making the lighting brighter, lighting cast left, and lighting cast right—as well as changing the texture wrapping the cabinet of the sliding door to be wood, metal, or tile. The R3M training and testing environments modify FrankaKitchen by randomizing the position of the kitchen, so we reimplement these distribution shifts in the R3M evaluation environment. Because the kitchen position is randomized, the task is much more difficult to solve using memorization. We expect replay data to perform much worse than in the original Kitchen Shift benchmark.

We compare a Vision Transformer trained with a shape-biased loss (DiNo) against three other visual representation learning approaches. In MVP we borrow the encoder from Radosavovic et al. [13] and finetune. MVP leverages a ViT trained with masked autoencoding (MAE) on a mixture of human interaction data including Ego-4D. We also compare against a frozen and finetuned model from R3M [11]. R3M utilizes a ResNet-50 architecture trained on top of Ego-4D.

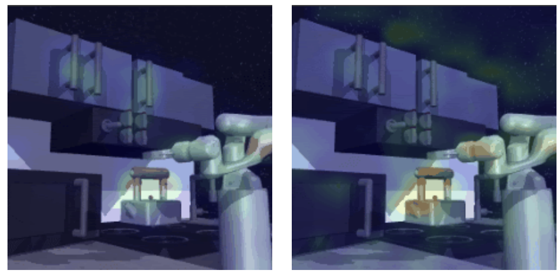
The zero-shot performance of each model across distribution shifts can be found in Figure 2. On the left, we present the performance of each model without any distribution shift. We then plot the performance the models by shift type and show the performance averaged across shifts on the right. For all of the shifts, we average results across level of demonstrations and camera angles and then take the average and standard error over seeds.

Visualizations of the attention heads after training are presented in Figure 1. Similar to DiNo, we visualize attention heads by mapping the weight at each head at the output of the last block to a heat map and smoothing the final map with bilinear interpolation. The finetuned MVP model is visualized on the top and the fine-tuned DiNo model is visualized on the bottom. Of the six heads in the last block, we select the head that best attends to the manipulation relevant objects by visual inspection. At the best head, fine-tuning with DiNo appears to give more manipulation-relevant and precise attention heads. Surprisingly, the same attention head is consistent across texture shifts and across viewpoints.

Leveraging attention for better policy learning. In this section, we study the question: are the attention head visualizations useful to the extent that they enable the development of a better performing policy? This is an important proof-of-concept that opens the door for future work to improve policy performance by leveraging attention heads that segment objects that are relevant to the desired manipulation task. For example, a practitioner could decide to mask a head that attends to the microwave if the policy needs to open the cabinet.

Heads Used	Success
All Heads	38.0 ± 6.0
Hand-Selected Head	43.3 ± 1.8
Random Head	32.7 ± 4.4

(a) Average success with different heads



(b) Hand-Selected Head (c) Random Head

Figure 3: (Left) If we mask out all but an intuitively-correct hand-selected head, we can boost average policy performance by 13%. (Middle) A visualization of an attention head before finetuning on the target task. Without any environment data, the attention head segments manipulation relevant objects. (Right) A visualization of an attention head selected at random. Compared to the hand selected head, the random head segments the background, which is irrelevant to the sliding door task.

We focus on the sliding door task trained with 5 human demonstrations and the left camera viewpoint. We visually inspect each of the 6 attention heads of a DiNo-pretrained vision transformer and select the head that segments the most task-relevant objects. We mask all but the hand-selected head and compare the success of training an MLP without finetuning after 1000 training steps. We present the average performance results with standard error across 3 seeds in Table 3a. For an additional baseline, we also report the results of masking all but a random head. Attention heads are masked by zeroing out the weights that map from input vectors to query, key, and value vectors. We only visualize and mask heads at the last attention block. After masking out the hand-selected head, success after 1000 training steps sees a modest performance improvement with reduced variance compared to using all heads.

4 Conclusion

In this paper we studied two questions related to visual representation learning for manipulation. First, we find that pre-training with a loss that induces a shape bias can provide strong performance gains when evaluating policies under visual distribution shift. Second, we present a proof of concept that leverages the insight that the attention heads of a DiNo-trained Vision Transformer segment task relevant objects. Our findings open up important questions for future work, such as: could training larger and more task-relevant datasets, such as Ego-4D, with a shape-biased loss further improve policy learning performance?

References

[1] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[2] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In *Conference on Robot Learning*, pages 1025–1037. PMLR, 2020.

[3] E. Xing, A. Gupta, S. Powers*, and V. Dean*. Kitchenshift: Evaluating zero-shot generalization of imitation-based policy learning under domain shifts. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. URL https://openreview.net/forum?id=DdglKo8hBq0.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. 2017. URL https://arxiv.org/pdf/1706.03762.pdf.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[6] N. Hansen, R. Jangir, Y. Sun, G. Alenyà, P. Abbeel, A. A. Efros, L. Pinto, and X. Wang. Self-supervised policy adaptation during deployment. In *International Conference on Learning Representations*, 2021.

[7] N. Hansen and X. Wang. Generalization in reinforcement learning by soft data augmentation. In *International Conference on Robotics and Automation*, 2021.

[8] L. Fan, G. Wang, D.-A. Huang, Z. Yu, L. Fei-Fei, Y. Zhu, and A. Anandkumar. Secant: Self-expert cloning for zero-shot generalization of visual policies. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3088–3099. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/fan21c.html.

[9] T. Yoneda, G. Yang, M. R. Walter, and B. Stadie. Invariance through latent alignment, 2021.

[10]

[11] S. Naie, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2022.

[12] K. Grauman, A. Westbury, E. Byrne, Z. Q. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. González, J. M. Hillis, X. Huang, Y. Huang, W. Jia, W. Y. H. Khoo, J. Kolár, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbeláez, D. J. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. A. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik. Ego4d: Around the world in 3,000 hours of egocentric video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990, 2022.

[13] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell. Real-world robot learning with masked visual pre-training. *CoRL*, 2022.