Modality-invariant Visual Odometry for Indoor Navigation

Marius Memmel*
Technical University of Darmstadt
marius.memmel@stud.tu-darmstadt.de

Amir Zamir
Swiss Federal Institute of Technology Lausanne (EPFL)
amir.zamir@epfl.ch

Abstract

Successful indoor navigation is a crucial skill for many robots. This fundamental ability has been extensively studied through the task of PointGoal navigation in simulated environments. With noisy observations and actuation, the setting becomes more realistic and previously successful agents fail dramatically. Visual Odometry has shown to be a practical substitute for GPS+compass and can effectively localize the agent from visual observations. With the availability of multiple sensors and estimators, the question naturally arises of how to make the most use of multiple input modalities. When having access to multiple modalities, the predictions of naive multi-modal approaches can be dominated by a single one, impeding overall robustness. Recent methods are modality-specific and can not deal with "privileged" modalities, e.g.*, irregular or no access to depth during test time. We propose the Visual Odometry Transformer, a novel approach to multi-modal Visual Odometry based on Vision Transformers that successfully replaces GPS+compass. Our experiments show that the model can deal with limited availability of modalities during test time by implicitly learning a representation invariant to the availability of input modalities.*

1. Introduction

One of the most fundamental skills embodied agents must learn is to effectively traverse the environment around them, allowing them to move past stationary tasks and provide services in multiple locations [11]. The ability of an agent to locate itself in an environment is vital to navigating it successfully [3, 17]. In a realistic PointGoal navigation setting, the agent does not have access to perfect `GPS+Compass` sensors [9]. Visual Odometry (VO) is one way to localize the agent from noisy `RGB` and `Depth` observations [1]. Deploying a separate VO model has shown to be beneficial when localizing the agent from visual observations only [3, 17] However, those methods are not robust to sensor failure or "privileged" modalities, i.e. when an input modality is only occasionally available at the test time. Access to multiple modalities has shown to be beneficial for many downstream tasks [12], and VO [10, 14, 18] in particular. Recent work has investigated learning representations from multiple ground truth modalities or pseudo labels [2,6,8] that could act as a better initialization for such approaches. However, the underlying Convolution Neural Network (ConvNet) architecture of recent methods [3, 17], assumes a constant channel size of the input. This reliance makes dealing with multiple modalities and privileged access to those sheer impossible. With its independence to the input size, multi-modal methods, therefore, naturally turn to the Vision Transformer (ViT) [2, 5, 6, 8]. In PointGoal navigation, the action space (move `fwd` $0.25m$, turn `left` and `right` by $30°$) causes large displacements in all pixels, requiring a larger receptive field to relate them and find correspondences. We assume that the attention mechanism can better exploit this property than locality-biased ConvNets. We find that ViTs indeed learn to focus on image regions that matter. To summarize, we propose the Visual Odometry Transformer (VOT), a multi-modal ViT for VO in indoor environments. With our work, we hope to emphasize the potential of multi-modal representations for embodied AI and navigation in particular. Code and visualization are available at github.com/memmelma/VO-Transformer.

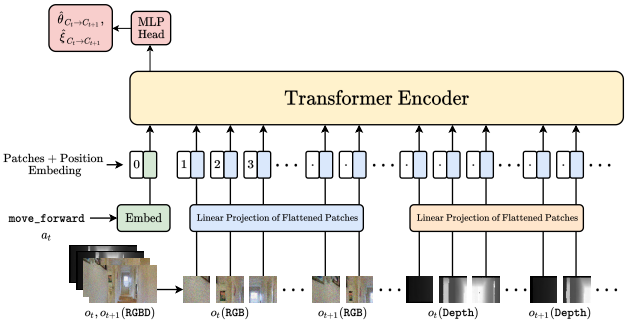

Figure 1. The Visual Odometry Transformer architecture with `RGB-D` input and based on the Multi-modal Multi-task Masked Autoencoder.

*work done while at EPFL

2. Proposed Method

Following [17] we update the agent's relative goal position by its coordinate transformation parameterized by rotation angle $\widehat{\beta}_{C_t \to C_{t+1}} \in [0, 2\pi)$ and translation vector $\widehat{\boldsymbol{\xi}}_{C_t \to C_{t+1}} \in \mathbb{R}^2$. To regress those VO parameters, we propose a novel architecture that leverages the multi-modal agent observations o_t (`RGB-D`) at time step t, and exploits the global property of the VO problem.

Dealing with all available modalities, the VOT follows the Multi-modal Multi-task Masked Autoencoder (MultiMAE) [2] based on the ViT-B/16 [5], and pre-trained on `RGB`, `Depth`, and semantic segmentation. We adopt the fixed positional embedding and separate linear projection layers for each modality from [2]. We keep the projections for `RGB` and `Depth`, and discard the one for semantic segmentation as the modality is not available in our setting. This design choice allows to easily extend the model to new modalities by adding linear projections and interpolating the positional embedding [2, 15]. An MLP-head estimates the VO parameters from a separate token passed to the model, similar to the class token in [4,5]. Figure 1 shows the presented architecture. We use the regression and geometric invariance losses of [17]. Following [17], we collect 250 k observation-transformation pairs from the training set of *Gibson-4+* [11], and augment `left` and `right` actions.

As training data is scarce in VO settings [5, 7, 13], and having a lot of it has shown to benefit ViT training [5, 16], we deploy various techniques to improve data-efficiency. During an extensive ablations study, we find the agent's action to be a strong prior on the estimation. To condition the model on it, we pass an embedding of the action as a separate token. Furthermore, we use a pre-trained MultiMAE [2] to speed up learning on less data.

3. Experimental Evaluation

We train our model on only `RGB`, only `Depth`, and `RGB-D` with results in Table 1, and report success S, Success weighted by (normalized inverse) Path Length (SPL) [1], and Soft Success Path Length (SSPL) [3] on the validation set of *Gibson-4+*. The `Depth` input emerges as the most informative modality due to its geometric properties. Training the VO model on `RGB` hurts the navigation performance as we find it to overfit the visual appearance of the scenes and not being able to generalize to unseen ones.

We also evaluate the models' invariance to privileged modalities by dropping access to one of the two modalities randomly. We find that even though the VOT was pre-trained and fine-tuned on `RGB-D`, it heavily relies on `Depth` input. However, VO models without access to `Depth` causes the agent to momentarily get stuck in narrow passages. This delay causes the agent to reach the goal slower than expected. In some cases, it even terminates the episode as the maximum number of steps is reached even though the agent still roughly follows the shortest path to the goal. Even though success rate and SPL appear to diminish drastically, the high SSPL indicates that the agent gets close to the goal. These results show that the model implicitly learns some invariance to the input modalities. However, it still relies on certain features of both input modalities, making it prone to small inconsistencies that get punished by the success-dependent metrics of PointGoal navigation [1].

Table 1. Results for different modality configurations (obs) and dropped modalities (drp). VOT strongly depends on `Depth` but the high SSPL shows an invariance to dropped modalities. The ConvNet approach [17] (unified estimator, ResNet-50 backbone, all geometric invariance losses) converges to a *blind* behavior when not all modalities are available during test time.

method	obs	drp	$S \uparrow$	SPL↑	SSPL↑
blind	–	–	0.00	0.00	5.40
oracle	–	–	97.89	74.80	73.10
[17]	`RGB-D`	–	64.50	48.90	65.40
[17]	`RGB-D`	`RGB`	0.00	0.00	5.40
[17]	`RGB-D`	`Depth`	0.00	0.00	5.40
VOT (ours)	`RGB`	–	59.30	45.40	66.70
VOT (ours)	`Depth`	–	93.30	71.70	72.00
VOT (ours)	`RGB-D`	–	88.20	67.90	71.30
VOT (ours)	`RGB-D`	`RGB`	75.90	58.50	69.90
VOT (ours)	`RGB-D`	`Depth`	26.10	20.00	58.70

For turning actions `left`, `right`, the model learns to focus on both modalities and on image regions that are present at both time steps. These findings suggest that the model is looking for corresponding features in the observations it can use to estimate the transformation between them. The assumption is strengthened as we find the model to attend primarily to the center regions of the image when estimating the parameters for a `fwd` action. When varying the action prior, *i.e.*, passing different actions for the same observation pair, these findings reappear, indicating that the model indeed learns to utilize the additional information.

4. Conclusions

Our work showcases the benefits of multi-modal ViT architectures for VO in indoor PointGoal navigation. Through our simple and easily expandable architecture, we hope to draw attention to applications of multi-modal ViTs in embodied AI. To overcome the performance gap between access to all modalities and "privileged" ones, we suggest to facilitate invariance during training by randomly dropping modalities. Further fine-tuning the policy might help adaptation to the VO inconsistencies. Future work may consider expanding the set of modalities and downstream tasks.

References

[1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. In *arXiv preprint arXiv:1807.06757*, 2018. 1, 2

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *arXiv preprint arXiv:2204.01678*, 2022. 1, 2

[3] Samyak Datta, Oleksandr Maksymets, Judy Hoffman, Stefan Lee, Dhruv Batra, and Devi Parikh. Integrating egocentric localization for more realistic point-goal navigation agents. In *Conference on Robot Learning*, 2021. 1, 2

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*, 2018. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 2

[6] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *arXiv preprint arXiv:2201.08377*, 2022. 1

[7] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *International Conference on Computer Vision*, 2015. 2

[8] Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. In *arXiv preprint arXiv:2111.12993*, 2021. 1

[9] Adithyavairavan Murali, Tao Chen, Kalyan Vasudev Alwala, Dhiraj Gandhi, Lerrel Pinto, Saurabh Gupta, and Abhinav Gupta. Pyrobot: An open-source robotics framework for research and benchmarking. In *arXiv preprint arXiv:1906.08236*, 2019. 1

[10] Noha Radwan, Abhinav Valada, and Wolfram Burgard. Vlocnet++: Deep multitask learning for semantic visual localization and odometry. In *IEEE Robotics and Automation Letters*, 2018. 1

[11] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[12] Alexander Sax, Jeffrey O. Zhang, Bradley Emi, Amir Zamir, Silvio Savarese, Leonidas Guibas, and Jitendra Malik. Learning to navigate using mid-level visual priors. In *Conference on Robot Learning*, 2020. 1

[13] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. In *arXiv preprint arXiv:2106.10270*, 2021. 2

[14] Abhinav Valada, Noha Radwan, and Wolfram Burgard. Deep auxiliary learning for visual localization and odometry. In *IEEE International Conference on Robotics and Automation*, 2018. 1

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2

[16] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021. 2

[17] Xiaoming Zhao, Harsh Agrawal, Dhruv Batra, and Alexander G Schwing. The surprising effectiveness of visual odometry techniques for embodied pointgoal navigation. In *International Conference on Computer Vision*, 2021. 1, 2

[18] Ran Zhu, Mingkun Yang, Wang Liu, Rujun Song, Bo Yan, and Zhuoling Xiao. Deepavo: Efficient pose refining with feature distilling for deep visual odometry. In *Neurocomputing*, 2022. 1