COMPUTER VISION

Does computer vision matter for action?

Brady Zhou[1]*, Philipp Krähenbühl[1,2], Vladlen Koltun[1]

Controlled experiments indicate that explicit intermediate representations help action.

Biological vision systems evolved to support action in physical environments (*1*, *2*). Action is also a driving inspiration for computer vision research. Problems in computer vision are often motivated by their relevance to robotics and their prospective utility for systems that move and act in the physical world. In contrast, a recent stream of research at the intersection of machine learning and robotics has demonstrated that models can be trained to map raw visual input directly to action (*3*–*6*). These models bypass explicit computer vision entirely. They do not incorporate modules that perform recognition, depth estimation, optical flow, or other explicit vision tasks. The underlying assumption is that perceptual capabilities will arise in the model as needed, as a result of training for specific motor tasks. This is a compelling hypothesis that, if taken at face value, appears to make much computer vision research obsolete. If any robotic system can be trained directly for the task at hand, with only raw images as input and no explicit vision modules, then what is the utility of further perfecting models for semantic segmentation, depth estimation, optical flow, and other vision tasks?

We report controlled experiments that assessed whether specific vision capabilities are useful in mobile sensorimotor systems that act in complex three-dimensional environments. To conduct these experiments, we used realistic three-dimensional simulations derived from immersive computer games. We instrumented the game engines to support controlled execution of specific scenarios that simulated tasks such as driving a car, traversing a trail in rough terrain, and battling opponents in a labyrinth. We then trained sensorimotor systems equipped with different vision modules and measured their performance on these tasks.

Our baselines were end-to-end pixels-to-actions models that were trained directly for the task at hand. These models did not rely on any explicit computer vision modules and embodied the assumption that perceptual capabilities will arise as needed, in the course of learning to perform the requisite sensorimotor task. To these we compared models that received as additional input the kinds of representations that are studied in computer vision research, such as semantic label maps, depth maps, and optical flow. We could therefore assess whether representations produced in computer vision are useful for sensorimotor challenges. In effect, we asked: What if a given vision task was solved? Would this matter for learning to act in complex three-dimensional environments?

Our first finding is that computer vision does matter. When agents were provided with representations studied in computer vision, they achieved higher performance in sensorimotor tasks. The effect is significant and consistent across simulation platforms and tasks.

We then examined in finer granularity how useful specific computer vision capabilities are in this context. Our second finding is that some computer vision capabilities appear to be more impactful for mobile sensorimotor operation than others. Specifically, depth estimation and semantic scene segmentation provided the highest boost in task performance among the individual capabilities we evaluated. Using all capabilities in concert was more effective still.

We also conducted supporting experiments that aimed to probe the role of explicit computer vision in detail. We found that explicit computer vision is particularly helpful in generalization by providing abstraction that helps the trained system sustain its performance in previously unseen environments. Last, we show that the findings hold even when agents predict the intermediate representations in situ with no privileged information.

RESULTS

We performed experiments using two simulation environments: the open-world urban and suburban simulation Grand Theft Auto V (*7*–*9*) and the ViZDoom platform for immersive three-dimensional battles (*5*, *10*). In these environments, we set up three tasks: urban driving, off-road trail traversal, and battle. The tasks are illustrated in Fig. 1A.

For each task, we trained agents that either acted on the basis of the raw visual input alone or were also provided with one or more of the following intermediate representations: semantic and instance segmentation, monocular depth and normals, optical flow, and material properties (albedo). The intermediate representations are illustrated in Fig. 1B. The environments, tasks, agent architectures, and further details are specified in the Supplementary Materials.

Figure 1C summarizes the main results. Intermediate representations clearly help. The Supplementary Materials reports additional experiments that examine these findings, possible causes, and alternative hypotheses.

Analysis

Our main results indicate that sensorimotor agents can greatly benefit from predicting explicit intermediate representations of scene content, as posited in computer vision research. Across three challenging tasks, an agent that saw not only the image but also the kinds of intermediate representations that were pursued in computer vision research learned significantly better sensorimotor coordination. Even when the intermediate representations were imperfectly predicted in situ by a small, light-weight deep network, the improvements were significant (Fig. 1D).

The benefits of explicit vision are particularly salient when it comes to generalization. Equipping a sensorimotor agent with explicit intermediate representations of the scene leads to more general sensorimotor policies. As reported in the Supplementary Materials, in urban driving, the performance of image-only and image-and-vision agents is nearly tied on the training set. However, when we tested generalization to new areas, the image-and-vision agent outperformed the image-only agent on the test set even with an order of magnitude less experience with the task during training.

[1]Intel Labs, Santa Clara, CA, USA. [2]University of Texas at Austin, Austin, TX, USA.
*Corresponding author. Email: brady.zhou@intel.com

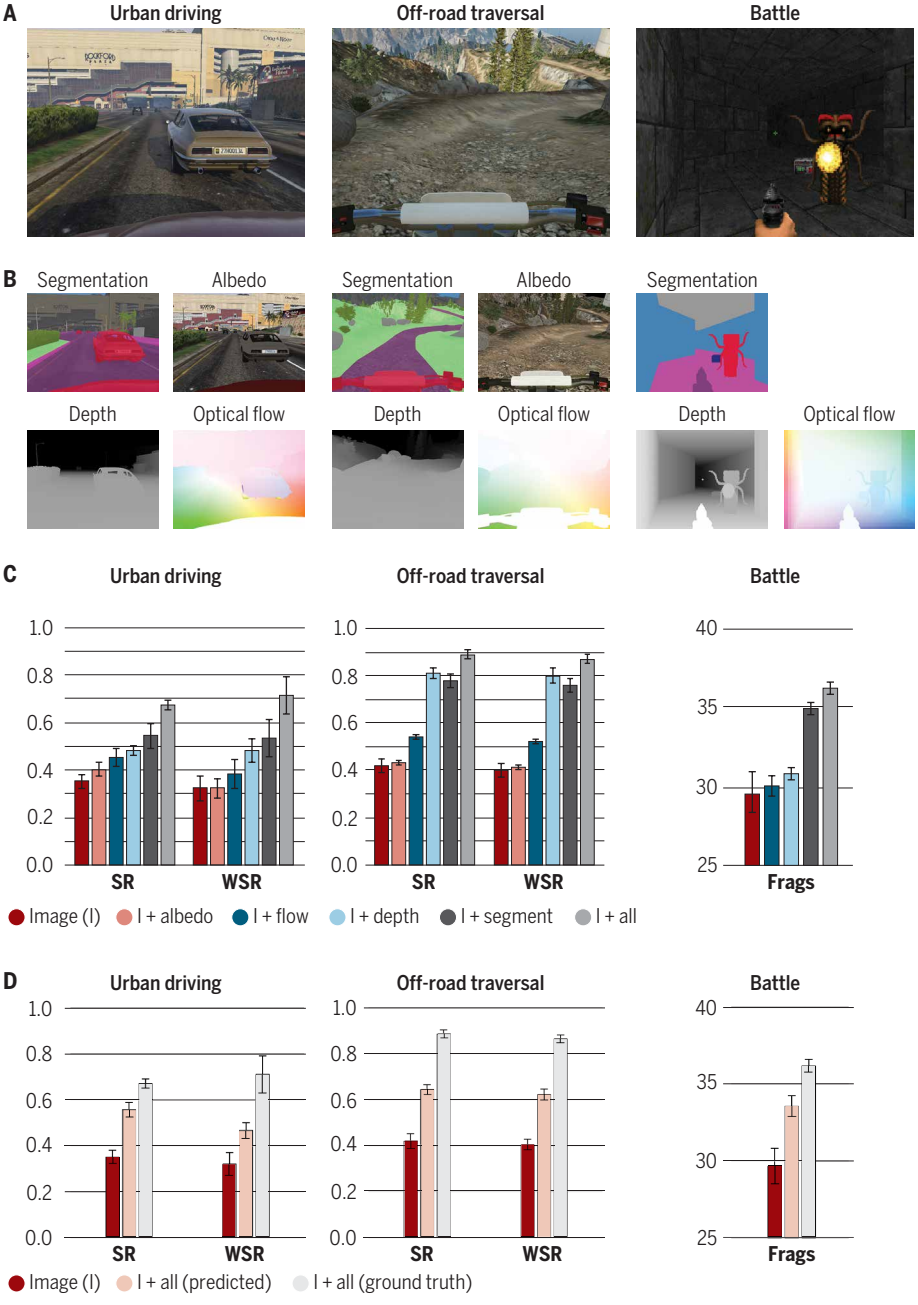

Fig. 1. Assessing the utility of intermediate representations for sensorimotor control. (**A**) Sensorimotor tasks. (**B**) Intermediate representations: semantic segmentation, intrinsic surface color (albedo), optical flow, and depth. (Albedo not used in battle.) (**C**) For each task, we compare an image-only agent with an agent that is also provided with ground-truth intermediate representations. The agent observes the intermediate representations during both training and testing. Success rate (SR; fraction of scenarios in which agent successfully reached target location), weighted success rate (WSR; weighted by track length), "frags" (number of enemies killed in a battle episode). Data are means and SD. (**D**) In "I + all (predicted)," the intermediate representations are predicted in situ by a convolutional network; the agent is not given ground-truth representations at test time. The results indicate that even predicted vision modalities confer a significant advantage.

This generalization was exhibited not only by agents equipped with ground-truth representations but also by agents that predicted the intermediate representations in situ with no privileged information at test time. An agent that explicitly predicted intermediate representations of the scene and used these explicit representations for control generalized better to previously unseen test scenarios than an end-to-end pixels-to-actions agent.

CONCLUSION

Computer vision produces representations of scene content. Much computer vision research is predicated on the assumption that these intermediate representations are useful for action. Recent work at the intersection of machine learning and robotics has called this assumption into question by training sensorimotor systems directly for the task at hand, from pixels to actions, with no explicit intermediate representations. Thus, the central question of our work: Does computer vision matter for action? Our results indicate that it does. Models equipped with explicit intermediate representations train faster, achieve higher task performance, and generalize better to previously unseen environments.

SUPPLEMENTARY MATERIALS

robotics.sciencemag.org/cgi/content/full/4/30/eaaw6661/DC1
Supplement S1. Synopsis of findings.
Supplement S2. Materials and methods.
Supplement S3. Experiments and analysis.
Fig. S1. Three sensorimotor tasks used in our experiments.
Fig. S2. Different computer vision modalities used in our experiments, illustrated on the urban driving task.
Fig. S3. Performance of agents equipped with different input representations.
Fig. S4. Performance of agents equipped with different predicted input representations.
Fig. S5. Performance of unsupervised, predicted, and ground-truth vision agents as a function of training set size.
Fig. S6. Performance of image-only and vision-equipped agents as a function of training set size on battle.
Fig. S7. Imitation learning agent architecture.
Fig. S8. Network architecture used to infer the vision modalities.
Table S1. Performance of image-only and vision-equipped agents on training tracks.
Movie S1. Summary.
References (*11*–*15*)

REFERENCES AND NOTES

1. J. J. Gibson, *The Ecological Approach to Visual Perception* (Houghton Mifflin, 1979).
2. P. S. Churchland, V. S. Ramachandran, T. J. Sejnowski, A critique of pure vision, in *Large-Scale Neuronal Theories of the Brain*, C. Koch, J. L. David, Eds. (MIT Press, 1994).
3. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
4. S. Levine, C. Finn, T. Darrell, P. Abbeel, End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.* **17**, 1–40 (2016).
5. A. Dosovitskiy, V. Koltun, Learning to act by predicting the future. *ICLR* (2017).
6. F. Codevilla, M. Müller, A. López, V. Koltun, A. Dosovitskiy, End-to-end driving via conditional imitation learning, in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2018), pp. 1–9.
7. S. R. Richter, V. Vineet, S. Roth, V. Koltun, Playing for data: Ground truth from computer games, in *Computer*

Vision ECCV 2016, B. Leibe, J. Matas, N. Sebe, M. Welling, Eds. (Springer International, 2016).
8. S. R. Richter, Z. Hayder, V. Koltun, Playing for benchmarks, in *2017 IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2017), pp. 2232–2241.
9. P. Krähenbühl, Free supervision from video games, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 2955–2964.
10. M. Kempka, M. Wydmuch, G. Runc, J. Toczek, W. Jaśkowski, ViZDoom: A Doom-based AI research platform for visual reinforcement learning, *IEEE Conference on Computational Intelligence and Games* (2016).
11. P. Anderson, A. X. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, A. R. Zamir, On evaluation of embodied navigation agents. arXiv:1807.06757 [cs.AI] (18 July 2018).
12. M. Lin, Q. Chen, S. Yan, Network in network, paper presented at the International Conference on Learning Representations (ICLR), 14 to 16 April 2014, Banff, Canada.
13. S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in *International Conference on Machine Learning* (2015), vol. 37, pp. 448–456.
14. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, poster presented at the International Conference on Learning Representations (ICLR), 7 to 9 May 2015, San Diego, CA.
15. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. Wells, A. Frangi, Eds. (vol. 9351 of Lecture Notes in Computer Science, Springer, 2015), pp. 234–241.

Author contributions: B.Z., P.K., and V.K. formulated the study, developed the methodology, and designed the experiments. B.Z. and P.K. developed the experimental infrastructure and performed the experiments. B.Z., P.K., and V.K. analyzed data and wrote the paper. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** For data not presented in this paper and/or in the Supplementary Materials, please contact the corresponding author. Code and data for reproducing the results will be released at https://github.com/intel-isl/vision-for-action upon publication.

10.1126/scirobotics.aaw6661

Citation: B. Zhou, P. Krähenbühl, V. Koltun, Does computer vision matter for action? *Sci. Robot.* **4**, eaaw6661 (2019).

Science Robotics

Does computer vision matter for action?

Brady Zhou, Philipp Krähenbühl, and Vladlen Koltun

Sci. Robot. **4** (30), eaaw6661. DOI: 10.1126/scirobotics.aaw6661

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.