From Static Domain Adaptation to State-Adaptive Perception in Embodied Agents

Yu Zhang
Boise State University
yzhang@boisestate.edu

Abstract

Fundamental vision tasks are increasingly important for robotics, from ground navigation to aerial monitoring. As robots are deployed in diverse scenarios, domain adaptation has been proposed to address the challenge of changing environments. However, current methods focus on bridging the gap between pre-defined, static domains such as synthetic-to-real or sunny-to-rainy. We argue that this static view of adaptation is insufficient for embodied agents. Instead, embodied agents should adapt to their own states, such as position, altitude, and orientation, to better handle changing environments and improve performance in core vision tasks. The goal should not be to merely cope with pre-defined shifts, but to enable systems to continuously adapt based on their operational status. Current models, despite their impressive performance, remain fundamentally unaware of their own states. We posit that the next generation of robust perception systems must be state-adaptive: dynamically modulating their internal processes in response to ever-changing conditions. This position paper calls for a paradigm shift from building generic, one-size-fits-all models toward adaptive systems that are intrinsically aware of their own states, paving the way for true domain robustness in robotic vision.

1 Introduction

In recent years, robotics has experienced transformative progress, much of it driven by breakthroughs in computer vision [7]. Embodied agents today can manipulate objects with remarkable precision [10, 10] and navigate complex environments as autonomous vehicles or drones [1, 3, 39], largely because they can perceive and interpret the world around them. This capability is grounded in a set of core vision tasks that convert raw sensor inputs into structured representations suitable for decision-making and control.

While a wide range of vision tasks support the development of effective robotic vision systems, here we highlight two representative tasks. **Semantic segmentation** [2, 14, 37] assigns a semantic label to every pixel of an image, enabling an embodied agent to build a detailed map of ***what*** is present in its environment. **Monocular depth estimation** [18] recovers geometric structure from a single camera view, informing the embodied agent ***where*** objects are located and supporting crucial reasoning about navigation and interaction [15, 29]. Together, these tasks give embodied agents both the semantic and geometric understanding needed for autonomous operation [17, 24].

Despite their importance, these models remain fragile outside controlled training settings. A recurring obstacle is **domain shift** [34], where a model trained under one set of conditions performs poorly when deployed in another. For instance, a segmentation model trained in clear weather may fail under rain [21, 31], and a depth estimator trained on outdoor driving scenes may yield unreliable results when applied to indoor navigation [16].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: SPACE in Vision, Language, and Embodied AI (SpaVLE).

To mitigate this problem, the community has largely pursued the paradigm of **unsupervised domain adaptation** (UDA)[4, 12]. For semantic segmentation, UDA research has focused on aligning feature distributions to bridge shifts in appearance and style, successfully adapting models from synthetic to real imagery [8] or from clear to adverse weather conditions [21, 31]. In depth estimation, the challenge is even sharper: the inherent scale ambiguity of inferring 3D from 2D images [9, 19] means that a model trained on long-range outdoor data such as KITTI [5] often produces arbitrarily scaled predictions when tested on short-range indoor datasets like NYUv2 [22].

As the demand for long-term autonomy grows [6, 33], researchers have also explored **continual learning** (CL)[13, 32]. Continual learning equips models to learn sequentially from ongoing data streams, incorporating new classes and domains without catastrophically forgetting previous knowledge [11, 13]. Continual learning offers a path for embodied agents to accumulate experience and expand their perceptual competence over a lifetime of operation [20].

However, both UDA and continual learning share a crucial limitation: they implicitly treat the embodied agent as a passive, disembodied observer. Adaptation is framed as a problem of distribution alignment across datasets, without explicitly considering the primary cause of those changes in a robotic context – ***the embodied agent's own state***. In practice, the robot's position, orientation, altitude, or motion directly shapes the data it observes [15, 30]. This embodiment is not a side effect but the root cause of many domain shifts. For example, in depth estimation, changes in altitude or viewpoint induce scale inconsistencies between training and deployment scenarios [9, 23]. In segmentation, that same geometric change alters class distributions: objects that once dominated the field of view may shrink into rare or hard-to-detect categories [28]. Traditional domain adaptation techniques, which focus only on image characteristics (e.g., synthetic vs. real), are blind to these state-dependent shifts [8, 27].

We therefore argue that vision models should explicitly adapt to the embodied agent's state. A model with an internal scaling mechanism linked to state information could address both challenges simultaneously: correcting metric scale for depth estimation and rebalancing class importance for segmentation. This reframes domain shift in robotics not merely as a visual discrepancy, but largely as a geometric consequence of embodiment. By making agents state-adaptive, we move closer to perception systems that are robust, generalizable, and ultimately capable of supporting long-term autonomy.

Using examples from recent advances in computer vision, this position paper argues that the primary bottleneck to achieving true domain robustness lies in the field's limited ability to adapt to the embodied agent's own state. Specifically, it issues a call to action for the computer vision research community:

- Move beyond adapting to static, appearance-based domains (e.g., synthetic-vs-real) and instead focus on adapting to the continuous, state-dependent domains generated by the embodied agent's own motion and interaction with the world.
- Prioritize research to develop new perception architectures with internal, dynamic mechanisms that are modulated by the embodied agent's state. This includes creating state-adaptive loss functions, attention mechanisms, and even architectures that can adapt their own computational graph based on context.
- Prioritize establishing new benchmarks that evaluate a model's ability to handle state-dependent domain shifts, unifying the evaluation of seemingly disparate problems like class imbalance in semantic segmentation and scale inconsistency in depth estimation under a common, robotics-centric framework.

2 The Real Gap: Perception Without State Awareness

The dominant paradigm for achieving robust perception is through data diversity. By training a single, large model on a vast and varied dataset encompassing numerous conditions (different weather, lighting, viewpoints, etc.), the goal is to learn a universal feature representation that generalizes across all scenarios. This approach, while powerful, implicitly models the world as an independent and identically distributed (i.i.d.) process. For an embodied agent like a robot, this assumption is perpetually violated. A robot's sensory stream is non-stationary and highly structured, with the statistical properties of the input data being a direct function of the agent's physical state. The failure

to account for this coupling between state and perception leads to a fragile "one-size-fits-all" model that may perform well on average but is unreliable in specific, state-dependent regimes. A notable example of this is the effect of visual scale, which is a direct consequence of the robot's distance from objects. This single state-dependent variable creates two well-known, yet typically disconnected, failure modes in vision:

Metric Scale Ambiguity in 3D Vision Monocular depth estimators are notoriously susceptible to domain shifts in depth range. A model trained on the outdoor KITTI dataset [5] (depths up to 80m) and tested on the indoor NYUv2 dataset [22] (depths up to 10m) will produce predictions that are not only metrically incorrect but also internally inconsistent, as the learned mapping from image cues to disparity is no longer valid [19]. This is a geometric problem, not a stylistic one; aligning image textures cannot resolve the fundamental ambiguity in metric scale.

Class Representation Imbalance in 2D Vision In semantic segmentation, visual scale, a direct function of the embodied agent's state, manifests as a severe class representation problem. For a drone, this is not a static property of a dataset but a dynamic, per-frame reality. Consider an autonomous drone performing a surveillance task. At a low altitude of 10 meters, a pedestrian is a salient object, occupying thousands of pixels. In this context, the class imbalance between the pedestrian and the sprawling *road* class is manageable. However, as the drone ascends to a cruising altitude of 80 meters to survey the area, its state changes, and the perception problem transforms entirely. The same pedestrian shrinks to a mere speck of 10-20 pixels. The instantaneous class imbalance within this high-altitude frame has now skyrocketed from perhaps 50:1 to over 1000:1. A model trained with a static balancing parameter is optimized for the average imbalance of its training set and is fundamentally unprepared for this extreme state-dependent shift. It will likely fail to classify the 20-pixel blob, treating it as noise, a safety-critical failure. This localized, state-dependent phenomenon explains the kind of large-scale performance deltas observed in recent research [25]. For instance, baseline models in continual learning often show a performance gap of over 25 mIoU points [25] between majority and minority classes on datasets like ADE20K [38]. This demonstrates that model performance already varies with object scale, which is itself dependent on the embodied agent's state. Current static balancing methods cannot dynamically adapt to this vulnerability.

These two issues mentioned above, one a 3D geometric failure, the other a 2D semantic failure, are two sides of the same coin. Both are symptoms of a state-agnostic model's inability to adapt to a change in its physical distance from the scene.

3 Our Position: Towards State-Adaptive Perception

Given the fundamental limitations of the state-agnostic paradigm, we posit that achieving the next level of robustness requires a paradigm shift. We must move beyond simply building larger, more diverse datasets and instead focus on creating perception systems that are intrinsically aware of their own states. This leads us to a new mandate for robotic vision research, a call to action centered on three core principles:

First, we must move beyond adapting to static, appearance-based domains and instead focus on adapting to the continuous, state-dependent domains generated by the embodied agent's own motion and interaction with the world. The current conception of a ***domain*** as a discrete category (e.g., synthetic, real, rainy) is a useful but ultimately coarse abstraction. For an embodied agent, the ***domain*** is not static but a continuous function of its state vector. The shift from a low-altitude to a high-altitude view is not a jump between two domains but a trajectory through a continuous space of visual scales. This perspective requires us to rethink our adaptation strategies. Rather than learning mappings between a few pre-defined points in domain space, we should develop methods that can smoothly and continuously adapt along these state-dependent axes. Some recent efforts address viewpoint changes in the context of segmentation [26, 36] and depth estimation [35], but they still frame the problem as adaptation among static datasets rather than as a continuous change in view and altitude. Therefore, research should therefore prioritize models that can generalize their adaptive behavior across the full spectrum of an embodied agent's operational states.

Second, we must prioritize research into new perception architectures with internal, dynamic mechanisms that are modulated by the embodied agent's state. A truly adaptive system cannot be a monolithic black box; it must possess explicit structures that allow its computational strategy to

change in response to its context. This is a call to explore a new class of architectures where the state is not just another input feature, but a control signal that governs the computation itself. Such a principle opens up numerous promising research directions, including:

- **state-adaptive Loss Functions**: We can design learning objectives that dynamically re-weight their own terms based on the current state. For example, a system could learn to infer its distance from a scene and use that inference to apply a more aggressive balancing factor to its loss function, forcing it to prioritize small, distant objects that are at high risk of being missed.
- **State-Modulated Attention**: We can develop attention mechanisms whose queries, keys, or values are conditioned on the embodied agent's state. This would allow a model to focus its capacity on different types of features based on its context—for instance, prioritizing fine-grained textures during a close-up inspection versus coarse object shapes during high-speed navigation.
- **Adaptive Computational Graphs**: We can explore models that can dynamically alter their own structure in a principled way, activating or deactivating certain layers or modules based on the mission phase or energy constraints.

Finally, we must establish new benchmarks and evaluation protocols that measure a model's ability to handle state-dependent domain shifts. Current benchmarks, which report a single, aggregate performance metric (e.g., mIoU), are not just insufficient; they are potentially misleading. Such metrics incentivize models that perform well on average, a strategy that can mask catastrophic failures in specific but common operational states, creating a false sense of security. To foster genuine progress, we must advocate for a move towards state-stratified evaluation. A benchmark for a drone perception model, for example, should not just report a single mIoU but should explicitly report performance binned by altitude, velocity, or viewpoint. Similarly, an autonomous vehicle benchmark should stratify results by the vehicle's speed and the distance to objects of interest (near, mid, and far-field performance). This granular approach provides a much clearer and more honest picture of a model's true robustness, revealing failure modes that are currently averaged away. More importantly, such a framework would force the community to confront the unified nature of these challenges, directly assessing both semantic failures (like class imbalance) and geometric failures (like scale inconsistency) under a common, robotics-centric lens, thereby encouraging the development of more holistic and truly robust perception systems.

These principles of state-adaptive perception extend far beyond the above examples. We believe this is a rich and vital area for robotics research. We encourage the community to explore **other** forms of state-adaptive perception, such as:

- **Motion-Adaptive Perception**: Models that dynamically adjust their temporal fusion strategies or allocate more capacity to de-blurring features when the embodied agent is moving at high speed.
- **Energy-Adaptive Perception**: For long-duration missions, models that can scale down their own computational graph and operate in a lower-power, lower-fidelity mode when the agent's battery is low.
- **Interaction-Adaptive Perception**: A manipulation embodied agent whose visual system changes its feature extraction strategy based on whether it is in a pre-grasp (object recognition) or post-grasp (slip detection) phase.

4 Conclusion

The current paradigm of training large, static models and hoping they generalize to all possible robotic scenarios has brought us far, but it is reaching its limits, especially for robot-deployable vision models. For embodied agents to become truly robust and trustworthy, they must be able to reason about their own state and adapt their perception strategies accordingly. We have argued for a shift towards state-adaptive perception and have presented a concrete, practical example of how these principles can be applied to solve a safety-critical problem in mobile robotics. By building models that are not just robust in general, but adaptive in particular, we posit that to move beyond current limitations in robotic robustness, the community should shift its focus from building static, universal models to creating state-adaptive perception systems.

References

[1] Aburaya, A., Selamat, H., Muslim, M.T.: Review of vision-based reinforcement learning for drone navigation. International Journal of Intelligent Robotics and Applications **8**(4), 974–992 (2024)

[2] Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)

[3] Chen, T., Shorinwa, O., Bruno, J., Swann, A., Yu, J., Zeng, W., Nagami, K., Dames, P., Schwager, M.: Splat-nav: Safe real-time robot navigation in gaussian splatting maps. IEEE Transactions on Robotics (2025)

[4] Du, Z., Li, X., Li, F., Lu, K., Zhu, L., Li, J.: Domain-agnostic mutual prompting for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23375–23384 (2024)

[5] Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)

[6] Guan, Y., Liao, H., Li, Z., Hu, J., Yuan, R., Zhang, G., Xu, C.: World models for autonomous driving: An initial survey. IEEE Transactions on Intelligent Vehicles (2024)

[7] Han, X., Chen, S., Fu, Z., Feng, Z., Fan, L., An, D., Wang, C., Guo, L., Meng, W., Zhang, X., et al.: Multimodal fusion and vision-language models: A survey for robot vision. Information Fusion p. 103652 (2025)

[8] Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning. pp. 1989–1998. Pmlr (2018)

[9] Hu, J., Fan, C., Zhou, L., Gao, Q., Liu, H., Lam, T.L.: Lifelong-monodepth: Lifelong learning for multidomain monocular metric depth estimation. IEEE Transactions on Neural Networks and Learning Systems (2023)

[10] Ji, Y., Tan, H., Shi, J., Hao, X., Zhang, Y., Zhang, H., Wang, P., Zhao, M., Mu, Y., An, P., et al.: Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1724–1734 (2025)

[11] Liang, Y.S., Li, W.J.: Inflora: Interference-free low-rank adaptation for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23638–23647 (2024)

[12] Liu, S., Lv, J., Kang, J., Zhang, H., Liang, Z., He, S.: Modfinity: Unsupervised domain adaptation with multimodal information flow intertwining. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5092–5101 (2025)

[13] Liu, W., Zhu, F., Wei, L., Tian, Q.: C-clip: Multimodal continual learning for vision-language model. In: The Thirteenth International Conference on Learning Representations (2025)

[14] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

[15] Matsuki, H., Murai, R., Kelly, P.H., Davison, A.J.: Gaussian splatting slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18039–18048 (2024)

[16] Park, H., Gupta, A., Wong, A.: Test-time adaptation for depth completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20519–20529 (2024)

[17] Pei, R., Deng, S., Zhou, L., Qin, H., Liang, Q.: Mcs-resnet: A generative robot grasping network based on rgb-d fusion. IEEE Transactions on Instrumentation and Measurement (2024)

[18] Rajapaksha, U., Sohel, F., Laga, H., Diepeveen, D., Bennamoun, M.: Deep learning-based depth estimation methods from monocular image and videos: A comprehensive survey. ACM computing surveys **56**(12), 1–51 (2024)

[19] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)

[20] Shaheen, K., Hanif, M.A., Hasan, O., Shafique, M.: Continual learning for real-world autonomous systems: Algorithms, challenges and frameworks. Journal of Intelligent & Robotic Systems **105**(1), 9 (2022)

[21] Shen, F., Zhou, L., Kuecuekaytekin, K., Eskandar, G.B.F., Liu, Z., Wang, H., Knoll, A.: W-controluda: Weather-controllable diffusion-assisted unsupervised domain adaptation for semantic segmentation. IEEE Robotics and Automation Letters (2025)

[22] Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)

[23] Sun, L., Bian, J.W., Zhan, H., Yin, W., Reid, I., Shen, C.: Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. IEEE transactions on pattern analysis and machine intelligence **46**(1), 497–508 (2023)

[24] Tong, L., Song, K., Tian, H., Man, Y., Yan, Y., Meng, Q.: A novel rgb-d cross-background robot grasp detection dataset and background-adaptive grasping network. IEEE Transactions on Instrumentation and Measurement **73**, 1–15 (2024)

[25] Truong, T.D., Prabhu, U., Raj, B., Cothren, J., Luu, K.: Falcon: Fairness learning via contrastive attention approach to continual semantic scene understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15065–15075 (2025)

[26] Truong, T.D., Prabhu, U., Wang, D., Raj, B., Gauch, S., Subbiah, J., Luu, K.: Eagle: Efficient adaptive geometry-based learning in cross-view understanding. Advances in Neural Information Processing Systems **37**, 137309–137333 (2024)

[27] Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017)

[28] Wang, Q., Dai, D., Hoyer, L., Van Gool, L., Fink, O.: Domain adaptive semantic segmentation with self-supervised depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8515–8525 (2021)

[29] Wei, S., Geng, H., Chen, J., Deng, C., Wenbo, C., Zhao, C., Fang, X., Guibas, L., Wang, H.: D$\ominus$3roma: Disparity diffusion-based depth sensing for material-agnostic robotic manipulation. In: ECCV 2024 Workshop on Wild 3D: 3D Modeling, Reconstruction, and Generation in the Wild (2024)

[30] Yan, C., Qu, D., Xu, D., Zhao, B., Wang, Z., Wang, D., Li, X.: Gs-slam: Dense visual slam with 3d gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19595–19604 (2024)

[31] Yang, X., Yan, W., Yuan, Y., Mi, M.B., Tan, R.T.: Semantic segmentation in multiple adverse weather conditions with domain knowledge retention. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6558–6566 (2024)

[32] Yang, Y., Zhou, J., Ding, X., Huai, T., Liu, S., Chen, Q., Xie, Y., He, L.: Recent advances of foundation language models-based continual learning: A survey. ACM Computing Surveys **57**(5), 1–38 (2025)

[33] Yin, P., Jiao, J., Zhao, S., Xu, L., Huang, G., Choset, H., Scherer, S., Han, J.: General place recognition survey: Towards real-world autonomy. IEEE Transactions on Robotics (2025)

[34] Zhang, W., Lv, Z., Zhou, H., Liu, J.W., Li, J., Li, M., Li, Y., Zhang, D., Zhuang, Y., Tang, S.: Revisiting the domain shift and sample uncertainty in multi-source active domain transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16751–16761 (2024)

[35] Zhang, Y., Rafique, M.U., Christie, G., Jacobs, N.: Crossadapt: cross-scene adaptation for multi-domain depth estimation. In: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. pp. 5328–5331. IEEE (2023)

[36] Zhang, Y., Rafique, M.U., Jacobs, N.: Crossseg: Cross-scene few-shot aerial segmentation using probabilistic prototypes. In: IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. pp. 5006–5009. IEEE (2023)

[37] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)

[38] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)

[39] Zhu, S., Mou, L., Li, D., Ye, B., Huang, R., Zhao, H.: Vr-robo: A real-to-sim-to-real framework for visual robot navigation and locomotion. IEEE Robotics and Automation Letters (2025)