Deep Learning for Active Robotic Perception

Nikolaos Passalis[a], Pavlos Tosidis, Theodoros Manousis and Anastasios Tefas[b]
Computational Intelligence and Deep Learning Group, AIIA Lab.,
Department of Informatics, Aristotle University of Thessaloniki, Greece

Keywords: Active Perception, Deep Learning, Active Vision, Active Robotic Perception.

Abstract: Deep Learning (DL) has brought significant advancements in recent years, greatly enhancing various challenging computer vision tasks. These tasks include but are not limited to object detection and recognition, scene segmentation, and face recognition, among others. DL's advanced perception capabilities have also paved the way for powerful tools in the realm of robotics, resulting in remarkable applications such as autonomous vehicles, drones, and robots capable of seamless interaction with humans, such as collaborative manufacturing. However, despite these remarkable achievements in DL within these domains, a significant limitation persists: most existing methods adhere to a static inference paradigm inherited from traditional computer vision pipelines. Indeed, DL models typically perform inference on a fixed and static input, ignoring the fact that robots possess the capability to interact with their environment to gain a better understanding of their surroundings. This process, known as ”active perception”, closely mirrors how humans and various animals interact and comprehend their environment. For instance, humans tend to examine objects from different angles, when being uncertain, while some animals have specialized muscles that allow them to orient their ears towards the source of an auditory signal. Active perception offers numerous advantages, enhancing both the accuracy and efficiency of the perception process. However, incorporating deep learning and active perception in robotics also comes with several challenges, e.g., the training process often requires interactive simulation environments and dictates the use of more advanced approaches, such as deep reinforcement learning, the deployment pipelines should be appropriately modified to enable control within the perception algorithms, etc. In this paper, we will go through recent breakthroughs in deep learning that facilitate active perception across various robotics applications, as well as provide key application examples. These applications span from face recognition and pose estimation to object detection and real-time high-resolution analysis.

1 INTRODUCTION

In recent years, Deep Learning (DL) has led to significant advancements in a range of challenging computer vision and robotic perception tasks (LeCun et al., 2015). These tasks encompass but are not restricted to object detection and recognition (Redmon et al., 2016), scene segmentation (Badrinarayanan et al., 2017), and face recognition (Wen et al., 2016), among others. DL's sophisticated perception capabilities have also yielded potent tools for diverse robotics applications, resulting in the emergence of impressive use cases, such as self-driving vehicles (Bojarski et al., 2016), unmanned aerial vehicles (drones) (Passalis et al., 2018), and robots capable of seamless interaction with humans, notably in collaborative manufacturing scenarios (Liu et al., 2019).

Despite the recent accomplishments of DL in these domains, a notable limitation plagues most existing approaches since they adhere to a static inference paradigm, which follows the traditional computer vision pipeline. Therefore, DL models perform inference on a fixed and static input, ignoring the ability of robots, as well as cyber-physical systems (Li, 2018; Loukas et al., 2017), to *interact* with their environment in order to enhance their perception. For example, we can consider the task of face recognition, where a robot captures a suboptimal profile view of a subject to be recognized. A conventional static perception-based DL model may struggle to identify the subject from a specific angle, particularly if it lacks training on profile face images for such angles. However, it is often feasible for the robot to attain a better and more distinguishing view by adjusting its position relative to the human subject. Consequently,

[a] https://orcid.org/0000-0003-1177-9139
[b] https://orcid.org/0000-0003-1288-3667

Passalis, N., Tosidis, P., Manousis, T. and Tefas, A.
Deep Learning for Active Robotic Perception.
DOI: 10.5220/0012295800003595
In *Proceedings of the 15th International Joint Conference on Computational Intelligence (IJCCI 2023)*, pages 15-21
ISBN: 978-989-758-674-3; ISSN: 2184-3236

in such scenarios, the same DL model will likely succeed in recognizing the subject after the robot repositions itself for a more suitable angle. This methodology, known as *active perception* (Aloimonos, 2013; Bajcsy et al., 2018; Shen and How, 2019), enables the manipulation of the robot or sensor to obtain a clearer and more informative view or signal, ultimately enhancing the perception capabilities and situational awareness of robotic systems. Note that this process closely mirrors how humans and various animals engage with and percept their surroundings. For instance, humans tend to explore different perspectives when processing complex visual stimuli, while many mammals possess specialized ear muscles that pivot their ears toward the source of an auditory signal in order to acquire a clearer version of the signal (Heffner and Heffner, 1992).

A number of recent, although relatively basic, approaches have illustrated that active perception can indeed enhance the perceptual capabilities of various models. For example, works such as (Ammirato et al., 2017) and (Passalis and Tefas, 2020), demonstrated that developing a deep learning system that predicts the next best move for a robot can significantly improve the accuracy of various perception tasks, such as object detection and face recognition, where the viewing angle, occlusions and the scale of each object can have a significant effect on the perception accuracy. Similar findings have also been documented in more recent research spanning a variety of domains (Han et al., 2019; Tosidis et al., 2022; Kakaletsis and Nikolaidis, 2023). It is also important to emphasize that active perception methodologies can also enable the development of less computationally intensive deep learning models. This occurs because these models are trained to address a less complex problem. For example, in (Passalis and Tefas, 2020), it is demonstrated that more lightweight face recognition models can be used when DL models can actively interact with the environment in order to acquire a more informative frontal view of the subjects.

However, training active perception models differs significantly compared to traditional static perception approaches, since models must learn also the dynamics of the perception process in order to provide control feedback. For example, an active face recognition model should also learn how perception accuracy varies as the robot moves around a subject, as well as the direction in which a robot should move in order to improve the accuracy of face recognition. Therefore, it becomes clear that training active perception models introduces additional challenges, both with respect to acquiring the necessary data for training, as well as for extending the traditional (usually supervised) learning pipelines to support such setups.

The main aim of this paper is to introduce the main active perception approaches used for training DL-based active perception models for different applications. To this end, we will first present and discuss the different options for acquiring the necessary data used for training active perception models. Next, we will present different training approaches that extend traditional supervised learning methods for active perception or employ reinforcement learning methods to provide active perception feedback. Finally, we will discuss applications in various fields related to robotics, as well as discuss implications and practical issues.

The rest of this paper is structured as follows. First, in Section 2 we present the different methodologies for acquiring data for active perception, while in Section 3 we present different learning approaches that are used for training active perception DL models. Then, in Section 4, we provide an overview of applications for different perception applications. Finally, Section 5 concludes this paper.

2 DATA FOR ACTIVE PERCEPTION

As discussed in Section 1, training active perception models requires a shift from traditional static perception methods, presenting a distinctive challenge. This distinction arises from the necessity for active perception models to not only grasp the static aspects of object recognition but also to encompass the dynamics inherent in the perception process, allowing them to generate control feedback. For instance, when considering an active face recognition model, the model should acquire knowledge concerning the optimal direction in which the robot should navigate to enhance the accuracy of face recognition. Unfortunately, a notable constraint emerges as a significant portion of the available datasets does not inherently facilitate the training of models for active perception tasks. Current literature can be roughly categorized into three distinct methodologies that can be used for getting data suitable for active perception: a) simulation-based training, b) multi-view dataset-based training, and c) on-demand (synthetic) data generation. An overview of the different approaches, among with benefits and drawbacks, is provided in Table 1.

Ideally, an active perception model would learn as it interacts with its environment. However, getting ground truth data in real-time is typically infeasible. Therefore, in most cases, active perception

Table 1: Comparing different approaches that can be used for acquiring data that can be used for training active perception models.

Approach	Benefits	Drawbacks	Examples
Simulation-based training	flexible, any movement can be simulated	computationally-demanding, sim-to-real gap	(Ginargyros et al., 2023; Tzimas et al., 2020; Tosidis et al., 2022)
Multi-view dataset-based training	real data used, no sim-to-real gap	limited flexibility, limited number of control actions, missing data	(Passalis and Tefas, 2020; Georgiadis et al., 2023)
On-demand (synthetic) data generation through manipulation	less susceptible to sim-to-real gap, faster than simulation-based training	less accurate simulation of control actions, perception dynamics might not be accurately modeled	(Dimaridou et al., 2023; Passalis and Tefas, 2021; Kakaletsis and Nikolaidis, 2023; Manousis et al., 2023; Bozinis et al., 2021)

models are trained in an offline fashion. The first category of methods employs realistic simulation environments, e.g., as in (Tosidis et al., 2022; Ginargyros et al., 2023), in order to simulate the effect of various movements and allow the agent to learn how perception accuracy varies when performing different actions. This approach provides great flexibility since any action can be simulated and the effect of the movement of a robot can be easily obtained. However, such approaches are computationally demanding, since they rely on realistic simulation environments and graphics engines, such as Webots (Michel, 2004) and Unity (Haas, 2014), slowing down the training process. Furthermore, these approaches are also hindered by the so-called "sim-to-real" gap (Zhao et al., 2020), since the agents are trained using data generated by a simulator.

The second category of approaches, called "multi-view dataset-based training" in this paper, employs datasets that contain multiple views of the same scene. In this way, the effect of various movements can be quantified by fetching the view that would correspond to the result of the said movement. For example, in (Passalis and Tefas, 2020), the multiple views around a person are used to simulate the effect of an agent moving around, enabling training active perception models that learn how to maximize face recognition accuracy. Such approaches can overcome the issues of computational complexity and the "sim-to-real" gap. However, they are often too restrictive, since the datasets should already contain the images that can be used for every possible action an agent can perform. This often leads to huge datasets, as well as to agents that can be trained for a limited number of control actions. Furthermore, such approaches often have to handle missing data, since, in many cases, there are missing data in the corresponding multi-view datasets.

Then, methods that generate "on-demand" data have also been proposed. Such approaches can try to simulate the effect of various movements starting from real data and then appropriately manipulating the data, e.g., simulating occlusions (Dimaridou et al., 2023). Another approach is to generate multiple views that can then be used either for deciding the best course of action or training the agent (Kakaletsis and Nikolaidis, 2023). These approaches fall in between simulation-based and multi-view dataset-based approaches since they employ real images that have been appropriately manipulated to simulate the effect of active perception. Therefore, even though they are less susceptible to the sim-to-real gap and they are typically faster, they often provide less accuracy in simulating the effect of active perception feedback, leading to models that might fail to capture all details of the dynamics of the active perception process.

3 LEARNING METHODOLOGIES FOR ACTIVE PERCEPTION

Training active perception models also departs from the typical supervised learning approach that is followed in many perception applications, such as face recognition (Wen et al., 2016), object detection (Redmon et al., 2016) and pose estimation (Zheng et al., 2023). Active perception models should not only analyze and understand their input but also provide some kind of control feedback, that can be then subsequently used for improving perception accuracy. Therefore, they tend to incorporate elements typically found in planning (Sun et al., 2021) and control (Tsounis et al., 2020) approaches used in robotics applications. The degree to which such elements are part of each model depends on the specific application requirements. In recent literature, two ap-

Table 2: Comparing different learning paradigms that can be used training active perception models.

Approach	Benefits	Drawbacks	Examples
Deep Reinforcement Learning	directly optimizes the active perception model	slow convergence, low sample efficiency, (usually) requires simulation environments	(Bozinis et al., 2021; Tzimas et al., 2020; Tosidis et al., 2022)
Supervised Learning	can work with any kind of data, easier and faster to train	requires carefully design heuristics to construct ground truth data	(Passalis and Tefas, 2020; Ginargyros et al., 2023; Dimaridou et al., 2023; Manousis et al., 2023)

proaches are prevalent: a) deep reinforcement learning (DRL)-based training and b) supervised training through carefully designed ground truth. An overview of these two different approaches, along with benefits and drawbacks, is provided in Table 2.

DRL has achieved remarkable progress in recent years, providing beyond human performance in many cases (Mnih et al., 2013). Such approaches naturally fit active perception, since they enable models to learn how to provide control feedback to maximize perception accuracy through the interaction with an environment. Such approaches almost exclusively require the use of a simulation environment to be trained. Even though DRL methods enable discovering complex policies that can directly optimize the objective at hand, i.e., perception accuracy, they suffer from low sample efficiency, long training times, and unstable convergence (Buckman et al., 2018). On the other hand, the supervised method typically follows an “imitation” learning training paradigm (Hua et al., 2021), where the best actions to be performed are found through an extensive search in the action space. This is better understood with the following example. A DRL-agent training to perform active face recognition, e.g., (Tosidis et al., 2022), would learn using the reward signal from the environment, e.g., confidence in correctly recognizing a person. On the other hand, a supervised approach, such as (Passalis and Tefas, 2020), would first require simulating the effect of various movements/actions and then provide ground truth data on which action should the agent perform at each step. This also enables supervised approaches to work with any kind of data available, since the actions to be evaluated can be dictated by the capacity of the dataset to support the corresponding action. Therefore, even though supervised approaches can provide more stable and faster convergence and typically do not require a complex simulation environment, they rely on hand-crafted heuristic-based approaches to constructing the ground truth data.

4 ACTIVE PERCEPTION FOR ROBOTIC APPLICATIONS

Several recent active perception approaches have been proposed for a variety of different applications. In the rest of this Section, we briefly overview methods proposed for different applications, as well as discuss practical issues that often arise in robotics. Among the most prominent applications of active perception is face recognition. Indeed, early DL-based approaches extended embedding-based active perception methods into active ones by including an additional head that predicts the next best movement that a robot should perform in order to increase face recognition confidence (Passalis and Tefas, 2020). This approach assumes that the robot moves on a predefined trajectory around the target in order to be compatible with the multi-view dataset employed. Then, the model is trained to both maximize face recognition confidence, following a constrastive learning objective, as well as to regress the direction of movement leading to the best face recognition accuracy. Note that this direction is calculated by leveraging the multiple views available in the dataset and then selecting the one that maximizes the confidence for the next active perception step. The experimental evaluation demonstrated the effectiveness of this approach over static perception for a variety of different active perception steps. However, this approach used a dataset with a small number of individuals and a relatively small number of possible control movements. Later methods, such as (Dimaridou et al., 2023), build upon this approach by a) simulating the effect of various occlusions on large-scale face recognition datasets, and b) regressing both the direction and distance the robot should move.

A simple DRL approach for training a DRL agent to perform drone control in order to acquire frontal views that can be used for face recognition was initially proposed in (Tzimas et al., 2020), highlighting the potential of DRL methods for active perception tasks. A more sophisticated approach was also recently proposed building upon DRL in (Tosidis et al., 2022). This approach leverages a realistic simulation environment, built using Webots (Michel, 2004), and directly trains a DRL agent to perform control in a drone that flies around humans in order to maximize face recognition confidence. The experimental evaluation demonstrated that the trained agent was able to perform control in a variety of different situations. However, the sim-to-real gap remains with DRL approaches, which can be a limiting factor in directly applying such approaches in real applications.

A supervised approach was also proposed for object detection in (Ginargyros et al., 2023), where a rich dataset for potential movements was built using a simulation environment. This approach enabled the models to learn the object detection confidence manifold for different types of objects, e.g., cars and humans, while taking into account possible occlusions, allowing them to perform control tailored to the unique characteristics of different cases. To this end, a separate navigation proposal network was trained according to the confidence manifold of each object, enabling the model to learn to propose trajectories that will maximize object detection confidence. At the same time, this paper also revealed limitations that are often intrinsic to the current state-of-the-art object detection models, since it provided a structured approach for revealing the confidence manifold of object detectors. A dataset that can support active vision for object detection was also proposed in (Ammirato et al., 2017), and a DRL agent was also trained and evaluated. This paper demonstrated that it is possible, given the appropriate dataset and annotations, to directly train DRL agents to perform control for active vision tasks.

Another line of research focuses on performing *virtual control*, i.e., not physically altering the position of a robot or the parameters of a physical sensor, but rather selectively analyzing specific parts of the input in order to improve perception accuracy, while reducing the computational load. Such approaches can be especially useful in cases where high-resolution input images must be analyzed, while the object of interest lies only in a small area within the input. An especially promising approach was presented in (Manousis et al., 2023), where the heat map extracted from a low-resolution version of a high-resolution image was used to drive the perception process. To this end, the proposed method first identified a region of interest in the original image by looking for potential activations (i.e., parts where the DL model detects something, but not necessarily with high confidence), in a low-resolution version of the input, and then performed targeted cropping into the high-resolution image in order to select the area that needs to be analyzed. The experimental evaluation demonstrated that significant accuracy and speed improvements can be acquired using the proposed method. However, a limitation of such approaches is that as the size of the region of interest grows, the performance benefit obtained using active perception is becoming smaller. It is worth noting that such approaches can be also easily adjusted to perform control of the parameters of a camera, e.g., physical zoom, in order to acquire signals that are easier to analyze.

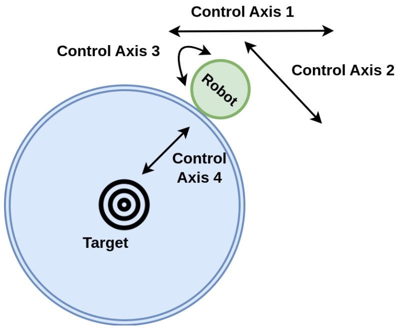

Figure 1: Active perception outputs can be represented in a homogeneous way using an application agnostic control specification defined by OpenDR (Passalis et al., 2022).

Another significant issue when implementing active perception models is the existence of a common way of expressing the outcomes of active perception. This is especially important, since in many robotics systems, different models might be employed for different perception tasks. Having to handle a completely different form of output for different models significantly complicates the development process. OpenDR toolkit (Passalis et al., 2022) has provided a common application agnostic control specification for standardizing such active perception outputs. This specification ensures that algorithms designed for active perception can effectively process the result. To this end, four control axes have been identified, as shown in Fig. 1. For all axes, it is assumed that the robot moves in a sphere and a real value from -1 to 1 is provided for the movement on each axis. Using this way of expressing the output of active perception approaches holds the credentials for simplifying the development of active perception-enable robotics systems, by enabling the the efficient re-use of components related to handling and executing the feedback provided by active perception algorithms.

5 CONCLUSIONS

DL has revolutionized computer vision and robotics by enabling remarkable advancements in perception tasks. However, as discussed in this paper, a significant limitation persists in many existing DL-based systems: the static inference paradigm. Most DL models operate on fixed, static inputs, neglecting the potential benefits of active perception – a process that mimics how humans and certain animals interact with their environment to better understand it. Active perception offers advantages in terms of accuracy and efficiency, making it a crucial area of exploration for enhancing robotic perception. While the incorporation of deep learning and active perception in robotics presents numerous opportunities, it also poses several challenges. Training often necessitates interactive simulation environments and more advanced approaches like deep reinforcement learning. Moreover, deployment pipelines need to be adapted to enable control within perception algorithms. These challenges highlight the importance of ongoing research and development in this field.

ACKNOWLEDGMENTS

This work was supported by the European Union's Horizon 2020 Research and Innovation Program (OpenDR) under Grant 871449. This publication reflects the authors' views only. The European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

Aloimonos, Y. (2013). *Active perception*. Psychology Press.

Ammirato, P., Poirson, P., Park, E., Košecká, J., and Berg, A. C. (2017). A dataset for developing and benchmarking active vision. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1378–1385.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495.

Bajcsy, R., Aloimonos, Y., and Tsotsos, J. K. (2018). Revisiting active perception. *Autonomous Robots*, 42(2):177–196.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Bozinis, T., Passalis, N., and Tefas, A. (2021). Improving visual question answering using active perception on static images. In *Proceedings of the International Conference on Pattern Recognition*, pages 879–884.

Buckman, J., Hafner, D., Tucker, G., Brevdo, E., and Lee, H. (2018). Sample-efficient reinforcement learning with stochastic ensemble value expansion. *Proceedings of the Advances in Neural Information Processing Systems*, 31.

Dimaridou, V., Passalis, N., and Tefas, A. (2023). Deep active robotic perception for improving face recognition under occlusions. In *Proceedings of the IEEE Symposium Series on Computational Intelligence (accepted)*, page 1.

Georgiadis, C., Passalis, N., and Nikolaidis, N. (2023). Activeface: A synthetic active perception dataset for face recognition. In *Proceedings of the International Workshop on Multimedia Signal Processing (accepted)*, page 1.

Ginargyros, S., Passalis, N., and Tefas, A. (2023). Deep active perception for object detection using navigation proposals. In *Proceedings of the IEEE Symposium Series on Computational Intelligence (accepted)*, page 1.

Haas, J. K. (2014). A history of the unity game engine.

Han, X., Liu, H., Sun, F., and Zhang, X. (2019). Active object detection with multistep action prediction using deep q-network. *IEEE Transactions on Industrial Informatics*, 15(6):3723–3731.

Heffner, R. S. and Heffner, H. E. (1992). Evolution of sound localization in mammals. In *The evolutionary biology of hearing*, pages 691–715.

Hua, J., Zeng, L., Li, G., and Ju, Z. (2021). Learning for a robot: Deep reinforcement learning, imitation learning, transfer learning. *Sensors*, 21(4):1278.

Kakaletsis, E. and Nikolaidis, N. (2023). Using synthesized facial views for active face recognition. *Machine Vision and Applications*, 34(4):62.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Li, J.-h. (2018). Cyber security meets artificial intelligence: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(12):1462–1474.

Liu, Q., Liu, Z., Xu, W., Tang, Q., Zhou, Z., and Pham, D. T. (2019). Human-robot collaboration in disassembly for sustainable manufacturing. *International Journal of Production Research*, 57(12):4027–4044.

Loukas, G., Vuong, T., Heartfield, R., Sakellari, G., Yoon, Y., and Gan, D. (2017). Cloud-based cyber-physical intrusion detection for vehicles using deep learning. *IEEE Access*, 6:3491–3508.

Manousis, T., Passalis, N., and Tefas, A. (2023). Enabling high-resolution pose estimation in real time using active perception. In *Proceedings of the IEEE International Conference on Image Processing*, pages 2425–2429.

Michel, O. (2004). Cyberbotics ltd. webots™: professional mobile robot simulation. *International Journal of Advanced Robotic Systems*, 1(1):5.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M.

(2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Passalis, N., Pedrazzi, S., Babuska, R., Burgard, W., Dias, D., Ferro, F., Gabbouj, M., Green, O., Iosifidis, A., Kayacan, E., et al. (2022). OpenDR: An open toolkit for enabling high performance, low footprint deep learning for robotics. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 12479–12484.

Passalis, N. and Tefas, A. (2020). Leveraging active perception for improving embedding-based deep face recognition. In *Proceedings of the IEEE International Workshop on Multimedia Signal Processing*, pages 1–6.

Passalis, N. and Tefas, A. (2021). Pseudo-active vision for improving deep visual perception through neural sensory refinement. In *Proceedings of the IEEE International Conference on Image Processing*, pages 2763–2767.

Passalis, N., Tefas, A., and Pitas, I. (2018). Efficient camera control using 2d visual information for unmanned aerial vehicle-based cinematography. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, pages 1–5.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788.

Shen, M. and How, J. P. (2019). Active perception in adversarial scenarios using maximum entropy deep reinforcement learning. In *Proceedings of the International Conference on Robotics and Automation*, pages 3384–3390. IEEE.

Sun, H., Zhang, W., Yu, R., and Zhang, Y. (2021). Motion planning for mobile robots—focusing on deep reinforcement learning: A systematic review. *IEEE Access*, 9:69061–69081.

Tosidis, P., Passalis, N., and Tefas, A. (2022). Active vision control policies for face recognition using deep reinforcement learning. In *Proceedings of the 30th European Signal Processing Conference*, pages 1087–1091.

Tsounis, V., Alge, M., Lee, J., Farshidian, F., and Hutter, M. (2020). Deepgait: Planning and control of quadrupedal gaits using deep reinforcement learning. *IEEE Robotics and Automation Letters*, 5(2):3699–3706.

Tzimas, A., Passalis, N., and Tefas, A. (2020). Leveraging deep reinforcement learning for active shooting under open-world setting. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 499–515.

Zhao, W., Queralta, J. P., and Westerlund, T. (2020). Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *Proceedings of the IEEE Symposium Series on Computational Intelligence*, pages 737–744.

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., and Shah, M. (2023). Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37.