Enhancing Robot Learning through Learned Human-Attention Feature Maps

Daniel Scheuchenstuhl* — *CPS Group*, *TU Wien*, Wien, Austria, 0009-0000-6080-2898

Stefan Ulmer* — *CPS Group*, *TU Wien*, Wien, Austria, 0009-0009-7654-7743

Felix Resch* — *CPS Group*, *TU Wien*, Wien, Austria, 0009-0004-3240-3725

Luigi Berducci — *CPS Group*, *TU Wien*, Wien, Austria, 0000-0002-3497-6007

Radu Grosu — *CPS Group*, *TU Wien*, Wien, Austria, 0000-0001-5715-2142

***Abstract*—Robust and efficient learning remains a challenging problem in robotics, in particular with complex visual inputs. Inspired by human attention mechanism, with which we quickly process complex visual scenes and react to changes in the environment, we think that embedding auxiliary information about focus point into robot learning would enhance efficiency and robustness of the learning process. In this paper, we propose a novel approach to model and emulate the human attention with an approximate prediction model. We then leverage this output and feed it as a structured auxiliary feature map into downstream learning tasks. We validate this idea by learning a prediction model from human-gaze recordings of manual driving in the real world. We test our approach on two learning tasks - object detection and imitation learning. Our experiments demonstrate that the inclusion of predicted human attention leads to improved robustness of the trained models to out-of-distribution samples and faster learning in low-data regime settings. Our work highlights the potential of incorporating structured auxiliary information in representation learning for robotics and opens up new avenues for research in this direction. All code and data are available online**[1]

***Index Terms*—Robot learning, Human attention**

I. Introduction

Robot learning has seen significant progress in recent years, developing systems able to perform increasingly complex tasks in a variety of challenging environments [1], [2]. However, the performance of the learning process often depends on the quality of representations, which retain the important features extracted from high-dimensional sensor data. Effective representation learning is therefore crucial for achieving high performance in robot learning tasks, and an increasing effort has been invested into this fundamental research area [3]–[5].

Most of the modern approaches to representation learning build on self-supervised learning and generative models [6]–[8], and have shown promising results in learning effective representations reusable in many different downstream tasks. However, we believe there is still much to be gained by taking inspiration from human behavior. Humans are able to process rich visual scenes and perform complex visual-motor tasks with great efficiency, due in part to the sophisticated attention mechanisms we employ [9]. Attention allows us to capture the most important features to accurately perform a task.

*Indicates authors with equal contributions

[1]Code/Data at https://github.com/CPS-TUWien/learning_human_attention

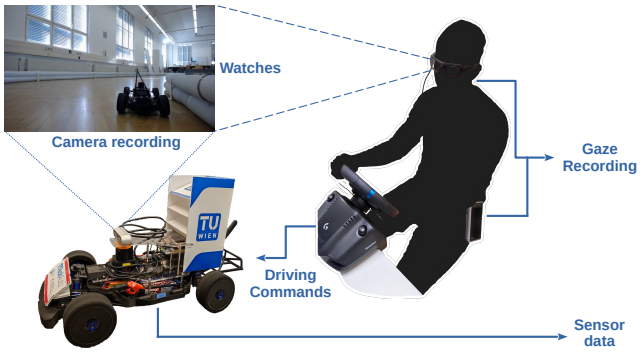

Fig. 1. Hardware setup to record human-gaze while manually driving a miniature racecar. It consists of a eye-tracking system, remote control of the vehicle and data acquisition system.

In this work, we build on this insight by developing a new model that mimics the human attention. However, instead of learning representations retaining the most salient features, we propose to enrich the input data with attentional feature maps.

To validate this idea, we collect real-world data of human gaze and train a model to distill human attention maps from a sequence of visual inputs. Having a model able to accurately predict attention maps, we can use them in unseen frame sequences, without any human in the loop.

Overall, our work proposes the following contributions:

- A novel model to predict human attention maps from visual input, trained on real-world data of human gaze collected during manual driving in scaled miniature cars.
- The integration of the learned representations based on human attention into two downstream robot-learning tasks of object detection and imitation learning.
- Experimental evaluation of the effectiveness of our approach by comparing it to existing methods that do not incorporate human-attention features.

In the rest of this paper, we will explain the proposed methodology and experimental results.

II. Related Works

In this section, we include the works related to our contributions. Considering our application in the context of au-

tonomous racing, we include a review of the existing work of learning-based approaches used in autonomous racing.

Attention models. Research on human attention dates back several decades and has been extensively studied due to its significant impact on learning and perception. In [9], the authors discuss the fundamental mechanism of attention and its advantageous effects on learning of humans. More recently, the notion of focus or attention has been adopted by the machine learning community [10], leading to many successful applications. Among them, notable mention is the self-attention mechanism adopted by transformer in natural language processing [11], emotion detection [12], or image recognition [13]. However, the connection between artificial attention mechanism commonly adopted in modern neural architectures and the human or biological attention is not clear. This motivates a new effort in research to explore how the two are related [14], [15]. In this direction, our contributions try to demonstrate the practical usability of human-based features.

Representation Learning. In recent years, there has been significant progress in the field of representation learning [16]. Most of the modern approaches rely on self-supervised learning and generative models [6]–[8], using variants of auto-encoders [17], [18] to learn a low-dimensional latent representation. The applications range from computer vision [19], [20], natural language [21], or multimodal inputs [22], [23]. Compared with these works, we do not use attention to provide a compact representation of the input, but instead propose to enrich it with additional attentional features.

Autonomous Racing. Considering our robotics experiments have been framed in the context of autonomous racing, we provide a review of existing approaches with focus on learning applications. A complete overview of this research field is provided in [24]. Despite the wide use of learning-based approaches for perception, even in racing competitions [25]–[27], most of the planning approaches currently deployed on hardware cars use either traditional control, such as model predictive control (MPC), or analytical approaches [28], [29]. However, an increase effort on learning-based control raised in the last years. Some works use learned models in conjunction with MPC [30], in an end-to-end fashion with imitation or reinforcement learning [31]–[35], or in a hierarchical framework [36]–[38] where a deep model generates trajectories that are then tracked with a low-level controller. In [39], MPC is combined with a novel attention mechanism that uses the generated trajectory to identify a region of interest on images.

III. Human-Attention Model

We first formalize the problem of reproducing the human-attention mechanism. Considering input images $x_t \in \mathbb{R}^{m \times n}$, we define an attention map $y_t \in \mathbb{R}^{m \times n}$ retaining all the focus points obtained by recording the human gaze.

The human-attention model M is then defined as

$$y_t = M(x_t)$$

Considering the sequential nature of the human-attention mechanism, which focuses on different areas of the image in a sequential way, and the limited capabilities of the recording system, which is able to capture up to a fixed number of points simultaneously, we propose an approximation of the attention maps with an exponential-decay time processing. Let $x_t \in \mathbb{R}^{m \times n}$ be the frame captured at time t, $P_t = \{p_{j,t}\}_j$ the set of focus points tracked with our system, and $h_t \in \mathbb{R}^{m \times n}$ the heatmap computed by centering a Gaussian distribution over each focus point $p \in P_t$. Since h_t only captures the instantaneous attention map of time step t, we aggregate it to produce y_t which resembles the complete attention focus as:

$$\begin{aligned} y_0 &= h_0 \\ y_t &= max(h_t, (1-r)y_{t-1}) \end{aligned}$$

where max refers to the pixel-wise max operation. In practice, we use $r = 0.17$ to cover all the focus point in the last second of recording, according to the frequency of our system. We normalize the attention maps to have likelihood values between 0 and 255.

Having formulated the input and output to model the human-attention mechanism, we now introduce the learning of an attention predictor that serves to exploit the use of attention beyond the data-collection setup. We frame the problem of learning an attention predictor as an image restoration problem which aims to produce approximated attention heatmaps. We build on top of the state-of-the-art model U-Net [40] and use a smooth L1 loss, as commonly adopted for image segmentation and related application domains. Specifically, we produce each attention map by centering an isotropic 2D Gaussian distribution with a zero mean and unit standard deviation over each focus point, and perform the aggregation described above. The use of Gaussian distribution is a smooth representation which resembles the likelihood of a driver's attendance to that location. The main advantage of it is to have a continuous output space where adjacent pixels are not independent.

Hardware Setup. In this section, we will provide a detailed description of the hardware setup used to collect a dataset of human attention while driving a miniature race car on various tracks. Since the primary objective of this experiment was to analyze human attention and behavior while driving, we mounted a camera on the car and asked human drivers to control the car through it by using either a steering wheel and pedals, or a game controller. To ensure that the driver's focus was entirely on the video stream captured by the camera and rely solely on it to control the car, we seated them away from the track.

To record the driver's gaze, we used a VPS 19 System eye-tracker. The eye-tracker recorded the driver's gaze points, which were then transformed into the camera's frame. All the gaze points outside the camera frame were discarded, to filter out when the drivers moved their head or looked around.

The miniature vehicle used in this experiment was an F1Tenth race car equipped with an inertial measurement unit, a 2D LiDAR, and an RGB camera. The LiDAR Hokuyo 10LX could sense distances up to 10 meters in a 270-degree field of view at a frequency of 40 Hz. To capture a diverse set

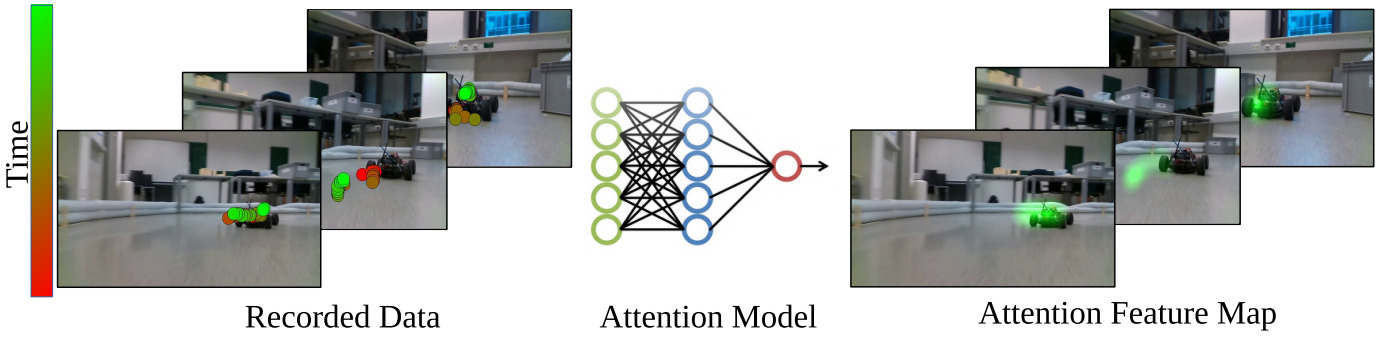

Fig. 2. The proposed human-attention model is trained to distill attention feature maps from visual input and learn the most-relevant focus area.

of images, we used two models of RGB camera, either a Logitech C930e or an Intel RealSense d435i. Both cameras had a resolution of 848×480 pixels and a frequency of 30 FPS. All the computation to control the car was carried out on a NVIDIA Jetson Xavier NX board. The board had six cores and a NVIDIA Volta GPU, making it powerful enough to run the deep models for inference later on. During the experiment, all the gaze data, sensor data, and video streams were recorded and synchronized with the first-person video for further processing.

Model training. The human-attention model introduced in the previous setting has been trained on a Tesla T4 GPU with 16 GiB VRAM for 60 epochs using a batch size of 16 and four workers. We compared various network architectures for semantic segmentation, like Attention U-Net [41], U-Net++ [42], Path Aggregation Network (PAN) [43] and DeepLabV3+ [44]. Based on the validation results, U-Net++ with Concurrent Spatial and Channel Excitation (scSE) [45] performed the best. During preprocessing, the upper third of the input frames is discarded to bias the learning to the track events. We use horizontal image flipping and colorspace augmentation, changing brightness, saturation and hue. The human-attention model is not biased towards out-of-distribution illumination perturbations, as the magnitude of change applied for colorspace augmentation is equal to YOLOv7. The optimization is carried on with AdamW optimizer with a Smooth L1 Loss and the hyperparameters tuned with PyHopper [46]. We finally select an initial learning rate of $3e^{-4}$ and weight decay of $5e^{-5}$, and apply automatic mixed precision and learning rate scheduling.

IV. Experiments with the Human-Attention Model

In the following, we describe the experiments to demonstrate the benefit of integrating human attention in autonomous racing tasks, respectively in object detection and imitation learning.

Object Detection. Considering object detection in the context of F1Tenth autonomous racing, we consider the detection of the most-common static and dynamic obstacles, respectively boxes and other cars. We collect a training dataset consisting of these object classes, and compare the performance of the state-of-the-art YOLOv7 [47] trained with the predicted attention (**Att-YOLOv7**) and without it (**YOLOv7**). While YOLOv7 expects a 3-channel input, we modified the Attention-YOLOv7 to additionally receive the attention map, resulting in 4-channels. We trained the models with SGD using learning rate $1e^{-2}$, momentum 0.937, and weight decay $5e^{-4}$.

To analyze the robustness of the trained model to out-of-distribution samples, we evaluate them on images with heavy-perturbed brightness, changing it between 75% and 185% of the original value. We consider brightness perturbations because vision models are sensitive to changes in illumination. While the data augmentation techniques applied during training (i.e., horizontal flipping, image translation ± 0.2, scaling ± 0.5, HSV, mosaic) make the model invariant to many transformations, we still observe high sensitivity to the perturbed inputs. To assess the impact of various augmentations, we conduct an ablation study on them and report the results for Mosaic in Figure 3. We choose Mosaic because it was the agumentation technique with largest impact on the model performance. For each of the two models, the performance of training with and without Mosaic ($\pm$**Mosaic**) is reported as *mAP*, a standard metric for this application.

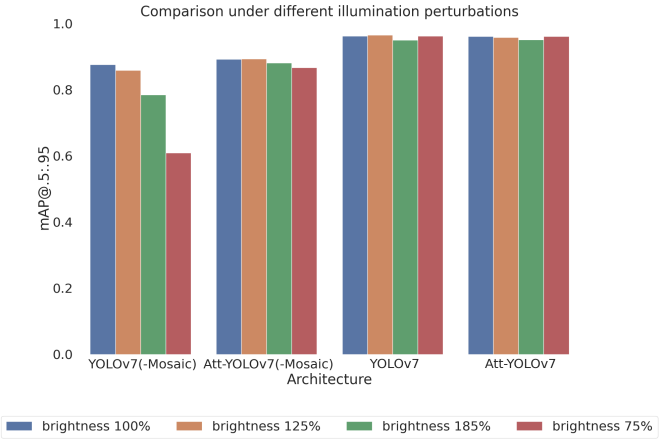

Fig. 3. Robustness evaluation under different brightness perturbations. The horizontal axis shows the various models and ablations, and the vertical axis reports the performance in term of mAP@.5:.95 score.

We observe that the introduction of predicted attention

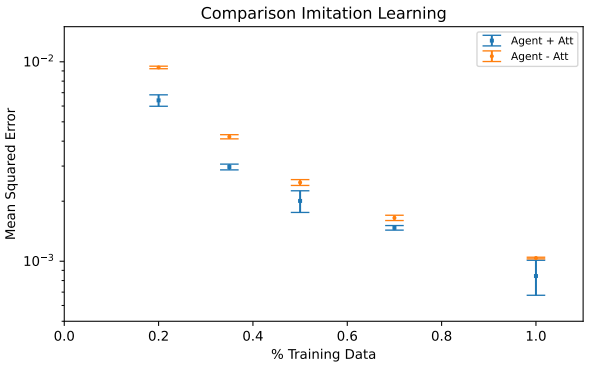

Fig. 4. Evaluation of prediction error for imitation learning under different training budgets (1.0=all available 16 000 samples). We report the average mean squared error over the last 50 epochs with relative error margins for the agent trained with (*Agent+Att*) and without (*Agent-Att*) attention features.

as additional input channel produces more robust models to brightness perturbations. The performance for models trained with Mosaic are equally good, but the gap is evident when trained without it. In particular, YOLO-v7 shows a drop of 10% and 25% for perturbations of 185% and 75% respectively. Conversely, the model trained with attention exhibits the same level of performance for all the perturbations. This result suggests that feeding human attention as additional input facilitates learning robust representations for object detection.

Imitation Learning. We consider the task of imitating the expert driver in controlling the F1Tenth racecar. The agent is an end-to-end model which receives RGB-images and predicts the driving commands for steering and velocity.

To experiment with a different integration of human attention in the input, we mark the attention points in the image, without adding it as additional channel. In this experiment, we use the recorded human attention, and give the marked input to the agent model which encodes it into an embedding with ResNet18. The representation is then feed into a series of fully-connected layers which predict the driving commands. We compare the agent model trained with and without attention and report the performance in term of prediction error. Figure 4 shows the evaluation for different training budgets, expressed as a fraction of the available training samples.

We observe a positive impact of marking the input frames with attention points, especially in low-data regimes. In fact, while using a large amount of data makes the performance comparable and the introduction of attention does not degraded the predictions, it results impactful when the training data are scarce. This result shows that the use of human attention has the potential for more-efficient learning.

V. Conclusions

In this study, we investigated the impact of human attention on enhancing robot learning in terms of both robustness and efficiency. To achieve this goal, we developed a data-collection pipeline to record data from human interaction in a driving task and propose a novel method to approximate human attention using a prediction model. This information is then used to enrich the visual input with attentional feature maps. We assess the impact in two specific learning tasks. In the object detection task, we observed a significant improvement in robustness to out-of-distribution samples when using attention data. In the imitation learning task, we observed lower prediction error and faster convergence, especially in low-data regimes.

These promising results highlight the potential of integrating human-based data into machine learning pipelines to improve robot learning. We intend to continue investigating this approach in future work to further understand its potential and explore other possible way to integrate human-based features into learning models. Overall, our findings suggest that the integration of human attention data can enhance the robustness and efficiency of machine learning models, which has significant implications for a range of real-world applications.

Acknowledgment

Luigi Berducci was supported by the Doctoral College Resilient Embedded System of TU Wien Informatics. Felix Resch was supported by the Horizon Europe Research and Innovation programme (Grant Agreement 101070679).

References

[1] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray *et al.*, "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.

[2] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International journal of robotics research*, vol. 37, no. 4-5, pp. 421–436, 2018.

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[4] V. Blukis, C. Paxton, D. Fox, A. Garg, and Y. Artzi, "A persistent spatial semantic representation for high-level natural language instruction execution," in *Conference on Robot Learning*. PMLR, 2022, pp. 706–717.

[5] N. Heravi, A. Wahid, C. Lynch, P. Florence, T. Armstrong, J. Tompson, P. Sermanet, J. Bohg, and D. Dwibedi, "Visuomotor control in multi-object scenes using object-aware representations," *arXiv preprint arXiv:2205.06333*, 2022.

[6] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto, "The surprising effectiveness of representation learning for visual imitation," *arXiv preprint arXiv:2112.01511*, 2021.

[7] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[9] A. Johnson and R. W. Proctor, *Attention: Theory and practice*. Sage, 2004.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," p. 11.

[12] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends," vol. 8, pp. 16 560–16 572, conference Name: IEEE Access.

[13] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 10 073–10 082. [Online]. Available: https://ieeexplore.ieee.org/document/9156532/

[14] G. W. Lindsay, "Attention in psychology, neuroscience, and machine learning," vol. 14, p. 29. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fncom.2020.00029/full

[15] G. W. Lindsay, D. B. Rubin, and K. D. Miller, "A unified circuit model of attention: Neural and behavioral effects." [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2019.12.13.875534

[16] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," vol. 35, no. 8, pp. 1798–1828, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[17] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," vol. 184, pp. 232–242. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231215017671

[18] C. Zhang, C. Zhang, J. Song, J. S. K. Yi, K. Zhang, and I. S. Kweon, "A survey on masked autoencoder for self-supervised learning in vision and beyond." [Online]. Available: http://arxiv.org/abs/2208.00173

[19] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, pp. 1691–1703, ISSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v119/chen20s.html

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale." [Online]. Available: http://arxiv.org/abs/2010.11929

[21] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training."

[22] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers." [Online]. Available: http://arxiv.org/abs/2106.08254

[23] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked autoencoders are scalable vision learners," pp. 16 000–16 009.

[24] J. Betz, H. Zheng, A. Liniger, U. Rosolia, P. Karle, M. Behl, V. Krovi, and R. Mangharam, "Autonomous vehicles on the edge: A survey on autonomous vehicle racing," vol. 3, pp. 458–488. [Online]. Available: http://arxiv.org/abs/2202.07008

[25] A. Dhall, D. Dai, and L. Van Gool, "Real-time 3d traffic cone detection for autonomous driving," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 494–501, ISSN: 2642-7214.

[26] K. Strobel, S. Zhu, R. Chang, and S. Koppula, "Accurate, low-latency visual perception for autonomous racing: Challenges, mechanisms, and practical solutions," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1969–1975, ISSN: 2153-0866.

[27] N. De Rita, A. Aimar, and T. Delbruck, "CNN-based object detection on low precision hardware: Racing car case study," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 647–652, ISSN: 2642-7214.

[28] R. C. Coulter, "Implementation of the pure pursuit path tracking algorithm," Carnegie-Mellon UNIV Pittsburgh PA Robotics INST, Tech. Rep., 1992.

[29] V. Sezer and M. Gokasan, "A novel obstacle avoidance algorithm:"follow the gap method"," *Robotics and Autonomous Systems*, vol. 60, no. 9, pp. 1123–1134, 2012.

[30] U. Rosolia and F. Borrelli, "Learning model predictive control for iterative tasks. a data-driven control framework," *IEEE Transactions on Automatic Control*, vol. 63, no. 7, pp. 1883–1896, 2017.

[31] K. Lee, Z. Wang, B. Vlahov, H. Brar, and E. A. Theodorou, "Ensemble bayesian decision making with redundant deep perceptual control policies," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 831–837.

[32] Y. Pan, C.-A. Cheng, K. Saigol, K. Lee, X. Yan, E. Theodorou, and B. Boots, "Agile autonomous driving using end-to-end deep imitation learning," in *Robotics: Science and Systems XIV*. Robotics: Science and Systems Foundation. [Online]. Available: http://www.roboticsproceedings.org/rss14/p56.pdf

[33] A. Brunnbauer, L. Berducci, A. Brandstätter, M. Lechner, R. Hasani, D. Rus, and R. Grosu, "Latent imagination facilitates zero-shot transfer in autonomous racing," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7513–7520.

[34] P. Cai, H. Wang, H. Huang, Y. Liu, and M. Liu, "Vision-based autonomous car racing using deep imitative reinforcement learning," vol. 6, no. 4, pp. 7262–7269, conference Name: IEEE Robotics and Automation Letters.

[35] L. Berducci, E. A. Aguilar, D. Ničković, and R. Grosu, "Hierarchical potential-based reward shaping from task specifications," *arXiv preprint arXiv:2110.02792*, 2021.

[36] T. Weiss and M. Behl, "Deepracing: A framework for autonomous racing," in *2020 Design, automation & test in Europe conference & exhibition (DATE)*. IEEE, 2020, pp. 1163–1168.

[37] S. N. Wadekar, B. J. Schwartz, S. S. Kannan, M. Mar, R. K. Manna, V. Chellapandi, D. J. Gonzalez, and A. E. Gamal, "Towards end-to-end deep learning for autonomous racing: On data collection and a unified architecture for steering and throttle prediction."

[38] Y. Mahmoud, Y. Okuyama, T. Fukuchi, T. Kosuke, and I. Ando, "Optimizing deep-neural-network-driven autonomous race car using image scaling," vol. 77, p. 04002. [Online]. Available: https://www.shs-conferences.org/10.1051/shsconf/20207704002

[39] K. Lee and V. Zakharov, "Perceptual attention-based predictive control."

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[41] O. Oktay, J. Schlemper, L. L. Folgoc, M. C. H. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *CoRR*, vol. abs/1804.03999, 2018. [Online]. Available: http://arxiv.org/abs/1804.03999

[42] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *CoRR*, vol. abs/1807.10165, 2018. [Online]. Available: http://arxiv.org/abs/1807.10165

[43] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," 2018.

[44] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *CoRR*, vol. abs/1802.02611, 2018. [Online]. Available: http://arxiv.org/abs/1802.02611

[45] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel squeeze & excitation in fully convolutional networks," *CoRR*, vol. abs/1803.02579, 2018. [Online]. Available: http://arxiv.org/abs/1803.02579

[46] M. Lechner, R. Hasani, P. Neubauer, S. Neubauer, and D. Rus, "Pyhopper – hyperparameter optimization," 2022. [Online]. Available: https://arxiv.org/abs/2210.04728

[47] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022. [Online]. Available: https://github.com/WongKinYiu/yolov7