In Search of Multi-Task Deep Architectures for Information Theoretic Field Exploration

Emanuele Masiero[1], Sathya Buršić[2], Vito Trianni[3], Giuseppe Vizzari[4] and Dimitri Ognibene[5]

Abstract— **Active vision can play a crucial role in navigating complex, unstructured environments like agricultural fields, where diverse scales, occlusions, and unknown element positions and numbers can hinder the perception of task-relevant information. Unlike traditional approaches that consider the overall environmental complexity or adopt simplistic sensing assumptions, this paper explores the efficacy of employing deep learning architectures to estimate information gain and expected loss in continuous, multidimensional observation spaces as those produced in the processing of sequential camera input of a small environment segment. In settings populated by numerous similar objects, as found in agricultural fields, such local estimations can be composed on-the-fly to address the challenge of predicting the contribution of successive viewpoints and drive active exploration strategies that can efficiently cover the entire environment and be adapted to diverse robotic architectures. By comparing multi-task architectures with various configurations of prediction heads for state estimation, information gain, expected loss, and best view prediction from an observation sequence, our findings test possible meta-learning gains and contribute to the development of computationally efficient and adaptable active sensing strategies, suitable for single or multi-robot systems. This approach not only aids in interpreting system decisions but also in formulating flexible strategies adaptable to changes in robot capabilities or task requirements.**

Active vision [1], [2], [3], [4], [5] is crucial for robotic systems operating in complex and unstructured environments, such as agricultural fields, where obstructions, varying scales, and the unknown distribution of elements can significantly hinder the acquisition of task-relevant information. This is particularly relevant in precision agriculture, a sector of growing importance [6].

Traditional information-theoretic approaches for active vision and perception have been successfully applied to a variety of settings where their analytic form can be exploited to deal with complexities like the presence of multiple instance of elements of the same class [7], [8], yet they face limitations in scalability and generalization to high-dimensional observations, that impede extracting an analytic representation (even if sampling solution may be adopted, e.g see [9]).

Conversely, recent (deep) reinforcement learning-based strategies for active vision offer higher flexibility in terms of observation processing but often at the cost of computational efficiency and flexibility [10], [11], [5]. In particular, they can extract policies that take into account the extensive structure of the environment, tasks, and robotic architecture and can be sensitive to changes of in all these factors, which is common in monitoring and exploration settings where not only the actual structure of the environment may change but is unknown and different robot fleets may be used (but see [12]).

Our work seeks to bridge this limitation by leveraging deep learning to perform efficient estimation of information gain and expected loss from multidimensional observations in environments characterized by the presence of numerous similar objects, such as those prevalent in agricultural fields. These evaluations can facilitate active exploration strategies that are both computationally efficient and adaptable to various robotic architectures and environmental conditions. By exploring and comparing multi-task deep learning architectures [13] that feature a variety of configurations for prediction heads, dedicated respectively to state estimation, information gain, expected loss, and best view prediction, all derived from sequences of observations, our study explores the potential meta-learning gains between apparently connected predictions in the active vision settings. Such an examination is critical for advancing our understanding of how to formulate computationally efficient and highly adaptable active sensing strategies, which are applicable to both single and multi-robot systems [6], [14].

In the following section we delve deeper into existing approaches, while Section II details our problem definition and proposed modelling framework. Section III presents the results and the analysis thereof. Finally, Section IV offers concluding remarks and future works.

I. Background and Related Works

Active vision, particularly in the context of robotics, has become an increasingly relevant area of research, driven by the need for systems that can autonomously explore

VT acknowledges support from the project E-CROPS (Technologies for Digital and Sustainable Agriculture, funded by the Italian Ministry of University and Research, Contract ARS01_01136) and from the project AGR-o-RAMA (Active monitoring for Agricultural Robots, POR FESR Lazio 2014-2020, Grant Agreement B85F21001360006).

[1] University of Milano-Bicocca, Department of Computer Science, Viale Sarca, 336, 20125 Milano (MI), Italy; e.masiero@campus.unimib.it

[2] Corresponding Author; University of Milano-Bicocca, Department of Computer Science, Viale Sarca, 336, 20125 Milano (MI), Italy; sathya.bursic@unimib.it

[3] Institute of Cognitive Sciences and Technologies, National Research Council, Via S. Martino della Battaglia, 44, 00185 Roma RM, Italy; vito.trianni@istc.cnr.it

[4] University of Milano-Bicocca, Department of Computer Science, Viale Sarca, 336, 20125 Milano (MI), Italy; giuseppe.vizzari@unimib.it

[5] Corresponding Author; University of Milano-Bicocca, Department of Computer Science, Viale Sarca, 336, 20125 Milano (MI), Italy; dimitri.ognibene@unimib.it

Generative AI language tools were used for light-editing of the author's original text, such as for spelling and grammar corrections and understand their environments and other agents in it [15], [16], [17], [18]. This paper aims to contribute to this field by focusing on evaluating a set of multi-task deep learning architectures to predict the informational gain of available actions when observing sequences of continuous multidimensional observations such as classification scores resulting from camera processing. Below, we outline some of the key related works that have shaped our approach and understanding.

After initial explorations on active vision and attention architectures [1], [2], [3], [4], [18], [19], information theory has assumed a central role in providing a conceptual framework for the field. Friston and colleagues [20] provide a review of different related frameworsk highlighting the role of epistemic value in exploration-exploitation dilemmas. This work underscores the importance of maximizing information gain to reduce uncertainty, a principle that guides our approach to active vision.

Ognibene and Demiris [7] develop a framework for active perception in dynamic environments, focusing on the importance of sensor control for information gain maximization and action anticipation. Their probabilistic generative framework utilizes a mixture of Kalman Filters for efficient recognition and exploration. They model sensors as detectors of object identity and position using Gaussian noise with distance dependent variance. Lee et al. [8] extend this information theoretic approach for active action perception to context with diverse actions and multiple actors using Probabilistic Context Free Grammars to represent activities and compute the information gain associated with observing each actor at its predicted position. [9] goes beyond myopic, single step, information gain by integrating its estimation in a Monte Carlo planning method to improve its environment exploration context in collaborative tasks.

Works like [21] introduce information theoretic metrics connected to information gain, i.e. mutual information, to drive the task of constructing an occupancy grid map with a binary Bayesian filter applied to a beam-based sensor model.

Palazzolo and colleagues [15] propose an exploration approach for construction of octrees-based voxelized 3D maps using applied to micro aerial vehicles that selects in real-time the next-best-view that maximizes the expected information gain estimated using Gaussian approximations.

Charrow et al. [22] propose an information-theoretic planning approach that efficiently guides robots to obtain measurements in uncertain regions of the map optimizing the Cauchy-Schwarz quadratic mutual information (CSQMI).

While these approaches require limited effort to account for the impact of active vision by relying on analytical formulations based on few parameters, they adopt simplistic assumptions on the complexity of visual stimuli and their interaction with observer configuration.

Other approaches have relied on reinforcment learning to develop active sensing and exploration skills.

Ognibene and Baldassere's work [5] presents a bioinspired approach to active vision integrating reinforcement learning and neural maps, emphasizing the need for systems that can adaptively focus their attention based on both top-down goals and bottom-up sensory inputs. Their architecture, BITPIC, demonstrates efficient task-solving in ecological conditions by integrating reinforcement learning, vision-manipulation coupling, and attention mechanisms.

Hausknecht and Stone [23] extended the Deep Q-Network (DQN) to operate in partially observable domains by incorporating a recurrent layer (LSTM) into the network architecture. This modification allows the network to maintain internal states that help infer missing information from the environment over time.

Grauman's group [10], [11] have developed deep reinforcement learning methods for visual exploration methods integrating also different heuristics to support exploration during the learning of complex visual exploration skills. Calafiore and colleagues [24] have applied similar methodologies to create active action perception systems in 3d environments and have shown that they present strategies comparable to those of human social perception.

Chaplot et al. [25] proposed Active Neural SLAM, a model for exploration in unknown environments that combines traditional SLAM (Simultaneous Localization and Mapping) techniques with deep reinforcement learning. The system actively decides where to move to build a map of the environment efficiently, effectively dealing with partial observability by deciding what to explore next.

Ammirato et al.'s [26] presents a dataset tailored for simulating robotic vision tasks, emphasizing the challenges posed by object scale, occlusion, and viewing direction. This dataset serves as a valuable resource for developing and testing our deep learning architectures. Ramakrishnan et al [10] proposes a framework to compare state of the art algorithms for exploration of photo-realistic environments.

While these approaches have shown strong performance in complex environments, they often rely on the learning of an environment and task dependent policy that may not easily adapt to tasks as exploration of novel environment with multiple similar elements, as in agricultural settings. For these condition we propose that advanced information gain estimation methods may enable active vision strategies that show high efficiency standing independent of the overarching environmental structure or specific task requirements, as demonstrated with some storng simplification by [6].

In summary, our work is deeply rooted in the advancements made in the fields of active vision, information gain estimation, and deep learning for autonomous exploration. By leveraging these insights, we aim to contribute to the development of more efficient and adaptive exploration strategies for robotic systems in complex and dynamic environments.

II. Problem and Model Definition

The task of estimating the uncertainty relating to estimated world states and the best sampling strategy is frequently not a simple task. Specifically, many real-world applications are characterized by high dimensionality, continuous variables, etc., making the exact estimation difficult. Approaches to circumvent this issue frequently include Monte Carlo methods.

An alternative approach is to create a model whose task is to estimate said uncertainty, which is the approach we take.

The problem we analyze is a subset of the task introduced by Carbone *et al.*[6] involving the monitoring and mapping of crop fields with unmanned aerial vehicle (UAV) swarms. The primary objective is to identify and map areas with high weed infestation in order to implement targeted weed control strategies, such as applying herbicide only where necessary. A typical example is detecting volunteer potatoes in sugarbeet fields, a common example in precision agriculture. The methodology involves utilizing automatic object classification from images captured by collaborative Unmanned Aerial Vehicles (UAVs) to generate and progressively refine a weed infestation map. Carbone *et al.* [6] approach the issue by developing a convolutional neural network (CNN) that given an image input from an UAV outputs a classification of weed-density for that particular perspective. This then forms the input to an exploration strategy utilizing reinforced random walks (RRW) and information gain (IG). They stop short of considering the correlations between the different point of view in computing the IG and adopt a simplified approach. For our experiments we focus on a similar experimental scenario that retains the relevant elements of the above problem but focus on estimating the IG taking into account the correlation between the successive observations from different perspective of the same cells. This is crucial to account for different phenomena such as full and partial occlusions.

A. Problem Definition

Consider a square grid of dimensions $C \times C$ where each cell contains cardboard boxes with a target representing the weed printed on one side. Targets are placed only on the side of the boxes to ensure the UAV agent needs to observe the boxes from specific points of view (POV) for accurate detection. Some boxes in the field do not have targets to represent false positives. The agent must detect only targets and consider the possibility that a target may not be visible from an inappropriate POV. Boxes in cells are positioned with targets on their side to make the target visible from certain POVs but not from others. Each cell can contain up to 9 boxes, with each box containing only one target. All targets must be observable from at least one POV for the counting CNN to provide an output resembling the ground truth count. See Fig. 1 for an example.

High-resolution images of the cells for all POVs are then used to train a CNN that outputs a probability distribution over the number of markers. This is then used as the input dataset for our experiments. Without loss of generality, let us consider only a single cell c.

Let $P_c^{(w)} = [p_c^{(w)}(0), \dots, p_c^{(w)}(N)]$ be the probability distribution over the number of markers in cell c from POV w, where $w = 1, 2, \dots, M$ represents the different POVs.

We define two key sets that will be crucial for our problem:

- **Observed POVs** (denoted W_o): This set contains all the viewpoints from which cell c has already been observed. In other words, for each POV $w \in W_o$, we have access to the corresponding probability distribution $P_c^{(w)}$.
- **Unobserved POVs** (denoted W_u): This set contains all the remaining viewpoints from which cell c has not yet been observed. These are the POVs that the agent might choose to observe next to refine its estimate of the number of markers. Therefore, $W_u \cup W_o = \{1, 2, \dots, M\}$ and $W_o \cap W_u = \emptyset$.

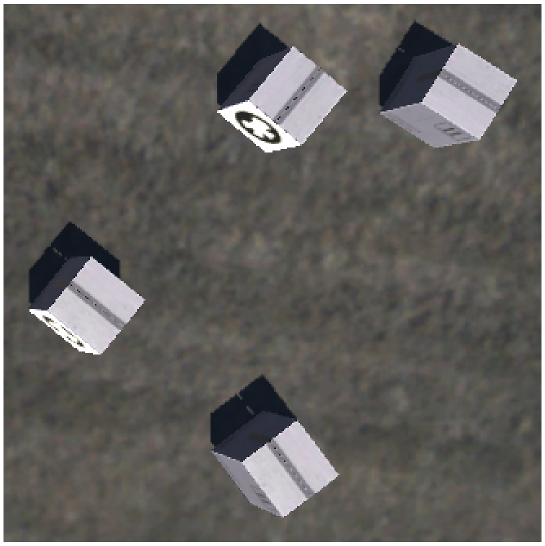

Fig. 1. An example of the simulated environment used for our experiment. Cardboard boxes with weed markers are visible.

Our problem can now be defined as estimating the updated probabilities for the total number of markers in cell c after observing some or all POVs. The updated probability distribution $\overline{P}_c$ can be expressed using conditional probabilities as:

$$\overline{P}_c(X \mid W_o) \tag{1}$$

where X represents the random variable of the number of markers in cell c after considering all POVs in the set of observed POVs W_o.

Furthermore, our interest also lies in sampling the environment in a ”cost-effective” way. We define this as, after having observed a set W_o of POVs, choosing a POV not yet observed $\overline{w}$ such that it offers the maximal reduction in entropy of the resulting estimation of the number of markers. Let the Shannon entropy for a discrete random variable X be defined as:

$$H(X) = -\sum_i P(x_i) \log_2 P(x_i) \tag{2}$$

Then, we are interested in the following optimization problem:

$$\underset{\overline{w} \in W_u}{\arg\min}\, H(P_c^{W_o \cup \{\overline{w}\}}) \tag{3}$$

This reasoning can be extended to any uncertainty measure Y, other than Shannon entropy.

There are many ways to approach this problem (see Section I for a discussion). We are interested in devising solutions that are immune to the rising complexity and dimensionality that characterize real-world applications. To this end, we choose to tackle the problem using deep neural networks (NN) for the estimation of values in equations 1 and 3.

Let f be the NN estimator function with parameters θ that takes the probability distributions from all POVs as input and outputs the updated probability distribution. Then we wish to approximate the following:

$$\overline{P}_c(X \mid W_o) = f_\theta(W_o) \tag{4}$$

Similarly, let g be an NN estimator function with parameters ϕ that given the same inputs and an uncertainty measure Y estimates the value of the measure upon successive sampling:

$$Y(P_c^{W_o \cup \{\overline{w}\}}) = g_\phi(W_o); \forall \overline{w} \in W_u \tag{5}$$

B. Model Definition

To approach the task at hand we devise two models: the first called the *base model* whose task is to estimate the world state; and the second that estimates the uncertainty related to where to sample next called the *IG model*, with IG standing for information gain. The base model, as any machine learning solution, suffers from its own prediction noise. Its output, representing a belief over the state of the world, then forms part of the inputs to the IG model, which will inherit this prediction noise. This indicates that one can have a well performing IG model only if the base model itself performs well.

To implement f_θ - the base model, and g_ϕ - the IG model, we use recurrent neural networks (RNN), specifically, we use a single LSTM layer followed by several fully connected layers. The use of a recurrent layer is aimed at handling the fact that we have a variable number of inputs and that the inputs may be arbitrarily ordered. Another prospective solution would be to use the transformer architecture. See tables I and II for model specification.

The base model takes as input a sequence of observed POVs for a cell and outputs the probability distribution over the true state of the marker count. It is trained with a cross-entropy loss. Instead, for the IG model we construct and test several hypotheses. First, we consider two different uncertainty measures: the Shannon entropy and the loss of the base model. While the choice of entropy is self-explanatory in this scenario, the loss of the base model as a learning objective may serve as a proxy for surprisal. If the IG model can predict where the base model might make the greatest prediction error, sampling that POV might give the greatest benefit. Second, we use two types of loss functions: the cross-entropy, which tries to predict the best course of action, i.e. the unobserved POV that optimizes it, and mean squared error (MSE) where the model predicts the values of the uncertainty measure over all POVs, including already observed ones. The two objectives are different, one is of classification and the other of regression. Regressions problems give far more information, but are more complex to train, and vice-versa. Finally, as a third research question we inquire in whether the training and test performance will be different if the same IG model is trained with different heads that predict more objectives, or separating the models during training and inference will improve results. We introduce maximally three objectives to the same model: predicting the marker count, same as in the base model; predicting the entropy; and predicting the base model loss. Having multiple heads might improve performance as knowledge is shared between the heads in the common layers. However, the same setup might potentially produce convergence issues and therefore decrease performance for one or more objectives.

TABLE I

BASE MODEL SPECIFICATION

Layer	Input Dimension	Output Dimension
LSTM	17	128
Linear	128	128
Dropout	-	-
Linear	128	128
Linear	128	128
Linear	128	9

TABLE II

IG MODEL SPECIFICATION

Layer	Input Dimension	Output Dimension
LSTM	17	128
Linear	128 Size	256
Dropout	-	-
Linear	256	256
Linear	256	256
Linear	256	9

III. RESULTS AND DISCUSSION

Tables III-VI present the accuracies of the various models tested compared to the random strategy, functioning as the baseline result. The random strategy chooses at random which POV to visit next, and is equivalent to the performance of the base model over the test set. The accuracies are presented on different sequence lengths over the test set. The total accuracy of the base model on the test set is 69.49%, averaged over all sequence lengths. As expected, the accuracies rise the more information is given to them, but with decreasing marginal improvements. Given only one POV, the mean accuracy is around 51%, rising monotonically up to around 81% for all POVs for a single cell. We suspect the reasons behind the accuracy for all available information for a single cell being capped at this level are twofold: the CNN that processes the images outputs a probability distribution over the marker count, reflecting its own uncertainty over the state; and secondly, the base model itself is a machine learning model prone to noise and other issues. This upper limit on accuracy presents also an upper limit obtainable by the IG model. Whichever sampling path over POVs we take, the maximum accuracy will respect this upper limit and won't surpass it by much. However, the goal of the IG model lies elsewhere. Its primary objective is to provide a sampling path that will give the maximal accuracy for a particular sequence length. In other words, we seek to optimize the marginal increment in accuracy at each sampling step. For each strategy, the tests are ran multiple times, and a paired t-test is performed between the results for that strategy and the corresponding random strategy to determine whether the two populations' means are statistically significant. If they are at a p-value level of 0.01, then the results are presented in boldface.

TABLE III
ENTROPY HEAD PERFORMANCE

Step	**Random**	**Entropy**	
		MSE	**CE**
1	51.50	**63.72**	**62.48**
2	61.09	**70.35**	**70.87**
3	67.03	**74.75**	**74.99**
4	72.25	**77.51**	**78.22**
5	75.41	**78.65**	**79.93**
6	77.70	**80.05**	**81.03**
7	80.09	**80.98**	**82.21**
8	81.71	81.99	**82.62**
Avg 4	62.97	71.58	71.64
Avg all	70.85	76.00	76.54

TABLE IV
LOSS PREDICTION HEAD PERFORMANCE

Step	**Random**	**Loss**	
		MSE	**CE**
1	51.50	**62.09**	**63.40**
2	61.09	**68.71**	**71.67**
3	67.03	**72.42**	**75.85**
4	72.25	**75.29**	**77.92**
5	75.41	**76.99**	**80.05**
6	77.70	**78.68**	**81.58**
7	80.09	80.72	**82.21**
8	81.71	81.93	**82.72**
Avg 4	62.97	69.63	72.21
Avg all	70.85	74.60	76.93

TABLE V
ENTROPY + LOSS PREDICTION HEADS PERFORMANCE

Step	**Random**	**Entropy**		**Loss**	
		MSE	**CE**	**MSE**	**CE**
1	51.50	**62.28**	**54.35**	51.15	**54.51**
2	61.09	**69.52**	**63.19**	60.78	**63.08**
3	67.03	**72.98**	68.41	66.67	68.48
4	72.25	**74.89**	**72.94**	71.55	72.64
5	75.41	**76.99**	76.46	76.03	76.71
6	77.70	78.31	78.72	77.69	78.89
7	80.09	80.31	80.29	79.96	80.22
8	81.71	81.60	82.02	81.46	81.95
Avg 4	62.97	69.92	64.72	62.54	64.68
Avg all	70.85	74.61	72.05	70.66	72.06

Table III presents the results of the IG model with only the entropy head trained with the MSE and CE losses, respectively. In both the cases, the entropy strategy beats the random strategy by a substantial margin, with statistical significance obtained at almost each step. The cross-entropy loss seems to surpass the MSE loss by a slight margin, and even beats random at the last step. Table IV presents instead the results for only the loss head. The MSE approach beats random up to step 6, while the CE approach beats both random and the entrpoy strategy. Table V contains results of the IG model where both the entropy and loss heads are trained simultaneously, and boasts inferior performance in relation to the previous results. More specifically, the loss head gives poor improvements over the random strategy, and the entropy strategy also gives inferior performance than evident in table III where only the entropy head was trained. Finally, training the marker head together with the entropy and loss heads with CE loss improves the outcomes for both strategies, as displayed in table VI. However, the entropy and loss heads respectively still reach better accuracies than when they are trained together. Interestingly, training under the MSE loss the same model (true only for the entropy and loss heads, markers are always trained with the CE loss) provides greater improvements in accuracy for the first few steps, but then struggles to keep up with the CE loss.

TABLE VI
MARKER+ENTROPY+LOSS PREDICTION HEADS PERFORMANCE

Step	**Random**	**Entropy**		**Loss**	
		MSE	**CE**	**MSE**	**CE**
1	51.50	**61.45**	**57.18**	**61.98**	**60.35**
2	61.09	**68.26**	**65.27**	**68.95**	**67.7**
3	67.03	**71.66**	**70.84**	**71.98**	**72.79**
4	72.25	**74.71**	**75.54**	**74.66**	**76.39**
5	75.41	**77.26**	**77.61**	**77.58**	**78.06**
6	77.70	**78.83**	**79.49**	79.09	**79.65**
7	80.09	80.73	81.05	81.06	**81.16**
8	81.71	81.28	81.59	81.29	**81.88**
Avg 4	62.97	69.02	67.21	69.39	69.31
Avg all	70.85	74.27	73.57	74.57	74.75

Several insights are offered by the above results. Firstly, the entropy-minimizing and the loss-maximizing strategy outperform the baseline random strategy by a significant margin, offering accuracy improvements of up to 11.9%. The loss strategy outperforms the entropy strategy by a slight margin, and the CE loss tends to outperform the MSE loss. We believe this to be due to a regression problem being inherently more difficult than a classification problem. Optimizing for classification conditions the model directly instead of indirectly to optimize for accuracy. Secondly, training multiple heads at once may present convergence issues. However, we do not believe this to be due to the model having an insufficient number of parameters or not being deep enough, as the task at hand is not particularly high-dimensional. Instead, we assume that this approach introduces more noise into the training process, creating a difficulty for the model to converge to the best solution. Furthermore, the loss functions are not of exactly the same magnitude, skewing the direction of the gradient and creating training instability. Finally, these results validate that a deep learning approach to active vision may produce good results. The fact that the model estimating uncertainty about world states depends on the model estimating the world states didn't seem to produce any performance issues, and may be a viable approach for similar tasks.

IV. Conclusions

In this paper, we searched for deep learning architectures that could estimate the quantities necessary for active visual classification, such as class estimations and next point of view dependent predictions of information gain and expected loss. In tiled environments, where such estimations can be performed locally for each part of the environment, separating these quantities from the decisions on the next action to perform can allow to efficiently devise effective strategies for exploration of the whole environment even in complex scenarios characterized by high dimensional continuous variable, with a focus on precision agriculture applications [6].

Our experimental results demonstrated the effectiveness of the entropy-minimizing and loss-maximizing strategies in outperforming random sampling by a significant margin, with accuracy improvements of up to 11.9%. Furthermore, our study highlighted the challenges of training multiple model heads simultaneously even if intuitively they should aligned and facilitate the learning process [13], suggesting potential convergence issues and training instability that depend on how to optimization problem is formulated, e.g. the specific loss function adopted for each head. In the future, adaptive multi-task training strategies [27] and the impact of dataset size will be explored.

Overall, our findings support the viability of deep learning approaches in addressing optimal sampling problems in complex environments. Investigating the integration of uncertainty estimation models with active vision systems could offer promising avenues for enhancing decision-making processes in various real-world applications.

References

[1] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," *International journal of computer vision*, vol. 1, pp. 333–356, 1988.

[2] M. Tistarelli and G. Sandini, "Dynamic aspects in active vision," *CVGIP: Image understanding*, vol. 56, no. 1, pp. 108–129, 1992.

[3] J. K. Tsotsos, "On the relative complexity of active vs. passive visual search," *International journal of computer vision*, vol. 7, no. 2, pp. 127–141, 1992.

[4] D. H. Ballard, "Animat vision," in *Computer vision: A reference guide*. Springer, 2021, pp. 52–57.

[5] D. Ognibene and G. Baldassare, "Ecological active vision: four bioinspired principles to integrate bottom–up and adaptive top–down attention tested with a simple camera-arm robot," *IEEE transactions on autonomous mental development*, vol. 7, no. 1, pp. 3–25, 2014.

[6] C. Carbone, D. Albani, F. Magistri, D. Ognibene, C. Stachniss, G. Kootstra, D. Nardi, and V. Trianni, "Monitoring and mapping of crop fields with uav swarms based on information gain," in *Distributed Autonomous Robotic Systems: 15th International Symposium*. Springer, 2022, pp. 306–319.

[7] D. Ognibene and Y. Demiris, "Towards active event recognition." in *IJCAI*, 2013, pp. 2495–2501.

[8] K. Lee, D. Ognibene, H. J. Chang, T.-K. Kim, and Y. Demiris, "Stare: Spatio-temporal attention relocation for multiple structured activities detection," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5916–5927, 2015.

[9] D. Ognibene, L. Mirante, and L. Marchegiani, "Proactive intention recognition for joint human-robot search and rescue missions through monte-carlo planning in pomdp environments," in *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11*. Springer, 2019, pp. 332–343.

[10] S. K. Ramakrishnan, D. Jayaraman, and K. Grauman, "An exploration of embodied visual exploration," *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1616–1649, 2021.

[11] D. Jayaraman and K. Grauman, "End-to-end policy learning for active visual categorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1601–1614, 2018.

[12] D. Ognibene, G. Pezzulo, and G. Baldassare, "Learning to look in different environments: An active-vision model which learns and readapts visual routines," in *From Animals to Animats 11: 11th International Conference on Simulation of Adaptive Behavior, SAB 2010, Paris-Clos Lucé, France, August 25-28, 2010. Proceedings 11*. Springer, 2010, pp. 199–210.

[13] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[14] R. Carli, G. Cavone, N. Epicoco, M. Di Ferdinando, P. Scarabaggio, and M. Dotoli, "Consensus-based algorithms for controlling swarms of unmanned aerial vehicles," in *International Conference on Ad-Hoc Networks and Wireless*. Springer, 2020, pp. 84–99.

[15] E. Palazzolo and C. Stachniss, "Effective exploration for mavs based on the expected information gain," *Drones*, vol. 2, no. 1, p. 9, 2018.

[16] D. Ognibene, T. Foulsham, L. Marchegiani, and G. M. Farinella, "Active vision and perception in human-robot collaboration," p. 848065, 2022.

[17] D. Kragic, "From active perception to deep learning," p. eaav1778, 2018.

[18] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *Autonomous Robots*, vol. 42, pp. 177–196, 2018.

[19] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz, "Appearance-based active object recognition," *Image and Vision Computing*, vol. 18, no. 9, pp. 715–727, 2000.

[20] K. Friston, F. Rigoli, D. Ognibene, C. Mathys, T. Fitzgerald, and G. Pezzulo, "Active inference and epistemic value," *Cognitive neuroscience*, vol. 6, no. 4, pp. 187–214, 2015.

[21] B. J. Julian, S. Karaman, and D. Rus, "On mutual information-based control of range sensing robots for mapping applications," *The International Journal of Robotics Research*, vol. 33, no. 10, pp. 1375–1392, 2014.

[22] B. Charrow, G. Kahn, S. Patil, S. Liu, K. Goldberg, P. Abbeel, N. Michael, and V. Kumar, "Information-theoretic planning with trajectory optimization for dense 3d mapping." in *Robotics: Science and Systems*, vol. 11, 2015, pp. 3–12.

[23] M. Hausknecht and P. Stone, "Deep recurrent q-learning for partially observable mdps," in *2015 aaai fall symposium series*, 2015.

[24] C. Calafiore, T. Foulsham, and D. Ognibene, "Humans select informative views efficiently to recognise actions," in *COGNITIVE PROCESSING*, vol. 22, no. SUPPL 1. SPRINGER HEIDELBERG TIERGARTENSTRASSE 17, D-69121 HEIDELBERG, GERMANY, 2021, pp. 48–48.

[25] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural slam," *arXiv preprint arXiv:2004.05155*, 2020.

[26] P. Ammirato, P. Poirson, E. Park, J. Košecká, and A. C. Berg, "A dataset for developing and benchmarking active vision," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1378–1385.

[27] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5824–5836, 2020.