Learning Priors of Human Motion With Vision Transformers

Placido Falqueto [ORCID]
Dept. of Information Engineering and Computer Science
University of Trento
Trento, Italy

Alberto Sanfeliu [ORCID]
Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
Universitat Politècnica de Catalunya
Barcelona, Spain

Luigi Palopoli [ORCID]
Dept. of Information Engineering and Computer Science
University of Trento
Trento, Italy

Daniele Fontanelli [ORCID]
Department of Industrial Engineering
University of Trento
Trento, Italy

Abstract—A clear understanding of where humans move in a scenario, their usual paths and speeds, and where they stop, is very important for different applications, such as mobility studies in urban areas or robot navigation tasks within human-populated environments. We propose in this article, a neural architecture based on Vision Transformers (ViTs) to provide this information. This solution can arguably capture spatial correlations more effectively than Convolutional Neural Networks (CNNs). In the paper, we describe the methodology and proposed neural architecture and show the experiments' results with a standard dataset. We show that the proposed ViT architecture improves the metrics compared to a method based on a CNN.

Index Terms—vision transformers, human motion prediction, semantic scene understanding, masked autoencoders, occupancy priors

I. INTRODUCTION

An essential requirement for a mobile robot to be able to move within a human-populated environment [1] is its ability to evaluate the human occupancy of the different areas of the environment and to foresee their most likely direction of motion in the near future. This information is reconstructed by humans by a quick sight of the scene and is instinctively used to identify the most convenient and efficient path to follow. Robots require a collection of sophisticated algorithms to accomplish the same results. In this paper, we will concentrate on the problem of understanding where people move in a scenario, which are their common trajectories and speeds and where they stop. This information can be mainly used to know the priors of human motion for different applications of robot navigation tasks, whilst its application is envisioned in many different fields of robotics. For instance, motion priors are of paramount importance in production plants when robots, most probably cobots, deal with cooperative and coordinated tasks with humans in modern robotics cells in order to improve simultaneously efficiency and safety.

Motivated by previous papers [2] on the importance of understanding human motion in shared spaces, our approach takes on the challenges of predicting occupancy priors for walking individuals in unfamiliar locations by relying solely on the semantic information of the area. Semantic maps allow us to break down the observed area into small parcels. The resulting network has a low complexity and is suitable for producing real–time predictions within a small time horizon. The price to pay is the loss of the "big picture", i.e., on how the motion between the different areas is related. Our proposed solution brings about an important advance in the state of the art proposing the adoption of novel architecture based on Vision Transformers (ViTs) to predict the occupancy distributions of walking individuals. The choice of ViTs is dictated by their well-known ability to extract effectively contextual information. This feature is exploited to understand the spatial relation between the different parcels, thereby enabling the network to reconstruct global information and learn how humans use the different areas (affordance). The resulting algorithm keeps the real–time computation cost within acceptable bounds, but it significantly improves the performance of the network even in the face of quick changes in the environment.

The simulation results unequivocally demonstrate that our ViT-based models outperform the baseline in terms of accuracy, reinforcing our belief that this solution can be a natural choice for real-world applications, in which mobile robots navigate across complex and dynamic environments.

The paper is organised as follows. In Section II, we offer a thorough review of the state-of-the-art on existing methodologies for inferring occupancy prior distributions in semantically rich urban environments. In Section III, we describe the key components of our proposed architecture, along with the proposed evaluation metrics and a description of the data set used in the training phase. In Section IV, we propose an ablation study to point out the impact of the different components in the architecture. In Section V, we illustrate our simulation results on known data set to show the improvement brought by our solution over the baselines. Finally, in Section VI, we offer our conclusions and announce future work directions.

II. RELATED WORK

In this section, we explore the existing body of research on human motion prediction, focusing particularly on the crucial role of map priors inference. We provide a succinct overview of various Neural Network architectures used in vision. Additionally, we scrutinize the methodologies and limitations of previous approaches, with a detailed examination of Rudenko et al.'s `semapp` [2]. Notably, we highlight the scarcity of literature addressing the direct prediction of priors from maps, a gap we aim to fill with our work.

A. Human Motion Prediction and Prior Occupancy Inference

Anticipating human motion intentions represents a long-standing challenge, demanding a nuanced comprehension of social dynamics [3]. As described in the survey [4], the modelling of human motion trajectories can be categorized through the representation of the underlying causes. Physics-based methods rely on explicit dynamical models derived from Newton's laws, either with a single model or a set of adaptive multi-models [5]–[7]. Pattern-based techniques learn motion patterns from observed data, either sequentially over time or non-sequentially considering the entire trajectory distribution [8]. Planning-based methods explicitly consider the agent's long-term goals [9], classifying into forward planning, assuming explicit optimality criteria, and inverse planning, estimating reward functions from observed trajectories. In [4], a significant increase in related works in this area is described, particularly in pattern-based methods.

Importance of Prior Occupancy Inference: Predicting prior occupancy distribution, rather than individual trajectories, proves valuable in extrapolating contextual information and enriching our understanding of a location. While the problems may appear similar, they represent distinct perspectives. The former focuses on dynamic predictions of individual or group actions within an environment, while the latter involves analyzing the environment itself, offering insights into typical human behaviours within that context. This differentiation enhances our ability to anticipate future events and make informed decisions based solely on environmental information [10]. Despite the evident importance of this approach, there is a distinct gap in existing literature dedicated to the direct prediction of priors from maps.

B. Neural Network Architectures

We'd like to highlight that our literature review will prominently showcase segmentation models, underscoring the extensive research in this domain. Unlike classification tasks, segmentation involves pixel-wise classification, where the goal is to assign a class label to each pixel in an image, effectively creating segments based on pixel content. As you explore further sections, you will observe our focus on a similar pixel-wise classification task, where the objective is to predict the likelihood of human presence in individual pixels. This focus aligns with cutting-edge approaches in segmentation, known for generating output tensors with the same dimension as the input. Consequently, our architecture is designed to meet the unique requirements of these tasks.

Convolutional Neural Network (CNN): CNNs have been instrumental in the advancements in computer vision, notably excelling in challenges such as the ImageNet classification challenge [11]. Their success in image classification and semantic segmentation has positioned them as a popular choice for various vision-related tasks. In spite of considerable efforts and substantial advancements in recent years, image segmentation remains a formidable challenge [12]. This difficulty arises from the intricate intra-class variations, contextual disparities, and ambiguities introduced by occlusions and low image resolution. All these limitations can be extended to the inference of map priors, where context is key for a good prediction. In [13], Doellinger et al. use the architecture introduced by Jegou et al. [14], which combines dense blocks [15] and fully CNNs, to predict average occupancy maps of walking humans even in environments where no human trajectory data is available, using the static grid maps as input. In [2], Rudenko et al. extend this architecture using as input semantic maps, instead of the plain grid map.

However, the landscape is undergoing a transformation, with a discernible shift towards transformer-based models. While CNNs have long dominated, the current trend indicates a growing adoption of transformer architectures in various vision-related tasks. This shift is clearly seen in the ImageNet classification challenge, where more and more people are opting for Vision Transformers (ViTs) [16], [17].

Vision Transformers: Vision Transformers (ViTs) [18] presents a cutting-edge approach to image processing by incorporating self-attention mechanisms to capture contextual information. While in image classification often an encoder structure to downsample features into a latent space and generate label predictions is enough, in image segmentation, we need to employ an encoder-decoder structure. In segmentation and reconstruction tasks, this structure upsamples the latent space to produce images with per-pixel class scores. To address the biases towards local interactions observed in convolutional architectures during segmentation tasks, Strudel et al. [19] propose a novel perspective. They formulate semantic segmentation as a sequence-to-sequence problem and adopt a transformer architecture to leverage contextual information throughout the entire model [20]. The authors claim to surpass all previous state-of-the-art convolutional approaches by a substantial margin of 5.3%. This notable improvement is attributed, in part, to the enhanced global context captured by their method at every layer of the model.

The potential of ViTs in tasks related to human motion prediction remains an area of exploration. In this paper, we delve into the capabilities of Vision Transformers (ViTs) to map priors inference, investigating their applicability and performance in this domain.

Masked Autoencoder: The success of masked language modelling, exemplified by BERT [21] and GPT [22] in NLP pre-training, lies in holding out portions of input sequences and training models to predict the missing content. This

method, proven to scale excellently, has demonstrated effective generalization to various downstream tasks. Inspired by these achievements, Masked Autoencoders (MAEs) [23] were developed to introduce a novel approach in computer vision, specifically addressing challenges related to latent representation learning. In contrast to traditional supervised learning in computer vision, which heavily depends on labeled datasets, Masked Autoencoders (MAEs) adopt a self-supervised approach for the classification task. Unlike conventional semantic segmentation based on ViTs, where images are decomposed into visual analogs of words, MAEs deviate by randomly removing patches during training. In essence, MAEs focus on reconstructing pixels, which are not inherently semantic entities. However, intriguingly, the MAE model demonstrates the ability to infer complex and holistic reconstructions, suggesting a learned understanding of various visual concepts and semantics. This behavior hints at the presence of a rich hidden representation within the MAE, leading to the hypothesis that the model captures diverse visual concepts through its self-supervised learning framework.

In this paper, we extend its exploration beyond pixel reconstruction. In particular, we delve into the performance of MAEs in the realm of map priors inference, investigating their ability to understand and interpret underlying visual concepts and semantics, and comparing its performance to the ViT. By scrutinizing the model's proficiency in this distinct task, we aim to unravel the extent to which MAEs can harness their learned representations for more advanced cognitive processes. This multifaceted analysis not only broadens our understanding of MAEs in computer vision but also provides valuable insights that can guide and inspire future research endeavors in the field.

To pay homage to the pioneering work in [2], we affectionately name our Vision Transformer-based approach "Semantic Map-Aware Pedestrian Prediction 2" (semapp2), described in the next section

III. Methodology

We start the description of the proposed solution by comparing the different metrics to compute the distance between probability distributions.

A. Metrics

1) Kullback-Leibler divergence: The Kullback-Leibler (KL) divergence [24] is a measure of how one probability distribution diverges from a second, expected probability distribution as

$$\mathrm{KL}(P_{GT} || Q_{pred}) = \sum_i P_{GT}(i) \log \left(\frac{P_{GT}(i)}{Q_{pred}(i)} \right),$$

i.e., the divergence of the probability distribution of the prediction Q_{pred} from the probability distribution of the target P_{GT}, over a discrete set of events indexed by i.

The KL divergence is not symmetric, meaning that $\mathrm{KL}(P || Q)$ is not necessarily equal to $\mathrm{KL}(Q || P)$. In the context of neural networks training, it is typically applied

in the direction of the predicted distribution (P) compared to the target distribution (Q). The reason for this choice is often related to the nature of the optimization problem. In tasks like probabilistic modeling or generative modeling, you want the predicted distribution to approach or match the target distribution. Minimizing the KL divergence in the direction of the predicted distribution helps achieve this goal.

However, in this specific application, it might be meaningful to also calculate the reverse KL divergence, i.e., $\mathrm{KL}(Q || P)$, contrary to the traditional machine learning approaches. This unconventional choice can be justified by examining the KL divergence formula: when $P_{GT}(i)$ is near zero, the KL divergence tends to be low, potentially masking issues in predictions, leading to misinterpretation of the model's performance. Instead, by calculating the reverse KL divergence $\mathrm{KL}(Q || P)$, the contribution to the divergence is weighted based on the prediction, ensuring that deviations in regions where the prediction is far from zero but the target is zero are appropriately penalized.

2) Earth Mover's Distance (EMD): The Earth Mover's Distance (EMD) [25], also known as Wasserstein distance or optimal transport distance, is a metric used to quantify the dissimilarity between two probability distributions. It provides a measure of the minimum amount of work required to transform one distribution into another. More in-depth, given two probability distributions P and Q representing the histograms of pixel intensities in the occupancy distributions, and a ground distance function $d(x, y)$ representing the cost of transporting mass from intensity x to intensity y, the EMD is defined as

$$\mathrm{EMD}(P, Q) = \min_{\gamma \in \Gamma(P,Q)} \sum_{(x,y) \in \mathrm{supp}(\gamma)} \gamma(x, y) \cdot d(x, y).$$

Here, $\Gamma(P, Q)$ represents the set of all possible joint distributions (couplings) of P and Q whose marginals are P and Q respectively. The minimization is over these couplings, while $\mathrm{supp}(\gamma)$ denotes the support of the coupling, i.e., the set of pairs (x, y) with non-zero probability.

Unlike the KL divergence, the Earth Mover's Distance is a metric that adheres to the triangle inequality and is symmetric. Its symmetry makes it particularly suitable for scenarios where a balanced evaluation of differences in both directions is desired. In the paper experiments, we will employ the forward KL divergence (KL-div), the reverse KL divergence (rKL-div) and the Earth Mover's Distance (EMD) to thoroughly assess the performance of our model in capturing the nuances of probability distributions.

B. Datasets

Our study builds upon the Stanford Drone Dataset (SDD) [26]. This extensive dataset captures images and videos featuring diverse agents like pedestrians, bicyclists, skateboarders, cars, buses, and golf carts navigating a real-world outdoor environment. It provides a comprehensive representation of human motion in shared spaces.

We utilized a subset of 20 maps from the Stanford Drone Dataset for training, including "bookstore", "coupa", "death

TABLE I
QUANTITATIVE EVALUATION OF 9 SEMANTIC CLASSES

	KL-div	rKL-div	EMD
semapp	0.66 ± 0.15	2.50 ± 1.51	40.18 ± 26.55
semapp2	0.49 ± 0.15	2.15 ± 1.20	34.24 ± 26.47

TABLE II
QUANTITATIVE EVALUATION OF 13 SEMANTIC CLASSES

	KL-div	rKL-div	EMD
semapp	0.58 ± 0.14	2.43 ± 1.24	41.16 ± 26.98
semapp2	0.46 ± 0.16	2.19 ± 1.50	27.65 ± 19.89

Fig. 1. The `semapp2` architecture.

Fig. 2. `semapp2` variation using a MAE autoencoder.

circle", "gates", "hyang", "little" and "nexus". Due to the limited availability of maps, we employed a cross-validation strategy, leaving one map out at a time for testing while training and validating on the remaining maps. This approach allowed us to maximize the use of the available data and ensure a robust evaluation of our model across various scenarios.

During preprocessing, the Stanford Drone Dataset (SDD) scenes were not only scaled but also manually segmented into refined semantic classes. All scenes were uniformly scaled to a resolution of 0.4 meters per pixel. Given that our network operates on map crops of fixed size ($64x64$ pixels), we adopted a strategy of decomposing the larger input images from the SDD into 500 random crops of appropriate size (like in [2]). Each crop in the training data is augmented 5 times by rotating and mirroring. The final distribution $p(s)$ for state s was reconstructed by averaging the predicted occupancy values of s across all crops containing that state. This approach, as stated in [2], enhances the robustness of our model's predictions by addressing potential artefacts associated with neighbouring crops.

To enhance prediction accuracy, we extend the semantic classes beyond those considered by Rudenko et al. [2]. In the paper, the authors use 9 semantic classes: pedestrian area, vehicle road, bicycle road, grass, tree foliage, building, entrance, obstacle and parking. We choose to add 4 more classes: sitting area, stairs, shaded area and intersection zone, reaching a total of 13 semantic classes. We find that using semantic classes that heavily influence human motion greatly affects the accuracy of the predictions. In Table I we compare the use of the 9 classes (pedestrian area, vehicle road, bicycle road, grass, tree foliage, building, entrance, obstacle and parking) and in Table II the complete model with all 13 classes (adding stairs, shaded area, intersection zone and sitting areas). Notably, semapp2 already exhibits notable advancements in prediction accuracy compared to semapp when restricted to the original 9 semantic classes, as evidenced by the metrics in Table I. In Table II, semapp2 consistently performs better than semapp across all evaluation metrics. The addition of the 4 new semantic classes refines the semantic understanding and contributes to improved accuracy in predicting occupancy distribution priors. 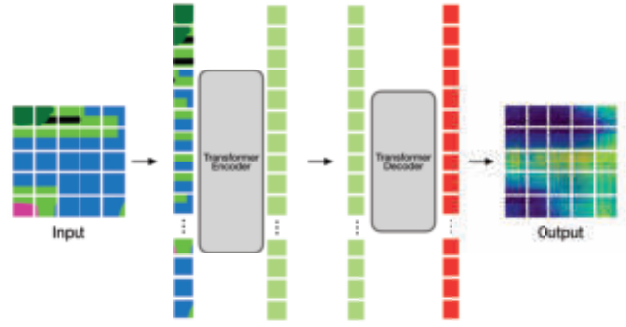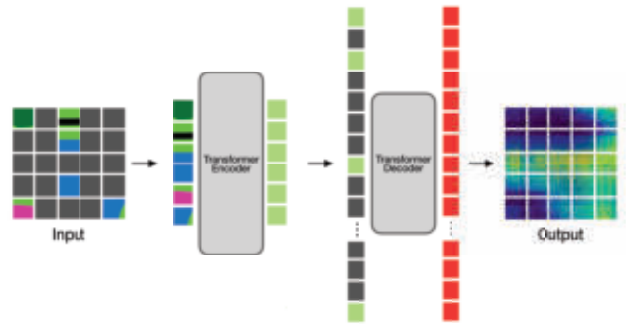This observation aligns with our goal of enhancing the model's capability to capture nuances in human-centric environments.

C. Proposed Vision Transformer architecture

The proposed `semapp2` consists of a ViT autoencoder designed to generate a prior prediction image from the input semantic map with multiple channels, each corresponding to different semantic classes (see Figure 1). We use a simple autoencoder architecture, where an encoder maps the observed signal (semantic map) to a latent representation, and a decoder predicts the prior from the latent representation.

In Figure 2 we provide a visual representation of the `MAE-semapp2`, an architecture variation of the `semapp2` that uses a MAE autoencoder, with 75% masking ratio. Note that the two architectures are the same: if we set a masking ratio of 0% on the `MAE-semapp2`, we obtain the same behaviour of a ViT-based `semapp2`. For simplicity, we will refer to a `semapp2` with a masking ratio of 75% as `MAE-semapp2` from now on.

Encoder: The encoder employs the ViT architecture, customized for semantic map processing. The input semantic map, with multiple channels representing various semantic classes, undergoes a linear projection with added positional embeddings. Subsequently, the resulting set of tokens is processed through a series of Transformer blocks. In the MAEs variation,

the encoder is identical to the ViT encoder, but it handles only the subset of unmasked patches of the semantic map.

Decoder: The decoder takes the full set of tokens consisting of encoded visible patches, and mask tokens. Each semantic class in the map is represented by a learned vector, and positional embeddings are added to all tokens in the set for reconstruction purposes. The decoder consists of a series of Transformer blocks designed to reconstruct the prior prediction image. Notably, the decoder architecture is independent of the encoder's design, providing flexibility.

D. Training and Evaluation

Our primary objective is predicting prior occupancy distribution based on semantic information, encompassing stop distribution and velocities heat map prediction. This task expands previous works, such as Rudenko et al.'s [2], which focus on occupancy distribution prediction only. We compare two main models in our study: our novel framework semapp2, which is based on ViTs, and Rudenko et al.'s semapp, based on CNNs. Additionally, we provide a concise comparison with a variation of our semapp2, based on the Masked Autoencoder.

The training process spans 100 epochs, employing the AdamW optimizer. We employ a mean squared error (MSE) loss per patch to calculate the prediction error. The training halts if the loss on the validation set shows no improvement for at least 15 consecutive epochs. A warmup cosine schedule is utilized for the learning rate, with a warmup period of 20 epochs. The base learning rate ($base\_lr$) is set to 1×10^{-4}, and the absolute learning rate ($absolute\_lr$) is calculated using the formula

$$absolute\_lr = base\_lr \times \frac{total\_batch\_size}{256} \ .$$

A weight decay of 0.3 is applied to the optimizer. Moreover, the training is conducted on two *NVIDIA RTX A5000* GPUs using PyTorch's *"Distributed Data Parallel"* to leverage distributed training. This configuration enhances the scalability and speed of our training process.

For cross-validation, we employ a leave-one-out strategy with semantic maps. Our dataset is divided into training and validation maps using an 80/20 split. During each iteration, we exclude one map and train the model on the training set, validating on the validation set and evaluating on the withheld map for assessment. This process is repeated for each map in the dataset. This strategy helps assess the generalization capability of our model across different semantic maps. Additionally, we experiment with various patches and crop sizes to identify optimal configurations.

IV. ABLATION STUDY

In order to methodically examine the effects of various elements in the suggested Semantic Map-Aware Pedestrian Prediction 2 (semapp2) model, we carry out an ablation study in this section. Our objective is to comprehend the role that each component plays in the overall performance and to choose the best performing model structure. To measure the impact of each ablation, we employ quantitative metrics such

TABLE III
IMPACT OF THE BACKBONE ON semapp2 MODEL

Backbone	KL-div	rKL-div	EMD
ViT-Base	0.42 ± 0.13	2.52 ± 1.88	54.53 ± 30.80
ViT-Large	0.34 ± 0.21	2.19 ± 1.84	45.77 ± 30.74
ViT-Huge	**0.31 ± 0.15**	**1.69 ± 1.11**	**39.64 ± 30.16**

TABLE IV
IMPACT OF THE CROP SIZE ON semapp2 MODEL

Crop Size	KL-div	rKL-div	EMD
32	0.62 ± 0.18	3.74 ± 1.13	54.56 ± 29.84
64	**0.34 ± 0.21**	**2.19 ± 1.84**	**45.77 ± 30.74**
100	0.56 ± 0.19	3.60 ± 1.77	117.38 ± 93.33

as KL divergence, reverse KL divergence and Earth Mover's Distance. These distances provide insights into the model's ability to accurately predict priors in semantic maps.

A. Architectural Components

To gauge the significance of specific architectural components, we conducted a series of experiments, systematically tweaking key elements within our semapp2 model based on Vision Transformer (ViT). First of all we need to choose the backbone for the architecture between ViT-Base, ViT-Large and ViT-Huge [18]. Then we explore variations on patch dimensions and crop size with the overarching goal of pinpointing the optimal configuration that strikes a balance between model complexity and predictive accuracy.

1) Backbone: We start by investigating the impact of different backbones on the MAE-semapp2 model, whose results are detailed in Table III. During the ablation tests, we keep unchanged the mask ratio of 75%, crop size of 64×64 pixels and patch size of 8×8 pixels. Our investigation revealed that the ViT-Huge backbone achieved the best results, demonstrating lower values across KL-divergence, reverse KL-divergence, and EMD metrics. Despite this superior performance, we opt for utilizing the ViT-Large backbone for practical considerations. The increment in performance with ViT-Huge is not significant, and it does not justify the significantly longer training times associated with its use. Thus, ViT-Large, being both proficient and quicker to train, emerges as the pragmatic choice for our MAE-semapp2 model. Moreover, as stated in [23], a single-block decoder can perform strongly and speed up training, for this reason we use a modified version of the ViT-Large changing the decoder layers depth to 1.

2) Crop Size: Delving into the impact of varying the size of the analyzed crop of the semantic map in our semapp2 model, we systematically adjusted the crop size, resulting the 64 the most promising (see IV).

3) Patch size: Examining the influence of patch size on the semapp2 model, we conducted experiments to observe variations in performance. Table V shows how a patch size of 8 fits the paper needs.

Patch Size	KL-div	rKL-div	EMD
8	**0.34 ± 0.21**	**2.19 ± 1.84**	**45.77 ± 30.74**
16	0.52 ± 0.10	2.43 ± 1.02	53.02 ± 34.39
32	0.60 ± 0.20	4.57 ± 1.79	46.69 ± 30.03

TABLE VI
IMPACT OF MASKING PERCENTAGE ON semapp2 MODEL

Masking Ratio	KL-div	rKL-div	EMD
0%	0.46 ± 0.16	2.19 ± 1.50	27.65 ± 19.89
25%	0.45 ± 0.17	2.32 ± 1.66	38.78 ± 31.72
50%	0.41 ± 0.11	2.36 ± 1.44	49.30 ± 29.92
75%	**0.34 ± 0.21**	**2.19 ± 1.84**	**45.77 ± 30.74**

4) MAE Masking Percentage: To investigate the impact of different masking percentages on the performance of the semapp2 model, we conducted ablation experiments by varying the masking percentage during training obtaining the results of Table VI. The masking percentage determines the proportion of patches excluded during the training process, influencing the model's ability to capture underlying patterns in the data.

V. RESULTS AND DISCUSSION

The evaluation metric involves computing the Kullback-Leibler (KL) divergence, Reverse KL divergence and Earth Mover's Distance (EMD) for all the leave-one-out maps, resulting in a mean metric value along with standard deviation, providing insights into the model's generalization across various semantic maps.

We present a qualitative comparative analysis in Figure 3 between semapp and semapp2. The top-left section depicts the semantic map and the top-right section represents the corresponding ground-truth occupancy distribution. In the bottom-left, predictions from semapp, while the bottom-centre and the bottom-right show predictions from semapp2 and MAE-semapp2. Moreover, in Figure 4, we compare the quality of the predictions of the model semapp2 using 9 semantic labels, shown on the left in the figure, or using 13 semantic labels, on the right. Visually the difference is barely noticeable, but quantitatively we have a slight improvement, as reported in Tables I and II.

A. Quantitative Evaluation

We provide in Table VII the mean and standard deviations of KL-divergences, reverse KL-divergences and EMDs for all three models (semapp, semapp2, MAE-semapp2) applied to the Stanford Drone Dataset using a cross-validation approach, as described in Section III-B. In the Stanford Drone Dataset, semapp2 shows competitive performance compared to semapp. This main comparison provides insights into the effectiveness of ViT in predicting occupancy priors based on semantic information.
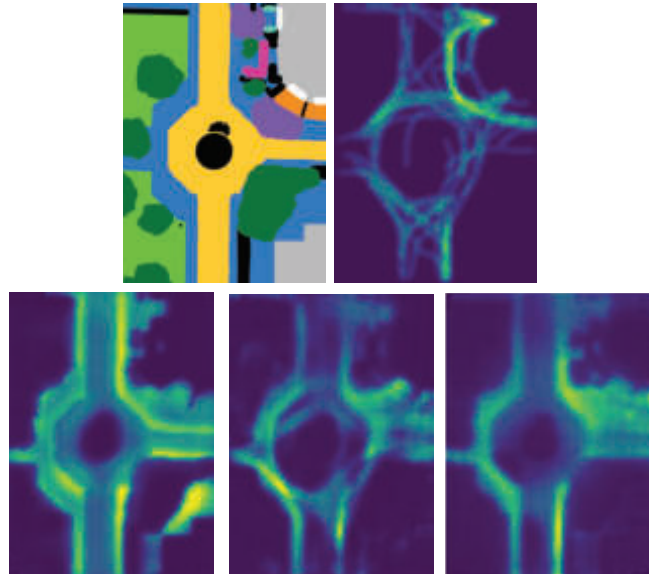

Fig. 3. Qualitative comparison of results in the Stanford Drone Dataset. Our ViT-based model showcases competitive performance compared to semapp (Rudenko et al. [2]), demonstrating the effectiveness of Vision Transformers in predicting occupancy priors. **Top left**: presents the original semantic map highlighting different classes, **Top right**: displays the ground-truth distribution of occupancies. **Bottom left**, **Bottom middle** and **Bottom right** showcase the predictions generated by semapp, semapp2 and MAE-semapp2, respectively.
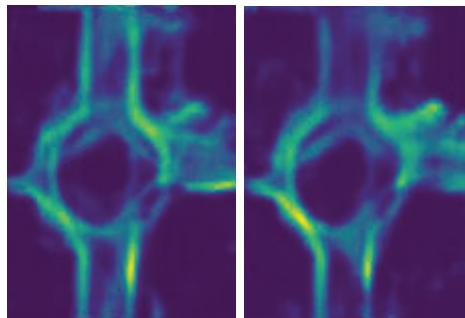

Fig. 4. Qualitative comparison between using 9 semantic classes (**Left**) and 13 semantic classes (**Right**)

semapp2 vs. MAE-semapp2: The four images in Fig 5 showcase different aspects of the prediction process using the MAE-based semapp2. The first *semantics* image represents the original crop of the semantic map, while the second *masked* image displays the same crop after the masking process, emphasizing the regions of interest during the model's inference. The *original* image presents the ground truth of the occupancy distribution, providing a reference for the expected outcome. Finally, the *prediction* image depicts the result of the MAE-based semapp2, illustrating the model's capability to anticipate and reproduce the occupancy distribution based on the masked semantic input.

The MAE-based semapp2 exhibits a notable level of generalization ability compared to the ViT-based semapp2.
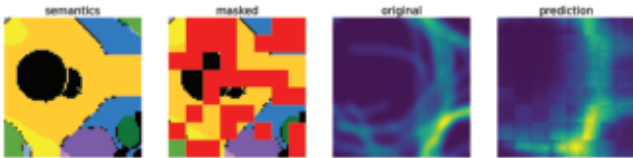
Fig. 5. Example of prediction using the MAE-based semapp2.

In [23], the authors demonstrate that masking patches in MAEs does not result in a decremental impact on reconstruction and classification, underscoring the significant data redundancy present in vision tasks. This observation suggests that the MAE model might be well-suited for learning underlying laws of social motion. Indeed, in Table VII, the `MAE-semapp2` variation shows slightly worse performance over the EMD metric compared to both `semapp` and `semapp2`. However, from a qualitative evaluation, the model seems to predict extremely well local variations of the occupancy distribution. Two examples are shown in Figure 6: the MAE-based `semapp2` is clearly superior at predicting the distribution. The low metric values could be due to a lack of trajectories in the timespan analysed in the SDD video. For this reason, the generalization ability of the MAE, especially in complex scenarios, warrants further exploration and investigation in future works. To assess the quality of the prediction, it could be necessary to collect more data on a specific location at different times in order to converge to a global probability distribution of the human occupancy, rather than the time-variant distribution that we obtain from the SDD videos.

B. Predicting Stops and Velocities

Furthermore, we delve into assessing the network's proficiency in predicting stops and velocities—priors that have been relatively underexplored in existing literature. This unique investigation holds significant implications for the field of mobility, where accurately anticipating stops and velocities could be crucial for enhancing planning tasks. Our exploration of these nuanced prediction tasks adds valuable insights to the broader understanding of Vision Transformers' capabilities in addressing complex aspects of occupancy prediction, particularly in real-world mobility scenarios. Figures 7 present the prediction of the velocity profile distribution and the stop distribution. Additionally, Table VIII provides a quantitative evaluation of the predictions of velocities and stops.

VI. CONCLUSION

The analysis of human occupancy of the different areas of an environment is essential to enable safe and efficient navigation of mobile robots. The past literature has shown that the use of semantic maps can significantly speed up the reconstruction of this information by breaking down the environment into several smaller parcels. The price to pay is a potential limitation of the accuracy due to the use of local information. We have proposed a solution that holds the promise to mitigate this problem by the use of a ViT backbone. Indeed the use of transformers allows the network to learn the spatial relation between nearby areas, hence reconstructing a global view of the environment. The results show that our solution significantly improves the prediction accuracy with a limited impact on the computation time, which remains acceptable for real–time applications of the solution. Many problems remain open and are reserved for future research activities. The first activity will be to test our method on a more complete dataset than the Stanford Drone dataset used for this paper, which contains more trajectories per map, to better evaluate the generalisation ability of the models. A second activity will be the integration of the module into a robot navigation framework. Third, we are considering a possible extension of the approach to predicting the motion of bicycles or cars, which would open other interesting application opportunities (i.e., autonomous driving). Finally, we are actually working to extend the same idea to cobots in production cells for product quality control and reworking of defected working pieces.

VII. ACKNOWLEDGMENTS

REFERENCES

[1] P. T. Singamaneni, P. Bachiller-Burgos, L. J. Manso, A. Garrell, A. Sanfeliu, A. Spalanzani, and R. Alami, "Advances and challenges in human-aware social robot navigation: A survey," *International Journal Robotics Research*, February 2024.

[2] A. Rudenko, L. Palmieri, J. Doellinger, A. J. Lilienthal, and K. O. Arras, "Learning occupancy priors of human motion from semantic maps of urban environments," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3248–3255, 2021.

[3] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh, "Core challenges of social robot navigation: A survey," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 3, pp. 1–39, 2023.

[4] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: a survey," *The International Journal of Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.
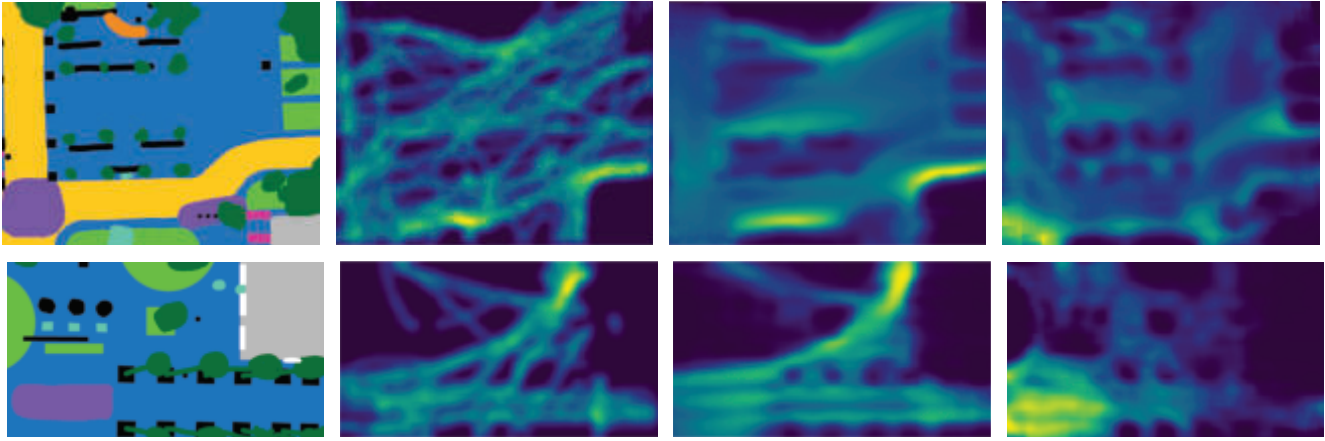
Fig. 6. Qualitative evaluation of the predictions of the `MAE-semapp2`. The **First column** shows the semantic maps, the **Second column** shows the groundtruths while the **Third column** shows the predictions using the MAE-based `semapp2`. The **Fourth column** shows the predictions using the ViT-based `semapp2`. (First row depicts the map bookstore, video 0, and the second row the map coupa, video 3)
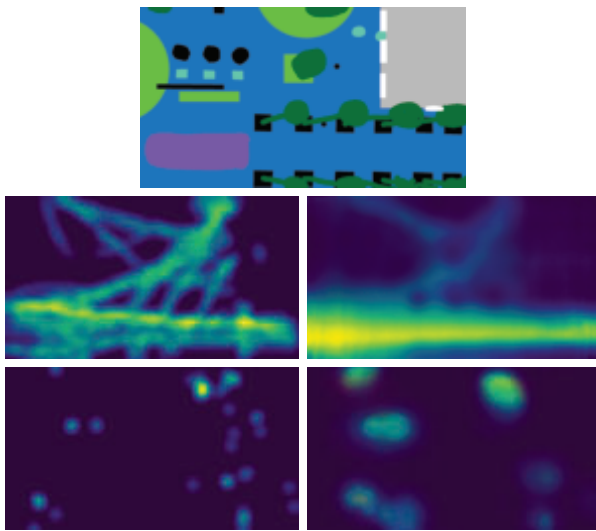

Fig. 7. Prediction of the velocity profile distribution and of the stops distribution in the coupa map.

[5] F. Farina, D. Fontanelli, A. Garulli, A. Giannitrapani, and D. Prattichizzo, "Walking Ahead: The Headed Social Force Model," *PLOS ONE*, vol. 12, no. 1, pp. 1–23, 1 2017.

[6] S. Zernetsch, S. Kohnen, M. Goldhammer, K. Doll, and B. Sick, "Trajectory prediction of cyclists using a physical model and an artificial neural network," in *2016 IEEE Intelligent Vehicles Symposium (IV)*, 2016, pp. 833–838.

[7] E. A. I. Pool, J. F. P. Kooij, and D. M. Gavrila, "Using road topology to improve cyclist path prediction," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 289–296.

[8] A. Vemula, K. Muelling, and J. Oh, "Modeling cooperative navigation in dense human crowds," 2017.

[9] A. Antonucci, G. R. Papini, P. Bevilacqua, L. Palopoli, and D. Fontanelli, "Efficient Prediction of Human Motion for Real-Time Robotics Applications with Physics-inspired Neural Networks," *IEEE Access*, vol. 10, pp. 144–157, December 2021.

[10] B. Kaleci, Ç. M. Şenler, H. Dutağacı, and O. Parlaktuna, "Semantic classification of mobile robot locations through 2d laser scans," *Intelligent Service Robotics*, vol. 13, no. 1, pp. 63–85, 2020.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[13] J. Doellinger, M. Spies, and W. Burgard, "Predicting occupancy distributions of walking humans with convolutional neural networks," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1522–1528, 2018.

[14] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," 2017.

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018.

[16] S. Srivastava and G. Sharma, "Omnivec: Learning robust representations with cross modal sharing," 2023.

[17] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, and L. Schmidt, "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," 2022.

[18] 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," 2021.

[19] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," 2021.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[22] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[23] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2021.

[24] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[25] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, pp. 99–121, November 2000.

[26] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory prediction in crowded scenes," in *European Conference on Computer Vision (ECCV)*, 2016.