IVOA: Introspective Vision for Obstacle Avoidance

Sadegh Rabiee[1] and Joydeep Biswas[1]

***Abstract*— Vision, as an inexpensive yet information rich sensor, is commonly used for perception on autonomous mobile robots. Unfortunately, accurate vision-based perception requires a number of assumptions about the environment to hold – some examples of such assumptions, depending on the perception algorithm at hand, include purely lambertian surfaces, texture-rich scenes, absence of aliasing features, and refractive surfaces. In this paper, we present an approach for *introspective vision* for obstacle avoidance (IVOA) – by leveraging a supervisory sensor that is occasionally available, we detect failures of stereo vision-based perception from divergence in plans generated by vision and the supervisory sensor. By projecting the 3D coordinates where the plans agree and disagree onto the images used for vision-based perception, IVOA generates a training set of reliable and unreliable image patches for perception. We then use this training dataset to learn a model of which image patches are likely to cause failures of the vision-based perception algorithm. Using this model, IVOA is then able to predict whether the relevant image patches in the observed images are likely to cause failures due to vision (both false positives and false negatives). We empirically demonstrate with extensive real-world data from both indoor and outdoor environments, the ability of IVOA to accurately predict the failures of two distinct vision algorithms.**

I. INTRODUCTION

Recent developments in 3D LiDAR technology have facilitated robust obstacle avoidance for mobile robots and autonomous vehicles. Yet, the high cost of such sensors makes them inaccessible for a wide range of robotic applications, leaving vision as the next best option. Vision systems, on the other hand, are prone to errors from various sources such as image saturation, blur, texture-less scenes, etc. This has motivated the question of whether we can develop competency-aware vision systems capable of predicting their own failures and estimating their level of uncertainty.

In this work we present an approach for introspective vision for obstacle avoidance (IVOA). The main idea behind IVOA is to equip one robot with a high-fidelity depth sensor and let the vision system use that as ground truth for learning of an introspection model that predicts failures of the stereo obstacle detection system. The introspection model can then be transferred to other agents, empowering them to predict failure cases of the vision-based obstacle detection, without having depth sensors for each robot. IVOA applies to any stereo vision-based obstacle detection system and provides the means for **self-supervised** training of an introspection model that predicts the probability of different types of failure (false positive and false negative) and pinpoints the location of the error on the input image. The proposed approach also provides a measure of uncertainty for its predictions. The benefits of such fine-grained reasoning about the performance of the vision are twofold. First, it provides planning and control modules with rich information that could be used for safe and optimal execution. Second, the extracted information can be used effectively to discover and categorize sources of errors for the vision system. While previous works on introspective vision systems [1], [2] output a single failure probability score for the whole input image, IVOA, to the best of our knowledge, is the first to digest the input image in detail and localize the potential sources of error and their type.

We implement and test IVOA on a real-world dataset collected with a ground robot in both indoor and outdoor environments. We demonstrate IVOA's capability to accurately predict both false positive and false negative failure cases of two different stereo obstacle detection systems. We also show how IVOA's extracted information can be used to categorize sources of error for a vision system.

[1]Sadegh Rabiee and Joydeep Biswas are with the College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, USA. Email: {srabiee, joydeepb}@cs.umass.edu

II. RELATED WORK

In recent years, there has been a rise in research on introspective vision systems. One line of work tries to use the inherent uncertainty measure provided by the vision model. Grimmet et al. [3] look into probabilistic function approximators such as Gaussian processes as well as boot strapped classifiers that use the consensus of an ensemble of models as a measure of confidence. They evaluate the inherent uncertainty measure of these models by inspecting their changes when the model is exposed to new unseen data. Hu et al. [4] tune the parameters of a localization algorithm by means of minimizing the inherent uncertainty measure of their perception model. Using only the underlying uncertainty of the perception model limits the introspective capacity of the system. A more rigorous approach is to train a second model, called the introspection model, to predict failure cases of the vision system. The introspection model does not need to know about the underlying details of the vision system. Instead it relies on the raw sensory input to predict the probability of failure of the vision system. Zhang et al. [1] use labeled data and train a binary classifier, which given an input image predicts the success or failure of a vision system. They test their method for two different tasks of image classification and image segmentation. Daftry et al. [2] train a convolutional neural network (CNN) that uses both still images and optical flow frames to predict the probability of failure for the navigation system of an actual UAV. A follow up work [5] trains an SVM classifier

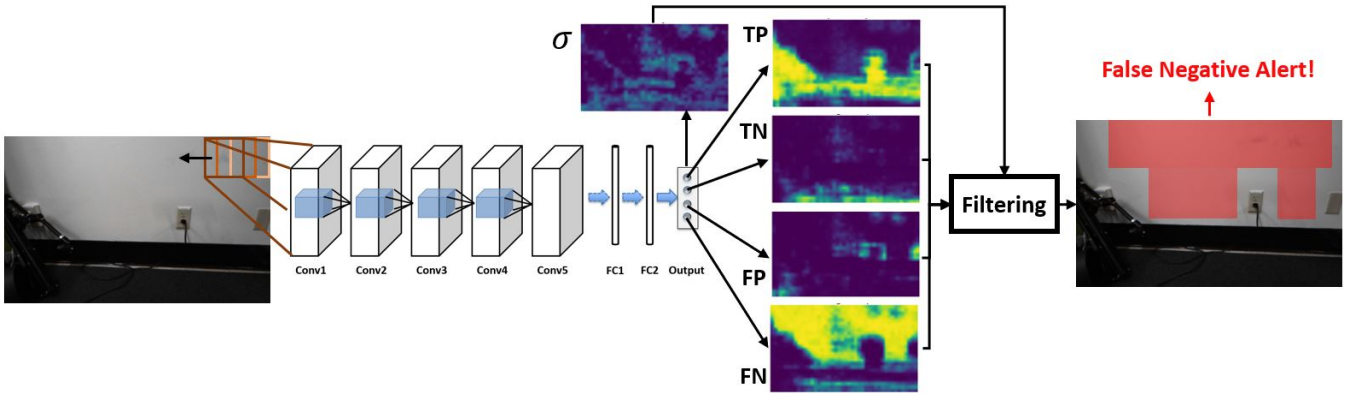


Fig. 1: Pipeline of the introspection model. Image patches in the areas of interest on the input image are passed through a CNN to generate probability scores for different classes of failure. Potential areas of failure are detected given the mean filtered probability scores and estimated uncertainty values.

to choose the best recovery action when the vision system has high uncertainty. While all the above works predict only the overall reliability of the vision system given an input image, Ramanagopa et al. [6] provides more detailed information, predicting and localizing false object detection instances on the input image for the use case of autonomous vehicles. They use stereo vision and leverage discrepancies between detected objects by each of the left and right cameras as cues for predicting failures. Their approach, however, is limited to predicting instances of false negatives and is specific to object detectors which do not suffice for safe obstacle avoidance.

Our work is similar to [2] in that it seeks introspective vision for safe robot navigation. However, it goes beyond answering the question of whether the vision system may fail at a specific time step. Instead, IVOA predicts where in the input space the failure will happen and what type the failure will be. It should be noted that IVOA is distinct from works such as [7], [8], [9] that use only a black box model for the purpose of obstacle avoidance. It instead relies on a model-based stereo obstacle detection at the core and accompanies that with a black box model that provides predictions of failure cases of the former along with an uncertainty estimate of the predictions. We believe that this approach allows for robust and long-term deployment of robots in the real-world, where the robot will experience previously unseen environment.

III. INTROSPECTIVE VISION SYSTEM

Architecture. In IVOA's architecture, the vision system consists of a perception and an introspection module. The perception module receives the raw sensory input and leverages its underlying model-based knowledge of the system, here stereo geometry, to provide the planning module with information about the surrounding of the robot. Unlike the perception module, the introspection is a black-box model and is responsible for assessing the reliability of the output of the perception given the same raw sensory input. In this work, we implement this vision system specifically for the purpose of obstacle avoidance for autonomous mobile robots. Our perception module is a stereo-vision based obstacle detector, that outputs an obstacle grid in front of the robot, where each cell in the grid is flagged as either obstacle free or occupied. The task of the introspection model is to predict, for each region on the input image, the probability of each of these four cases happening with regards to the perception module: 1-wrongly detecting an obstacle (false positive), 2-wrongly not detecting an obstacle (false negative), 3-correctly detecting an obstacle (true positive), 4-correctly detecting no obstacles (true negative).

Perception Model. The perception model can be any stereo vision-based obstacle detection algorithm. The only requirement for the model is to be able to check the traversability of a point $(x, y, 0)$ in the reference frame of the robot, assuming it is in the field of view of the cameras. In this work we mainly use a version of the Joint Perception and Planning (JPP) [10] algorithm. JPP utilizes a fast and computationally efficient method for detecting obstacles using stereo vision. Instead of creating a full dense reconstruction of the scene, it samples points of interest (x, y, z) in the reference frame of the robot and projects them to the image planes of both cameras. Matching pairs of projections in the two image planes signal the existence of an object at the query point. We use JPP-C, which is JPP with the assumption of a convex world, i.e. obstacles are assumed to be on the ground (not hanging). We also test our approach on an implementation of dense stereo reconstruction with ELAS [11]. This method creates a full 3D reconstruction of the scene via performing stereo matching for all pixels on the image and uses that for detection of obstacles.

Introspection Model. We implement the introspection model as a multi-class classification convolutional neural network (CNN). We use the same layer architecture as the well known AlexNet [12], i.e. 5 convolution layers followed by 3 fully connected layers. The outputs of the last layer are passed through a softmax layer to provide normalized probability scores in the range $[0, 1]$. The model uses the image stream from only one of the cameras. Each 960×600

image obtained from the camera is sliced into overlapping 100×100 patches with a stride of 20 pixels. Each patch is separately fed to the network and the outputs of network are scalar probability scores for each of the 4 classes of false positive (FP), false negative (FN), true positive (TP), and true negative (TN). The output class probabilities of all patches are arranged in the original patches' configuration to form 49×26 heat maps of the probability of each class over all the input image. We want the introspection model to not only predict probability values for different classes of failure, but also to provide the degree of confidence it has in its prediction. We realize this by means of using two dropout layers before the first two fully connected layers during inference. Dropouts are mainly used for preventing overfitting in neural networks [13] during the training phase by randomly dropping units. Recent research, however, has shown that the same technique could be used during the inference phase to provide an estimate of uncertainty of the network [14]. We employ this technique in our network as following: each input image patch is passed through the network multiple times (we pick 20), and at each pass different neurons are randomly dropped with a probability of 0.5 at the dropout layers. The variance of the output of the network over these passes is taken as a measure of the introspection model's uncertainty for the given input image patch. In other terms, each input image is treated as a set of particles that pass through a stochastic model. The mean and variance of the output particles define the output of the model. The last stage of the introspection model is the post processing of the obtained probability scores. A mean filter is applied to the output probability heat maps of each class, and then at regions where the uncertainty is lower than some safety threshold classes with highest probability scores are announced as predictions. Fig. 1 shows the pipeline of the introspection model.

Training. We automate the training process for the introspection model by adding a high fidelity 3D depth sensor to the system. This sensor provides ground truth information for the monitoring module which in turn compares the depth sensor output with that of the perception module to generate labeled training data. Algorithm 1 outlines the training data generation procedure. It should be noted that the depth sensor is only used for training. This training scheme helps reduce cost of large-scale robot deployments. Only a few of them need to be equipped with the costly monitoring depth sensor and the trained introspection model will be transferred to all robots. The ideal depth sensors to use in this system are 3D Lidars, which provide accurate depth readings of the surrounding environment upto long ranges and in various weather conditions. For a low-cost implementation of the system, however, we use a Kinect sensor to obtain ground truth depth readings. This limits us to training in indoor environments and outdoor environments only when there is not much sunlight as it interferes with the IR camera of the Kinect. Fig. 2 illustrates the diagram of the navigation stack of the robot during training.

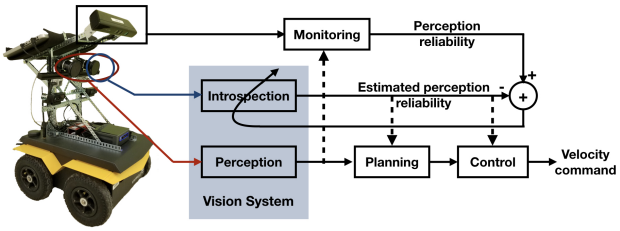


Fig. 2: Training scheme for the introspection model. Depth readings by the Kinect sensor are used as the ground truth for automated generation of the training data.

Algorithm 1 TRAIN DATA GENERATION

Input: Left camera image I_l, right camera image I_r, depth image I_d
Output: Set of image patches $\{I_k\}_{k=1}^{N}$, set of ground truth labels $\{d_k\}_{k=1}^{N}$
$r \leftarrow$ safety radius
$\mathbf{P} \leftarrow \{(x', y', 0) : X_{min} \leq x' \leq X_{max} \wedge Y_{min} \leq y' \leq Y_{max}\}$
$N \leftarrow |\mathbf{P}|$
for $k \leftarrow 1$ to N **do**
    $O_S \leftarrow$ ISOBSTACLEFREESTEREO(I_l, I_r, p_k, r)
    $O_M \leftarrow$ ISOBSTACLEFREEMONITOR(I_d, p_k, r)
    $(u, v) \leftarrow$ PROJECTTOLEFTCAM(p_k)
    $I_k \leftarrow$ EXTRACTPATCH(u, v)
    if $O_M \wedge O_S$ **then** $d_k \leftarrow TN$
    else if $O_M \wedge \neg O_S$ **then** $d_k \leftarrow FP$
    else if $\neg O_M \wedge O_S$ **then** $d_k \leftarrow FN$
    else $d_k \leftarrow TP$
    end if
end for

IV. EXPERIMENTAL RESULTS

A. Evaluation Dataset

We use the Clearpath Jackal, a mobile robot with a skid-steer drive system, for data collection. The robot is equipped with a stereo pair of Point Grey cameras that record 960×600 images at a rate of $30Hz$. Obstacle detection ground truth is provided by a Kinect depth sensor that is mounted on the robot and captures depth images at a rate of $30Hz$. The cameras and the Kinect are extrinsically calibrated with respect to each other. The robot is driven using a joystick and RGB and depth images are logged at full frame rate. The data is then processed offline: the depth images are converted to pointclouds and synchronized with the stereo camera images. For each set of synchronized images the perception module and the Kinect-based monitoring system are queried to determine whether a set of $(x, y, 0)$ points on a 2D grid in front of the robot and on the ground plane are obstacle free or not within a radius of $r = 10cm$. The corresponding pixel coordinates of the query points on the left camera's image plane along with the obstacle detection results are stored to form the dataset. The indoor dataset

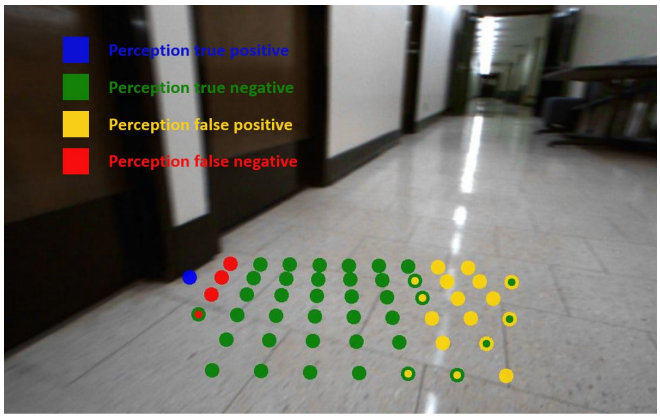


Fig. 3: Comparing the ground truth label of the points on a 2D grid with the predicted labels by the introspection model. For each point on the grid, the inner circle's color represents the prediction of the introspection model and the outer one indicates the ground truth result of the perception system.

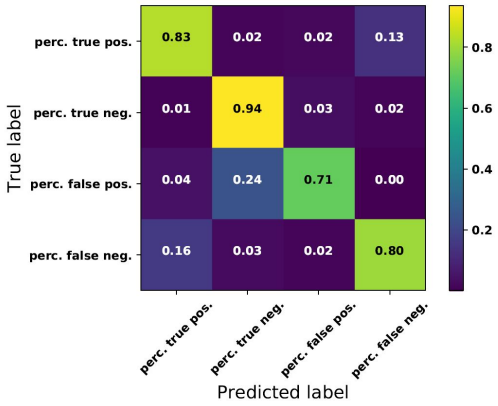


Fig. 4: Prediction results of the introspection model on the indoor dataset.

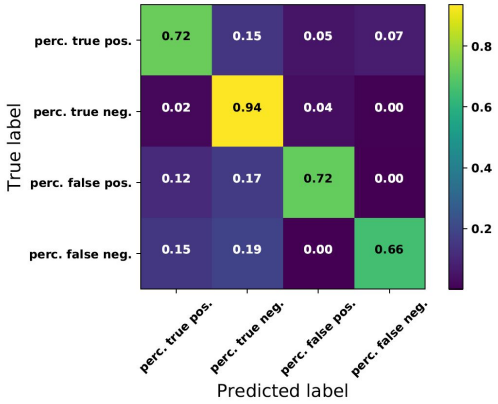


Fig. 5: Prediction results of the introspection model on the outdoor dataset.

spans multiple buildings with different types of tiling and carpet. The outdoor dataset is also collected on different surfaces such as asphalt, concrete, and tile in both dry and wet conditions. The total dataset of more than 1.4km traversed by the robot includes about 6 million extracted image patches and 120k full image frames. The data is split into train and test datasets, each composed of separate full robot deployment sessions, such that both train and test sets include data from all types of terrain.

B. Evaluation Metric

The performance of the introspection model is evaluated based on its ability to predict the behavior of the perception model. For each image the introspection model is queried with the same points on the image, for which we have the prediction result of the perception system as one of the four classes of FP, FN, TP, and TN. The accuracy of the model in predicting each of these classes is assessed as a measure of its performance. Fig. 3 denotes an example of comparing the output of the introspection model against the ground truth.

C. Model Accuracy Results

We train the introspection model on the train dataset and then test the model separately on the indoor and outdoor portions of the test dataset. Please note that for the rest of the paper until the end of Section IV-E, the reported results correspond to IVOA using JPP-C as the perception model. In Section IV-F we show results of IVOA trained on ELAS.

In this section, we present the results with the uncertainty-based filtering of the introspection model turned off, i.e. the model classifies all data points as one of the four classes even if the uncertainty level is high. We analyze the effect of model uncertainty in the next section. The results are demonstrated in Fig. 4,5. The introspection model is able to catch a significant portion of the failure cases and predict their type correctly for both indoor and outdoor datasets. It is interesting to note that even in cases when the introspection model is not able to predict a failure, it still correctly detects the existence of an obstacle. Fig. 6 shows the detailed outputs of the introspection model for an example input image.

D. Effect of Model Uncertainty

As explained in section III, our proposed introspection model provides an estimate of its inherent uncertainty. In this section, we analyze the importance of the uncertainty measure in the reliability and performance of the system. We run the introspection model on the whole test dataset. We then sort the data based on the introspection model's uncertainty score in ascending order. We start removing data points from the bottom of the list, whose uncertainty score is higher than an uncertainty threshold value. The accuracy of the introspection model is then calculated on the retained data points as the mean of the prediction accuracy values for each classe. Fig. 7a illustrates that the accuracy increases monotonically with decrease in the uncertainty threshold. At the point, when still 70% of the data is retained, it reaches an accuracy of more than 96% with a 6% improvement compared to not using the uncertainty measure. Fig. 7b also shows the percentage of the retained data for each class and over the same range of uncertainty thresholds. As can be seen in the figure, the rate of dropping data is roughly the same for all 4 classes. The results prove the correctness

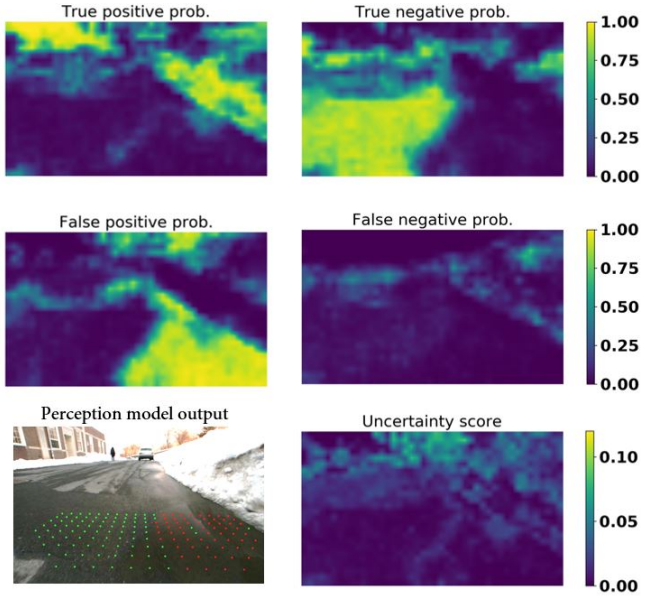


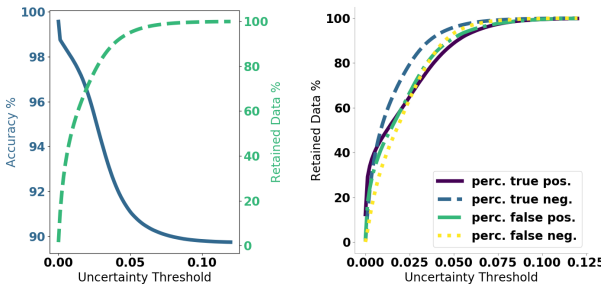
Fig. 6: Example output of the introspection model. The bottom left image shows the estimated obstacle grid in front of the robot by the perception model, where the red and green dots denote the detected occupied and obstacle free cells respectively. The introspection model correctly predicts water reflections to cause false positives for the perception model (middle left image).

of the estimated uncertainty measure, in that it is inversely correlated to the accuracy of model. It should be noted that such uncertainty measure is of paramount importance especially for a failure detection system that is based on a black-box model. Deep learning models are prone to making false predictions when exposed to unseen and totally new inputs. Using an uncertainty estimation, however, reduces such failures and makes neural networks suitable for use in real-world applications such as robotics.

E. Categorizing Sources of Error

As mentioned earlier, one of the motivations of IVOA from performing a fine-grained failure detection is to behave as an assisting tool for debugging of vision systems. In order to test this hypothesis, we try clustering the detected instances of failure. From all instances of false positive and false negative, detected by the introspection model and on the test dataset, we pick the top 50% in terms of the confidence of the predictions. Then for each corresponding image patch $\mathbf{I}_i$, the normalized output of the second fully connected layer of the introspection model $\mathbf{x}_i \in \mathbb{R}^{256}$ is extracted as an embedding.

In order to decide on the number of clusters, we first visualize a 2D representation of the data. PCA is performed to reduce the dimension of embeddings by a factor of 10, and then t-SNE [15], a nonlinear dimensionality reduction approach well-suited for visualization of high dimensional data, is applied to obtain a 2D representation of the samples. Based on the result of the visualization, a cluster number of 2 is chosen and k-means clustering is applied to the data in



(a) (b)

Fig. 7: The effect of introspection model's uncertainty threshold. (a) The solid graph shows the total accuracy of the introspection model over all classes after retaining only part of the data for which the model uncertainty is less than the threshold. The dashed graph shows the percentage of retained data for different uncertainty thresholds. (b) The percentage of retained data vs. uncertainty threshold for different classes of data points.

the original embedding space. Fig. 8 illustrates the resultant clusters projected down to the 2D space as well as sampled image patches from each cluster. The result shows that IVOA the dark edges at the bottom of the walls and reflection/glare to be the most dominant sources of error for the perception model under test.

F. Adaptability to Different Perception Systems

Our proposed architecture of the introspective vision system as explained in section III is agnostic to the perception model. Any obstacle detection system can be used in place of JPP-C, and the training and inference of the system will remain intact. In order to test this feature, we trained the introspection model for an obstacle detection system based on stereo dense reconstruction of the scene using the ELAS [11] stereo matching technique. The resulting introspection model was able to learn failure cases of the new perception model. Fig. 9 compares the output of the two different perception systems alongside the introspection model's prediction of their performance for the same scene. Both JPP-C and ELAS wrongly detect the reflection on the tile as an obstacle (hole in the ground). Also, JPP-C fails to detect the texture-less wall, while ELAS is able to correctly detect it. As shown in the figure, the introspection model correctly predicts the behavior of both models.

V. CONCLUSION

In this paper, we introduced IVOA: an architecture for competency-aware stereo vision-based obstacle avoidance systems capable of predicting their failures, while distinguishing between false positive and false negative instances. We demonstrate IVOA's ability to accurately predict failures of the vision on a real-world dataset in both indoor and outdoor environments. As future work, we would like to integrate IVOA with planning and control to leverage its detailed estimate of the reliability of vision for safe and optimal navigation of mobile robots.

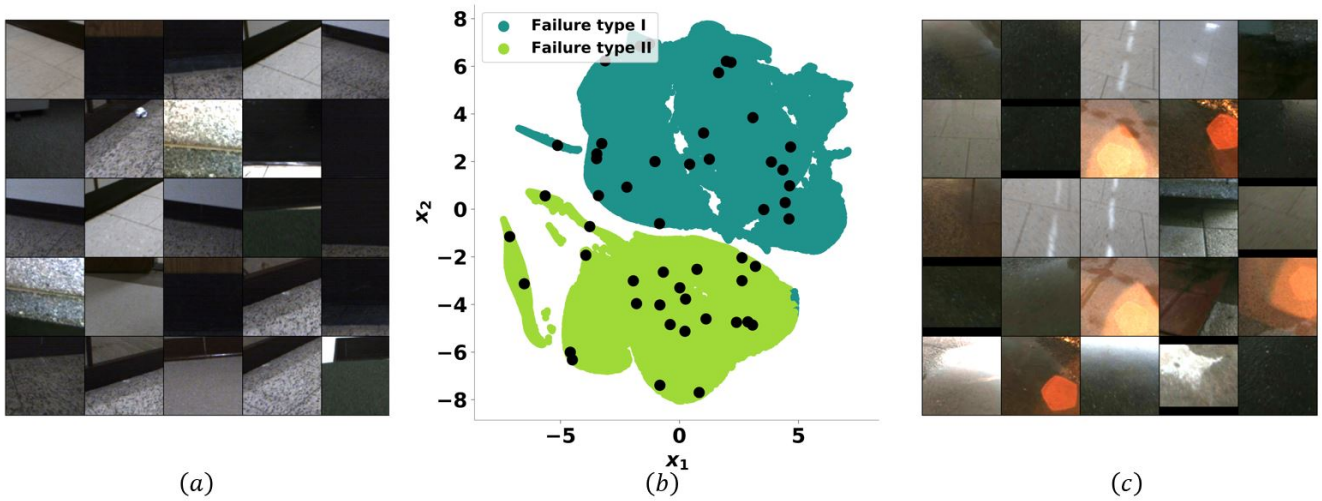


Fig. 8: Clustering predicted sources of error. b) Two extracted clusters of perception failures in the embedding space projected down to a $2D$ space via dimensionality reduction. a) Randomly sampled image patches from failure type II cluster, shown as black dots on (b). c) Image patches randomly sampled from failure type I cluster. IVOA detects reflection and glare (type I) and the dark stripes at the bottom of the walls (type II) to be the most dominant sources of error for the perception model under test.

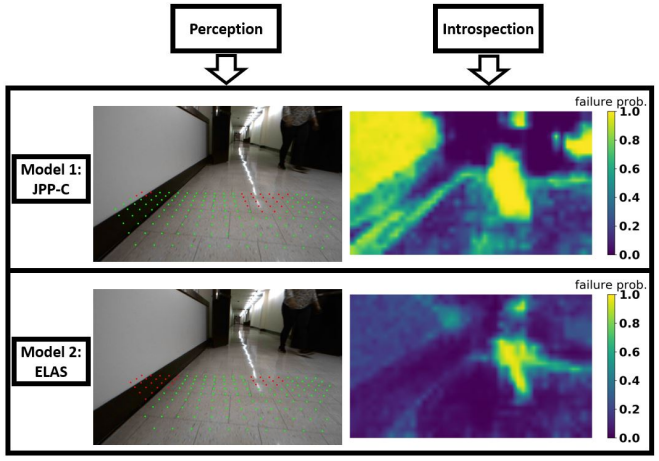


Fig. 9: Introspection model's prediction for different perception models. On the left column, the produced obstacle grids by each of the two perception models are visualized on the input image. The green and red dots represent the detected obstacle free and occupied cells respectively. On the right, output of the introspection model is shown as the total probability of failure $\Pr(FP) + \Pr(FN)$ for each perception model. The introspection model correctly predicts the reflection to cause false positives for both models and the wall to cause false negatives for only one of them.

References

[1] P. Zhang, J. Wang, A. Farhadi, M. Hebert, and D. Parikh, "Predicting failures of vision systems," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3566–3573.

[2] S. Daftry, S. Zeng, J. A. Bagnell, and M. Hebert, "Introspective perception: Learning to predict failures in vision systems," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1743–1750.

[3] H. Grimmett, R. Triebel, R. Paul, and I. Posner, "Introspective classification for robot perception," *The International Journal of Robotics Research*, vol. 35, no. 7, pp. 743–762, 2016.

[4] H. Hu and G. Kantor, "Introspective evaluation of perception performance for parameter tuning without ground truth." in *Robotics: Science and Systems*, 2017.

[5] D. M. Saxena, V. Kurtz, and M. Hebert, "Learning robust failure response for autonomous vision based flight," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 5824–5829.

[6] M. S. Ramanagopal, C. Anderson, R. Vasudevan, and M. Johnson-Roberson, "Failing to learn: autonomously identifying perception failures for self-driving cars," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3860–3867, 2018.

[7] P. Ross, A. English, D. Ball, B. Upcroft, and P. Corke, "Online novelty-based visual obstacle detection for field robotics," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 3935–3940.

[8] N. Hirose, A. Sadeghian, M. Vázquez, P. Goebel, and S. Savarese, "Gonet: A semi-supervised deep learning approach for traversability estimation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3044–3051.

[9] L. Tai, S. Li, and M. Liu, "A deep-network solution towards model-less obstacle avoidance," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 2759–2764.

[10] S. Ghosh and J. Biswas, "Joint perception and planning for efficient obstacle avoidance using stereo vision," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1026–1031.

[11] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Asian conference on computer vision*. Springer, 2010, pp. 25–38.

[12] 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.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[14] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[15] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.