

Online Self-supervised Long-range Scene Segmentation for MAVs

Shreyansh Daftry, Yashasvi Agrawal and Larry Matthies

Abstract—Recently, there have been numerous advances in the development of payload and power constrained lightweight Micro Aerial Vehicles (MAVs). As these robots aspire for high-speed autonomous flights in complex dynamic environments, robust scene understanding at long-range becomes critical. The problem is heavily characterized by either the limitations imposed by sensor capabilities for geometry-based methods, or the need for large-amounts of manually annotated training data required by data-driven methods. This motivates the need to build systems that have the capability to alleviate these problems by exploiting the complimentary strengths of both geometry and data-driven methods. In this paper, we take a step in this direction and propose a generic framework for adaptive scene segmentation using self-supervised online learning. We present this in the context of vision-based autonomous MAV flight, and demonstrate the efficacy of our proposed system through extensive experiments on benchmark datasets and real-world field tests.

I. INTRODUCTION

Micro Aerial Vehicles have built a formidable résumé by making themselves useful in a number of important applications, from disaster scene surveillance and package delivery to robots used in aerial imaging, architecture and construction. The most important benefit of using such lightweight MAVs is that it allows the capability to fly at high speeds in space-constrained environments. However, in order to function in such unstructured environments with complete autonomy, it is essential that they are able to see and interpret the scene, and navigate robustly.

In recent years, autonomous capabilities of resource-constrained autonomous aerial vehicles have seen considerable progress [1]. However, the sensor characteristics of typically used active and passive range sensors on such agile MAVs severely limit the range at which reliable 3D perception can be done. For a typical MAV, this range is usually in the order of 10 ~ 20m; at longer distances, range data becomes too sparse or noisy for reliable scene understanding from geometry. This presents a challenge, as it makes the system inherently myopic.

In contrast, humans effortlessly navigate through most environments - observing, understanding and planning around distant obstacles even in new, previously unseen environments. Illumination changes, parallax, scale ambiguity - none of these affect our ability to perceive our environment and strategically plan decisions based on only visual information. Such human visual abilities are not solely due to better stereo perception; rather, humans are excellent at reasoning from monocular images. Therefore, in this work we are interested

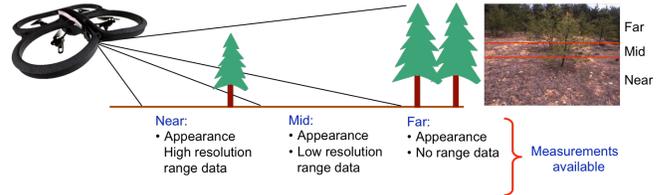


Fig. 1. Near-to-far paradigm. Reliable range information is only available in the near-range, where as appearance information is available even at far-range. We want to exploit the relationship between these disparate modalities, to build a more holistic perception system.

in methods for extending the look-ahead distance of perception systems (See Fig. 1) beyond where geometry can be used, by learning to reason from contextual information in monocular imagery.

The problem of visual recognition has been well studied in general [2]. Recently, learning-based methods have proven to be more competitive than traditional geometry-based methods on solving complex vision tasks, including image-based scene segmentation [3]. However, beyond benchmarks and new end-to-end learning applications, they have yet to become the go-to solution for vision-based autonomous navigation. This can be attributed to several reasons: First, vision algorithms have to contend with continuously evolving, unstructured sensor data during long-term operations. As a result, the performance of data-driven methods do not necessarily translate to real-world scenarios [4]. In particular, the problem aggravates for far-away scenes where the visual appearance is dependent on the locale (including the time of day and viewing angle) and is not easily generalized. Second, most learning methods rely on strongly annotated pixel-accurate data that is highly time-consuming to collect, and often even infeasible. Lastly, the resource-constrained budget of MAVs do not allow for real-time deployment of computationally intense state-of-the-art methods. It is the above considerations that motivate our contribution.

In this paper, we advocate that continuous online learning of scene segmentation would allow the system to constantly adapt to its local, and avoids the need to learn a universal scene classifier. This requires a method to automatically generate large amounts of training labels in real-time. Therefore, as proxy, we propose to exploit readily available geometric cues from the near-range data to generate segmentation labels in real-time, which can then be used to adaptively train the long-range classifier online. Learning from such self-supervision would sidestep the annotation cost to scale up learning performance, and mitigate the challenges of learning algorithms towards long-range perception. We call this self-supervision ability as near-to-far learning.

All authors are members or alumni of the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA. Corresponding Email: Shreyansh.Daftry@jpl.nasa.gov

II. RELATED WORK

Our work addresses an issue that has received attention in various communities. In particular, similar related work can be found in the literature for autonomous self-driving cars, where the need for faster driving motivated similar approaches that attempted to extend the appearance of the road out to longer ranges [5], [6]. Other approaches have used learning-based methods to map traversability information to color histograms, textured pixels or geometric point clouds [7], [8], [9]. Another interesting method for extending the range of perception that has been studied in literature involves using overhead imagery in combination with local data [10]. All the above approaches either relied on strong priors [11], had assumptions on vehicle kinematics [12], or used fairly simple features and classifiers [13] that do not generalize well to changing environments.

Recently, neural-networks based deep learning methods have achieved human-level performance in the task of scene understanding [14], [15], with increased adoption in real-world autonomous systems [16]. Furthermore, there has also been a growing interest in exploiting end-to-end frameworks; such methods argue against learning mid-level perceptual representations, and propose to directly regress to control commands [17]. However, learning such end-to-end frameworks require considerable amount of training data. This reliance on full supervision is a major limitation on the scalability of these methods to robotics applications [18]. In contrast, weakly- or unsupervised training methods [19], [20] side-step the need for large amounts of labelled data, by leveraging a limited set of related annotations for training. While this line of research has shown potential, they are currently still outperformed by fully supervised methods. In the paper, we advocate for the middle-ground via self-supervised training.

III. APPROACH

Near-to-Far learning is a self-supervised learning paradigm that uses visual representations from estimated near-range geometric cues, and thereby learns to reason about far-range scenes. In this section, we describe our proposed approach for scene segmentation using this near-to-far learning paradigm. Fig. 2 provides a schematic overview of the proposed framework, which is explained below.

A. Self-supervision from Geometry

Our system takes as input a RGB image, a registered depth map, which can be obtained either using an active or passive range sensor, and the estimated pose of the vehicle. We segment the near-range 3D information into obstacle and free-space regions based on the typical application of ground-plane estimation. Ground-plane estimation entails finding the largest planar region in the 3D scene. Assuming the free-space in front of the MAV is locally flat, it is assumed that this plane corresponds locally to the ground. Any points that are outliers with respect to the current ground plane model are then distinguished and assumed to be obstacles.

Given a point x_w on the ground plane in the Euclidean coordinates, referenced to an arbitrary frame of reference, a plane in this space can be parametrized by a vector $c_w^T x_w = 1$. It can be shown that under the projective transformation of a pinhole camera model, every such ground plane corresponds to a linear model in image coordinates and disparity [8], and can be written as:

$$\alpha_0 x + \alpha_1 y + \alpha_3 = d \quad (1)$$

where (x, y) are the image coordinates of the pixel with the observed disparity d . Given enough number of points on the ground, it is possible to estimate the model parameters in a robust least squares sense using RANSAC.

Once the ground plane has been estimated using the above described method, we use it to segment each pixel for which there is a valid depth information as either obstacle or free-space. This is accomplished using a simple threshold on the residual of the fit for each observation, i.e. the i^{th} pixel is labeled ground if $|\alpha_1 x_i + \alpha_2 y_i + \alpha_3 - d_i| \leq d_g$ (where d_g is a constant threshold), and is labeled an obstacle if the residual exceeds another threshold. This near-range segmentation is then used as a self-supervised training set for the appearance-based learning algorithm.

B. Image-based Scene Segmentation

The problem of learning far-range scene segmentation from appearance can be defined as associating each pixel of an input image to one of the semantic classes. Here, we reduce the semantic scene understanding problem to learning a two-class classification: obstacle and free-space. This is the minimum classification required by an autonomous navigation system onboard an MAV to plan collision-free trajectories. This approach can easily be extended to tackle multi-class problems, so as to provide more semantically rich outputs if required.

Inspired by the recent progress with deep learning, we approach the problem with a CNN that can be trained end-to-end to predict a map of class-labels. Neural network models based-on Fully Convolutional Network [3] or SegNet [21], have shown exceptional performance on pixel-level segmentation tasks. However, most of these networks typically employ a VGG-16 [22] architecture (or similar), which is a very large model, originally designed for multi-class classification. These networks have huge numbers of parameters and long inference times, making them infeasible for robotics applications [23], which require processing images in real-time on low-latency embedded devices.

In this work, we advocate that reducing the computational burden of semantic segmentation is essential towards making them feasible for deployment on embedded systems for real-world robotics applications. Thus, we design a network based on the E-Net architecture [24], that is optimized for both fast inference and high accuracy. E-Net introduces a deep convolutional encoder-decoder model with a bottleneck structure, motivated by ResNets [25], to build an efficient network architecture that is $18\times$ faster, has $79\times$ less parameters but still achieves similar accuracy to prior models.

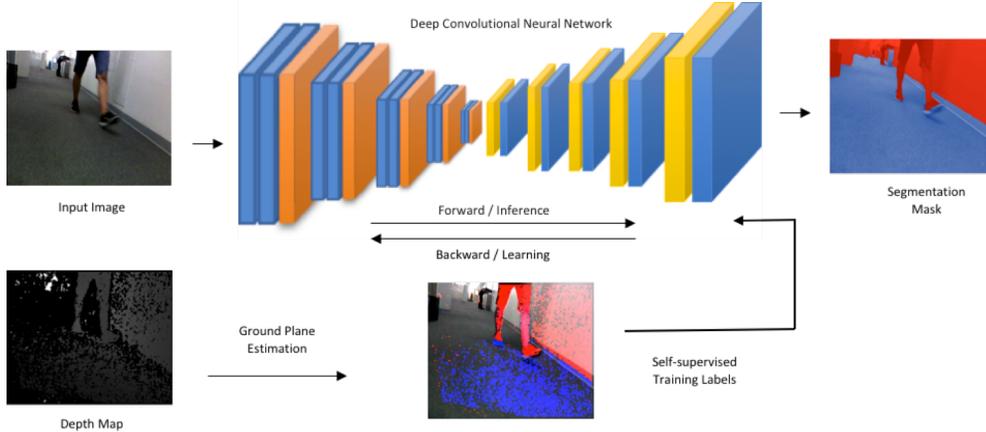


Fig. 2. Schematic overview of the near-to-far scene segmentation approach. Geometric information from the near-range via ground plane estimation is used to generate self-supervised training labels in real-time. These labels are then used to adaptively train a classifier for appearance-based scene segmentation.

The network architecture consists of encoder (initial and stage 1–3) and decoder segments (stage 4–5) with initial, down-sample, up-sample, and bottleneck modules. The encoder maps the input to a low resolution representation, while the decoder maps the low resolution feature maps to high-resolution segmentation output. The initial layer consists of a 3x3 convolutional layer. In parallel to this, a max-pooling layer outputs 3 feature-maps - one for each color channel of an RGB image. The outputs of these layers are then concatenated, making 16 feature-maps. Rest of the network consists of bottleneck modules.

The bottleneck module (as shown in Fig. 3) has an architecture of a single main branch and a separated branch with 3 convolutional filters: a 1x1 projection for dimensionality reduction, a main convolutional layer (conv) and a 1x1 projection. The conv layer is either a regular, dilated or asymmetric convolution. Further, between all convolutions, a batch normalization and a PReLU activation layer is placed, and spatial dropout layer is included for regularization in the bottleneck modules. If the bottleneck is down-sampling, a max-pooling layer is added to the main branch, the 1x1 projection is replaced with a 2x2 convolution with stride of 2 in both dimensions. Max-pooling is replaced with max unpooling in the decoder, and zero-padding is replaced with spatial convolution without bias. We refer readers to [24] for more details regarding the design choices of the network.

C. Online Training

Let us represent our network model as $f(x, \gamma)$, that maps an input image x to the target segmentation. The model is described by the network parameters γ and is learned online by minimizing its error output for an instance x_i given an output ground-truth label y_i :

$$\gamma = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^N L(f(x_i, \gamma), y_i) \quad (2)$$

where N is the self-supervised training set generated from near-range geometry-based segmentation, and L is the cross-entropy (softmax) loss. During inference, the softmax is replaced by an argmax function, so as to provide a single

output per pixel. The outputs of the network are scores for each of the learned categories.

Training a neural network typically requires a large amount of data samples for convergence of the parameters. However, several strategies have been adopted in literature to handle this. One of the popular methods being to initialize the network parameters with a network trained on a larger generic dataset in an unsupervised manner [26]; to apply it to a new task, one can simply fine-tune the last layers of the network. This exploits the observation that these pre-trained networks are a compact and yet rich representation of the images in general. We perform training in a similar multi-stage fashion.

In the offline step, we initialized the encoder part of the network with pre-trained weights from a generic dataset (different from the test dataset). The decoder part is initialized using Xavier initialization, and fine-tuned online in real-time by back-propagation using a stochastic gradient descent in a sliding window fashion, where each mini-batch consists of the last N frames. Furthermore, to improve the runtime efficiency of the network, we modified the network hyper-parameters; used higher learning rate, from $1e^{-10}$ to $1e^{-9}$ and lower momentum of 0.90 instead of 0.99. We also changed the fixed learning rate (L_r) with a poly-learning policy

$$L_r = L(1 - i/\max_i)^p \quad (3)$$

where L is the base learning rate, i is the learning step and p is the power index.

The proposed modifications of network parameters makes our approach more run-time efficient. From such a reduction, one might expect a drop in segmentation accuracy; however, it is to be noted that there is an inherent trade-off between segmentation quality and run-time that needs to be optimized for real-world applications. Our current work is not meant to offer an exhaustive test on optimizing the network architecture or the hyper parameters of the training process. We acknowledge the fact that our results may be improved by investigating that more properly, but the focus in this paper is to show the feasibility of online self-supervised near-to-far learning paradigm in the context of scene segmentation.

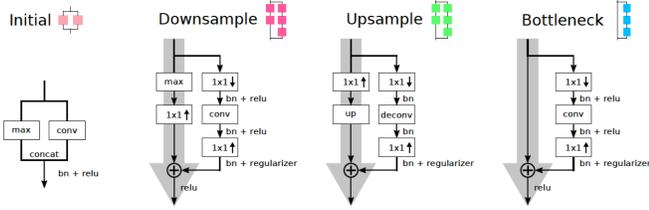


Fig. 3. E-Net modules used in our network. max: maxpooling layer with non-overlapping 2x2 windows, up: upsample layer by a factor of 2, conv: either a regular, dilated, or asymmetric convolution layer, bn: batch normalization, regularizer:spatial dropout, 1x1 with down or up arrow: 1x1 convolution to reduce or expand channels.

IV. EXPERIMENTS AND RESULTS

In this section we analyze the qualitative and quantitative performance of our proposed method on benchmark datasets, and demonstrate its efficacy through real-world flight experiments on a MAV.

A. Performance on Benchmark Datasets

Datasets: Accurately measuring scene segmentation performance is a challenging task. Ground truth labels for far-range scene is difficult and tedious to create, particularly in natural unstructured terrain, and running the entire system enough times, and over enough courses, to statistically show improved performance in the context of MAVs is extremely time consuming. Furthermore, at the time of writing this paper, the authors were not aware of any real-world large-scale publicly available MAV datasets with dense annotations for scene segmentation. Recently, high-fidelity simulators [27] have been proposed as a proxy to real-world datasets, to enable development and testing of autonomous MAVs. However, in our experience visual learning using simulations do not directly translate to real-world performance.

To overcome these issues, we take advantage of the recent developments in large scale datasets for semantic scene understanding for self-driving cars and autonomous ground vehicles. We use 2 datasets, Freiburg Forest [28] and Cityscapes [29], containing diverse environments ranging from urban environments to outdoor natural landscapes. Freiburg forest dataset contains multimodal images of forest environments and varying conditions such as low-lighting, snow, glare and motion blur, and semantic labels of 6 classes: Sky, Obstacles, Road, Grass, Vegetation, Background and Void. We use RGB and depth modalities as the input to our system and combine the class labels into a yield a binary obstacle vs free-space category; for example, grass and road belongs to free-space, and the rest as obstacles. In comparison, the Cityscapes dataset contains RGB and depth images from over 50 cities with varying seasons, time of the day and weather conditions. We only use near-range depth data ($\leq 15m$), to simulate the scenario of low-SWaP sensors on a MAV.

Baseline: To evaluate the system’s capability in real-world applications, we envision scenarios where an end-user would typically train a network on publicly available datasets and use it on new tasks or systems, without the

TABLE I
NEURAL NETWORK ARCHITECTURE

| Name | Type | Output Size |
|--|--------------|-------------|
| Initial | | 16x256x256 |
| bottleneck1.0 | downsampling | 64x128x128 |
| 4x bottleneck1.x | | 64x128x128 |
| bottleneck2.0 | downsampling | 128x64x64 |
| bottleneck2.1 | | 128x64x64 |
| bottleneck2.2 | diluted 2 | 128x64x64 |
| bottleneck2.3 | asymmetric 5 | 128x64x64 |
| bottleneck2.4 | dilated 4 | 128x64x64 |
| bottleneck2.5 | | 128x64x64 |
| bottleneck2.6 | dilated 8 | 128x64x64 |
| bottleneck2.7 | asymmetric 5 | 128x64x64 |
| bottleneck2.8 | diluted 16 | 128x64x64 |
| <i>Repeat section 2, without bottleneck2.0</i> | | |
| bottleneck4.0 | upsampling | 64x128x128 |
| bottleneck4.1 | | 64x128x128 |
| bottleneck4.2 | | 64x128x128 |
| bottleneck5.0 | upsampling | 16x256x256 |
| bottleneck5.1 | upsampling | 16x256x256 |
| fullconv | | 2x512x512 |

TABLE II
COMPARISON TO GROUND TRUTH

| Test Dataset | Training Dataset | Method | IoU | AP |
|--------------|------------------|---------------------|--------------|--------------|
| Freiburg | Cityscapes | FCN-8s | 76.98 | 87.36 |
| Freiburg | Cityscapes | SegNet | 72.12 | 82.90 |
| Freiburg | Cityscapes | E-Net | 71.37 | 81.21 |
| Freiburg | Cityscapes | Ours (E-Net+Online) | 87.11 | 91.12 |
| Cityscape | Freiburg | FCN-8s | 71.87 | 77.31 |
| Cityscape | Freiburg | SegNet | 69.43 | 74.67 |
| Cityscape | Freiburg | E-Net | 68.25 | 72.11 |
| Cityscape | Freiburg | Ours (E-Net+online) | 80.36 | 82.21 |

need to get dense annotation again. To emulate this, we train and test our system on similar but unrelated datasets - for e.g. train on Cityscapes and test on Freiburg Forest, and vice versa. As baseline, we compare the performance of our approach to standard semantic segmentation methods (FCN-8s, SegNet and E-Net) trained in a supervised manner, with no online fine-tuning. For our proposed approach, we train only the encoder segment of the network, initialize the decoder weights using Xavier initialization and fine-tune it using online updates in real-time.

Quantitative Comparison to Ground Truth: We benchmark against standard image segmentation metrics such as Mean Intersection-over-Union (IoU) and Average precision (AP) for both the above datasets. Table II shows the performance of our proposed algorithm on the above metrics, and Fig. 4 illustrates some of the corresponding qualitative results. Table III shows the runtime performance of the different methods on a Nvidia Jetson TX1 processor.

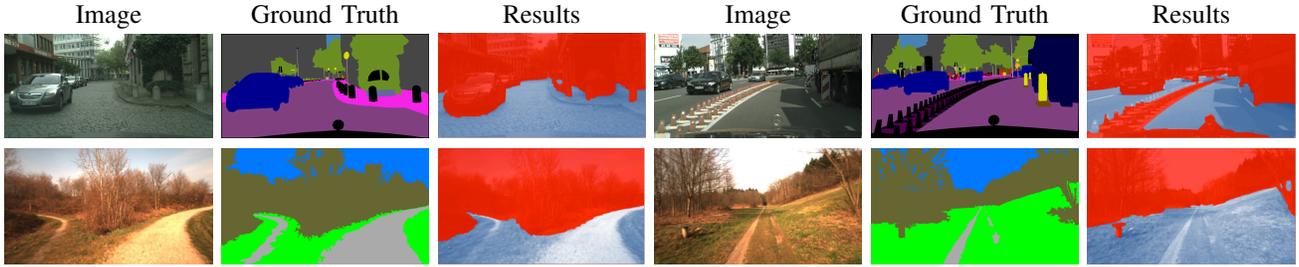


Fig. 4. Qualitative results scene segmentation on (Row-1) Cityscapes and (Row-2) Freiburg Forest datasets. Note: Red is obstacle, blue is free-space.

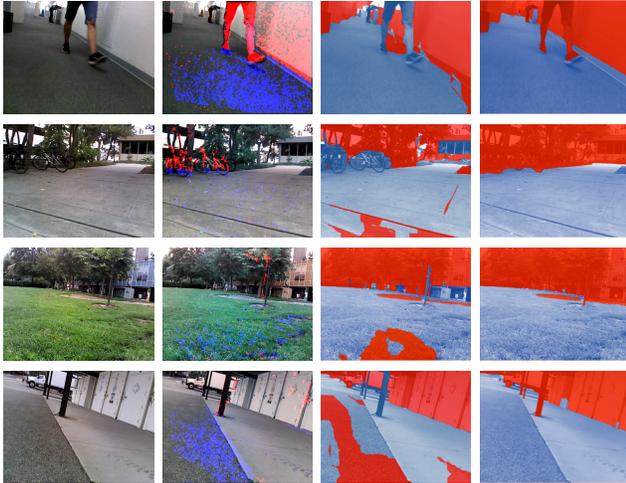


Fig. 5. Qualitative results of near-to-far scene segmentation from MAV flight tests in different indoor and outdoor environments. (Col. 1) Input RGB image, (Col. 2) Near-range geometry-based segmentation overlaid on the input image; this is further used for online training of the classifier (Col. 3) Overlaid segmentation mask for the baseline algorithm (without online learning component) and (Col. 4) Overlaid segmentation mask for the proposed approach. Note: Red is obstacle, blue is free-space.

Our method achieves an IoU of 87.11 and 80.36 on the Freiburg forest and Cityscapes datasets, respectively; outperforming the other approaches by a significant margin. This improvement can be attributed to the highly representational features adaptively learned by our model in real-time. Computationally, the geometry-based near-range segmentation module runs at ~ 10 hz, and the image-based segmentation module runs at ~ 1 hz; this involves both the inference, and the online update step. These run-time are sufficient for our application of high-speed autonomous navigation, where we envision a slower long-range strategic planning layer running in tandem with a faster near-range geometry-based reactive avoidance system.

B. Performance on Real-world MAV Flights

Setup: We evaluate our proposed approach through real-world flights tests in a varied set of indoor and outdoor environments, ranging from untextured building corridors to natural cluttered scenes. We use a modified version of the Asctec Hummingbird quad-copter platform. Our flight computer is a Nvidia Jetson TX1 board, which incorporates a quad-core ARM processor and embedded GPU with 256 Cuda cores. Onboard, we have a downward looking camera

TABLE III
RUNTIME PERFORMANCE ON NVIDIA JETSON TX1

| Method | Runtime (fps) |
|---------------------|---------------|
| FCN-8s | 0.43 |
| SegNet | 1.32 |
| E-Net | 19.71 |
| Ours (E-Net+online) | 1.12 |

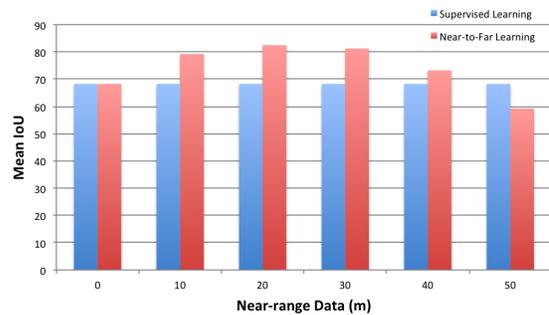


Fig. 6. Performance of our approach on Cityscapes dataset as a function of near-range depth information.

for state-estimation, and a front-facing Intel RealSense R200 structured light sensor that provides us with visual and range data. Tensorflow deep learning library with CuDNN backend was used for our implementation, and all the processing is done onboard. During the experiments, the MAV was manually flown by a pilot, while the near-to-far segmentation module was run in real-time as software-in-the-loop. We decoupled the segmentation module from the control and planning layer, so as to independently evaluate the performance of the proposed approach. Closing the loop with planning and control to show system-wide advantage of the proposed algorithm is beyond the scope of this paper, and will be addressed in future work.

Qualitative Results: Fig. 5 shows the qualitative results from some of the flight experiments. Similar to performance evaluation on benchmark datasets, we compare our proposed approach to a system trained only in the supervised manner, without near-to-far online component. We used the Cityscapes dataset for the off-line supervised training. It can be seen that using self-supervised learning in a continuous fashion performs better, especially in scenarios where the environment changes dynamically. In the first example, we notice that as soon as the person appears in near-field, the

perception algorithm is able to adapt its learning to classify people as potential obstacles; this is completely missed in the supervised scenario. Similar behavior is observed in outdoor environments, in the context of narrow tree trunks as shown in the third row. Further, the ability of the near-to-far scene segmentation to learn from its current 'local' information is evident in the last example where it can successfully segment the planar surface, even though they have a considerably different texture and visual appearance.

C. Sensitivity Analysis

Performance vs. Depth-range: Our algorithm is highly dependent on the fidelity of the self-supervised training, since sparse depth information may only be representative for only a part of the scene. Thus we analyze the performance of our algorithm as a function of the range of self-labelled data available. As shown in Fig. 6, the performance of the algorithm gets better to a certain range but degraded beyond that. This can be explained due to the fact that the ground-plane assumptions no longer hold at far-range, and thus the labels are more noisy.

V. CONCLUSION

In this paper, we introduced the concept of near-to-far perception system - an online self-supervised learning based approach to enable long-range scene segmentation, and support this claim through qualitative and quantitative results. We studied this in the context of autonomous flights for resource constrained MAVs. However, our approach and findings equally apply to other autonomous systems with similar sensing modalities. Finally with this work we hope to bridge the gap between geometry-based and data-driven approaches, by taking a step in the direction of building systems that exploit the complimentary benefits of both world. In ongoing and future work, we plan to close the loop on perception and planning, so as to allow end-to-end system evaluation.

VI. ACKNOWLEDGEMENT

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

REFERENCES

- [1] D. Dey, K. S. Shankar, S. Zeng, R. Mehta, M. T. Agcayazi, C. Eriksen, S. Daftry, M. Hebert, and J. A. Bagnell, "Vision and learning for deliberative monocular cluttered flight," in *In Proceedings of the International Conference on Field and Service Robotics (FSR)*, 2015.
- [2] L.-J. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2036–2043.
- [3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [4] S. Daftry, S. Zeng, J. A. Bagnell, and M. Hebert, "Introspective perception: Learning to predict failures in vision systems," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 1743–1750.
- [5] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. R. Bradski, "Self-supervised monocular road detection in desert terrain." in *Robotics: science and systems*, vol. 38. Philadelphia, 2006.
- [6] D. Lieb, A. Lookingbill, and S. Thrun, "Adaptive road following using self-supervised learning and reverse optical flow." in *Robotics: Science and Systems*, 2005, pp. 273–280.
- [7] M. Bajracharya, B. Tang, A. Howard, M. Turmon, and L. Matthies, "Learning long-range terrain classification for autonomous navigation," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE, 2008, pp. 4018–4024.
- [8] P. Vernaza, B. Taskar, and D. D. Lee, "Online, self-supervised terrain classification via discriminatively trained submodular markov random fields," in *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*. IEEE, 2008, pp. 2750–2757.
- [9] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuoglu, U. Muller, and Y. LeCun, "Learning long-range vision for autonomous off-road driving," *Journal of Field Robotics*, vol. 26, no. 2, pp. 120–144, 2009.
- [10] A. Hajjam, "A near-to-far learning framework for terrain characterization using an aerial/ground-vehicle team," 2016.
- [11] D. Kim, J. Sun, S. M. Oh, J. M. Rehg, and A. F. Bobick, "Traversability classification using unsupervised on-line visual learning for outdoor robot navigation," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE, 2006, pp. 518–525.
- [12] B. Lee, K. Daniilidis, and D. D. Lee, "Online self-supervised monocular visual odometry for ground vehicles," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5232–5238.
- [13] B. Sofman, E. Lin, J. A. Bagnell, J. Cole, N. Vandapel, and A. Stentz, "Improving robot navigation through self-supervised online learning," *Journal of Field Robotics*, vol. 23, no. 11-12, pp. 1059–1075, 2006.
- [14] G. L. Oliveira, W. Burgard, and T. Brox, "Efficient deep models for monocular road segmentation," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 4885–4891.
- [15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.
- [16] M. Trembl, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich *et al.*, "Speeding up semantic segmentation for autonomous driving," in *MLITS, NIPS Workshop*, 2016.
- [17] S. Daftry, J. A. Bagnell, and M. Hebert, "Learning transferable policies for monocular reactive mav control," in *International Symposium on Experimental Robotics*. Springer, 2016, pp. 3–11.
- [18] S. Daftry, Y. Agrawal, and L. Matthies, "Online self-supervised scene segmentation for micro aerial vehicles," *arXiv preprint arXiv:1806.05269*, 2018.
- [19] D. Pathak, E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional multi-class multiple instance learning," *arXiv preprint arXiv:1412.7144*, 2014.
- [20] W. P. Sanberg, G. Dubbleman *et al.*, "Free-space detection with self-supervised and online trained fully convolutional networks," *Electronic Imaging*, vol. 2017, no. 19, pp. 54–61, 2017.
- [21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.
- [22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [23] A. Mkhitarian, "Pixel-wise semantic segmentation for low-power devices."
- [24] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [26] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [27] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*. Springer, 2018, pp. 621–635.
- [28] A. Valada, G. Oliveira, T. Brox, and W. Burgard, "Deep multispectral semantic scene understanding of forested environments using multi-

modal fusion,” in *The 2016 International Symposium on Experimental Robotics (ISER 2016)*, Tokyo, Japan, Oct. 2016.

- [29] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.