Learning Reliable and Scalable Representations Using Multimodal Multitask Deep Learning

Abhinav Valada, and Wolfram Burgard
Department of Computer Science, University of Freiburg, Germany

I. Introduction

> Fifties - in 5 years robots would be everywhere.
> Sixties - in 10 years robots would be everywhere.
> Seventies - in 20 years robots would be everywhere.
> Eighties - in 40 years robots would be everywhere.
> *-Marvin Minsky*

Those were the words from one of the pioneers of AI when asked to comment on the progress of robotics in the twentieth century. This shows the high expectations and unforeseen challenges that we are faced with for deploying robots in complex real-world environments. One of the primary impediments has been the robustness of scene understanding models as it is a prerequisite for any action execution or planning. The tremendous progress made in machine learning in the last decade has enabled us to learn representations from raw sensor data, rather than relying on hand-engineered features. However, these models still perform inconsistently, especially in challenging weather conditions. The current dominant paradigms rely on camera images or depth data. However, alternate modalities such as infrared [1] and sound [2] need to be exploited for learning the most comprehensive information about the scene that will enable us to reduce perceptual ambiguity in challenging conditions.

State-of-art deep learning models rely on thousands to millions of annotated training data and acquiring this data is an arduous task, if not impossible for every foreseeable scenario and for every task. Moreover, learning task-specific models on task-specific datasets, limits the overall learning ability of the robot as most models are trained in a supervised fashion and independently, therefore they have no ability to share cross-domain information and exploit training signals from complementary tasks. In order to address this limitation, our models should be able to learn representations across different modalities as well as reuse and share the learned representations across different tasks.

My work enables models to effectively learn fused representations from multiple modalities and across tasks, exploiting complementary features and cross-modal interdependencies. Despite that fact that we are still far away from creating robots with human-level intelligence, equipping them with these basic capabilities will enable robots to learn new tasks from limited amount of data by leveraging transfer learning which facilitates self-supervised model adaptation. Advancing self-supervised learning techniques to autonomous learning is a strong starting point that will enable robots to continuously learn from what it experiences and perceives in the real-world.

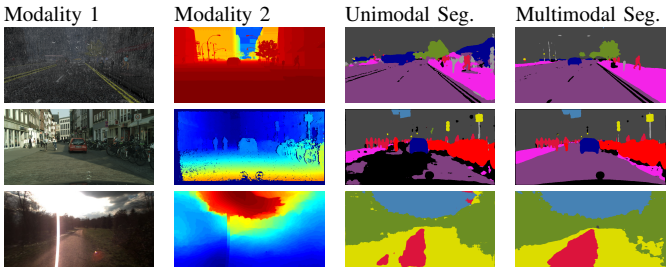


Fig. 1. Comparison of segmentation obtained using only RGB versus our multimodal fusion [8] on the datasets: Synthia, Cityscapes and Freiburg Forest.

II. Robust Scene Understanding

In the last decade, there has been a sharp transition in semantic segmentation approaches from employing hand engineered features with flat classifiers such as Support Vector Machines [3] or Random Forests [4, 5], to end-to-end Deep Convolutional Neural Network (DCNN) based approaches [6, 7]. However, a big drawback in employing the top performing DCNN approaches is the computational complexity and substantially large inference time despite using modern GPUs that hinder them from being deployed in robots. In my work, I developed solutions for fully-convolutional semantic segmentation architectures tailored for real-world robotic perception that enable them to achieve state-of-the-art performance without compromising on speed or memory requirements so that they can be efficiently deployed on embedded GPUs. Some of these improvements include multi-scale residual skip layers [8], pyramid decomposition to enable faster inference, and multistage refinement for high-resolution segmentation [1].

In an effort to improve robustness as well as granularity of the segmentation, novel approaches that exploit features from alternate modalities have been proposed [9, 10, 1]. However most of these approaches naively combine feature maps from modality-specific networks at various stages by element-wise concatenation or summation which inhibits learning complementary features. I introduced a fusion scheme [11, 8] that empowers the network with the ability to choose class-specific features based on the scene condition, followed by learning deeper representations from the mixture of fused kernels. More specifically, the proposed framework consists of three components: modality-specific networks that map the representation of the input to corresponding segmentation outputs, an adaptive gating fusion layer that acts like a multiplexer which maps outputs of expert networks to a probabilistically fused representation, and finally a fusion segment that further learns complementary fused kernels. Extensive experiments on publicly available datasets including Cityscapes [12], Synthia [13]

(a) Input frame (b) Ego-motion subtracted flow (c) Semantic motion segmentation

Fig. 2. Illustration of semantic motion segmentation using the SMSnet [14] architecture on the Cityscapes dataset. (Blue=Static Car, Green=Moving Car)

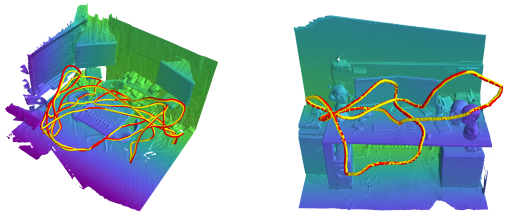
and Freiburg Forest [1], demonstrated that my proposed fusion approach exceeds the performance of existing fusion techniques and more importantly yields an accurate representation of scene in adverse conditions including rainfall and snow. Fig 1 shows qualitative comparisons of segmentation using unimodal RGB and the proposed multimodal fusion scheme.

For autonomous robots navigating in urban environments, it is not only imperative to understand the distinction between different objects in the scene such as cars, roads, and buildings, but also be able to distinguish between a moving and a static car so that it can plan paths based on this joint knowledge. There are several challenges that make this problem inherently hard including the ego-motion of the camera, lighting changes between consecutive frames and varying pixel displacements due to motion with different velocities. To this end, I proposed a DCNN architecture [14] that learns to predict both the semantic category and motion status of each pixel from a pair of consecutive monocular images. The network builds upon the aforementioned segmentation architecture [8] and fuses semantic features with learned motion features from generated optical flow maps to yield pixel-wise semantic motion segmentation. This work demonstrated the utility of jointly learning both these tasks as the features learned to distinguish object classes help infer motion labels for the corresponding pixels and motion in the image improves the inference of object distinction. This work is currently the state-of-the-art for semantic motion segmentation on the Cityscapes [12] and KITTI [15] datasets. Fig. 2 shows an input frame, the generated optical flow image with the ego-motion subtracted and the semantic motion segmentation output.

III. Geometry and Structure-Aware Localization

Visual localization is one of the fundamental enablers of robot autonomy as it offers navigation capabilities using low-cost sensors. It has mostly been tackled using local feature-based pipelines [16, 17] that efficiently encode knowledge about the environment and the underlying geometrical constraints. While, DCNN-based approaches for pose regression [18, 19] have shown considerable robustness in the context of significant perceptual changes, repeating structures and texture-less regions, they have been unable to match the performance of state-of-the-art local feature-based localization methods. In my work, I have proposed deep multitask learning models that overcome these limitations by simultaneously embedding geometric and semantic knowledge of the world into the pose regression network [20, 21].

The proposed architecture [21] consists of four CNN streams: a global pose regression stream, a semantic segmentation stream and Siamese-type double stream for visual odometry estimation. The framework aims to exploit the inter-task relationship between learning semantics, regressing 6-DoF global pose and odometry, for the mutual benefit of each of these tasks. I



(a) Chess (b) Heads

Fig. 3. Qualitative localization results of VLocNet++ [21] on the Microsoft 7 Scenes dataset depicting the estimated global pose (yellow) versus the ground-truth pose (red) plotted with respect to the 3D scene model for visualization

proposed a new loss function for global pose regression that incorporates the predicted relative motion information from the odometry stream during training and enforces the predicted poses to be geometrically consistent with respect to the true motion model. In order to instill structural cues about the environment into the global pose regression stream, I proposed an adaptive weighted fusion layer that fuses semantic features into the pose regression stream based on region activations. To further exploit the inter-task relationship, I proposed a self-supervised warping technique that uses the relative motion from the odometry stream to warp intermediate network representations in the segmentation stream for learning consistent semantics. This model is currently the state-of-the-art on the challenging Microsoft 7-Scenes benchmark, outperforming existing learning-based competitors and local feature-based pipelines.

IV. Future Work

In the short term, my focus will be on combining the aforementioned semantic localization framework with the semantic motion segmentation network. Subtracting the ego-motion from the optical flow prediction plays a crucial role in the estimation of motion features. Moreover, a large number of dynamic objects in the scene quickly degrades the localization performance. By learning both these models jointly, the semantic motion estimation network benefits from utilizing the learned relative motion and simultaneously, the localization network can benefit from focusing its attention on regions in the image that do not have dynamic objects. Furthermore, incorporating the predicted pixel-wise ephemerality mask [22] from an auxiliary stream can further benefit both the tasks.

The overall goal of my work is to eventually be able to equip robots with models that are able to aid in lifelong learning. The two main directions that I wish to pursue towards this goal are: self-supervised as well as unsupervised learning, and uncertainty-aware multitask learning. As we train models to perform more and more complex tasks jointly, it will no longer be feasible to annotate data for a multitude of tasks. I aim to make progress towards self-supervised and unsupervised techniques that use intrinsic signals within the data and that leverage prior knowledge across tasks. Estimating the uncertainty of predictions is crucial for making decisions on top of the model estimates. Further building upon the unsupervised learning scenario, I plan to develop learning approaches for uncertainty-aware predictions [23] that are capable of performing self-calibration based on the uncertainty. This is the most exciting time to be working on artificial intelligence (and hopefully for more than 40 years) and I believe the best is yet to come.

References

[1] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, *Deep Multispectral Semantic Scene Understanding of Forested Environments Using Multimodal Fusion*. Cham: Springer International Publishing, 2016, pp. 465–477.

[2] A. Valada and W. Burgard, "Deep spatiotemporal models for robust proprioceptive terrain classification," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1521–1539, 2017.

[3] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proceedings of the International Conference on Computer Vision*, 2009.

[4] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2008.

[5] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Proceedings of the European Conference on Computer Vision*, D. Forsyth, P. Torr, and A. Zisserman, Eds., 2008.

[6] 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.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2017.

[8] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2017.

[9] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian Conference on Computer Vision*, 2016, pp. 213–228.

[10] D.-K. Kim, D. Maturana, M. Uenoyama, and S. Scherer, "Season-invariant semantic segmentation with a deep multimodal network," in *Field and Service Robotics*, 2018, pp. 255–270.

[11] A. Valada, A. Dhall, and W. Burgard, "Convoluted mixture of deep experts for robust semantic segmentation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems Workshop, State Estimation and Terrain Perception for All Terrain Mobile Robots*, 2016.

[12] 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 Conference on Computer Vision and Pattern Recognition*, 2016.

[13] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2016.

[14] J. Vertens, A. Valada, and W. Burgard, "Smsnet: Semantic motion segmentation using deep convolutional neural networks," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.

[15] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2012.

[16] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 9, pp. 1744–1756, 2017.

[17] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in rgb-d images," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2930–2937.

[18] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2017.

[19] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala, "Camera relocalization by computing pairwise relative poses using convolutional neural network," *arXiv preprint arXiv:1707.09733*, 2017.

[20] A. Valada, N. Radwan, and W. Burgard, "Deep auxiliary learning for visual localization and odometry," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018.

[21] N. Radwan, A. Valada, and W. Burgard, "Vlocnet++: Deep multitask learning for semantic visual localization and odometry," *arXiv preprint arXiv:1804.08366*, 2018.

[22] D. Barnes, W. Maddern, G. Pascoe, and I. Posner, "Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments," *arXiv preprint arXiv:1711.06623*, 2017.

[23] E. Ilg, Ö. Çiçek, S. Galesso, A. Klein, O. Makansi, F. Hutter, and T. Brox, "Uncertainty estimates for optical flow with multi-hypotheses networks," *arXiv preprint arXiv:1802.07095*, 2018.