# Incorporating Semantic and Geometric Priors in Deep Pose Regression

Abhinav Valada* Noha Radwan* Wolfram Burgard
Department of Computer Science, University of Freiburg, Germany

***Abstract*—Deep learning has enabled recent breakthroughs across a wide spectrum of scene understanding tasks, however, its applicability to camera pose regression has been unfruitful due to the direct formulation that renders it incapable of encoding scene-specific constrains. In this work, we propose the VLocNet++ architecture that overcomes this limitation by simultaneously embedding geometric and semantic knowledge of the world into the pose regression network. We employ a multitask learning approach 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. Furthermore, in order to enforce global consistency during camera pose regression, we propose the novel Geometric Consistency Loss function that leverages the predicted relative motion estimated from odometry to constrict the search space while training. Extensive experiments on the challenging Microsoft 7-Scenes benchmark and our DeepLoc dataset demonstrate that our approach exceeds the state-of-the-art outperforming local feature-based methods while simultaneously performing multiple tasks and exhibiting substantial robustness in challenging scenarios.**

## I. Introduction

Visual localization is a fundamental transdisciplinary problem and a crucial enabler for numerous robotics as well as computer vision applications, including autonomous navigation, simultaneous localization and mapping, structure-from-motion and augmented reality. Recently, deep learning-based localization approaches [1, 2, 3, 4, 5] have shown considerable robustness in the context of significant perceptual changes, repeating structures and textureless regions. However, their performance has been subpar in comparison to state-of-the-art local feature-based pipelines [6, 7] as they perform direct pose regression from image embeddings using naive loss functions.

In this work, we propose a principled approach to simultaneously embed geometric and semantic knowledge of the world into the pose regression model, complemented with a novel loss function that enforces the predicted poses to be geometrically consistent with respect to the true motion model. To achieve this, we approach this problem from a multitask learning (MTL) perspective and propose a framework [8] that jointly learns semantic segmentation, visual localization and odometry from consecutive monocular images. Our network utilizes our proposed Geometric Consistency loss function [9] that incorporates relative motion information from a shared auxiliary odometry stream to learn a model that is globally consistent. As our network also needs to effectively utilize the learned motion specific features from the previous timestep, we introduce an adaptive weighting technique to aggregate motion-specific temporal information in the global pose regression network.

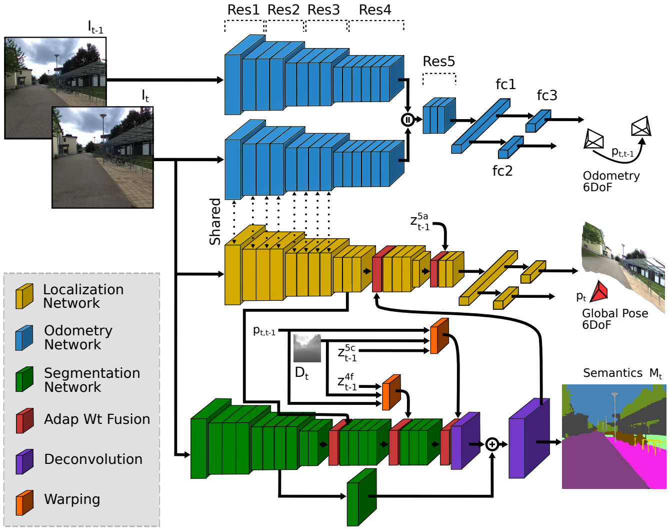


Fig. 1. Schematic representation of our proposed VLocNet++ architecture. The network takes two consecutive monocular images as input and simultaneously predicts the global 6-DoF pose, odometry and semantics of the scene.

Existing semantics-aware localization techniques extract predefined stable features, emphasize [10] or combine them with local features [11] but often fail when the predefined structures are occluded or not visible in the scene. In contrast, our approach is robust to such situations as we use the proposed adaptive weighting layer to selectively fuse learned relevant features not only based on the semantic category but also the activations in the region. Moreover, by jointly estimating the semantics, we instil structural cues about the environment into the pose regression network and implicitly pull the attention towards more informative regions in the scene. Inspired by early cognitive studies in humans showing the importance of learning self-motion for acquiring basic perceptual skills [12], we also propose a novel self-supervised semantic context aggregation technique leveraging the predicted relative motion from the odometry stream. This enables our semantic segmentation network to aggregate more scene-level context, thereby improving the performance and leading to faster convergence.

## II. Technical Approach

Our architecture, depicted in Fig. 1 consists of four CNN streams; a global pose regression stream, a semantic segmentation stream and a Siamese-type double stream for odometry estimation. Given a pair of consecutive monocular images $I_{t-1}, I_t \in \mathbb{R}^{\rho}$, the pose regression stream predicts the global pose $\mathbf{p}_t = [\mathbf{x}_t, \mathbf{q}_t]$ for image $I_t$, where $\mathbf{x} \in \mathbb{R}^3$ denotes the translation and $\mathbf{q} \in \mathbb{R}^4$ denotes the rotation in quaternion representation, while the semantic stream predicts a pixel-wise segmentation

*These authors contributed equally.

mask $M_t$ mapping each pixel $u$ to one of the $C$ semantic classes, and the odometry stream predicts the relative motion $\mathbf{p}_{t,t-1} = [\mathbf{x}_{t,t-1}, \mathbf{q}_{t,t-1}]$ between consecutive input frames.

### A. Network Architecture

We base each stream of our network on the ResNet-50 [13] architecture as it offers a good trade-off between learning highly discriminative deep features and the computational complexity required. For both the camera pose regression and visual odometry streams, we add a global average pooling layer after the fifth residual block, followed by three inner-product layers *fc1, fc2 and fc3* of dimensions 1024, 3 and 4 respectively, where *fc2* and *fc3* regress the translational $\mathbf{x}$ and rotational $\mathbf{q}$ components of the pose. Additionally, we use ELU [14] for the activation function as it helps in learning representations that are more robust to noise and also leads to faster convergence. In order to estimate the odometry, we adopt a Siamese-type double stream architecture, were we maintain separate streams upto the last downsampling stage (*Res4*), after which the feature maps are concatenated and convolved through the last residual block (*Res5*), followed by the regressors. While, for learning the semantics, we build upon our AdapNet [15] architecture which follows the general encoder-decoder design principle. The encoder incorporates our multi-scale residual units [15] which have dilated convolutions [16] parallel to the $3 \times 3$ convolutions for aggregating features from different spatial scales without increasing the number of parameters. The output of the encoder is 16-times downsampled with respect to the input dimensions, therefore our decoder consisting of deconvolution layers and skip refinement stages, upsamples the downscaled feature maps back to the input resolution.

We incorporate geometric knowledge into the global pose regression stream as three-folds: a) we employ hybrid hard parameter sharing between the camera pose regression stream and the odometry stream that both take the image from the current timestep as input. This exploits the task-specific similarities among both tasks, as well as influences the shared weights of the camera pose regression stream to integrate motion-specific features due to the inductive bias from odometry estimation, while effectuating implicit attention on regions that are more informative for relative motion estimation. b) As opposed to naively minimizing the Euclidean loss between the groundtruth and predicted poses, we employ our proposed Geometric Consistency Loss (Sec. II-B), which in addition to minimizing the Euclidean loss, adds another loss term to constrain the current pose prediction by minimizing the relative motion error between the ground truth and the estimated motion obtained from the odometry stream. c) Finally, in order to incorporate the relative motion information into the global pose regression stream, we integrate the intermediate representation from the last downsampling stage (*Res5a*) of the previous timestep into the current timestep. As opposed to naively concatenating these feature maps, which often accumulates irrelevant information, we utilize our proposed adaptive weighted fusion layer that learns the optimal element-wise weightings for the fusion based on the activations in the region, followed by a non-linear feature pooling over the weighted tensors. We formulate the output of our proposed fusion layer with respect to two activation maps $z^a$ and $z^b$ from layers $a$ and $b$ as follows

$$\hat{z}_{fuse} = \max\left(\mathbf{W} * \left(\left(w^a \odot z^a\right) \oplus \left(w^b \odot z^b\right)\right) + \mathbf{b}, 0\right), \qquad (1)$$

where $w^a$ and $w^b$ are learned weightings having the same dimensions as $z^a$ and $z^b$; $\mathbf{W}$ and $\mathbf{b}$ are the parameters of the non-linear feature pooling; with $\odot$ and $\oplus$ representing per-channel scalar multiplication and concatenation across the channels; and $*$ representing the convolution operation.

Incorporating semantic knowledge of the environment into the pose regression stream enables the network to focus its attention on areas of the image that are more informative for estimating the current pose. In order to identify and fuse only the semantically relevant information, we utilize our adaptive fusion layer to fuse semantic feature maps into the global pose regression stream at *Res4c*, as shown by the red block in Fig. 1. Concurrently, knowledge of the camera poses can be used to learn a globally consistent semantic representation of the scene. In order to facilitate this action, we leverage the relative motion information from the odometry stream to warp intermediate feature maps of the segmentation stream from the previous timestep into the current view using a predicted depth map obtained from a CNN [17]. We then fuse the warped feature maps with the intermediate representations using the adaptive fusion layer at the end of *Res3* and *Res4* blocks. This does not require any pre-computation as it is fully differentiable. Moreover, by incorporating feature maps from multiple views and resolutions using the representational warping concept from multi-view geometry, we enable our model to be robust to camera angle deviations, object scale and frame-level distortions, while implicitly introducing feature augmentation which facilitates faster convergence.

### B. Loss Function

In this section, we first detail the loss functions that we use for training the task-specific networks, followed by the joint loss function for training the multitask model. For training the semantic segmentation network, we use the cross-entropy loss function to minimize the Kullback-Leibler divergence between the predicted and the groundtruth pixel labels. We define a set of training images $\mathcal{T} = \{(I_n, M_n) \mid n = 1, \ldots, N\}$, where $I_n = \{u_r \mid r = 1, \ldots, \rho\}$ denotes the input frame and the corresponding ground truth mask $M_n = \{m_r^n \mid r = 1, \ldots, \rho\}$, where $m_r^n \in \{1, \ldots, C\}$ is the set of semantic classes. We define $\theta$ as the network parameters. Using the classification scores $s_j$ at each pixel $u_r$, we obtain the probabilities $\mathbf{P} = (p_1, \ldots, p_C)$ with the $\mathsf{softmax}$ function $\sigma(.)$ such that $p_j(u_r, \theta \mid I_n) = \sigma\left(s_j\left(u_r, \theta\right)\right) = \frac{exp\left(s_j(u_r,\theta)\right)}{\sum_k^C exp(s_k(u_r,\theta))}$ denotes the probability of pixel $u_r$ being classified with label $j$. $\theta$ is estimated by minimizing

$$\mathscr{L}_{seg}(\mathcal{T}, \theta) = -\sum_{n=1}^{N} \sum_{r=1}^{\rho} \sum_{j=1}^{C} \delta_{m_r^n, j} \log p_j(u_r, \theta \mid I_n), \qquad (2)$$

for $(I_n, M_n) \in \mathcal{T}$, where $\delta_{m_r^n, j}$ is the Kronecker delta. For training the odometry network, we utilize the loss function shown in Eq. (3) which minimizes the Euclidean distance between the groundtruth and predicted relative motion.

$$\begin{aligned} \mathscr{L}_{vo}\left(f\left(\theta \mid I_t, I_{t-1}\right)\right) := {} & \mathscr{L}_x\left(f\left(\theta \mid I_t, I_{t-1}\right)\right) \exp(-\hat{s}_{x_{vo}}) \qquad (3) \\ & + \hat{s}_{x_{vo}} + \mathscr{L}_q\left(f\left(\theta \mid I_t, I_{t-1}\right)\right) \exp(-\hat{s}_{q_{vo}}) + \hat{s}_{q_{vo}}, \end{aligned}$$

$$\mathcal{L}_x\left(f\left(\theta \mid I_t, I_{t-1}\right)\right) := \left\|\mathbf{x}_{t,t-1} - \hat{\mathbf{x}}_{t,t-1}\right\|_2 \qquad (4)$$
$$\mathcal{L}_q\left(f\left(\theta \mid I_t, I_{t-1}\right)\right) := \left\|\mathbf{q}_{t,t-1} - \hat{\mathbf{q}}_{t,t-1}\right\|_2 .$$

where $\mathcal{L}_x$ and $\mathcal{L}_q$ refers to the translational and rotational components respectively. We also employ learnable weighting parameters, $\hat{s}_{x_{vo}}, \hat{s}_{q_{vo}}$, to balance the scale between components.

In order to learn geometrically consistent poses, we employ the proposed Geometric Consistent Loss function, which in addition to minimizing the Euclidean loss, adds another loss term to constrain the current pose prediction by minimizing the relative motion error between the groundtruth and the estimated relative pose. By utilizing the predictions of the network from the previous timestep along with the current prediction, the relative motion loss term $\mathcal{L}_{Rel}\left(f\left(\theta \mid I_t\right)\right)$ can be computed as a weighted summation of the translational and rotational errors. Eq. (5) details this loss term, in which we assume that the quaternion output of the network has been normalized a priori

$$\mathcal{L}_{Rel}\left(f\left(\theta \mid I_t\right)\right) = \mathcal{L}_{x_{Rel}}\left(f\left(\theta \mid I_t\right)\right)\exp(-\hat{s}_{x_{Rel}}) + \hat{s}_{x_{Rel}} \qquad (5)$$
$$+ \mathcal{L}_{q_{Rel}}\left(f\left(\theta \mid I_t\right)\right)\exp(-\hat{s}_{q_{Rel}}) + \hat{s}_{q_{Rel}}$$
$$\mathcal{L}_{x_{Rel}}\left(f\left(\theta \mid I_t\right)\right) := \left\|\mathbf{x}_{t,t-1} - \left(\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t-1}\right)\right\|_2$$
$$\mathcal{L}_{q_{Rel}}\left(f\left(\theta \mid I_t\right)\right) := \left\|\mathbf{q}_{t,t-1} - \left(\hat{\mathbf{q}}_{t-1}^{-1}\hat{\mathbf{q}}_t\right)\right\|_2 .$$

Following the notation, the Euclidean loss can be defined as

$$\mathcal{L}_{Euc}\left(f\left(\theta \mid I_t\right)\right) = \mathcal{L}_x\left(f\left(\theta \mid I_t\right)\right)\exp(-\hat{s}_x) \qquad (6)$$
$$+ \hat{s}_x + \mathcal{L}_q\left(f\left(\theta \mid I_t\right)\right)\exp(-\hat{s}_q) + \hat{s}_q .$$

The final loss term to be minimized is

$$\mathcal{L}_{loc}\left(f\left(\theta \mid I_t\right)\right) := \mathcal{L}_{Euc}\left(f\left(\theta \mid I_t\right)\right) + \mathcal{L}_{Rel}\left(f\left(\theta \mid I_t\right)\right) . \qquad (7)$$

By minimizing the aforementioned loss function, our network learns a model that is geometrically consistent with respect to the motion. Moreover, by employing the adaptive fusion layer to aggregate motion specific features temporally, we enable the Geometric Consistency Loss to efficiently leverage this information. In order to jointly learn all the tasks, we minimize the following loss function:

$$\mathcal{L}_{multi} := \mathcal{L}_{loc}\exp(-\hat{s}_{loc}) + \hat{s}_{loc} + \mathcal{L}_{vo}\exp(-\hat{s}_{vo}) + \hat{s}_{vo} \qquad (8)$$
$$+ \mathcal{L}_{seg}\exp(-\hat{s}_{seg}) + \hat{s}_{seg},$$

where $\mathcal{L}_{loc}$ is the global pose regression loss as per Eq. (7); $\mathcal{L}_{vo}$ is the visual odometry loss from Eq. (3), and $\mathcal{L}_{seg}$ is the cross-entropy loss for semantic segmentation from Eq. (2).

## III. Experimental Results and Conclusions

We evaluate the performance of our proposed approach on the indoor Microsoft 7-Scenes benchmark [6] and the outdoor DeepLoc dataset [8]. In Fig. 2, we present the localization accuracy of our proposed approach on the 7-Scenes dataset using the median localization error metric and the percentage of poses for which the error is below 5cm and 5°. From the results presented in Fig. 2, we see that our single-task VLocNet++ model achieves an accuracy of 96.4%, improving over the state-of-the-art [5] by 20.3% and by over an order of magnitude compared to the other deep learning approaches [1, 3, 2]. Moreover, by employing our proposed multitask framework, VLocNet++ further improves on the performance and achieves an accuracy of 99.2%, setting the new state-of-the-art on this benchmark. Furthermore, for the task of visual odometry estimation, VLocNet++ outperforms end-to-end learning approaches [18, 19, 20, 9] by 25.8% and 24.8% in the translational and rotational components respectively.

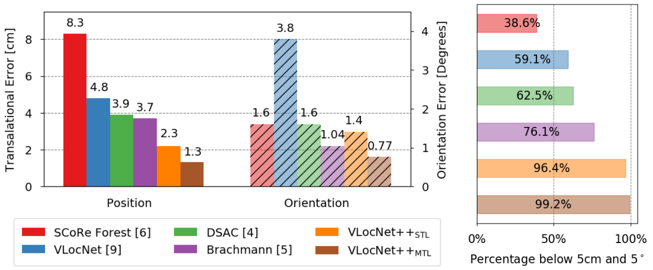


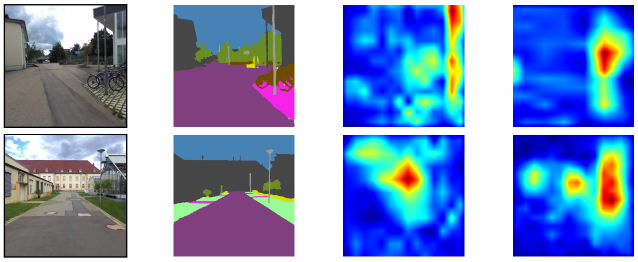
Fig. 2. Benchmarking 6DoF localization on the entire 7-Scenes dataset.

TABLE I
Median localization error on the DepLoc dataset.

| PoseNet [21] | B-PoseNet [22] | SVS [23] | VLocNet [9] | VLocNet++ |
|---|---|---|---|---|
| 2.42m, 3.66° | 2.24m, 4.31° | 1.61m, 3.52° | 0.68m, 3.43° | **0.32**m, **1.48**° |

Tab. I shows the median localization error for the DeepLoc dataset, where VLocNet++ achieves almost half the localization error as previous methods. Moreover, despite the difficulty of accurately estimating ego-motion in outdoor environments due to the more apparent motion parallax, VLocNet++ surpasses the accuracy of end-to-end approaches by 20.0% in the translational and 40.0% in the rotational components. While, for the task of semantic segmentation on the DeepLoc dataset, VLocNet++ consistently outperforms all baselines [24, 25, 26, 27, 28, 15], achieving a mean IoU score of 80.44%. In an effort to investigate the effect of incorporating semantic information into the global pose regression stream, we visualize the regression activation maps of the network for both the single-task and multitask variants of VLocNet++ using Grad-CAM++ [29]. In Fig. 3 we show two example scenes that contain glass facades and optical glare. Despite their challenging nature, our model is able to accurately segment both the scenes with high granularity. As we compare the activation maps of our single-task and multitask models, we observe that the multitask activation maps have less noisy activations focusing on multiple structures to yield an accurate pose estimate.

To summarize, we presented a deep learning approach to address the problem of camera pose regression. Experimental evaluations show that by integrating both the motion prior and the semantic knowledge, our network is able to accuartly estimate the pose while being robust to motion blur and perceptual aliasing. Comprehensive evaluations demonstrate that VLocNet++ sets the new state-of-the-art on the Microsoft 7-Scenes and DeepLoc datasets. More extensive evaluations [8] and a live demo is presented at http://deeploc.cs.uni-freiburg.de.



(a) Input Image (b) Semantic Output (c) ST Activation (d) MT Activation

Fig. 3. Qualitative analysis of the segmentation output along with a visualization of the regression activation maps [29] on the DeepLoc dataset.

# References

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

[2] F. Walch, C. Hazirbas, L. Leal-Taxe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using lstms for structured feature correlation," in *Proceedings of the International Conference on Computer Vision*, 2017.

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

[4] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother, "DSAC - differentiable RANSAC for camera localization," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2017.

[5] E. Brachmann and C. Rother, "Learning less is more - 6d camera localization via 3d surface regression," *arXiv preprint arXiv:1711.10228*, 2017.

[6] 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 International Conference on Computer Vision and Pattern Recognition*, 2013.

[7] T. Sattler, W. Madden, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla, "Benchmarking 6dof urban visual localization in changing conditions," *arXiv preprint arXiv:1707.09092*, 2017.

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

[9] 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.

[10] N. Kobyshev, H. Tiemenschneider, and L. V. Gool, "Matching features correctly through semantic understanding," in *2nd International Conference on 3D Vision*, 2014.

[11] G. Singh and J. Košecká, "Semantically guided geolocation and modeling in urban environments," *Large-Scale Visual Geo-Localization*, 2016.

[12] N. Rader, M. Bausano, and J. Richards, "On the nature of the visual-cliff-avoidance response in human infants," *Child Development*, vol. 51, no. 1, pp. 61–68, 1980.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2015.

[14] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[15] 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.

[16] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proceedings of the International Conference on Learning Representations*, 2016.

[17] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2016.

[18] A. Nicolai, R. Skeele, C. Eriksen, and G. A. Hollinger, "Deep learning for laser based odometry estimation," in *RSS workshop Limits and Potentials of Deep Learning in Robotics*, 2016.

[19] V. Mohanty, S. Agrawal, S. Datta, A. Ghosh, V. D. Sharma, and D. Chakravarty, "Deepvo: A deep learning approach for monocular visual odometry," *arXiv preprint arXiv:1611.06069*, 2016.

[20] I. Melekhov, J. Kannala, and E. Rahtu, "Relative camera pose estimation using convolutional neural networks," *arXiv preprint arXiv:1702.01381*, 2017.

[21] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the International Conference on Computer Vision*, 2015.

[22] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," *Proceedings of the IEEE International Conference on Robotics and Automation*, 2016.

[23] T. Naseer and W. Burgard, "Deep regression for monocular camera-based 6-dof global localization in outdoor environments," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2015.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture," *arXiv preprint arXiv: 1511.00561*, 2015.

[26] G. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox, "Deep learning for human part discovery in images," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2016.

[27] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv: 1506.04579*, 2015.

[28] 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.

[29] A. Chattopadhyay, A. Sarkar, and V. B. Prantik Howlader, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," *arXiv preprint arXiv:1710.11063*, 2017.