VLocNet++: Deep Multitask Learning for Semantic Visual Localization and Odometry

Noha Radwan* Abhinav Valada* Wolfram Burgard

***Abstract*—Semantic understanding and localization are fundamental enablers of robot autonomy that have for the most part been tackled as disjoint problems. While deep learning has enabled recent breakthroughs across a wide spectrum of scene understanding tasks, its applicability to state estimation tasks has been limited due to the direct formulation that renders it incapable of encoding scene-specific constrains. In this work, we propose the VLocNet++ architecture that employs 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. Our network overcomes the aforementioned limitation by simultaneously embedding geometric and semantic knowledge of the world into the pose regression network. We propose a novel adaptive weighted fusion layer to aggregate motion-specific temporal information and to fuse semantic features into the localization stream based on region activations. Furthermore, we propose a self-supervised warping technique that uses the relative motion to warp intermediate network representations in the segmentation stream for learning consistent semantics. Finally, we introduce a first-of-a-kind urban outdoor localization dataset with pixel-level semantic labels and multiple loops for training deep networks. 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.**

***Index Terms*—Deep Learning in Robotics and Automation; Visual Learning; Localization**

I. Introduction

AUTONOMOUS robots today are a complex ensemble of modules each of which specializes in a particular domain such as state estimation and scene understanding. While significant strides have been made considering these domains separately [1], [2], very little progress has been made towards exploiting the relationship between them. In this work, we focus on jointly learning three diverse vital tasks that are crucial for robot autonomy, namely, semantic segmentation, visual localization and odometry from consecutive monocular images. We approach this problem from a multitask learning (MTL) perspective with the goal of learning more accurate localization and semantic segmentation models by leveraging the predicted ego-motion. This problem is extremely challenging as it involves simultaneously learning cross-domain tasks that perform pixel-wise classification and regression with different units and scales. However, this joint formulation enables inter-task learning which improves both generalization capabilities and alleviates the problem of requiring vast amounts of labeled training data which is especially hard to obtain in the robotics domain. Moreover, as robots are equipped with limited resources, a joint model is more efficient for deployment and enables real-time inference on a consumer grade GPU.

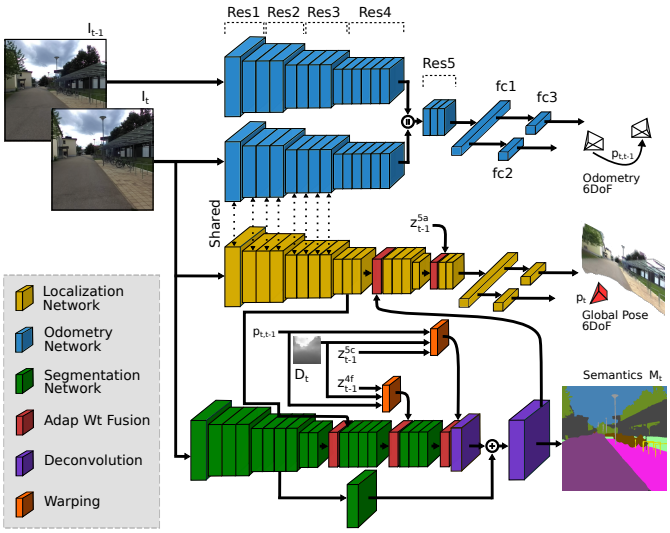


Fig. 1. Schematic representation of our proposed VLocNet++ architecture. The network takes two consecutive monocular images (I_t, I_{t-1}) as input and simultaneously predicts the global 6-DoF pose p_t, odometry $p_{t,t-1}$ and semantics M_t of the scene. z^l_{t-1} denotes the feature maps of layer l from the previous timestep and D_t denotes a predicted depth map that is used for representational warping in the semantic stream.

Most existing CNN-based metric localization approaches [3], [4] perform direct pose regression from image embeddings using naive loss functions. In order to more effectively encode knowledge about the environment, we propose a principled approach to embed geometric and semantic knowledge into the pose regression model. Our network utilizes our Geometric Consistency loss function [5] that incorporates relative motion information to learn a model that is globally consistent. Firstly, unlike the previous approach [5], to efficiently utilize the learned motion specific features from the previous timestep, we employ an adaptive weighting technique to aggregate motion-specific temporal information. Secondly, by jointly estimating the semantics, we instill structural cues about the environment into the pose regression network and implicitly pull the attention towards more informative regions in the scene. Existing semantics-aware localization techniques extract predefined stable features, emphasize or combine them with local features but often fail when the predefined structures are occluded or not visible in the scene. Our approach is robust to such situations as it uses our proposed adaptive fusion layer to fuse learned relevant features not only based on the semantic category but also the activations in the region.

Predicting consistent semantics is a critical prerequisite

Manuscript received: May, 02, 2018; Revised July, 31, 2018; Accepted September, 03, 2018.
This paper was recommended for publication by Editor Tamim Asfour upon evaluation of the Associate Editor and Reviewers' comments.
*These authors contributed equally. All authors are with the Department of Computer Science, University of Freiburg, Germany.
Digital Object Identifier (DOI): see top of this page.

for semantic visual localization. Inspired by early cognitive studies in humans showing the importance of learning self-motion for acquiring basic perceptual skills [6], we propose a novel self-supervised semantic context aggregation technique leveraging the predicted relative motion from the odometry stream of our network. Using pixel-wise depth predictions from a CNN [7] and differential warping, we fuse intermediate network representations from the previous timestep into the current frame using our proposed adaptive weighted fusion layer. This enables our semantic segmentation network to aggregate more scene-level context, thereby improving the performance and leading to faster convergence.

In summary, the primary contributions of this paper are as follows: (i) A novel MTL framework for jointly learning semantics, visual localization and odometry from consecutive monocular images. (ii) A CNN architecture for pose regression that significantly outperforms state-of-the-art approaches on the challenging Microsoft 7-Scenes benchmark. (iii) A self-supervised context aggregation technique based on differential warping that improves semantic segmentation and reduces the training time by half. (iv) A novel adaptive weighted fusion layer for element-wise fusion of feature maps based on region activations to exploit inter/intra task dependencies. (v) Finally, to facilitate this work, we introduce a first-of-a-kind outdoor dataset consisting of multiple loops with pixel-level semantic labels and localization ground truth. It contains repetitive, translucent and reflective surfaces, weakly textured regions and low-lighted scenes with shadows, thereby making it extremely challenging for benchmarking a variety of tasks.

II. Related Works

Over the past decade there has been a gradual shift from employing traditional handcrafted pipelines to learning-based methods particularly for perception related tasks. In this section, we discuss recent learning-based approaches for multitask learning, pose regression and semantic segmentation.

Multitask Learning can be defined as an inductive transfer mechanism that improves generalization by leveraging domain specific information from related tasks. It has been applied to a wide range of tasks [8]. Bilen *et al.* propose the use of an instance normalization layer to train a network that recognizes objects across multiple visual domains including digits, signs and faces [9]. In [10], the authors introduce a model with a sparsely-gated mixture of experts layer for the task of language modeling and machine translation. Multinet [8] proposes a unified architecture consisting of a shared encoder and task-specific decoders for classification, detection and segmentation. For combining different loss functions in a multitask model, Kendall *et al.* [11] propose a loss function based on maximizing the Gaussian likelihood using homoscedastic task uncertainty. While the aforementioned approaches mostly have shared parts of the network that learn low-level features followed by individual task-specific branches, we propose a novel adaptive weighted fusion layer that learns the most favorable weighting of feature maps for the mutual benefit of the tasks.

Visual Localization has been addressed by a variety of approaches ranging from image retrieval, local feature-based pipelines, to end-to-end learning methods [12]. Recently, pre-trained DCNNs designed for classification have been successfully adapted for pose regression. Kendall *et al.* proposed PoseNet [13], an end-to-end approach for directly regressing the 6-DoF camera pose from a monocular image using a DCNN. Since then several improvements have been proposed in terms of incorporating Long-Short Term Memory (LSTM) units for dimensionality reduction [14], symmetric encoder-decoder architecture for regression [15] and an improved loss function based on scene geometry [3]. NNnet [16] employs a hybrid approach where a DCNN trained on relative camera pose estimation is employed to extract features for identifying the nearest neighbors of a query image among the database images. Recently, Brachmann *et al.* proposed a differentiable version of RANSAC (DSAC) [12] and a successor version [17] for camera localization. Currently this approach [17] achieves state-of-the-art performance on the Microsoft 7-Scenes benchmark. More recently, VLocNet [5] presented the Geometric Consistency Loss function that constricts the search space with the relative motion information during training to obtain pose estimates that are consistent with the true motion model. The focus of this paper is to improve VLocNet's performance by adaptively fusing semantic features and aggregated motion-specific information from the previous timestep into the localization network.

Visual Odometry: Another similar line of work is to estimate the incremental change in position from images. Nicolai *et al.* [18] employed a simple Siamese architecture with alternating convolution and pooling layers to estimate the transforms from consecutive point clouds. Konda *et al.* [19] proposed an end-to-end architecture for learning ego-motion from a sequence of RGB-D images using a prior set of discretized velocities and directions. DeepVO [20] presents an AlexNet-based Siamese architecture for odometry estimation from monocular images, in which they also experiment with appending FAST features along with the images as input to the network. Melekhov *et al.* [2] propose a CNN architecture that incorporates spatial pyramid pooling and demonstrates improved performance compared to local feature-based approaches that utilize SIFT or ORB.

Semantic Segmentation: Fully Convolutional Neural Networks (FCNs) [21] first proposed the encoder-decoder model replacing inner-product layers with convolution layers to enable pixel-wise classification. Several networks built upon FCNs by introducing more refinement stages [22], efficient non-linear upsampling schemes [23] and adding global context [24]. Yu *et al.* [25] proposed a context module the uses dilated convolutions to enlarge the receptive field. DeepLab [26] proposed using multiple parallel dilated convolutions with different sampling rates for multi-scale learning in addition to using CRFs for post-processing. AdapNet [1] introduced multi-scale residual blocks with dilated convolutions as parallel convolutions to enable faster inference without compromising the performance. For learning consistent semantics in VLocNet++, we build upon AdapNet's model and propose a self-supervised warping technique for scene-level context aggregation. We warp feature maps of the preceding frame and fuse them using the proposed adaptive fusion layer which leads to improved accuracy and faster convergence.

III. Technical Approach

In this section, we detail our MTL framework for jointly estimating global pose, odometry and semantic segmentation from consecutive monocular images. While the approach

presented in this paper focuses on joint learning of the aforementioned tasks, each of the task-specific models can be deployed independently during test-time. We propose a novel strategy for encoding geometric and structural constraints into the pose regression network, namely by incorporating information from the previous timesteps to accumulate motion specific information and by adaptively fusing semantic features based on the activations in the region using our proposed fusion scheme. As being able to predict robust semantics is an essential prerequisite for the proposed fusion, we present a new self-supervised warping technique for aggregating scene-level context in the semantic segmentation model. 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 visual 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 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. z^l denotes the feature maps from layer l of a particular stream. In the remainder of this section, we describe the constituting network components and our MTL scheme.

A. Geometrically Consistent Pose Regression

Our model for regressing the global pose is based on the recently proposed VLocNet [5] architecture. It has five residual blocks that downsample the feature maps by half at each block, similar to the full preactivation ResNet-50 architecture, but replaces the conventional Rectified Linear Units (ReLUs) activation function with Exponential Linear Units (ELUs), which help learning representations that are more robust to noise and also lead to faster convergence. 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. Unlike VLocNet which fuses the previous predicted pose directly using inner-product layers, we adopt a more methodological approach to provide the network with this prior. Fusing the previous prediction directly inhibits the network from being able to correlate motion specific spatial relations crucial for this task, with that of the previous timestep as the network does not retain these features thereafter. In this work, we integrate the network's intermediate representation z^{5a}_{t-1} from the last downsampling stage (*Res5a*) of the previous timestep using our proposed adaptive weighted fusion layer detailed in Sec. III-D. Our fusion scheme learns the most favorable element-wise weighting for this fusion, and when trained end-to-end with the Geometric Consistency Loss, enables aggregation of motion-specific features across the temporal dimension. We denote the aforementioned architecture as VLocNet++$_{\text{STL}}$ in our experiments.

As opposed to naively minimizing the Euclidean loss between the predicted poses and the ground truth, we employ the Geometric Consistency 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 ground truth and the estimated motion from the odometry stream. By utilizing the predictions of the network from the previous timestep along with the current timestep, the relative motion loss term $\mathscr{L}_{Rel}(f(\theta \mid I_t))$ can be computed as a weighted summation of the translational and rotational errors, where θ is defined to be the parameters of the network, and $f(\theta \mid I_t)$ denotes the predicted output of the network for image I_t. Eq. (1) details the relative motion loss term, in which we assume that the quaternion output of the network has been normalized a priori for ease of notation, and $\hat{s}_{x_{Rel}}, \hat{s}_{q_{Rel}}$ denote the learnable weighting variables for the translational and rotational components [3].

$$\begin{aligned}\mathscr{L}_{Rel}(f(\theta \mid I_t)) &= \mathscr{L}_{x_{Rel}}(f(\theta \mid I_t))\exp(-\hat{s}_{x_{Rel}}) + \hat{s}_{x_{Rel}} \\ &\quad + \mathscr{L}_{q_{Rel}}(f(\theta \mid I_t))\exp(-\hat{s}_{q_{Rel}}) + \hat{s}_{q_{Rel}} \\ \mathscr{L}_{x_{Rel}}(f(\theta \mid I_t)) &:= \|\mathbf{x}_{t,t-1} - (\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t-1})\|_2 \\ \mathscr{L}_{q_{Rel}}(f(\theta \mid I_t)) &:= \left\|\mathbf{q}_{t,t-1} - \left(\hat{\mathbf{q}}_{t-1}^{-1}\hat{\mathbf{q}}_t\right)\right\|_2 .\end{aligned} \tag{1}$$

Following the aforementioned notation, the Euclidean loss term can be defined as

$$\begin{aligned}\mathscr{L}_{Euc}(f(\theta \mid I_t)) &= \mathscr{L}_x(f(\theta \mid I_t))\exp(-\hat{s}_x) \\ &\quad + \hat{s}_x + \mathscr{L}_q(f(\theta \mid I_t))\exp(-\hat{s}_q) + \hat{s}_q \\ \mathscr{L}_x(f(\theta \mid I_t)) &:= \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_2 \\ \mathscr{L}_q(f(\theta \mid I_t)) &:= \|\mathbf{q}_t - \hat{\mathbf{q}}_t\|_2 .\end{aligned} \tag{2}$$

The final loss term to be minimized is

$$\mathscr{L}_{loc}(f(\theta \mid I_t)) := \mathscr{L}_{Euc}(f(\theta \mid I_t)) + \mathscr{L}_{Rel}(f(\theta \mid I_t)) . \tag{3}$$

By minimizing the aforementioned loss function, our network learns a model that is geometrically consistent with respect to the motion. Moreover, by employing a mechanism to aggregate motion specific features temporally, we enable the Geometric Consistency Loss to efficiently leverage this information.

B. Learning Visual Odometry

Our proposed architecture for relative pose estimation takes a pair of consecutive monocular images (I_{t-1}, I_t) as input and yields an estimate of ego-motion $\mathbf{p}_{t,t-1} = [\mathbf{x}_{t,t-1}, \mathbf{q}_{t,t-1}]$. We employ a dual-stream architecture in which each of the streams is identically similar in structure and is based on the full preactivation ResNet-50 model. We concatenate the feature maps of the individual streams before the last downsampling stage (end of *Res4*) and convolve them through the last residual block, followed by an inner-product layer and two regressors for estimating the pose components. During training, we optimize the following loss function by minimizing the Euclidean error between the ground truth and the predicted motion.

$$\begin{aligned}\mathscr{L}_{vo}(f(\theta \mid I_t, I_{t-1})) &:= \mathscr{L}_x(f(\theta \mid I_t, I_{t-1}))\exp(-\hat{s}_{x_{vo}}) \\ &\quad + \hat{s}_{x_{vo}} + \mathscr{L}_q(f(\theta \mid I_t, I_{t-1}))\exp(-\hat{s}_{q_{vo}}) + \hat{s}_{q_{vo}},\end{aligned} \tag{4}$$

where $\mathscr{L}_x$ and $\mathscr{L}_q$ refers to the translational and rotational components respectively. We also employ learnable weighting parameters to balance the scale between the translational and rotational components in the loss term. As shown in Fig. 1, the dual odometry streams have an architecture similar to the global pose regression network. In order to enable the inductive transfer of information between both networks, we share parameters between the odometry stream taking the current image I_t and the global pose regression network as detailed in Sec. III-D.

C. Learning Semantics

Our model for learning consistent semantics has two variants: a single-task base architecture that takes a monocular image as input and predicts a pixel-wise segmentation mask (green and purple blocks in Fig. 1) and a multitask architecture built upon the base model that incorporates our proposed self-supervised warping and adaptive fusion layers (orange and red blocks).

Network Architecture: For the single-task base model, we adopt the AdapNet [1] architecture which follows the general encoder-decoder design principle. Similar to the localization network, the encoder is based on the ResNet-50 model which includes skip connections and batch normalization layers that enable training such deep architectures by alleviating the vanishing gradient problem. The encoder learns highly discriminative semantic features and yields an output 16-times downsampled with respect to the input dimensions. While the decoder consists of two deconvolution layers and a skip convolution from the encoder for fusing high resolution feature maps and upsampling the downscaled feature maps back to the input resolution. The architecture also incorporates multi-scale ResNet blocks which have dilated convolutions parallel to the 3×3 convolutions for aggregating features from different spatial scales, concurrently maintaining fast inference times.

Following the notation convention, 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 θ as the network parameters. Using the classification scores s_j at each pixel u_r, we obtain a probabilities $\mathbf{P} = (p_1, \ldots, p_C)$ with the $\mathsf{softmax}$ function $\sigma(.)$ such that

$$p_j(u_r, \theta \mid I_n) = \sigma(s_j(u_r, \theta)) = \frac{exp(s_j(u_r, \theta))}{\sum_k^C exp(s_k(u_r, \theta))} \quad (5)$$

denotes the probability of pixel u_r being classified with label j. The optimal θ is estimated by minimizing

$$\mathscr{L}_{seg}(\mathcal{T}, \theta) = -\sum_{n=1}^{N} \sum_{r=1}^{\rho} \sum_{j=1}^{C} \delta_{m_r^n, j} \log p_j(u_r, \theta \mid I_n), \quad (6)$$

for $(I_n, M_n) \in \mathcal{T}$, where $\delta_{m_r^n, j}$ is the Kronecker delta.

Self-Supervised Warping: In order to aggregate scene-level context for learning consistent semantics, we first leverage the estimated relative pose from the odometry stream to warp feature maps from the previous timestep into the current view using a predicted depth map. We then fuse the warped feature maps with the intermediate network representations of the current timestep. 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, frame-level distortions and implicitly introduce feature augmentation which facilitates faster convergence. We utilize DispNet [7] to obtain the depth map D_t and fuse warped feature maps as described in Sec. III-D. We introduce the warping and fusion layers (red and orange blocks in Fig. 1) at *Res4f* and *Res5c* to fuse the corresponding feature maps z_{t-1}^{4f} and z_{t-1}^{5c} from the previous timestep into the network. As the warping is fully differentiable, our approach does not require any pre-computation for training and runs online. Moreover, our self-supervised warping adds minimal overhead as we only calculate the warping grid once at the input resolution in terms of pixels u_r and employ average pooling to apply the grid at multiple scales for transforming the feature maps z_{t-1} to its warped current view representation $\hat{z}_{t-1}$. In order to facilitate computation of gradients necessary for back-propagation, we use bilinear interpolation as a sampling mechanism for warping. Utilizing the relative pose, a depth map D_t of the image, and the projection function π, we formulate the warping as

$$\hat{u}_r := \pi\left(T(\mathbf{p}_{t,t-1})\, \pi^{-1}(u_r, D_t(u_r))\right). \quad (7)$$

Given a previous image I_{t-1} and the relative motion between the images $\mathbf{p}_{t,t-1}$, we can project each pixel u_r from I_{t-1} to I_t as per Eq. (7). The warped pixel $\hat{u}_r$ is obtained using the depth information $D_t(u_r)$ and the relative pose $\mathbf{p}_{t,t-1}$, where the function $T(\mathbf{p}_{t,t-1})$ denotes the homogenous transformation matrix of $\mathbf{p}_{t,t-1}$, π denotes the projection function transforming from world to camera coordinates such that $\pi : \mathbb{R}^3 \mapsto \mathbb{R}^2$ and π^{-1} denotes the transformation from camera to world coordinates using a depth map $D_t(u_r)$.

D. Deep Multitask Learning

Our main motivation for jointly learning semantics, global pose regression and odometry is twofold: to enable inductive transfer by leveraging domain specific information while simultaneously exploiting complementary features, and to enable the global pose regression network to encode geometric and semantic knowledge of the environment while training. To achieve this goal, we structure our multitask framework to be interdependent on the outputs as well as intermediate representations of each of these tasks. Specifically, as shown in Fig. 1, we employ hybrid hard parameter sharing until the end of the *Res3* block between the global pose regression stream and the odometry stream that both receive the image from the current timestep. This exploits the task-specific similarities among these pose regression tasks and influences the shared weights of global pose regression network to integrate motion-specific features due to inductive bias from odometry estimation, in addition to effectuating implicit attention on regions that are more informative for relative motion estimation.

A common practice employed for combining features from multiple layers or multiple networks is to perform concatenation of the tensors or element-wise addition/multiplication. Although this might be effective when both tensors contain sufficient relevant information, it often accumulates irrelevant feature maps and its effectiveness highly depends upon the intermediate stages of the network where the fusion is performed. One of the key components of our multitask learning framework is the proposed adaptive weighted fusion layer which learns the most favorable element-wise weighting for the fusion based on the activations in the region, followed by a non-linear feature pooling over the weighted tensors. Pooling in the feature space (as opposed to spatial pooling) is a form of coordinate-dependent transformation which yields the same number of filters as the input tensor. For ease of notation, we formulate the mathematical representation of the adaptive weighted fusion layer with respect to two activation maps z^a and z^b from layers a and b, while extending the notation to multiple activation maps is straightforward. The activation maps can be from layers in the same network or from different task-specific networks. The output of the adaptive weighted fusion layer can be formulated as

$$\hat{z}_{fuse} = \max\left(\mathbf{W} * \left((w^a \odot z^a) \oplus \left(w^b \odot z^b\right)\right) + \mathbf{b}, 0\right), \quad (8)$$

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. In other words, each channel of the activation map z^a is first weighted, then linearly combined with the corresponding weighted channels of the activation map z^b. Non-linear feature pooling is then applied, which can be easily realized with existing layers in the form of a 1×1 convolution with a non-linearity such as ReLUs. As shown in Fig. 1, we incorporate the adaptive fusion layers (red blocks) at *Res4c* to fuse semantic features into the global pose regression stream. In addition, we also employ them to fuse warped semantic feature maps from the previous timestep into the segmentation stream at the end of *Res3* and *Res4* blocks. We denote this architecture as VLocNet++$_{\text{MTL}}$ in our experiments. Moreover, in Sec. IV-C, we demonstrate that over simple concatenation, our adaptive weighted fusion learns what features are relevant for both inter-task and intra-task fusion. In order to jointly learn all tasks, we minimize the loss function below:

$$\mathscr{L}_{multi} := \mathscr{L}_{loc} \exp(-\hat{s}_{loc}) + \hat{s}_{loc} + \mathscr{L}_{vo} \exp(-\hat{s}_{vo}) + \hat{s}_{vo} + \mathscr{L}_{seg} \exp(-\hat{s}_{seg}) + \hat{s}_{seg}, \quad (9)$$

where $\mathscr{L}_{loc}$ is the global pose regression loss as per Eq. (3); $\mathscr{L}_{vo}$ is the visual odometry loss from Eq. (4), and $\mathscr{L}_{seg}$ is the cross-entropy loss for semantic segmentation from Eq. (6). Due to the inherent nature of the diverse tasks at hand, each of the associated task-specific loss terms has a different scale. If the task-specific losses were to be naively combined, the task with the highest scale would dominate during training and there would be little if no gain for any of the other tasks. To counteract this problem, we use learnable scalar weights $\hat{s}_{loc}, \hat{s}_{vo}, \hat{s}_{seg}$ to balance the scale of each of the loss terms.

E. Datasets and Augmentation

Supervised learning techniques such as DCNNs require a large amount of training data with ground truth annotations which is laborious to acquire. This becomes even more critical for jointly learning multiple diverse tasks which necessitate individual task-specific labels. Although there are publicly available task-specific datasets for visual localization and semantic segmentation, to the best of our knowledge there is a lack of a large enough dataset that contains both semantic and global localization ground truth with multiple loops in the same scene. To this end, we introduce the challenging *DeepLoc* dataset containing RGB-D images tagged with 6-DoF poses and pixel-level semantic labels of an outdoor urban scene that we make publicly available. In addition to our new dataset, we also benchmark the performance of our localization network (without joint semantics learning) on the challenging Microsoft 7-Scenes dataset. We chose these datasets based on the criteria of having diversity in scene structure and environment as well as the medium with which the images were captured.

We do not perform any pose augmentations [4] as our initial experiments employing them did not demonstrate any improvement in performance in the aforementioned datasets. However for learning semantics, we randomly apply image augmentations including rotation, translation, scaling, skewing, cropping, flipping, contrast and brightness modulation.

Microsoft 7-Scenes dataset [27] is a widely used dataset for camera relocalization and tracking. It contains RGB-D image sequences tagged with 6-DoF camera poses of 7 different indoor environments. The data was captured with a Kinect camera at a resolution of 640×480 pixels and ground truth poses were generated using KinectFusion [27]. Each of the sequences contains about 500 to 1000 frames. This dataset is very challenging as it contains textureless surfaces, reflections, motion blur and perceptual aliasing due to repeating structures.

DeepLoc: We introduce a large-scale urban outdoor localization dataset collected around the university campus, which we make publicly available [1]. The dataset was collected using our robot platform equipped with a ZED stereo camera, an XSens IMU, a Trimble GPS Pathfinder Pro and several LiDARs. RGB and depth images were captured at a resolution of 1280×720 pixels, at 20Hz. The dataset was collected in an area spanning 110×130m, that the robot traverses multiple times with different driving patterns. We use the LiDAR-based SLAM system from Kümmerle *et al.* [28] to compute the ground truth pose labels.

Furthermore, for each image we provide pixel-level semantic segmentation annotations for ten categories: *Background, Sky, Road, Sidewalk, Grass, Vegetation, Building, Poles & Fences, Dynamic and Other*. To the best of our knowledge, this is the first publicly available dataset containing images tagged with 6-DoF poses and pixel-level semantic segmentation labels for an entire scene with multiple loops. We divide the dataset into a train and a test split such that the training set consists of seven loops with alternating driving styles amounting to 2737 images, while the test set consists of three loops with a total of 1173 images. This dataset can be very challenging for vision based applications such as global localization, camera relocalization, semantic segmentation, visual odometry and loop closure detection, as it contains substantial lighting, weather changes, motion blur and perceptual aliasing due to similar buildings and glass structures. We hope that this dataset enables future research in multitask and multimodel learning.

IV. Experimental Evaluation

In order to quantify the performance of VLocNet++, we first compare our single-task models against other deep learning based methods in each corresponding task in Sec. IV-A, followed by a more comprehensive comparison against the state-of-the-art in Sec. IV-B and with multitask variants in Sec. IV-C. Furthermore, we present extensive qualitative experiments and an ablation study in Sec. IV-D which demonstrates the efficacy of our approach and provides insights on the representations learned by our network. For all the experiments, we train our models from random crops of the image and test on the center crop. We initialize the five residual blocks of our tasks-specific networks with weights from the ResNet-50 model trained on the ImageNet dataset and the other layers with Xavier initialization. We use the Adam solver for optimization with $\beta_1 = 0.9, \beta_2 = 0.999$ and $\varepsilon = 10^{-10}$. We employ a multi-stage training procedure and first train task-specific models individually using an initial learning rate of $\lambda_0 = 10^{-3}$ with a mini-batch size of 32 and a dropout probability of 0.2. Using transfer learning, we initialize the joint MTL architecture with weights from the best performing single-task models and train with a lower learning rate of $\lambda_0 = 10^{-4}$. We use TensorFlow for the implementation and training the network on a single

[1] VLocNet++ live demo and dataset are publicly available at: http://deeploc.cs.uni-freiburg.de

TABLE I
MEDIAN LOCALIZATION ERROR ON THE 7-SCENES DATASET.

Scene	PoseNet2 [3]	NNnet [16]	VLocNet [5]	VLocNet++$_{STL}$ (Ours)
Chess	0.13m, 4.48°	0.13m, 6.46°	0.036m, 1.71°	**0.023**m, **1.44**°
Fire	0.27m, 11.3°	0.26m, 12.72°	0.039m, 5.34°	**0.018**m, **1.39**°
Heads	0.17m, 13.0°	0.14m, 12.34°	0.046m, 6.64°	**0.016**m, **0.99**°
Office	0.19m, 5.55°	0.21m, 7.35°	0.039m, 1.95°	**0.024**m, **1.14**°
Pumpkin	0.26m, 4.75°	0.24m, 6.35°	0.037m, 2.28°	**0.024**m, **1.45**°
RedKitchen	0.23m, 5.35°	0.24m, 8.03°	0.039m, 2.20°	**0.025**m, **2.27**°
Stairs	0.35m, 12.4°	0.27m, 11.82°	0.097m, 6.48°	**0.021**m, **1.08**°
Average	0.23m, 8.12°	0.21m, 9.30°	0.048m, 3.80°	**0.022**m, **1.39**°

TABLE II
MEDIAN LOCALIZATION ERROR ON THE DEPLOC DATASET.

PoseNet [13]	Bayesian PoseNet [29]	SVS-Pose [4]	VLocNet [5]	VLocNet++$_{STL}$ (Ours)
2.42m, 3.66°	2.24m, 4.31°	1.61m, 3.52°	0.68m, 3.43°	**0.37**m, **1.93**°

TABLE III
6DOF VISUAL ODOMETRY ON THE 7-SCENES DATASET [%, deg/m].

Scene	LBO [18]	DeepVO [20]	cnnBspp [2]	VLocNet [5]	VLocNet++ (Ours)
Chess	1.69, 1.13	2.10, 1.15	1.38, 1.12	1.14, 0.75	**0.99**, **0.66**
Fire	3.56, 1.42	5.08, 1.56	2.08, 1.76	1.81, 1.92	**0.99**, **0.78**
Heads	14.43, 2.39	13.91, 2.44	3.89, 2.70	1.82, 2.28	**0.58**, **1.59**
Office	3.12, 1.92	4.49, 1.74	1.98, 1.52	1.71, 1.09	**1.32**, **1.01**
Pumpkin	3.12, 1.60	3.91, 1.61	1.29, 1.62	1.26, 1.11	**1.16**, **0.98**
RedKitchen	3.71, 1.47	3.98, 1.50	1.53, 1.62	1.46, **1.28**	**1.26**, 1.52
Stairs	3.64, 2.62	5.99, 1.66	2.34, 1.86	**1.28**, 1.17	1.55, **1.10**
Average	4.75, 1.79	5.64, 1.67	2.07, 1.74	1.51, 1.45	**1.12**, **1.09**

TABLE IV
6DOF VISUAL ODOMETRY ON THE ON THE DEEPLOC DATASET [%, deg/m].

LBO [18]	DeepVO [20]	cnnBspp [2]	VLocNet [5]	VLocNet++ (Ours)
0.41, 0.053	0.33, 0.052	0.35, 0.049	0.15, 0.040	**0.12**, **0.024**

NVIDIA Titan X GPU takes 23 hours for the model to converge.

A. Comparison with the State-of-the-art

In this section, we show empirical evaluations comparing each of the single-task models VLocNet++$_{STL}$ with other CNN-based methods for each of the corresponding tasks.

Evaluation of Visual Localization: As a primary evaluation criteria, we first report results in comparison to deep learning-based approaches on both the publicly available Microsoft 7-Scenes (indoor) and DeepLoc (outdoor) datasets. We analyze the performance in terms of the median translation and orientation errors for each scene using the original train and test splits provided by the datasets. Tab. I shows the results for the 7-Scenes dataset, for which VLocNet++$_{STL}$ achieves an overall improvement of 54.17% in translation and 63.42% in rotation, thereby substantially outperforming existing CNN-based approaches. The largest improvement was obtained in the perceptually hardest scenes that contain textureless regions and repeating structures such as in the stairs scene shown in Fig. 5(a). In this scene, we achieve an improvement of 78.35% in translation and 83.34% in rotation over the previous state-of-the-art. Tab. II shows the results on the DeepLoc dataset, for which we obtain almost half the error as previous methods. This demonstrates that VLocNet++$_{STL}$ performs equally well in outdoor environments where there is a significant amount of perceptual aliasing as well as in indoor textureless environments.

Evaluation of Visual Odometry: We evaluate the performance of VLocNet++ for 6-DoF visual odometry estimation and show quantitative results in Tab. III for the 7-Scenes dataset and in Tab. IV for the DeepLoc dataset. We report the average translational and rotational errors relative to the sequence length. On the 7-Scenes dataset VLocNet++ outperforms end-to-end approaches, achieving a translational error of 1.12% and rotational error of 1.09 deg/m. While on the outdoor DeepLoc dataset, accurately estimating ego-motion is a rather challenging task due to the more apparent motion parallax and dynamic lighting changes. Despite this fact, VLocNet++ surpasses the accuracy of competitors with a translational error of 0.12% and a rotational error of 0.024 deg/m.

Evaluation of Semantic Segmentation: We present comprehensive evaluations of VLocNet++ for semantic segmentation on the DeepLoc dataset and report the Intersection over Union (IoU) score for each of the individual categories as well as the mean IoU. As shown in Tab. V, VLocNet++ achieves a mean IoU of 80.44%, consistently outperforming the baselines in all the categories. This improvement can be attributed to both the self-supervised warping as well the inductive transfer that occurs from the training signals of the localization network, as the AdapNet model which we build upon achieves a lower performance without our proposed improvements. In addition, this enables the model to converge in about 26k iterations, whereas Adapnet requires 120k iterations to converge.

B. Benchmarking on Microsoft 7-Scenes Dataset

We benchmark the performance of our single-task VLocNet++$_{STL}$ model and the multitask variant VLocNet++$_{MTL}$ on the Microsoft 7-Scenes dataset by comparing against both local feature-based pipelines and learning-based techniques. We present our main results in Fig. 2 using the median localization error metric and the percentage of poses for which the error is below 5cm and 5°. While VLocNet [5] was the first deep learning-based approach to yield an accuracy comparable to local feature-based pipelines achieving higher performance than SCoRe Forests [30] in terms of number of images with pose error below 5cm and 5°, it was recently outperformed by the approach of Brachmann *et al.* [17] which is the current state-of-the-art.

From the results presented in Fig. 2, we see that our VLocNet++$_{STL}$ model achieves a localization accuracy of 96.4%, improving over the accuracy of Brachmann *et al.* [17] by 20.3% and by over an order of magnitude compared to the other deep learning approaches [3], [16]. Moreover, by employing our proposed multitask framework, VLocNet++$_{MTL}$ further improves on the performance and achieves an accuracy of 99.2%, setting the new state-of-the-art on this benchmark. Furthermore, VLocNet++ only requires 79ms for a forward-pass on a single consumer grade GPU versus the 200ms required by the previous state-of-the-art [17]. It is important to note that other than VLocNet [5], the competitors shown in Fig. 2 rely on a 3D scene model and hence require RGB-D data, whereas VLocNet++ only utilizes monocular images. DSAC [12] and its variant [17] that utilize only RGB images demonstrate a lower performance than the results shown in Fig. 2. The improvement achieved by VLocNet++ shows that the apt combination of employing the Geometric Consistency Loss and the adaptive weighted fusion layer enables the network to efficiently leverage the motion-specific and semantic features

TABLE V
COMPARISON OF SEMANTIC SEGMENTATION PERFORMANCE WITH STATE-OF-THE-ART APPROACHES ON OUR DEEPLOC DATASET.

Approach	Sky	Road	Sidewalk	Grass	Vegetation	Building	Poles	Dynamic	Other	Mean IoU
FCN-8s [21]	94.65	98.98	64.97	82.14	84.47	87.68	45.78	66.39	47.27	69.53
SegNet [23]	93.42	98.57	54.43	78.79	81.63	84.38	18.37	51.57	33.29	66.05
UpNet [22]	95.07	98.05	63.34	81.56	84.79	88.22	31.75	68.32	45.21	72.92
ParseNet [24]	92.85	98.94	62.87	81.61	82.74	86.28	27.35	65.44	45.12	71.47
DeepLab v2 [26]	93.39	98.66	76.81	84.64	88.54	93.07	20.72	66.84	52.70	67.54
AdapNet [1]	94.65	98.98	64.97	82.14	84.48	87.68	45.78	66.40	47.27	78.59
VLocNet++ (ours)	**95.84**	**98.99**	**80.85**	**88.15**	**91.28**	**94.72**	**45.79**	**69.83**	**58.59**	**80.44**

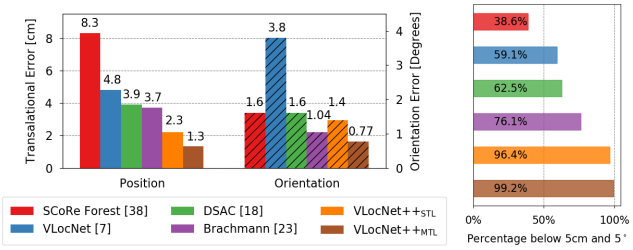


Fig. 2. Benchmarking 6DoF localization on the entire 7-Scenes dataset. We compare against state-of-the-art approaches that utilize RGB or RGB-D data and even with approaches that depend on a 3D model, VLocNet++ only uses RGB images. We report the performance as median localization errors (left) and the percentage of test images with a pose error below 5cm and 5° (right).

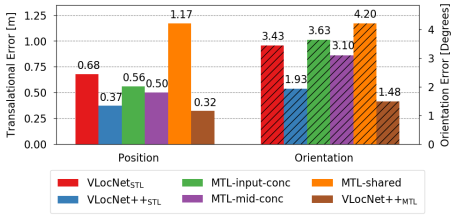


Fig. 3. Localization error of various multitask models in comparison to our proposed VLocNet++ incorporating our novel adaptive weighted fusion layer for fusing semantic features into the localization stream.

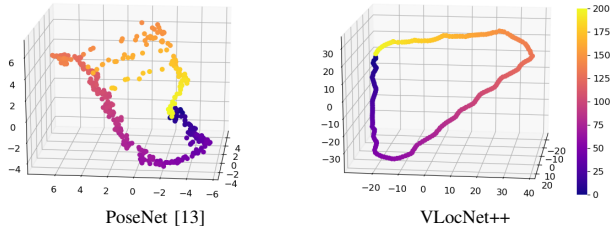


Fig. 4. 3D multi-dimensional scaling (MDS) of features from the penultimate layer of PoseNet [13] and VLocNet++ trained on the DeepLoc dataset. Inputs are images from the testing seq-01 loop and the points shown are chronologically coloured. Features learned by VLocNet++ show precise correlation with the trajectory (Fig. 5(b)), whereas PoseNet fails to capture the distribution especially for the poses near the glass buildings.

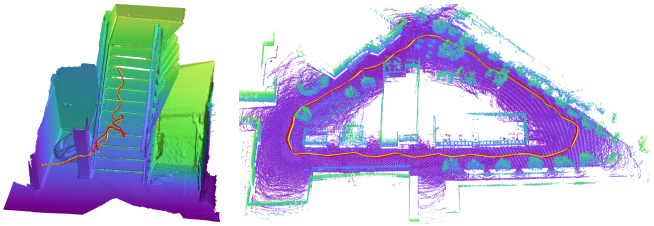
in order to learn a geometrically consistent motion model. More extensive evaluations are shown in the supplementary material and a live demo at http://deeploc.cs.uni-freiburg.de.

C. Multitask Learning

In this section, we primarily investigate the effectiveness of employing our proposed adaptive fusion layer for encoding semantic information and aggregating motion-specific information into the global localization stream. We compare the localization accuracy of VLocNet++$_{\text{MTL}}$ which incorporates our fusion scheme against the performance of single-task models and three competitive multitask baseline approaches. A rather simple and naive approach to fuse semantic features learned by the segmentation stream into the global localization stream would be to concatenate the predicted segmentation mask as a fourth channel to the input image, which we refer to as *"MTL-input-conc"*. As a second baseline, we concatenate intermediate feature maps of the segmentation stream with the corresponding intermediate feature maps in the global localization stream. We exhaustively evaluated various intermediate stages to do this fusion and in our setting we obtained the best accuracy when concatenating the feature maps of *Res5c* from the segmentation with *Res4f* of the global localization stream, which we denote as *"MTL-mid-conc"*. For the third baseline, we share the latent space of both the networks as a variant of the approach proposed in [31], which we denote in our experiments as *"MTL-shared"*. More details on these baseline architectural topologies as well as extended ablation studies on the effect of semantic fusion and representational warping are shown in the supplementary material. Fig. 3 shows the results from this experiment on the DeepLoc dataset. VLocNet++$_{\text{MTL}}$ achieves the highest performance with an improvement of 36% in translational and 53.87% in rotational components of the pose, compared to the best performing MTL-input-conc baseline. While in comparison to our single-task VLocNet++$_{\text{STL}}$ model we achieve an improvement of 13.51% and 23.32% in the translation and rotation components respectively, demonstrating that our network is able to learn the most favorable weighting for fusion based on region activations in the feature maps. We visualize these activation maps in Fig. 6.

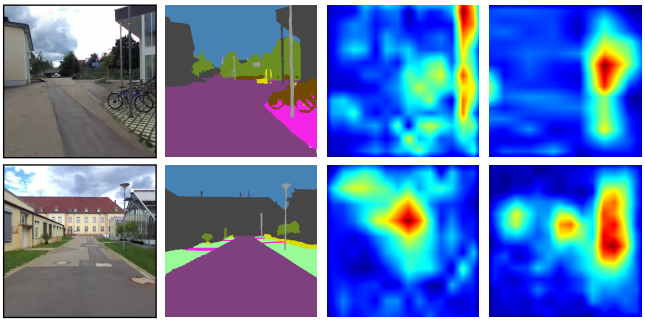
D. Ablation Study and Qualitative Analysis

Despite the recent surge in applying deep learning approaches to various domains, there is still a lack of fundamental knowledge regarding what kind of representations are learned by the networks, which is primarily due to their high dimensionality. To aid in this understanding, feature visualization and dimensionality reduction techniques can provide helpful insights when applied thoughtfully. Such techniques transform the data from high dimensional spaces to one of smaller dimensions by obtaining a set of principle values. For the task of localization, techniques that preserve the global geometry of the features such as Multi-Dimensional Scaling (MDS) are more meaningful to employ than approaches that find clusters and subclusters in the data such as the t-Distributed Stochastic Neighbor Embedding (t-SNE). Therefore, we apply 3D metric MDS to the features learned by the penultimate layer of our VLocNet++ model to visualize the underlying distribution. Fig. 4 displays the down-projected features obtained after applying MDS for VLocNet++ and PoseNet [13] on the DeepLoc dataset. Unlike PoseNet, the features learned in VLocNet++ directly correspond to the ground truth trajectory



(a) Stairs (b) DeepLoc

Fig. 5. Qualitative localization results of one test loop depicting the estimated global pose (yellow trajectory) versus the ground truth pose (red trajectory) plotted with respect to the 3D scene model for visualization. VLocNet++ accurately estimates the pose in both indoor (a) and outdoor (b) environments while being robust to textureless regions, repetitive and reflective structures in the environment where local feature-based pipelines perform poorly.



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

Fig. 6. Qualitative analysis of the predicted segmentation output along with a visualization of the regression activation maps [32] for both the single-task (ST), and multitask (MT) variant of VLocNet++ on the DeepLoc dataset.

shown in Fig. 5(b) (red trajectory), whereas PoseNet fails to capture the pose distribution in some areas of the dataset. Furthermore, in Fig. 5 we show the plot of the ground truth and the estimated poses as trajectories within the 3D model of the scenes for visualization. Using our proposed VLocNet++ the estimated poses are visually indistinguishable from the ground truth demonstrating the efficacy of our approach.

In an effort to investigate the effect of incorporating semantic information on the features learned by the localization stream, we visualize the regression activation maps of the network for both the single-task and multitask variants of VLocNet++ using Grad-CAM++ [32]. In Fig. 6 we show two example scenes that contain glass facades and optical glare. Despite their challenging nature, our model is able to segment both scenes with high granularity. As we compare the activation maps of our single-task and multitask models, we observe that multitask activation maps have less noisy activations focusing on multiple structures to yield an accurate pose estimate.

V. Conclusion

In this paper, we proposed a novel multitask learning framework for 6-DoF visual localization, semantic segmentation and odometry estimation, with the goal of exploiting interdependencies within these tasks for their mutual benefit. We presented a strategy for simultaneously encoding geometric and structural constraints into the the pose regression network by temporally aggregating learned motion specific information and adaptively fusing semantic features. To this end, we proposed an adaptive weighted fusion layer that learns the most favorable weighting for fusion based on region activations. In addition, we proposed a self-supervised warping technique for scene-level context aggregation in semantic segmentation networks that improves performance and adds minimal computational overhead while substantially decreasing the training time. Furthermore, we introduced a large-scale outdoor localization dataset with multiple loops and pixel-level semantic ground truth for training multitask deep networks. Comprehensive evaluations on benchmark datasets demonstrate that VLocNet++ exceeds the state-of-the-art by 67.5% in the translational and 25.9% in the rotational components of the pose, while being 60.5% faster and simultaneously performing multiple tasks.

References

[1] A. Valada, J. Vertens, *et al.*, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *ICRA*, 2017.

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

[3] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," *CVPR*, 2017.

[4] T. Naseer and W. Burgard, "Deep regression for monocular camera-based 6-dof global localization in outdoor environments," in *IROS*, 2017.

[5] A. Valada, N. Radwan, and W. Burgard, "Deep auxiliary learning for visual localization and odometry," in *ICRA*, 2018.

[6] N. Rader *et al.*, "On the nature of the visual-cliff-avoidance response in human infants," *Child Development*, vol. 51, no. 1, pp. 61–68, 1980.

[7] N.Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *CVPR*, 2016.

[8] M. Teichmann *et al.*, "Multinet: Real-time joint semantic reasoning for autonomous driving," *arXiv preprint arXiv:1612.07695*, 2016.

[9] H. Bilen and A. Vedaldi, "Universal representations: The missing link between faces, text, planktons, and cat breeds," *arXiv:1701.07275*, 2017.

[10] N. Shazeer *et al.*, "Outrageously large neural networks: The sparsely-gated mixture-of-experts," *arXiv preprint arXiv:1701.06538*, 2017.

[11] A. Kendall *et al.*, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," *arXiv:1705.07115*, 2017.

[12] E. Brachmann, A. Krull, S. Nowozin, *et al.*, "DSAC - differentiable RANSAC for camera localization," in *CVPR*, 2017.

[13] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *ICCV*, 2015.

[14] F. Walch, C. Hazirbas, *et al.*, "Image-based localization using lstms for structured feature correlation," in *ICCV*, 2017.

[15] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu, "Image-based localization using hourglass networks," *arXiv:1703.07971*, 2017.

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

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

[18] A. Nicolai *et al.*, "Deep learning for laser based odometry estimation," in *RSSws Limits and Potentials of Deep Learning in Robotics*, 2016.

[19] K. R. Konda and R. Memisevic, "Learning visual odometry with a convolutional network," in *VISAPP*, 2015.

[20] V. Mohanty *et al.*, "Deepvo: A deep learning approach for monocular visual odometry," *arXiv preprint arXiv:1611.06069*, 2016.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[22] G. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox, "Deep learning for human part discovery in images," in *ICRA*, 2016.

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

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

[25] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Int. Conf. on Learning Representations*, 2016.

[26] L. Chen, , *et al.*, "Semantic image segmentation with deep convolutional nets, atrous convolution, and crfs," *arXiv: 1606.00915*, 2016.

[27] J. Shotton *et al.*, "Scene coordinate regression forests for camera relocalization in rgb-d images," in *CVPR*, 2013.

[28] R. Kümmerle *et al.*, "Autonomous robot navigation in highly populated pedestrian zones," *JFT*, vol. 32, no. 4, pp. 565–589, 2015.

[29] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," *ICRA*, 2016.

[30] J. Shotton and others., "Scene coordinate regression forests for camera relocalization in rgb-d images," in *CVPR*, 2013.

[31] A. H. Abdulnabi *et al.*, "Multi-task cnn model for attribute prediction," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1949–1959, 2015.

[32] A. Chattopadhyay *et al.*, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," *arXiv:1710.11063*, 2017.