Neural vision-based semantic 3D world modeling

Sotirios Papadopoulos Ioannis Mademlis Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
AUTH Campus, Thessalonki, Greece
papadops@csd.auth.gr imademlis@csd.auth.gr pitas@csd.auth.gr

Abstract

Scene geometry estimation and semantic segmentation using image/video data are two active machine learning/computer vision research topics. Given monocular or stereoscopic 3D images, depicted scene/object geometry in the form of depth maps can be successfully estimated, while modern Deep Neural Network (DNN) architectures can accurately predict semantic masks on an image. In several scenarios, both tasks are required at once, leading to a need for combined semantic 3D world mapping methods. In the wake of modern autonomous systems, DNNs that simultaneously handle both tasks have arisen, exploiting machine/deep learning to save up considerably on computational resources and enhance performance, as these tasks can mutually benefit from each other. A great application area is 3D road scene modeling and semantic segmentation, e.g., for an autonomous car to identify and localize in 3D space visible pavement regions (marked as "road") that are essential for autonomous car driving. Due to the significance of this field, this paper surveys the state-of-the-art DNN-based methods for scene geometry estimation, image semantic segmentation and joint inference of both.

1. Introduction

Autonomous/robotic systems (e.g., autonomous cars [16], drones [38, 53, 39], etc.) are characterized by their ability to navigate an area on their own, by exploiting sensor data acquired on-the-fly and AI algorithms. Knowing the geometry of a depicted scene/object is a prerequisite for understanding its surroundings and, thus, safely navigate in its vicinity. Additionally, an autonomous system has to know the semantics of its environment as well, e.g., in order to differentiate between what part of the visible scene is "road" and what is "pedestrian" (in the case of a car). Visual sensors (monocular/stereoscopic 3D RGB cameras, RGB-D sensors or LiDARs) are the most important data sources for both semantics and geometry estimation.

In recent years, it has been shown that concurrent execution of these AI tasks provides important synergy benefits [37, 6]. For instance, in cases where there is no evident *stereoscopic disparity* [51, 11] between a homogeneous segment of the same object appearing in the images of an RGB stereo pair, semantic segmentation can possibly inform us about this object's shape. Similarly, geometric information may guide semantic segmentation in problematic scenarios. Examples of inference on both tasks can be seen in Figure 1.

Geometry has been long known to provide insightful information on several computer vision tasks. Geometry can be described by various formats such as 3D polygon mesh representations [76], 3D voxel representations [36], point clouds [15], surface normal maps [12] and depth maps [13]. Depth maps, in particular, are widely utilized as priors in computer vision applications such as pose estimation [5], object detection [41, 59], instance segmentation [31] and image semantic segmentation [26] in the form of RGB-D input data, as they can provide cues about shape, texture and distance from the image plane. Traditional geometric computer vision were typically used up until recently for scene geometry estimation, such as stereo estimation algorithms [19], Structure-from-Motion (SfM) [79], etc. However, modern machine learning approaches, typically relying on feed-forward Deep Neural Networks (DNNs), have proven to be more robust as scene geometry estimation methods, if enough training data are available.

Image semantic segmentation is the task of processing an input image in order to extract a spatially aligned 2D map (in pixel coordinates) marking all semantic classes of interest. Each one expresses how likely it is for every image pixel to belong to a specific class; in effect it is per-pixel classification. Point cloud segmentation may be needed in certain variants, a scenario where classification occurs at 3D point level, instead of pixel-level. Modern semantic segmentation approaches typically rely on Convolutional Neural Networks (CNNs), which have shown excellent performance on most computer vision tasks.

Since geometry and semantics estimation are two tasks frequently needed together, it is plausible to unify them under a common framework. This framework should be able to extract both outputs from image data, as well as provide a fertile breeding ground for task collaboration. This can be achieved with the use of multitask DNNs [57, 73, 6], specifically designed for simultaneously predicting semantic segmentation and depth estimation from single monocular images. A successful estimation of the depth map corresponding to a given image, can be used to project every image pixel back to its 3D coordinates, provided that the camera intrinsic parameters are known. Then, the problem of classifying every 3D point becomes trivial, given that the semantic map is already estimated.

This paper is a survey of learning-based geometry estimation, image semantic segmentation and joint inference of both, *focusing only on DNN methods*. Special emphasis is given on unsupervised monocular depth map estimation, where scene geometry is derived from typical RGB camera data.

2. Scene geometry estimation

Scene geometry estimation is a traditional computer vision problem, originating from photogrammetry [69]. Older methods, relying on image processing and direct multi-view geometry, tend nowadays to be replaced by machine learning approaches (mostly DNN-based) exploiting geometric insights. Several subtasks can be identified within the larger scene geometry estimation problem, which are presented below.

2.1. Depth map estimation

Traditionally, DNN-based depth map estimation consisted in extracting a disparity map from an input stereo image pair and then converting it to an approximate depth map by exploiting camera parameters. Disparity estimation proceeded by matching features across the left/right image, either by employing DNNs to learn a good match between image patches [89], or implicitly match learned image features [56]. Lately, relevant research has moved towards inferring depth maps from monocular input. Although initial attempts [64, 45, 47, 13] treated depth map estimation as a strictly supervised task, requiring large datasets annotated with ground-truth (obtainable only by expensive sensors), more convenient unsupervised methods have recently emerged.

In one of the first unsupervised DNN approaches [18], the depth map is estimated from the right image of an input stereo pair, but no ground-truth is required for training. The employed loss function converts each pixel's predicted depth value into a disparity value (using known camera parameters) and uses it to warp the input right image into a reconstruction of the known left image. Aggregate photometric difference between the original left image and its reconstruction (synthesized using the predicted depth map) serves as the training loss. The method exploits the fact that a pixel's disparity value on the left image $\mathbf{I}_l$ directly indicates its apparent position on the right image. In this case, assuming rectified stereo pairs, pixel $\mathbf{p} \in \mathbb{R}^2$ of $\mathbf{I}_l$ should appear displaced in $\mathbf{I}_r$, in position $\mathbf{p}' \in \mathbb{R}^2$:

$$\mathbf{p}' = \mathbf{p} + \frac{fT}{\mathbf{D}(\mathbf{p})}, \tag{1}$$

where f is the focal length, T the stereo baseline and $\mathbf{D}$ the depth map corresponding to $\mathbf{I}_l$. In this way, $\mathbf{I}_l$ can be recreated by warping $\mathbf{I}_r$ to form $\mathbf{I}'_l$, such that:

$$\mathbf{I}_l(\mathbf{p}) \approx \mathbf{I}'_l(\mathbf{p}) = \mathbf{I}_r(\mathbf{p}'). \tag{2}$$

The DNN input during training is monocular, since the complementary view of the image pair is only exploited for computing the loss at each training iteration. As a result, the trained model only requires single-view input during inference. The network architecture follows the CNN encoder-decoder paradigm, with skip connections between the two subnetworks used for retaining high-resolution fine spatial details. The network output is a depth map that is gradually optimized during training by a variant of gradient descent on the following photometric loss:

$$L_{photo} = \sum_{\mathbf{p} \in \Omega} \|\mathbf{I}_r(\mathbf{p}') - \mathbf{I}_l(\mathbf{p})\|^2. \tag{3}$$

Going a step further, [21] directly infers both left-to-right and right-to-left disparity maps during training, using only the left image as CNN input, and the loss function enforces them to be consistent with each other. During inference on the trained model, each predicted left disparity map is converted to a depth map by trivial post-processing, using the known camera parameters, while the right disparity map is ignored.

Depth estimation from monocular videos in an unsupervised manner has been actively explored in recent years, using a neural SfM approach [93, 74, 83, 2, 23, 73, 50, 25], where, instead of stereo image pairs, the desired multiview nature of the data is derived from consecutive video frames of a sequence $\mathcal{I} = \{\mathbf{I}_0, ..., \mathbf{I}_t, \mathbf{I}_{t+1}, ...\}$ coming from a moving camera. These methods exploit the ability to estimate the relative camera pose $\mathbf{T}_{t \to t+1}$ between $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$ and knowledge of the intrinsic camera parameters $\mathbf{K}$. Thus, correspondences between pixels $\mathbf{p}_t \in \mathbf{I}_t$ and $\mathbf{p}_{t+1} \in \mathbf{I}_{t+1}$ can be easily found:

$$\mathbf{p}_{t+1} \approx \mathbf{K}\mathbf{T}_{t \to t+1}\mathbf{D}(\mathbf{p}_t)\mathbf{K}^{-1}\mathbf{p}_t, \tag{4}$$

where $\mathbf{D}$ is a depth map predicted by a CNN. Based on this, an approximation $\mathbf{I}'_t$ of $\mathbf{I}_t$ can be found by warping $\mathbf{I}_{t+1}$ via

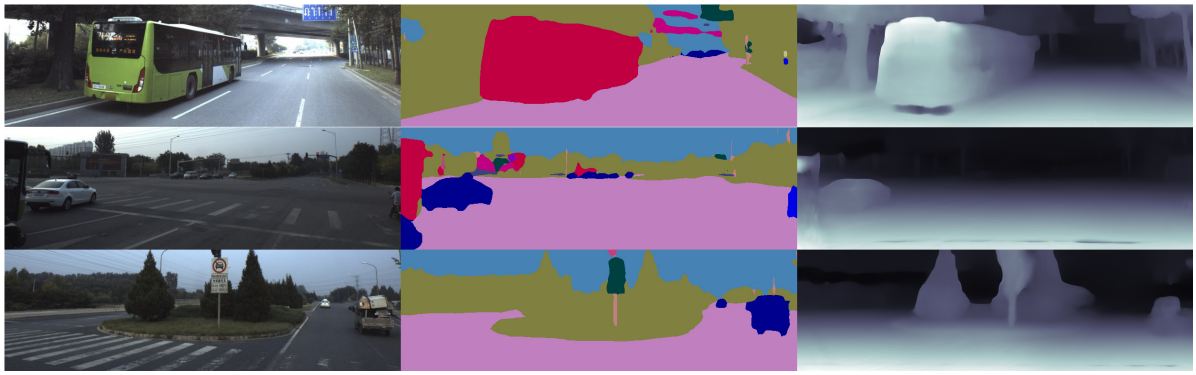

Figure 1: Examples of DNN-predicted segmentation maps (center column, using method [85]) and depth maps (right column, using method [2], on the Apolloscape dataset [32]. The left column contains the respective RGB image inputs fed to the trained DNNs.

differentiable bilinear interpolation proposed in [35], and then, the photometric loss between $\mathbf{I}'_t$ and $\mathbf{I}_t$ is minimized.

The basic loss can be enriched with additional loss terms meant to handle problematic cases, namely occlusions[62, 2], complex illumination changes[2], differences between the estimations of each view [2, 50], homogeneous regions [62, 2], among others. Thus, depth map estimation per video frame can be learnt, using a monocular video frame sequence and known (or estimated [23, 9, 73]) camera intrinsic parameters as input.

Due to depth estimation being an ill-posed problem in the case of SfM, mostly due to the appearance of multiple independently moving objects in the scene, rigid and non-rigid, some methods handle moving objects via estimating optical flow or semantic segmentation [62, 95, 83] or via observing inconsistencies between estimated depth maps of neighboring video frames [2].

Exploiting stereo video sequences has been investigated as well. For instance, [90] uses an encoder-decoder CNN architecture (ResNet50 with with half filters in the encoder part) and employs both spatial (through stereo) and temporal (forward-backward) photometric warp error as loss terms during training, thus constraining the scene depth and camera motion to be in a common, real world scale. Come inference time, the network is able to estimate depth using single-view input.

2.2. Point cloud generation

There have been numerous works on 3D point cloud estimation [78, 87]. For instance, [15] uses a CNN that generates 3D point cloud coordinates given a single image as input. The proposed network has an encoder stage and a predictor stage with skip connections between them. The encoder predicts embeddings from images paired with a random vector. The predictor outputs a matrix, each row containing the coordinates of a point. An improved version of the predictor employs two parallel branches, one that predicts points as before, and another one that predicts a 3-channel image, of which the three values at each pixel are the coordinates of a point, giving a new set of points. Their predictions are later merged together to form the whole set of points in the matrix. To optimize the network, two novel supervised loss functions for point cloud data are proposed. [80] generates new views of a scene from a single image. The model reasons about the 3D structure without 3D supervision, trained end-to-end on image pairs. The input image is projected to a point cloud of learned feature vectors, which are rendered from the target view using a novel differentiable point cloud renderer and passed to a CNN to generate the final image.

2.3. Mesh representation estimation

[76] predicts triangular meshes from single images. It progressively deforms an ellipsoid using a Graph CNN. Similarly, [20] builds on a popular instance segmentation network [30] and extends it to infer triangular meshes. Given an input image, all objects' 2D instance boxes, masks and their 3D object shapes are inferred in an end-to-end manner. [24] deforms a set of squares to cover the surface of a 3D shape, taking images and point cloud data as inputs. All these works rely on the presence of ground truth 3D data for training.

2.4. Voxel and octree representation estimation

[81] predicts a voxel representation of an object, given a single-view depth map as input. In [36] a 3D CNN is used to generate a 3D model in voxel format, with a set of multiview images plus the corresponding camera parameters as inputs. [10] employs a 3D Recurrent Reconstruction Neural Network to map an image to a voxel representation. The

learned representation is incrementally refined as the network sees more views of the object. Ground truth 3D data is again necessary for training all of the aforementioned methods. For higher resolution, without further memory needs, octree representations have been explored [28, 63, 70].

2.5. Other geometry-related problems

Since it is feasible to infer 3D scene structure from a single image, matching 3D models extracted from multiple images of the same scene in order to reconstruct the latter would be desirable. [22] proposes a new 3D data representation called smoothed density value (SDV) voxelization that can be handled with regular CNNs. It also proposes a Siamese network able to learn 3D local feature descriptors that are rotation invariant and compact to search for correspondences. Point cloud ground truth is required for training. In [14], a large-scale point cloud and a close-proximity scanned point cloud are matched, providing a localization solution, with no need of supervision during training.

Long before the advent of modern CNNs, SfM had been extended in robotics towards Visual Simultaneous Localization and Mapping (vSLAM) [52]. In effect, vSLAM systems integrate real-time, on-the-fly SfM methods with additional modules, such as ones for place recognition (which is a semantics problem) and loop closure detection (which can be considered a geometry-related task) [44, 72]. Thus, [17] uses a modified stacked denoising autoencoder to extract meaningful features from image patches, which are then compared to extracted features from other images using a similarity score. A loop closure is detected when two images resemble sufficiently.

[75] uses a Siamese encoder-decoder structured network trained with a novel Gauss-Newton loss to perform relocalization. The features extracted from the network are shown to be invariant of different weather conditions. [82] is a vSLAM system based on ORB-SLAM2 [58] that handles dynamic objects in dynamic environments by training an SSD visual object detector [48] to identify them. Since dynamic objects such as moving cars and pedestrians can negatively affect the SLAM procedure, [82] treats them as outliers.

[3] tries to minimize the size complexity of dense geometric representations (compared to sparse ones) and employs an autoencoder architecture in order to compress the information. Thus, it uses a U-Net [93] to decompose an intensity image into convolutional features. These features are then fed into the depth autoencoder (variational autoencoder) in its intermediate levels. A variational component in the bottleneck of the depth autoencoder is composed of two fully connected layers followed by the computation of the mean and variance, from which the latent space is then sampled. The network, instead of predicting just raw depth values, it predicts a mean μ and an uncertainty b for every depth pixel.

3. Semantic image segmentation

As in most computer vision subfields, CNNs turned out to be a revolution regarding semantic segmentation accuracy. Semantic segmentation CNNs are typically composed of an encoding and a decoding subnetwork, arranged in a consecutive fashion. The encoder extracts semantic features from the input lowering spatial resolution progressively, while the decoder receives the final encoder output and upsamples it. The final output of the decoder is a semantic map, having the same spatial resolution as the input and as many channels as the supported number of discrete classes, thus performing per-pixel classification. Training is performed in a typical supervised manner.

Fully convolutional networks [49] and dilated convolutions [86] are typically used, for upsampling the computed abstract feature maps and for enlarging neuronal receptive fields, respectively. Large receptive fields significantly aid semantic segmentation by enriching local-scale image representation with task-relevant wider region semantic context, for more accurate per-pixel classification. Grasping global image context while retaining spatial detail is an important consideration in relevant research, with current CNNs attempting to explicitly capture it and properly enhance local image representation, using multi-scale [92], attention-based relational [88], or network branching approaches [85].

PSPNet [92] offers a good balance between speed and accuracy, forming also the backbone of more recent real-time segmentation networks such as ICNet [91]. Its main novelty is a PPM (Pyramid Pooling Module) decoder, able to enrich local image representation with more global context information from larger image regions of various scales. Semantic information from each image region is aggregated using global average pooling within the region, separately for each tensor channel. DeepLabV3+ [7] has a structure similar to PSPNet, but relies on the so-called ASPP module instead of the PPM; the former employs multiple dilated convolutions with different dilation factors, for achieving the same purpose as the latter. The advantage compared to PPM is that fine spatial information is not lost, as is the case with global average pooling. Aiming towards accurate real-time semantic segmentation, [85] employs two separate network branches, one shallow branch (Spatial path) that extracts low level image features to preserve spatial details, and a deep lightweight feature extractor (Context path) to obtain a large receptive field for high level context. The two branches are later concatenated and fed to a shallow CNN module for the final prediction. [88] presents an object-contextual representation (OCR) approach. Instead of naively using dilated convolutions to model the context of a pixel, OCR learns object region representations; the context of a pixel is the object that it belongs to.

[68] observes that texture and shape should not be processed together in the same network stream. Thus, two separate streams are proposed, namely "Regular" stream for producing dense pixel features, and a parallel "Shape" stream that processes image gradients (image edges) and Regular steam features in order to produce semantic boundaries. The two predictions are later fused together using an ASPP module to come up with the final predicted semantic map. During training, the Shape stream is trained on semantic boundaries extracted from the segmentation ground truth using binary cross entropy loss. The output semantic maps are optimized on a regular cross entropy loss. The two tasks constrain each other as well, using a dual task regularizer that penalizes differences between predicted semantic map boundaries with boundaries predicted by the Shape stream. [55] proposes an end-to-end trainable model for semantic segmentation with a built-in awareness of semantic boundaries. [4] introduces a loss function that encourages the DNN to predict segments with correct boundaries, but the method is limited to binary semantic segmentation problems (class of interest and "background").

[43] achieves real-time semantic segmentation on high resolution images by proposing a novel deep feature aggregation strategy. [33] introduces the "criss-cross" attention mechanism, which adaptively captures contextual information for each pixel on the vertical and horizontal axes. [94] uses video prediction models to both get new image frames and their respective semantic labels (label propagation). A new boundary label relaxation technique is presented, to alleviate errors along the object's borders on the semantic maps during label propagation.

A different approach [54] aiming towards faster inference (e.g., for real-time execution on embedded computing boards that are typically found in autonomous systems), employs non-uniform content-aware input image downsampling, instead of typical uniform input downsampling. The transformation parameters are predicted by an additional CNN that learns from a non-uniform sample geometric model driven by semantic boundaries.

4. Joint scene geometry estimation and semantic segmentation

A vast amount of research [27, 66, 65] has treated depth maps as a given input source in computer vision tasks, either by making depth an additional channel to RGB images (RBG-D inputs), or by exploiting geometry in the 3D domain. [6] exploits depth by training a multitask network that jointly learns to infer depth and semantic segmentation. It also introduces feature concatenation, in which depth features are concatenated with RGB features, thus incorporating depth on feature level. In [46, 57] single-view depth and semantic segmentation estimated by either a multitask network [57] or two separate ones [46] are fed to fully connected Conditional Random Fields (CRF) to further refine the predicted semantic maps. [29] makes use of RGB-D data by having two encoders, one for processing the RGB image and one for the depth channel. The two encoders are later fused together. In the encoder section each block of the depth encoder is fed as an input to the corresponding block of the segmentation encoder. [1] uses a single encoder-decoder architecture for both tasks. Predicted depth is treated as an additional output channel to be estimated along with semantic segmentation. It uses a Huber regression loss [34] (less sensible to outliers than mean squared error) to train both tasks, as it proves that it can successfully handle semantic segmentation as well, although being a classification problem.

While most multitask networks for depth and semantic maps estimation are trained using a simple weighted averaging of the respective losses, [40] proposes a way to make loss term coefficients learnable, using homoscedastic uncertainty, leading to easy tuning of the training process.

[60] performed 2D semantic image segmentation using a 3D Graph Neural Network built on top of 3D points with both color intensities and depth, extracted from RGB-D data. [77] introduces depth-wise convolution in semantic segmentation. It simply adds a similarity term inside the convolution operation via multiplication, so that pixels that have similar depth values to the central pixel (pixel that corresponds to the center of the kernel) contribute more to the total convolution summation than those of dissimilar depth values. In the same sense, depth-aware average pooling is proposed, in which the same similarity term is used as well. Both depth-aware convolution and depth-aware average pooling can enable any segmentation network to incorporate depth information without additional parameters.

[73] proposes a multi-stage network architecture that can predict depth, semantic maps, optical flow, per-pixel motion probabilities and motion masks from monocular video. It manages to infer all these tasks in real time, plus achieve state-of-the art accuracy for self-supervised monocular depth estimation, optical flow estimation among monocular multi-task frameworks and motion segmentation.

In [37] depth information is also learned along with semantic segmentation in a supervised multitask manner. The learned features of both tasks are later merged in a Geometry Aware propagation block, to further enhance semantic segmentation performance. [67] uses RGB-D cameras to model the background of an image before using it for foreground segmentation.

[61] trains a multitask network for both semantic segmentation and supervised depth estimation. For consistency between the two tasks, a Cross-Domain Discontinuity Term is proposed based on the observation that depth discontinuities are likely to co-occur with semantic boundaries.

Figure 2: Example of semantic segmentation performed on drone-captured RGB video data.

This term detects discontinuities between semantic labels encoded by the sign of the absolute value of the gradients in the semantic map. The idea behind this loss is that there should be a gradient peak between adjacent pixels belonging to different classes. [8] formulates a similar smoothness loss term that regularizes the smoothness of depth values within each segmentation mask.

CNN-predicted dense depth maps are fused together with depth measurements obtained from direct monocular SLAM in [71]. This integrates SLAM with monocular depth prediction. It is also shown that CNN-predicted semantic segmentation can be coherently fused with the global reconstruction model. The depth prediction architecture used here is derived from [42]. It is able to overcome problems such as good estimation of the absolute scale, depth prediction in textureless areas, etc.

In [84] RGB sequences are fed to a segmentation network to segment each video frame in real-time. It also uses non-neural optical flow estimations to detect inconsistently moving points. Then, using depth maps that correspond to these input RGB images and the predicted semantic maps, a semantic octree map is built, with non-static objects removed.

5. Conclusions

The evident progress of autonomous systems technology makes semantic 3D world modelling algorithms relying on simple sensors (such as RGB cameras) a necessity. AI methods hold the promise of achieving high-accuracy, on-the-fly 3D scene perception, with Deep Neural Networks being at the forefront of relevant research. This paper surveyed the current state-of-the-art in this very active area, with the aim to facilitate further progress in the field.

Acknowledgment

This work has received funding from the European Union's Seventh Horizon 2020 research and innovation programme under grant agreement number 871479 (AERIAL-CORE). This publication reflects the authors' views only. The European Commission is not responsible for any use that may be made of the information it contains

References

[1] M. Aladem and S.A. Rawashdeh. A single-stream segmentation and depth prediction CNN for autonomous driving. *IEEE Intelligent Systems*, 2020.

[2] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M. Cheng, and I. Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Proceedings of Advances in neural information processing systems (NIPS)*, 2019.

[3] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A.J. Davison. CodeSLAM—learning a compact, optimisable representation for dense visual SLAM. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[4] A. Bokhovkin and E. Burnaev. Boundary loss for remote sensing imagery semantic segmentation. In *International Symposium on Neural Networks*. Springer, 2019.

[5] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6D object pose estimation using 3D

object coordinates. In *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2014.

[6] Y. Cao, C. Shen, and H. T. Shen. Exploiting depth from single monocular images for object detection and semantic segmentation. *IEEE Transactions on Image Processing*, 26(2), 2017.

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[8] P.Y. Chen, A. H Liu, Y.C. Liu, and Y.C.F Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Y. Chen, C. Schmid, and C. Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2019.

[10] C.B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2016.

[11] S. Delis, I. Mademlis, N. Nikolaidis, and I. Pitas. Automatic detection of 3D quality defects in stereoscopic videos using binocular disparity. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(5), 2016.

[12] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2015.

[13] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of Advances in neural information processing systems (NIPS)*, 2014.

[14] G. Elbaz, T. Avraham, and A. Fischer. 3D point cloud registration for localization using a deep neural network auto-encoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] H. Fan, H. Su, and L.J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.

[16] R. Fan, U. Ozgunalp, B. Hosking, M. Liu, and I. Pitas. Pothole detection based on disparity transformation and road surface modeling. *IEEE Transactions on Image Processing*, 29, 2019.

[17] X. Gao and T. Zhang. Unsupervised learning to detect loops using deep neural networks for visual SLAM system. *Autonomous robots*, 41(1), 2017.

[18] R. Garg, V.K. Bg, G. Carneiro, and I. Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *Proceedings of European conference on computer vision (ECCV)*. Springer, 2016.

[19] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In *Proceedings of the Asian conference on computer vision (ACCV)*. Springer, 2010.

[20] G. Gkioxari, J. Malik, and J. Johnson. Mesh R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[21] C. Godard, O. Mac Aodha, and G.J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] Z. Gojcic, C. Zhou, J.D. Wegner, and A. Wieser. The perfect match: 3D point cloud matching with smoothed densities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[23] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[24] T. Groueix, M. Fisher, V.G. Kim, B.C. Russell, and M. Aubry. A papier-mâché approach to learning 3D surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[25] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon. 3D packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[26] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik. Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation. *International Journal of Computer Vision*, 112(2), 2015.

[27] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2014.

[28] C. Häne, S. Tulsiani, and J. Malik. Hierarchical surface prediction for 3D object reconstruction. In *Proceedings of the 2017 International Conference on 3D Vision (3DV)*. IEEE, 2017.

[29] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers. FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In *Proceedings of the Asian conference on computer vision (ACCV)*. Springer, 2016.

[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017.

[31] J. Hou, A. Dai, and M. Nießner. 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[32] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[33] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu. CCNet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[34] P. J. Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*. Springer, 1992.

[35] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Proceedings of Advances in neural information processing systems (NIPS)*, 2015.

[36] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang. SurfaceNet: An end-to-end 3D neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[37] J. Jiao, Y. Wei, Z. Jie, H. Shi, R.WH. Lau, and T.S. Huang. Geometry-aware distillation for indoor semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[38] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas. UAV cinematography constraints imposed by visual target tracking. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2018.

[39] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas. Shot type constraints in UAV cinematography for autonomous target tracking. *Information Sciences*, 506, 2020.

[40] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[41] J. Lahoud and B. Ghanem. 2D-driven 3D object detection in RGB-D images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[42] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *Proceedings of the 2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016.

[43] H. Li, P. Xiong, H. Fan, and J. Sun. DFANet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[44] S. Li, T. Zhang, X. Gao, D. Wang, and Y. Xian. Semi-direct monocular visual and visual-inertial SLAM with loop closure detection. *Robotics and Autonomous Systems*, 112, 2019.

[45] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015.

[46] J. Liu, Y. Wang, Y. Li, J. Fu, J. Li, and H. Lu. Collaborative deconvolutional neural networks for joint depth estimation and semantic segmentation. *IEEE transactions on neural networks and learning systems*, 29(11), 2018.

[47] M. Liu, M. Salzmann, and X. He. Discrete-continuous depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[48] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single-shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016.

[49] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[50] X. Luo, J. Huang, R. Szeliski, K. Matzen, and J. Kopf. Consistent video depth estimation. *arXiv preprint arXiv:2004.15021*, 2020.

[51] I. Mademlis, N. Nikolaidis, and I. Pitas. Stereoscopic video description for key-frame extraction in movie summarization. In *Proceedings of the EURASIP European Signal Processing Conference (EUSIPCO)*. IEEE, 2015.

[52] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina. Autonomous unmanned aerial vehicles filming in dynamic unstructured outdoor environments. *IEEE Signal Processing Magazine*, 36(1), 2018.

[53] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and Al. Messina. Autonomous UAV cinematography: a tutorial and a formalized shot-type taxonomy. *ACM Computing Surveys (CSUR)*, 52(5), 2019.

[54] D. Marin, Z. He, P. Vajda, P. Chatterjee, S. Tsai, F. Yang, and Y. Boykov. Efficient segmentation: Learning downsampling near semantic boundaries. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[55] D. Marmanis, K. Schindler, J.D. Wegner, S. Galliani, M. Datcu, and U. Stilla. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135, 2018.

[56] N. Mayer, E. Ilg, P. Hausser, D. Fischer, P.and 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 IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.

[57] A. Mousavian, H. Pirsiavash, and J. Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 2016.

[58] R. Mur-Artal and J. D. Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5), 2017.

[59] T. Ophoff, K. Van Beeck, and T. Goedemé. Exploring RGB+Depth fusion for real-time object detection. *Sensors*, 19(4), 2019.

[60] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun. 3D graph neural networks for RGBD semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[61] P. Z. Ramirez, M. Poggi, F. Tosi, S. Mattoccia, and L. Di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2018.

[62] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M.J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019.

[63] G. Riegler, A. Osman Ulusoy, and A. Geiger. OctNet: Learning deep 3D representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[64] A. Roy and S. Todorovic. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.

[65] M. Schwarz, A. Milan, A.S. Periyasamy, and S. Behnke. RGB-D object detection and semantic segmentation for autonomous manipulation in clutter. *The International Journal of Robotics Research*, 37(4-5), 2018.

[66] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *Proceedings of the European conference on computer vision (ECCV)*. Springer, 2012.

[67] Y. Sun, M. Liu, and M.Q-H. Meng. Active perception for foreground segmentation: An RGB-D data-based background modeling method. *IEEE Transactions on Automation Science and Engineering*, 16(4), 2019.

[68] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler. Gated-SCNN: Gated shape CNNs for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[69] C. V. Tao and Y. Hu. 3D reconstruction methods. *Photogrammetric Engineering & Remote Sensing*, 68(7), 2002.

[70] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[71] K. Tateno, F. Tombari, I. Laina, and N. Navab. CNN-SLAM: Real-time dense monocular slam with learned depth prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[72] Y. Tian, K. Khosoussi, and J.P. How. A resource-aware approach to collaborative loop closure detection with provable performance guarantees. *arXiv preprint arXiv:1907.04904*, 2019.

[73] F. Tosi, F. Aleotti, P. Z. Ramirez, M. Poggi, S. Salti, L. Di Stefano, and S. Mattoccia. Distilled semantics for comprehensive scene understanding from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[74] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017.

[75] L. von Stumberg, P. Wenzel, Q. Khan, and D. Cremers. GN-Net: The gauss-newton loss for multi-weather relocalization. *IEEE Robotics and Automation Letters*, 5(2), 2020.

[76] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y. Jiang. Pixel2Mesh: Generating 3D mesh models from single RGB images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[77] W. Wang and U. Neumann. Depth-aware CNN for RGB-D segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[78] Y. Wang, Y. Sun, Z. Liu, S.E. Sarma, M.M. Bronstein, and J.M. Solomon. Dynamic graph CNN for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5), 2019.

[79] M.J. Westoby, J. Brasington, N.F. Glasser, M.J. Hambrey, and J.M. Reynolds. 'structure-from-motion'photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179, 2012.

[80] Ol. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson. SynSin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[81] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015.

[82] L. Xiao, J. Wang, X. Qiu, Z. Rong, and X. Zou. Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment. *Robotics and Autonomous Systems*, 117, 2019.

[83] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[84] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei. DS-SLAM: A semantic visual SLAM towards dynamic environments. In *Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018.

[85] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[86] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[87] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert. PCN: Point completion network. In *Proceedings of 2018 International Conference on 3D Vision (3DV)*. IEEE, 2018.

[88] Y. Yuan, X. Chen, and J. Wang. Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019.

[89] J. Zbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015.

[90] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[91] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. ICNet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[92] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[93] T. Zhou, M. Brown, N. Snavely, and D.G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*, 2017.

[94] Y. Zhu, K. Sapra, F.A. Reda, K.J. Shih, S. Newsam, A. Tao, and B. Catanzaro. Improving semantic segmentation via

video propagation and label relaxation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[95] Y. Zou, Z. Luo, and J. Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.