Distilled Semantics for Comprehensive Scene Understanding from Videos

Fabio Tosi* Filippo Aleotti* Pierluigi Zama Ramirez*
Matteo Poggi Samuele Salti Luigi Di Stefano Stefano Mattoccia
Department of Computer Science and Engineering (DISI)
University of Bologna, Italy
*{fabio.tosi5, filippo.aleotti2, pierluigi.zama}@unibo.it

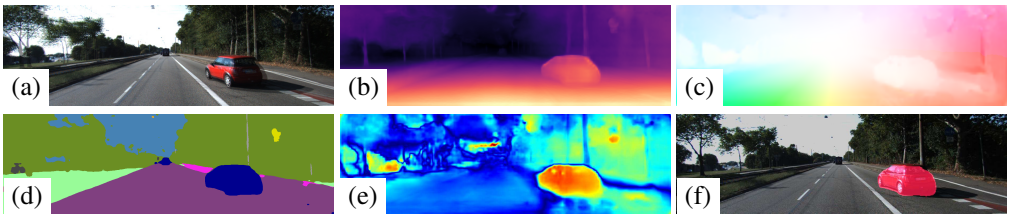

Figure 1. Given an input monocular video (a), our network can provide the following outputs in real-time: depth (b), optical flow (c), semantic labels (d), per-pixel motion probabilities (e), motion mask (f).

Abstract

Whole understanding of the surroundings is paramount to autonomous systems. Recent works have shown that deep neural networks can learn geometry (depth) and motion (optical flow) from a monocular video without any explicit supervision from ground truth annotations, particularly hard to source for these two tasks. In this paper, we take an additional step toward holistic scene understanding with monocular cameras by learning depth and motion alongside with semantics, with supervision for the latter provided by a pre-trained network distilling proxy ground truth images. We address the three tasks jointly by a) a novel training protocol based on knowledge distillation and self-supervision and b) a compact network architecture which enables efficient scene understanding on both power hungry GPUs and low-power embedded platforms. We thoroughly assess the performance of our framework and show that it yields state-of-the-art results for monocular depth estimation, optical flow and motion segmentation.

1. Introduction

What information would an autonomous agent be keen to gather from its sensory sub-system to tackle tasks like navigation and interaction with the explored environment? It would need to be informed about the geometry of the surroundings and the type of objects therein, and likely better know which of the latter are actually moving and how they do so. What if all such cues may be provided by as simple a sensor as a single RGB camera?

Nowadays, deep learning is advancing the state-of-the-art in classical computer vision problems at such a quick pace that single-view holistic scene understanding seems to be no longer out-of-reach. Indeed, highly challenging problems such as monocular depth estimation and optical flow can nowadays be addressed successfully by deep neural networks, often through unified architectures [88, 3, 96]. Self-supervised learning techniques have yielded further major achievements [95, 58] by enabling effective training of deep networks without annotated images. In fact, labels are hard to source for depth estimation due to the need of active sensors and manual filtering, and are even more cumbersome in the case of optical flow. Concurrently, semi-supervised approaches [90, 16] proved how a few semantically labelled images can improve monocular depth estimation significantly. These works have also highlighted how, while producing per-pixel class labels is tedious yet feasible for a human annotator, manually endowing images with depth and optical flow ground-truths is prohibitive.

In this paper, we propose the first-ever framework for comprehensive scene understanding from monocular videos. As highlighted in Figure 1, our multi-stage network architecture, named ΩNet, can predict depth, semantics, optical flow, per-pixel motion probabilities and motion masks. This comes alongside with estimating the pose between adjacent frames for an uncalibrated camera, whose intrinsic parameters are also estimated. Our training methodology

*Joint first authorship.

leverages on self-supervision, knowledge distillation and multi-task learning. In particular, peculiar to our proposal and key to performance is distillation of proxy semantic labels gathered from a state-of-the-art pre-trained model [52] within a self-supervised and multi-task learning procedure addressing depth, optical flow and motion segmentation. Our training procedure also features a novel and effective self-distillation schedule for optical flow mostly aimed at handling occlusions and relying on tight integration of rigid flow, motion probabilities and semantics. Moreover, ΩNet is lightweight, counting less than 8.5M parameters, and fast, as it can run at nearly 60 FPS and 5 FPS on an NVIDIA Titan Xp and a Jetson TX2, respectively. As vouched by thorough experiments, the main contributions of our work can be summarized as follows:

- The first real-time network for joint prediction of depth, optical flow, semantics and motion segmentation from monocular videos
- A novel training protocol relying on proxy semantics and self-distillation to effectively address the self-supervised multi-task learning problem
- State-of-the-art self-supervised monocular depth estimation, largely improving accuracy at long distances
- State-of-the-art optical flow estimation among monocular multi-task frameworks, thanks to our novel occlusion-aware and semantically guided training paradigm
- State-of-the-art motion segmentation by joint reasoning about optical-flow and semantics

2. Related Work

We review previous works relevant to our proposal.

Monocular depth estimation. At first, depth estimation was tackled as a supervised [24, 49] or semi-supervised task [48]. Nonetheless, self-supervision from image reconstruction is now becoming the preferred paradigm to avoid hard to source labels. Stereo pairs [25, 28] can provide such supervision and enable scale recovery, with further improvements achievable by leveraging on trinocular assumptions [64], proxy labels from SGM [76, 80] or guidance from visual odometry [2]. Monocular videos [95] are a more flexible alternative, although they do not allow for scale recovery and mandate learning camera pose alongside with depth. Recent developments of this paradigm deal with differentiable direct visual odometry [77] or ICP [57] and normal consistency [87]. Similarly to our work, [88, 96, 17, 3, 86, 56] model rigid and non-rigid components using the projected depth, relative camera transformations, and optical flow to handle independent motions, which can also be estimated independently in the 3D space [9, 83]. In [30], the authors show how to learn camera intrinsics together with depth and egomotion to enable training on any unconstrained video. In [29, 94, 6], reasoned design choices such as a minimum reprojection loss between frames, self-assembled attention modules and auto-mask strategies to handle static camera or dynamic objects proved to be very effective. Supervision from stereo and video have also been combined [91, 29], possibly improved by means of proxy supervision from stereo direct sparse odometry [84]. Uncertainty modeling for self-supervised monocular depth estimation has been studied in [63]. Finally, lightweight networks aimed at real-time performance on low-power systems have been proposed within self-supervised [62, 61] as well as supervised [81] learning paradigms.

Semantic segmentation. Nowadays, fully convolutional neural networks [55] are the standard approach for semantic segmentation. Within this framework, multi-scale context modules and proper architectural choices are crucial to performance. The former rely on spatial pyramid pooling [31, 93] and atrous convolutions [14, 13, 15]. As for the latter, popular backbones [47, 74, 32] have been improved by more recent designs [34, 18]. While for years the encoder-decoder architecture has been the most popular choice [70, 4], recent trends in Auto Machine Learning (AutoML) [52, 12] leverage on architectural search to achieve state-of-the-art accuracy. However, these latter have huge computational requirements. An alternative research path deals with real-time semantic segmentation networks. In this space, [60] deploys a compact and efficient network architecture, [89] proposes a two paths network to attain fast inferences while capturing high resolution details. DABNet [50] finds an effective combinations of depth-wise separable filters and atrous-convolutions to reach a good trade-off between efficiency and accuracy. [51] employs cascaded sub-stages to refine results while FCHardNet [11] leverages on a new harmonic densely connected pattern to maximize the inference performance of larger networks.

Optical flow estimation. The optical flow problem concerns estimation of the apparent displacement of pixels in consecutive frames, and it is useful in various applications such as, *e.g.*, video editing [10, 43] and object tracking [82]. Initially introduced by Horn and Schunck [33], this problem has traditionally been tackled by variational approaches [8, 7, 69]. More recently, Dosovitskiy *et al.* [21] showed the supremacy of deep learning strategies also in this field. Then, other works improved accuracy by stacking more networks [38] or exploiting traditional pyramidal [65, 75, 35] and multi-frame fusion [67] approaches. Unfortunately, obtaining even sparse labels for optical flow is extremely challenging, which renders self-supervision from images highly desirable. For this reason, an increasing number of methods propose to use image reconstruction and spatial smoothness [41, 68, 73] as main signals to guide the training, paying particular attention to occluded regions [58, 85, 53, 54, 40, 37].

Semantic segmentation and depth estimation. Monocular depth estimation is tightly connected to the

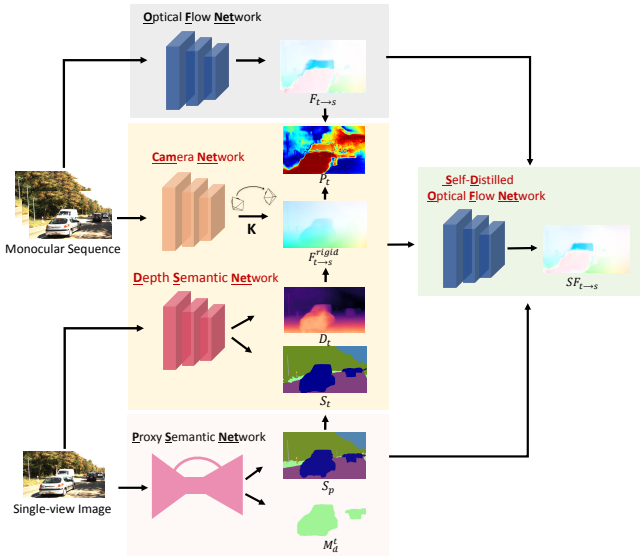

Figure 2. Overall framework for training ΩNet to predict depth, camera pose, camera intrinsics, semantic labels and optical flow. In **red** architectures composing ΩNet.

semantics of the scene. We can infer the depth of a scene by a single image mostly because of context and prior semantic knowledge. Prior works explored the possibility to learn both tasks with either full supervision [78, 23, 59, 45, 92, 44, 22] or supervision concerned with semantic labels only [90, 16]. Unlike previous works, we propose a compact architecture trained by self-supervision on monocular videos and exploiting proxy semantic labels.

Semantic segmentation and optical flow. Joint learning of semantic segmentation and optical flow estimation has been already explored [36]. Moreover, scene segmentation [72, 5] is required to disentangle potentially moving and static objects for focused optimizations. Differently, [66] leverages on optical flow to improve semantic predictions of moving objects. Peculiarly w.r.t. previous work, our proposal features a novel self-distillation training procedure guided by semantics to improve occlusion handling.

Scene understanding from stereo videos. Finally, we mention recent works approaching stereo depth estimation with optical flow [1] and semantic segmentation [42] for comprehensive scene understanding. In contrast, we are the first to rely on monocular videos to this aim.

3. Overall Learning Framework

Our goal is to develop a real-time comprehensive scene understanding framework capable of learning strictly related tasks from monocular videos. Purposely, we propose a multi-stage approach to learn first geometry and semantics, then elicit motion information, as depicted in Figure 2.

3.1. Geometry and Semantics

Self-supervised depth and pose estimation. We propose to solve a self-supervised single-image depth and pose estimation problem by exploiting geometrical constraints in a sequence of N images, in which one of the frames is used as the target view I_t and the other ones in turn as the source image I_s. Assuming a moving camera in a stationary scene, given a depth map D_t aligned with I_t, the camera intrinsic parameters K and the relative pose $T_{t\to s}$ between I_t and I_s, it is possible to sample pixels from I_s in order to synthesise a warped image $\widetilde{I}_t$ aligned with I_t. The mapping between corresponding homogeneous pixels coordinates $p_t \in I_t$ and $p_s \in I_s$ is given by:

$$p_s \sim K T_{t\to s} D_{p_t} K^{-1} p_t \tag{1}$$

Following [95], we use the sub-differentiable bilinear sampler mechanism proposed in [39] to obtain $\widetilde{I}_t$. Thus, in order to learn depth, pose and camera intrinsics we train two separate CNNs to minimize the photometric reconstruction error between $\widetilde{I}_t$ and I_t, defined as:

$$\mathcal{L}_{ap}^{D} = \sum_{p} \psi(I_t(p), \widetilde{I}_t(p)) \tag{2}$$

where ψ is a photometric error function between the two images. However, as pointed out in [29], such a formulation is prone to errors at occlusion/disocclusion regions or in static camera scenarios. To soften these issues, we follow the same principles as suggested in [29], where a minimum per-pixel reprojection loss is used to compute the photometric error, an automask method allows for filtering-out spurious gradients when the static camera assumption is violated, and an edge-aware smoothness loss term is used as in [28]. Moreover, we use the depth normalization strategy proposed in [77]. See supplementary material for further details.

We compute the rigid flow between I_t and I_s as the difference between the projected and original pixel coordinates in the target image:

$$F_{t\to s}^{rigid}(p_t) = p_s - p_t \tag{3}$$

Distilling semantic knowledge. The proposed distillation scheme is motivated by how time-consuming and cumbersome obtaining accurate pixel-wise semantic annotations is. Thus, we train our framework to estimate semantic segmentation masks S_t by means of supervision from cheap proxy labels S_p distilled by a semantic segmentation network, pre-trained on few annotated samples and capable to generalize well to diverse datasets. Availability of proxy semantic labels for the frames of a monocular video enables us to train a single network to predict jointly depth and semantic labels. Accordingly, the joint loss is obtained

by adding a standard cross-entropy term $\mathcal{L}_{sem}$ to the previously defined self-supervised image reconstruction loss $\mathcal{L}_{ap}^{D}$. Moreover, similarly to [90], we deploy a cross-task loss term, $\mathcal{L}_{edge}^{D}$ (see supplementary), aimed at favouring spatial coherence between depth edges and semantic boundaries. However, unlike [90], we do not exploit stereo pairs at training time.

3.2. Optical Flow and Motion Segmentation

Self-supervised optical flow. As the 3D structure of a scene includes stationary as well as non-stationary objects, to handle the latter we rely on a classical optical flow formulation. Formally, given two images I_t and I_s, the goal is to estimate the 2D motion vectors $F_{t\to s}(p_t)$ that map each pixel in I_t into its corresponding one in I_s. To learn such a mapping without supervision, previous approaches [58, 54, 88] employ an image reconstruction loss $\mathcal{L}_{ap}^{F}$ that minimizes the photometric differences between I_t and the back-warped image $\widetilde{I}_t$ obtained by sampling pixels from I_s using the estimated 2D optical flow $F_{t\to s}(p_t)$. This approach performs well for non-occluded pixels but provides misleading information within occluded regions.

Pixel-wise motion probability. Non-stationary objects produce systematic errors when optimizing $\mathcal{L}_{ap}^{D}$ due to the assumption that the camera is the only moving body in an otherwise stationary scene. However, such systematic errors can be exploited to identify non-stationary objects: at pixels belonging to such objects the rigid flow $F_{t\to s}^{rigid}$ and the optical flow $F_{t\to s}$ should exhibit different directions and/or norms. Therefore, a pixel-wise probability of belonging to an object independently moving between frames s and t, P_t, can be obtained by normalizing the differences between the two vectors. Formally, denoting with $\theta(p_t)$ the angle between the two vectors at location p_t, we define the per-pixel motion probabilities as:

$$P_t(p_t) = \max\{\frac{1-\cos\theta(p_t)}{2}, 1-\rho(p_t)\} \tag{4}$$

where $\cos\theta(p_t)$ can be computed as the normalized dot product between the vectors and evaluates the similarity in direction between them, while $\rho(p_t)$ is defined as

$$\rho(p_t) = \frac{\min\{\|F_{t\to s}(p_t)\|_2, \|F_{t\to s}^{rigid}(p_t)\|_2\}}{\max\{\|F_{t\to s}(p_t)\|_2, \|F_{t\to s}^{rigid}(p_t)\|_2\}}, \tag{5}$$

i.e. a normalized score of the similarity between the two norms. By taking the maximum of the two normalized differences, we can detect moving objects even when either the directions or the norms of the vectors are similar. A visualization of $P_t(p_t)$ is depicted in Fig. 3(d).

Semantic-aware Self-Distillation Paradigm. Finally, we combine semantic information, estimated optical flow, rigid flow and pixel-wise motion probabilities within a final training stage to obtain a more robust self-distilled optical flow network. In other words, we train a new instance of the model to infer a self-distilled flow $\text{SF}_{t\to s}$ given the estimates $\text{F}_{t\to s}$ from a first self-supervised network and the aforementioned cues. As previously discussed and highlighted in Figure 3(c), standard self-supervised optical flow is prone to errors in occluded regions due to the lack of photometric information but can provide good estimates for the dynamic objects in the scene. On the contrary, the estimated rigid flow can properly handle occluded areas thanks to the minimum-reprojection mechanism [29]. Starting from these considerations, our key idea is to split the scene into stationary and potentially dynamics objects, and apply on them the proper supervision. Purposely, we can leverage several observations:

1. **Semantic priors.** Given a semantic map S_t for image I_t, we can binarize pixels into static M_t^s and potentially dynamic M_t^d, with $M_t^s \cap M_t^d = \emptyset$. For example, we expect that points labeled as *road* are static in the 3D world, while pixels belonging to the semantic class *car* may move. In M_t^d, we assign 1 for each potentially dynamic pixel, 0 otherwise, as shown in Figure 3(e).

2. **Camera Motion Boundary Mask.** Instead of using a backward-forward strategy [96] to detect boundaries occluded due to the ego-motion, we analytically compute a binary boundary mask M_t^b from depth and ego-motion estimates as proposed in [57]. We assign a 0 value for out-of-camera pixels, 1 otherwise as shown in Figure 3(f).

3. **Consistency Mask.** Because the inconsistencies between the rigid flow and $\text{F}_{t\to s}$ are not only due to dynamic objects but also to occluded/inconsistent areas, we can leverage Equation (4) to detect such critical regions. Indeed, we define the consistency mask as:

$$M_t^c = P_t < \xi, \xi \in [0,1] \tag{6}$$

This mask assigns 1 where the condition is satisfied, 0 otherwise (*i.e.* inconsistent regions) as in Figure 3(g).

Finally, we compute the final mask M, in Figure 3(h), as:

$$M = \min\{\max\{M_t^d, M_t^c\}, M_t^b\} \tag{7}$$

As a consequence, M will effectively distinguish regions in the image for which we can not trust the supervision sourced by $F_{t\to s}$, *i.e.* inconsistent or occluded areas. On such regions, we can leverage our proposed self-distillation mechanism. Then, we define the final total loss for the self-distilled optical flow network as:

$$\mathcal{L} = \sum \alpha_r \phi(SF_{t\to s}, F_{t\to s}^{rigid}) \cdot (1-M) + \alpha_d \phi(SF_{t\to s}, F_{t\to s}) \cdot M + \psi(I_t, \widetilde{I}_t^{SF}) \cdot M \tag{8}$$

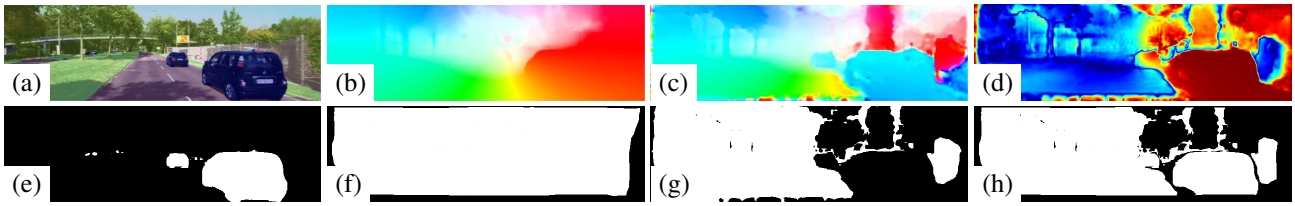

Figure 3. Overview of our semantic-aware and self-distilled optical flow estimation approach. We leverage semantic segmentation S_t (a) together with rigid flow $F^{rigid}_{t\to s}$ (b), teacher flow $F_{t\to s}$ (c) and motion probabilities P_t (d), the warmer the higher. From a) we obtain semantic priors M^d_t (e), combined with boundary mask M^b_t (f) and consistency mask M^c_t (g) derived from (d) as in Eq. 6, in order to obtain the final mask M (h) as in Eq. 7.

where ϕ is a distance function between two motion vectors, while α_r and α_d are two hyper-parameters.

3.3. Motion Segmentation

At test time, from pixel-wise probability P_t computed between $SF_{t\to s}$ and $F^{rigid}_{t\to s}$, semantic prior M^d_t and a threshold τ, we compute a motion segmentation mask by:

$$M^{mot}_t = M^d_t \cdot (P_t > \tau), \tau \in [0, 1] \tag{9}$$

Such mask allows us to detect moving objects in the scene independently of the camera motion. A qualitative example is depicted in Figure 1(f).

4. Architecture and Training Schedule

In this section we present the networks composing ΩNet (highlighted in red in Figure 2), and delineate their training protocol. We set $N = 3$, using 3-frames sequences. The source code is available at https://github.com/CVLAB-Unibo/omeganet.

4.1. Network architectures

We highlight the key traits of each network, referring the reader to the supplementary material for exhaustive details.

Depth and Semantic Network (DSNet). We build a single model, since shared reasoning about the two tasks is beneficial to both [90, 16]. To achieve real-time performance, DSNet is inspired to PydNet [62], with several key modifications due to the different goals. We extract a pyramid of features down to $\frac{1}{32}$ resolution, estimating a first depth map at the bottom. Then, it is upsampled and concatenated with higher level features in order to build a refined depth map. We repeat this procedure up to half resolution, where two estimators predict the final depth map D_t and semantic labels S_t. These are bi-linearly upsampled to full resolution. Each conv layer is followed by batch normalization and ReLU, but the prediction layers, using reflection padding. DSNet counts 1.93M parameters.

Camera Network (CamNet). This network estimates both camera intrinsics and poses between a target I_t and some source views $I_s (1 \leq s \leq 3, s \neq t)$. CamNet differs from previous work by extracting features from I_t and I_s independently with shared encoders. We extract a pyramid of features down to $\frac{1}{16}$ resolution for each image and concatenate them to estimate the 3 Euler angles and the 3D translation for each I_s. As in [30], we also estimate the camera intrinsics. Akin to DSNet, we use batch normalization and ReLU after each layer but for prediction layers. CamNet requires 1.77M parameters for pose estimation and 1.02K for the camera intrinsics.

Optical Flow Network (OFNet). To pursue real-time performance, we deploy a 3-frame PWC-Net [75] network as in [54], which counts 4.79M parameters. Thanks to our novel training protocol leveraging on semantics and self-distillation, our OFNet can outperform other multi-task frameworks [3] built on the same optical flow architecture.

4.2. Training Protocol

Similarly to [88], we employ a two stage learning process to facilitate the network optimisation process. At first, we train DSNet and CamNet simultaneously, then we train OFNet by the self-distillation paradigm described in 3.2. For both stages, we use a batch size of 4 and resize input images to 640×192 for the KITTI dataset (and to 768×384 for pre-training on Cityscapes), optimizing the output of the networks at the highest resolution only. We also report additional experimental results for different input resolutions where specified. We use the Adam optimizer [46] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. As photometric loss ψ, we employ the same function defined in [28]. When training our networks, we apply losses using as I_s both the previous and the next image of our 3-frame sequence. Finally, we set both τ and ξ to be 0.5 in our experiments.

Depth, Pose, Intrinsics and Semantic Segmentation. In order to train DSNet and CamNet we employ sequences of 3 consecutive frames and semantic proxy labels yielded by a state-of-the art architecture [12] trained on Cityscapes with ground-truth labels. We trained DSNet and CamNet for 300K iterations, setting the initial learning rate to 10^{-4}, manually halved after 200K, 250K and 275K steps. We apply data augmentation to images as in [28]. Training takes

$\sim$ 20 hours on a Titan Xp GPU.

Optical Flow. We train OFNet by the procedure presented in 3.2. In particular, we perform 200K training steps with an initial learning rate of 10^{-4}, halved every 50K until convergence. Moreover, we apply strong data augmentation consisting in random horizontal and vertical flip, crops, random time order switch and, peculiarly, time stop, replacing all I_s with I_t to learn a zero motion vector. This configuration requires about 13 hours on a Titan Xp GPU with the standard 640×192 resolution. We use an L1 loss as ϕ. Once obtained a competitive network in non-occluded regions we train a more robust optical flow network, denoted as SD-OFNet, starting from pre-learned weights and the same structure of OFNet by distilling knowledge from OFNet and rigid flow computed by DSNet using the total mask M and 416×128 random crops applied to $F_{t\to s}$, $F^{rigid}_{t\to s}$, M and RGB images. We train SD-OFNet for $15K$ steps only with a learning rate of 2.5×10^{-5} halved after 5K, 7.5K, 10K and 12.5K steps, setting α_r to 0.025 and α_d to 0.2. At test-time, we rely on SD-OFNet only.

5. Experimental results

Using standard benchmark datasets, we present here the experimental validation on the main tasks tackled by ΩNet.

5.1. Datasets.

We conduct experiments on standard benchmarks such as KITTI and Cityscapes. We do not use feature extractors pre-trained on ImageNet or other datasets. For the sake of space, we report further studies in the supplementary material (*e.g.* results on pose estimation or generalization).

KITTI (K) [27] is a collection of 42,382 stereo sequences taken in urban environments from two video cameras and a LiDAR device mounted on the roof of a car. This dataset is widely used for benchmarking geometric understanding tasks such as depth, flow and pose estimation.

Cityscapes (CS) [19] is an outdoor dataset containing stereo pairs taken from a moving vehicle in various weather conditions. This dataset features higher resolution and higher quality images. While sharing similar settings, this dataset contains more dynamics scenes compared to KITTI. It consists of 22,973 stereo pairs with 2048×1024 resolution. 2,975 and 500 images come with fine semantic annotations for training and validation, respectively.

5.2. Monocular Depth Estimation

In this section, we compare our results to other state-of-the-art proposals and assess the contribution of each component to the quality of our monocular depth predictions.

Comparison with state-of-the-art. We compare with state-of-the-art self-supervised networks trained on monocular videos according to the protocol described in [24]. We follow the same pre-processing procedure as [95] to remove static images from the training split while using all the 697 images for testing. LiDAR points provided in [27] are re-projected on the left input image to obtain ground-truth labels for evaluation, up to 80 meters [25]. Since the predicted depth is defined up to a scale factor, we align the scale of our estimates by multiplying them by a scalar that matches the median of the ground-truth, as introduced in [95]. We adopt the standard performance metrics defined in [24]. Table 1 reports extensive comparison with respect to several monocular depth estimation methods. We outperform our main competitors such as [88, 96, 17, 3] that solve multi-task learning or other strategies that exploit additional information during the training/testing phase [9, 83]. Moreover, our best configuration, *i.e.* pre-training on CS and using 1024×320 resolution, achieves better results in 5 out of 7 metrics with respect to the single-task, state-of-the-art proposal [29] (and is the second best and very close to it on the remaining 2) which, however, leverages on a larger ImageNet pre-trained model based on ResNet-18. It is also interesting to note how our proposal without pre-training obtains the best performance in 6 out of 7 measures on 640×192 images (row 1 vs 15). These results validate our intuition about how the use of semantic information can guide geometric reasoning and make a compact network provide state-of-the-art performance even with respect to larger and highly specialized depth-from-mono methods.

Ablation study. Table 2 highlights how progressively adding the key innovations proposed in [30, 29, 77] contributes to strengthen ΩNet, already comparable to other methodologies even in its baseline configuration (first row). Interestingly, a large improvement is achieved by deploying joint depth and semantic learning (rows 5 vs 7), which forces the network to simultaneously reason about geometry and content within the same shared features. By replacing DSNet within ΩNet with a larger backbone [88] (rows 5 vs 6) we obtain worse performance, validating the design decisions behind our compact model. Finally, by pre-training on CS we achieve the best accuracy, which increases alongside with the input resolution (rows 8 to 10).

Depth Range Error Analysis. We dig into our depth evaluation to explain the effectiveness of ΩNet with respect to much larger networks. Table 3 compares, at different depth ranges, our model with more complex ones [29, 88]. This experiment shows how ΩNet superior performance comes from better estimation of large depths: ΩNet outperforms both competitors when we include distances larger than 8 m in the evaluation, while it turns out less effective in the close range.

5.3. Semantic Segmentation

In Table 4, we report the performance of ΩNet on semantic segmentation for the 19 evaluation classes of CS ac-

					Lower is better				Higher is better		
Method	M	A	I	CS	Abs Rel	Sq Rel	RMSE	RMSE log	δ <1.25	$\delta < 1.25^2$	$\delta < 1.25^3$
Godard *et al.* [29]					0.132	1.044	5.142	0.210	0.845	0.948	0.977
Godard *et al.* [29] (1024 × 320)			✓		**0.115**	0.882	4.701	0.190	**0.879**	**0.961**	0.982
Zhou *et al.* [94]			✓		0.121	0.837	4.945	0.197	0.853	0.955	0.982
Mahjourian *et al.* [57]				✓	0.159	1.231	5.912	0.243	0.784	0.923	0.970
Wang *et al.* [77]				✓	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Bian *et al.* [6]				✓	0.128	1.047	5.234	0.208	0.846	0.947	0.970
Yin *et al.* [88]	✓			✓	0.153	1.328	5.737	0.232	0.802	0.934	0.972
Zou *et al.* [96]	✓			✓	0.146	1.182	5.215	0.213	0.818	0.943	0.978
Chen *et al.* [17]	✓		✓		0.135	1.070	5.230	0.210	0.841	0.948	0.980
Luo *et al.* [56]	✓				0.141	1.029	5.350	0.216	0.816	0.941	0.976
Ranjan *et al.* [3]	✓				0.139	1.032	5.199	0.213	0.827	0.943	0.977
Xu *et al.* [83]		✓	✓		0.138	1.016	5.352	0.217	0.823	0.943	0.976
Casser *et al.* [9]		✓			0.141	1.026	5.290	0.215	0.816	0.945	0.979
Gordon *et al.* [30]	✓	✓			0.128	0.959	5.230	-	-	-	-
ΩNet(640 × 192)	✓	✓			0.126	0.835	4.937	0.199	0.844	0.953	0.982
ΩNet(1024 × 320)	✓	✓			0.125	0.805	4.795	0.195	0.849	0.955	0.983
ΩNet(640 × 192)	✓	✓		✓	0.120	0.792	4.750	0.191	0.856	0.958	0.984
ΩNet(1024 × 320)	✓	✓		✓	0.118	**0.748**	**4.608**	**0.186**	0.865	**0.961**	**0.985**

Table 1. Depth evaluation on the Eigen split [24] of KITTI [26]. We indicate additional features of each method. M: multi-task learning, A: additional information (*e.g.* object knowledge, semantic information), I: feature extractors pre-trained on ImageNet [20], CS: network pre-trained on Cityscapes [19].

									Lower is better		Higher is better		
Resolution	Learned Intr. [30]	Norm. [77]	Min. Repr. [29]	Automask [29]	Sem. [12]	Pre-train	Abs Rel	Sq Rel	RMSE	RMSE log	δ <1.25	$\delta < 1.25^2$	$\delta < 1.25^3$
640 × 192	-	-	-	-	-	-	0.139	1.056	5.288	0.215	0.826	0.942	0.976
640 × 192	✓	-	-	-	-	-	0.138	1.014	5.213	0.213	0.829	0.943	0.977
640 × 192	✓	✓	-	-	-	-	0.136	1.008	5.204	0.212	0.832	0.944	0.976
640 × 192	✓	✓	✓	-	-	-	0.132	0.960	5.104	0.206	0.840	0.949	0.979
640 × 192	✓	✓	✓	✓	-	-	0.130	0.909	5.022	0.207	0.842	0.948	0.979
640 × 192 †	✓	✓	✓	✓	-	-	0.134	1.074	5.451	0.213	0.834	0.946	0.977
640 × 192	✓	✓	✓	✓	✓	-	0.126	0.835	4.937	0.199	0.844	0.953	0.980
416 × 128	✓	✓	✓	✓	✓	✓	0.126	0.862	4.963	0.199	0.846	0.952	0.981
640 × 192	✓	✓	✓	✓	✓	✓	0.120	0.792	4.750	0.191	0.856	0.958	0.984
1024 × 320	✓	✓	✓	✓	✓	✓	0.118	0.748	4.608	0.186	0.865	0.961	0.985

Table 2. Ablation study of our depth network on the Eigen split [24] of KITTI. †: our network is replaced by a ResNet50 backbone [88].

Method	Cap (m)	Abs Rel	Sq Rel	RMSE	RMSE log
Godard *et al.* [29]	0-8	**0.059**	**0.062**	0.503	**0.082**
ΩNet†	0-8	0.060	0.063	**0.502**	**0.082**
ΩNet	0-8	0.062	0.065	0.517	0.085
Godard *et al.* [29]	0-50	0.125	0.788	3.946	0.198
ΩNet†	0-50	0.127	0.762	4.020	0.199
ΩNet	0-50	**0.124**	**0.702**	**3.836**	**0.195**
Godard *et al.* [29]	0-80	0.132	1.044	5.142	0.210
ΩNet†	0-80	0.134	1.074	5.451	0.213
ΩNet	0-80	**0.126**	**0.835**	**4.937**	**0.199**

Table 3. Depth errors by varying the range. †: our network is replaced by a ResNet50 backbone [88].

Method	Train	Test	mIoU Class	mIoU Cat.	Pix.Acc.
DABNet [50]	CS(S)	CS	69.62	87.56	94.62
FCHardNet [11]	CS(S)	CS	**76.37**	**89.22**	**95.35**
ΩNet	CS(P)	CS	54.80	82.92	92.50
DABNet [50]	CS(S)	K	35.40	61.49	80.50
FCHardNet [11]	CS(S)	K	**44.74**	68.20	72.07
ΩNet	CS(P)	K	43.80	**74.31**	**88.31**
ΩNet	CS(P) + K(P)	K	46.68	75.84	88.12

Table 4. Semantic segmentation on Cityscapes (CS) and KITTI (K). S: training on ground-truth, P: training on proxy labels.

cording to the metrics defined in [19, 4]. We compare ΩNet against state-of-the art networks for real-time semantic segmentation [11, 50] when training on CS and testing either on the validation set of CS (rows 1-3) or the 200 semantically annotated images of K (rows 4-6). Even though our network is not as effective as the considered methods when training and testing on the same dataset, it shows greater generalization capabilities to unseen domains: it significantly outperforms other methods when testing on K for $\text{mIoU}_{\text{category}}$ and pixel accuracy, and provides similar results to [11] for $\text{mIoU}_{\text{class}}$. We relate this ability to our training protocol based on proxy labels (P) instead of ground truths (S). We validate this hypothesis with thorough ablation studies reported in the supplementary material. Moreover, as we have already effectively distilled the knowledge from DPC [12] during pre-training on CS, there is only a slight benefit in training on both CS and K (with proxy labels only) and testing on K (row 7). Finally, although achieving 46.68 mIoU on fine segmentation, we obtain 89.64 mIoU for the task of segmenting static from potentially dynamic classes, an important result to obtain accurate motion masks.

5.4. Optical Flow

In Table 5, we compare the performance of our optical flow network with competing methods using the KITTI 2015 stereo/flow training set [26] as testing set, which con-

		train			test
Method	Dataset	Noc	All	F1	F1
Meister*et al*. [58] - C	SYN + K	-	8.80	28.94%	29.46%
Meister *et al*. [58] - CSS	SYN + K	-	8.10	23.27%	23.30%
Zou *et al*. [96]	SYN + K	-	8.98	26.0%	25.70%
Ranjan *et al*. [3]	SYN + K	-	5.66	20.93%	25.27%
Wang *et al*. [79] **	K	-	5.58	-	18.00%
Yin *et al*. [88]	K	8.05	10.81	-	-
Chen *et al*. [17] †	K	5.40	8.95	-	-
Chen *et al*. [17] (online) †	K	4.86	8.35	-	-
Ranjan *et al*. [3]	K	-	6.21	26.41%	-
Luo *et al*. [56]	K	-	5.84	-	21.56%
Luo *et al*. [56] *	K	-	5.43	-	20.61%
ΩNet (Ego-motion)	K	11.72	13.50	51.22%	-
OFNet	K	3.48	11.61	25.78%	-
SD-OFNet	K	**3.29**	**5.39**	**20.0%**	**19.47%**

Table 5. Optical flow evaluation on the KITTI 2015 dataset. †: pre-trained on ImageNet, SYN: pre-trained on SYNTHIA [71], *: trained on stereo pairs, **: using stereo at testing time.

Method	Pixel Acc.	Mean Acc.	Mean IoU	f.w. IoU
Yang *et al*. [86] *	0.89	0.75	0.52	0.87
Luo *et al*. [56]	0.88	0.63	0.50	0.86
Luo *et al*. [56] *	0.91	0.76	0.53	0.87
Wang *et al*. [79] (Full) **	0.90	0.82	0.56	0.88
Ranjan *et al*. [3]	0.87	0.79	0.53	0.85
ΩNet	**0.98**	**0.86**	**0.75**	**0.97**
ΩNet (Proxy [12])	0.98	0.87	0.77	0.97

Table 6. Motion segmentation evaluation on the KITTI 2015 dataset. *: trained on stereo pairs, **: using stereo at testing time.

Device	Watt	D	DS	OF	Cam	Ω
Jetson TX2	15	12.5	10.3	6.5	49.2	4.5
i7-7700K	91	5.0	4.2	4.9	31.4	2.4
Titan XP	250	170.2	134.1	94.1	446.7	57.4

Table 7. Runtime analysis on different devices. We report the power consumption in Watt and the FPS. D: Depth, S: Semantic, OF: Optical Flow, Cam: camera pose, Ω: Overall architecture.

tains 200 ground-truth optical flow measurements for evaluation. We exploit all the raw K images for training, but we exclude the images used at testing time as done in [96] , to be consistent with experimental results of previous self-supervised optical flow strategies [88, 96, 17, 3]. From the table, we can observe how our self-distillation strategy allows SD-OFNet to outperform by a large margin competitors trained on K only (rows 5-11), and it even performs better than models pre-initialized by training on synthetic datasets [71]. Moreover, we submitted our flow predictions to the online KITTI flow benchmark after retraining the network including images from the whole official training set. In this configuration, we can observe how our model achieves state-of-the-art $F1$ performances with respect to other monocular multi-task architectures.

5.5. Motion Segmentation

In Table 6 we report experimental results for the motion segmentation task on the KITTI 2015 dataset, which provides 200 images manually annotated with motion labels for the evaluation. We compare our methodology with respect to other state-of-the-art strategies that performs multi-task learning and motion segmentation [3, 56, 79] using the metrics and evaluation protocol proposed in [56]. It can be noticed how our segmentation strategy outperforms all the other existing methodologies by a large margin. This demonstrates the effectiveness of our proposal to jointly combine semantic reasoning and motion probability to obtain much better results. We also report, as upper bound, the accuracy enabled by injecting semantic proxies [12] in place of ΩNet semantic predictions to highlight the low margin between the two.

5.6. Runtime analysis

Finally, we measure the runtime of ΩNet on different hardware devices, *i.e*. a Titan Xp GPU, an embedded NVIDIA Jetson TX2 board and an Intel i7-7700K@4.2 GHz CPU. Timings averaged over 200 frames at 640×192 resolution. Moreover, as each component of ΩNet may be used on its own, we report the runtime for each independent task. As summarized in Table 7, our network runs in real-time on the Titan Xp GPU and at about 2.5 FPS on a standard CPU. It also fits the low-power NIVIDA Jetson TX2, achieving 4.5 FPS to compute all the outputs. Additional experiments are available in the supplementary material.

6. Conclusions

In this paper, we have proposed the first real-time network for comprehensive scene understanding from monocular videos. Our framework reasons jointly about geometry, motion and semantics in order to estimate accurately depth, optical flow, semantic segmentation and motion masks at about 60 FPS on high-end GPU and 5FPS on embedded systems. To address the above multi-task problem we have proposed a novel learning procedure based on distillation of proxy semantic labels and semantic-aware self-distillation of optical-flow information. Thanks to this original paradigm, we have demonstrated state-of-the-art performance on standard benchmark datasets for depth and optical flow estimation as well as for motion segmentation.

As for future research, we find it intriguing to investigate on whether and how would it be possible to self-adapt ΩNet on-line. Although some very recent works have explored this topic for depth-from-mono [9] and optical flow [17], the key issue with our framework would be to conceive a strategy to deal with semantics.

Acknowledgement. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

[1] Filippo Aleotti, Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning end-to-end scene flow by distilling single

tasks knowledge. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020. 3

[2] Lorenzo Andraghetti, Panteleimon Myriokefalitakis, Pier Luigi Dovesi, Belen Luque, Matteo Poggi, Alessandro Pieropan, and Stefano Mattoccia. Enhancing self-supervised monocular depth estimation with traditional visual odometry. In *7th International Conference on 3D Vision (3DV)*, 2019. 2

[3] Ranjan Anurag, Varun Jampani, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 5, 6, 7, 8

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2, 7

[5] Min Bai, Wenjie Luo, Kaustav Kundu, and Raquel Urtasun. Exploiting semantic information and deep matching for optical flow. In *European Conference on Computer Vision*, pages 154–170. Springer, 2016. 3

[6] Jia-Wang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Thirty-third Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 2, 7

[7] Michael J Black and Paul Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding*, 63(1):75–104, 1996. 2

[8] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European conference on computer vision*, pages 25–36. Springer, 2004. 2

[9] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019. 2, 6, 7, 8

[10] Ya-Liang Chang, Zhe Yu Liu, and Winston Hsu. Vornet: Spatio-temporally consistent video inpainting for object removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[11] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. Hardnet: A low memory traffic network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3552–3561, 2019. 2, 7

[12] Liang-Chieh Chen, Maxwell Collins, Yukun Zhu, George Papandreou, Barret Zoph, Florian Schroff, Hartwig Adam, and Jon Shlens. Searching for efficient multi-scale architectures for dense image prediction. In *Advances in Neural Information Processing Systems*, pages 8699–8710, 2018. 2, 5, 7, 8

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[14] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[15] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[16] Po-Yi Chen, Alexander H. Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3, 5

[17] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *ICCV*, 2019. 2, 6, 7, 8

[18] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 2

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6, 7

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[21] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 2

[22] Pier Luigi Dovesi, Matteo Poggi, Lorenzo Andraghetti, Miquel Martí, Hedvig Kjellström, Alessandro Pieropan, and Stefano Mattoccia. Real-time semantic stereo matching. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2020. 3

[23] David Eigen and Rob 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*, pages 2650–2658, 2015. 3

[24] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 2, 6, 7

[25] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016. 2, 6

[26] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 7

[27] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012. 6

[28] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 2, 3, 5

[29] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 2, 3, 4, 6, 7

[30] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *ICCV*, 2019. 2, 5, 6, 7

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 2

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[33] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2

[34] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[35] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[36] Junhwa Hur and Stefan Roth. Joint optical flow and temporally consistent semantic segmentation. In *European Conference on Computer Vision*, pages 163–177. Springer, 2016. 3

[37] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5754–5763, 2019. 2

[38] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[39] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 3

[40] Joel Janai, Fatma G"uney, Anurag Ranjan, Michael J. Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *European Conference on Computer Vision (ECCV)*, volume Lecture Notes in Computer Science, vol 11220, pages 713–731. Springer, Cham, Sept. 2018. 2

[41] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[42] Huaizu Jiang, Deqing Sun, Varun Jampani, Zhaoyang Lv, Erik Learned-Miller, and Jan Kautz. Sense: A shared encoder network for scene-flow estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 3

[43] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[44] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 53–69, 2018. 3

[45] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weight losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018. 3

[46] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[47] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[48] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[49] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016. 2

[50] Gen Li and Joongkyu Kim. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. In *British Machine Vision Conference*, 2019. 2, 7

[51] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. Dfanet: Deep feature aggregation for real-time semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9522–9531, 2019. 2

[52] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *CVPR*, 2019. 2

[53] Pengpeng Liu, Irwin King, Michael R. Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *AAAI*, 2019. 2

[54] Pengpeng Liu, Michael R. Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *CVPR*, 2019. 2, 4, 5

[55] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[56] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *PAMI*, 2019. 2, 7, 8

[57] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4, 7

[58] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, New Orleans, Louisiana, Feb. 2018. 1, 2, 4, 8

[59] Arsalan Mousavian, Hamed Pirsiavash, and Jana Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 611–619. IEEE, 2016. 3

[60] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016. 2

[61] Valentino Peluso, Antonio Cipolletta, Andrea Calimera, Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Enabling energy-efficient unsupervised monocular depth estimation on ARMv7-based platforms. In *Design, Automation & Test in Europe Conference & Exhibition, DATE 2019, Florence, Italy, March 25-29, 2019*, pages 1703–1708, 2019. 2

[62] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. Towards real-time unsupervised monocular depth estimation on CPU. In *IEEE/JRS Conference on Intelligent Robots and Systems (IROS)*, 2018. 2, 5

[63] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[64] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *6th International Conference on 3D Vision (3DV)*, 2018. 2

[65] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[66] Hazem Rashed, Senthil Yogamani, Ahmad El-Sallab, Pavel Krizek, and Mohamed El-Helw. Optical flow augmented semantic segmentation networks for automated driving. *arXiv preprint arXiv:1901.07355*, 2019. 3

[67] Zhile Ren, Orazio Gallo, Deqing Sun, Ming-Hsuan Yang, Erik Sudderth, and Jan Kautz. A fusion approach for multi-frame optical flow estimation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2077–2086. IEEE, 2019. 2

[68] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Bin, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017. 2

[69] J. Revaud, P Weinzaepfel, Z. Harchaoui, and C. Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *CVPR*, 2019. 2

[70] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[71] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 8

[72] Laura Sevilla-Lara, Deqing Sun, Varun Jampani, and Michael J Black. Optical flow with semantic segmentation and localized layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3889–3898, 2016. 3

[73] Wei-Shi Zheng Shuosen Guan, Haoxin Li. Unsupervised learning for optical flow estimation using pyramid convolution lstm. In *Proceedings of IEEE International Conference on Multimedia and Expo(ICME)*, 2019. 2

[74] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 2

[75] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 5

[76] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[77] Chaoyang Wang, Jose Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 6, 7

[78] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2809, 2015. 3

[79] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8071–8081, 2019. 8

[80] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *ICCV*, 2019. 2

[81] Wofk, Diana and Ma, Fangchang and Yang, Tien-Ju and Karaman, Sertac and Sze, Vivienne. FastDepth: Fast Monocular Depth Estimation on Embedded Systems. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019. 2

[82] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE international conference on computer vision*, pages 4705–4713, 2015. 2

[83] Haofei Xu, Jianmin Zheng, Jianfei Cai, and Juyong Zhang. Region deformer networks for unsupervised depth estimation from unconstrained monocular videos. In *IJCAI*, 2019. 2, 6, 7

[84] Nan Yang, Rui Wang, Jörg Stückler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *European Conference on Computer Vision*, pages 835–852. Springer, 2018. 2

[85] Wang Yang, Yi Yang, Zhenheng Yang, Liang Zhao, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[86] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding. In *The European Conference on Computer Vision (ECCV) Workshops*, September 2018. 2, 8

[87] Zhenheng Yang, Peng Wang, Wang Yang, Wei Xu, and Nevatia Ram. Lego: Learning edge with geometry all at once by watching videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[88] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 4, 5, 6, 7, 8

[89] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 325–341, 2018. 2

[90] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantic for semi-supervised monocular depth estimation. In *14th Asian Conference on Computer Vision (ACCV)*, 2018. 1, 3, 4, 5

[91] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[92] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018. 3

[93] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[94] Junsheng Zhou, Yuwang Wang, Naiyan Wang, and Wenjun Zeng. Unsupervised high-resolution depth learning from videos with dual networks. In *Inter. Conf. on Computer Vision*. IEEE, IEEE, October 2019. 2, 7

[95] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 3, 6

[96] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 4, 6, 7, 8