MonoDVPS: A Self-Supervised Monocular Depth Estimation Approach to Depth-aware Video Panoptic Segmentation

Andra Petrovai
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
andra.petrovai@cs.utcluj.ro

Sergiu Nedevschi
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
sergiu.nedevschi@cs.utcluj.ro

Abstract

Depth-aware video panoptic segmentation tackles the inverse projection problem of restoring panoptic 3D point clouds from video sequences, where the 3D points are augmented with semantic classes and temporally consistent instance identifiers. We propose a novel solution with a multi-task network that performs monocular depth estimation and video panoptic segmentation. Since acquiring ground truth labels for both depth and image segmentation has a relatively large cost, we leverage the power of unlabeled video sequences with self-supervised monocular depth estimation and semi-supervised learning from pseudo-labels for video panoptic segmentation. To further improve the depth prediction, we introduce panoptic-guided depth losses and a novel panoptic masking scheme for moving objects to avoid corrupting the training signal. Extensive experiments on the Cityscapes-DVPS and SemKITTI-DVPS datasets demonstrate that our model with the proposed improvements achieves competitive results and fast inference speed.

1. Introduction

Environment perception is a fundamental component of autonomous systems such as automated vehicles. Traditionally, in order to achieve robust perception, multi-modal sensors such as LiDARs and cameras scan the environment and their output is either fused or processed independently by algorithms in order to detect, track and classify the objects in the environment. While specialized sensors such as LiDARs provide precise depth measurements, they have at the same time a high cost and reduced output density. In a multi-modal sensory setup, further challenges have to be addressed, such as sensor synchronization and fusion. Images can be used to infer both semantics and depth, and as a result, perception using cameras only is attractive due to the simple setup and low cost.

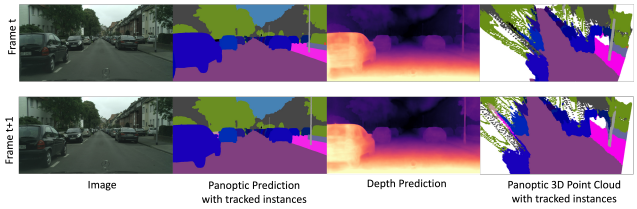

Figure 1: **Depth-aware Video Panoptic Segmentation.** We generate temporally consistent panoptic 3D point clouds from monocular sequences. Our multi-task network predicts video panoptic segmentation with tracked instances (same instance identifier and color in consecutive frames) and monocular depth. Depth is trained in a self-supervised regime, while video panoptic segmentation is trained in a semi-supervised regime, on both human annotated labels and pseudo-labels.

Depth estimation from monocular cameras is a long-lasting research field of computer vision. With the advent of deep learning, it has seen major leaps in performance, especially in the supervised setting. However, large-scale acquisition of depth ground truth has a prohibitively large cost, which led to the emergence of self-supervised monocular depth estimation (SSMDE) methods where ground truth is not employed. The idea behind these approaches is that an image synthesis formulation using 3D reprojection models can be used to jointly learn depth and ego motion. SSMDE is usually less accurate than supervised methods, however the gap can be bridged by leveraging large-scale datasets with unlabeled video sequences.

Panoptic segmentation [17] provides a rich 2D environment representation by performing pixel-level semantic and instance-level segmentation. Video panoptic segmentation [16] extends the task to video and requires temporally consistent instance predictions. To obtain a holistic 3D representation of the environment, depth-aware video panoptic segmentation (DVPS) [22] is introduced as the combination of monocular depth estimation [24] and video panoptic seg-

mentation. ViP-DeepLab [22] proposes a strong baseline for the task, with a network trained in a supervised regime.

With this work, we aim to reduce the ground truth dependency and leverage large amounts of unlabeled video sequences for improved depth-aware video panoptic segmentation (DVPS). Therefore, we propose **MonoDVPS** a novel multi-task network for the DVPS task based on self-supervised monocular depth estimation and semi-supervised video panoptic segmentation. For video panoptic segmentation, we train on both labeled images and pseudo-labels. In the complex multi-task training setting, we aim to improve the performance of all sub-tasks and propose several techniques for this purpose. We investigate loss balancing to increase the accuracy of all sub-tasks and leverage panoptic guidance to reduce the depth error. Since the self-supervised depth estimation relies on the assumption that the scene is static, moving objects corrupt the training signal and introduce high photometric errors. To overcome this problem, we propose a novel moving objects masking based on panoptic segmentation maps from consecutive frames and remove those pixel locations from the photometric loss computation. To further improve the depth prediction, we introduce three loss terms based on the observation that depth discontinuities occur at panoptic edges: panoptic-guided smoothness loss [23] to ensure depth smoothness of neighboring pixels inside panoptic segments, panoptic-guided edge discontinuity loss to enforce large depth difference at panoptic contour and finally we adapt the semantic-guided triplet loss [15] into the panoptic domain. We perform extensive experiments on the Cityscapes-DVPS [22] and SemKITTI-DVPS [22] datasets and demonstrate the effectiveness of our approach.

2. Related Work

Depth-aware Video Panoptic Segmentation. ViP-DeepLab [22] introduces the task as well as the baseline network. ViP-DeepLab processes concatenated image pairs and extends Panoptic DeepLab [6] with a next-frame instance center offset decoder and a monocular depth estimation decoder. The main differences between our network and ViP-DeepLab are related to the depth estimation and tracking sub-tasks. We employ self-supervised depth estimation, which requires no ground truth and is based on geometric projections that allow view-synthesis of adjacent frames. On the other hand, ViP-DeepLab trains the depth in a fully supervised regime with a regression loss. For improved depth results, we adopt multi-scale depth prediction at four scales during training, while ViP-DeepLab uses a single scale. For instance tracking, we use self-supervised optical flow estimation to warp the current panoptic prediction into the next and match the instance IDs based on their mask overlap. On the other hand, ViP-DeepLab tracks instances by predicting pixel-wise offsets to the previous image instance centers, after which it uses the same instance matching algorithm.

Video Panoptic Segmentation. Kim *et al.* has recently introduced the task in [16] along with the baseline VPSNet network. VPSNet is built on top of the proposal-based two-stage panoptic segmentation network UPSNet [31]. In order to improve the current prediction, a pixel-level fusion module gathers features from the previous and next five frames, which are further aligned with optical flow and fused with spatio-temporal attention. Our network, on the other hand, does not employ temporal aggregation and operates in an online fashion, it processes only the current frame, which makes the inference faster. For the tracking functionality, VPSNet employs a MaskTrack head [32], which learns an affinity matrix between RoI proposals, while our network learns optical flow and uses mask overlapping for instance ID propagation. SiamTrack [30] improves VPSNet by designing novel learning objectives that learn segment-wise and dense temporal associations in a contrastive learning framework. In contrast to VPSNet and SiamTrack, our network is box-free and uses the paradigm *segment then group thing pixels into instances*, which makes our network much faster. VPS-Transformer [21] proposes a hybrid architecture derived from Panoptic DeepLab [6] with a focus on both efficiency and performance. A video module based on the Transformer block [27], equipped with attention mechanisms, models spatio-temporal relations between features from consecutive frames for enhanced feature representations. We do not explicitly encode video correlations between frames, but we employ semi-supervised learning with pseudo-labels for improved video panoptic predictions.

Semantically-guided SSMDE. The SFMLearner [35] is the first solution for self-supervised depth estimation which jointly trains an ego motion and a depth estimation network. Recent works tackle various problems that arise from the self-supervised problem formulation. For example, Monodepth2 [12] solves the stationary camera problem by introducing an auto-masking of stationary pixels and the occlusion problem with a minimum reprojection loss. Potentially moving objects break the rigid scene assumption and corrupt the training signal. Semantic and instance information can be used to detect moving objects. Casser *et al.* [3] introduces a 3D object motion network that processes images filtered by instance masks. SGDepth [18] detects frames with moving objects based on semantic segmentation and removes them from the training set. [26] segments the object motion with semantic knowledge distillation. Guizilini *et al.* [13] enhances the feature representation with semantic guidance from a fixed teacher segmentation network using pixel-adaptive convolutions [25]. The tight relation between semantic segments and depth has been exploited in several works by introducing semantic-guided loss functions: semantics-guided smoothness loss [5], which en-

forces depth smoothness for neighboring pixels in a semantic segment, cross-domain discontinuity loss [33], which is based on the assumption that pixels across the semantic edges should have large depth differences. A panoptic-guided smoothness loss leveraging ground truth or panoptic predictions is introduced in [23]. To improve depth alignment to semantic segmentation, the semantics-guided triplet loss [15] with a patch-based sampling technique along semantic edges is proposed. While these networks solely aim to improve the training of a depth estimation network, we train the depth in a multi-task network and we aim to improve all sub-tasks. Compared to semantic edges, panoptic edges, which separate instances but also *stuff* segments, are better aligned to depth edges, where depth discontinuities occur. Therefore, we introduce panoptic-guided loss functions to improve depth training and extend the above semantic-guided losses [5, 33, 15] to the panoptic domain.

3. Method

We propose MonoDVPS, a novel depth-aware video panoptic segmentation network that performs panoptic segmentation, instance tracking and monocular depth estimation. In this section we provide details about the network architecture and training framework, illustrated in Figure 2.

3.1. Video Panoptic Segmentation

Baseline Network. We build our solution on top of the panoptic image segmentation network Panoptic DeepLab [6] which we extend to video. This network predicts semantic segmentation and groups pixels into instances to obtain the panoptic prediction. Panoptic DeepLab has a shared backbone and dual decoders for semantic and instance segmentation. The instance segmentation branch predicts the instance centers and the offsets from each instance pixel to its corresponding center. In this way, the grouping operation can be easily achieved by assigning *thing* pixels to the closest center based on the predicted offsets. Panoptic DeepLab is supervised by the semantic segmentation loss, instance center regression loss and instance offset regression loss. We extend the network with an optical flow decoder [21] for instance tracking, which is simultaneously trained with the rest of the network in a self-supervised regime by minimizing the photometric loss between the current and warped previous frame. Tracking is achieved by matching the current instance predictions with the warped instance masks from the previous frame using the predicted optical flow.

Semi-supervised Training with Pseudo-Labels. We employ the Panoptic DeepLab [6] image panoptic segmentation network with HRNet-W48 [28] to generate pseudo-labels for the unlabeled data on the Cityscapes-DVPS train set. The initial train set provides human annotated labels for every 5th frame in a 30 frame video sequence. Following Naive-Student [4], we use test-time augmentations, such as horizontal flips and multi-scale inputs at scales 0.5:2:0.25 in order to improve the pseudo-labels predictions. For Cityscapes-DVPS images, the ego-car pixels are labeled as void in order to be ignored during training. We also generate dense pseudo-labels for SemKITTI-DVPS images.

3.2. Self-Supervised Monocular Depth Estimation

We extend the semantic decoder with a depth prediction head that has a $[5 \times 5, 64]$ depthwise separable convolution, followed by bilinear interpolation, concatenation with low-level features and $[5 \times 5, 32]$ and $[1 \times 1, 1]$ convolutions. We adopt multi-scale depth prediction and image reconstruction at four scales with output stride 2, 4, 8 and 16 relative to the original image resolution. In practice, the network learns the disparity, the inverse of depth, as it is more robust [12].

The goal of self-supervised monocular depth estimation is to predict the depth map for a single image, while no ground truth is employed in the training phase. The self-supervised depth estimation paradigm is based on geometric projections and is formulated as the minimization of reprojection errors between synthesized adjacent frames and the current frame. During training sequential triplets of frames are required, while during inference a single frame is processed. The mechanism assumes that the scene is static, the camera is moving and all image regions can be reconstructed from neighboring frames. For camera motion, a separate camera pose estimation network is jointly trained.

Let I^t be target frame for which the depth is predicted, and I^s the adjacent frames, where $s = \{t-1, t+1\}$, captured by a moving camera. The camera pose network estimates the ego motion $M_{t\to s}$, that is the 3D translation $T_{t\to s}$ and rotation $R_{t\to s}$ between consecutive 3D positions.

$$M_{t\to s} = \begin{bmatrix} R_{t\to s} & T_{t\to s} \\ 0 & 1 \end{bmatrix} \tag{1}$$

Let K be the intrinsic matrix that defines the focal length and principal point, which are known for a specific dataset. For a pixel p in the target frame we compute its corresponding 3D point x by backprojection using the predicted target depth map D_t. The 3D point is then displaced by the predicted ego motion $M_{t\to s}$ to the source 3D position. Its location in the source frame can be obtained by reprojection:

$$p' = \begin{bmatrix} K|0 \end{bmatrix} M_{t\to s} \begin{bmatrix} D_t(p) K^{-1} p \\ 1 \end{bmatrix} \tag{2}$$

Finally, the target image can be synthesized from the source image by bilinear interpolation [11, 12] $I_{s\to t} = I_s\langle p' \rangle$. During training, the photometric loss between the synthesized images and the target frame is minimized, which is computed as the weighted sum between structural similarity SSIM [29] and L1 loss:

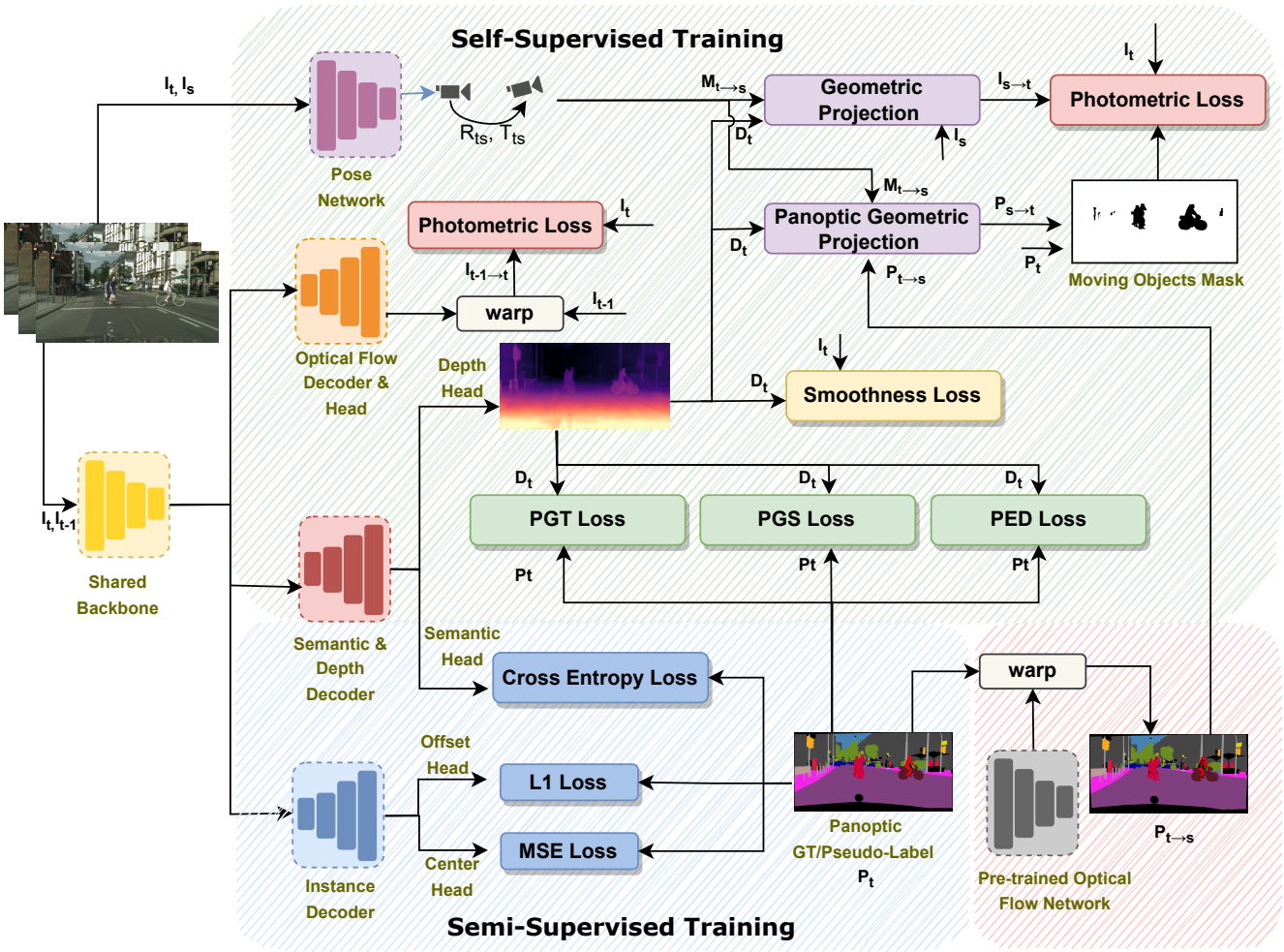

Figure 2: **Our MonoDVPS Depth-aware Video Panoptic Segmentation Network.** We employ a mixed training regime, where the depth, optical flow and ego motion are trained in a self-supervised manner. Panoptic segmentation is semi-supervised with ground truth and pseudo-labels. We introduce several loss functions, panoptic-guided triplet loss (PGT), panoptic-guided smoothness loss (PGS) and panoptic-guided edge discontinuity loss (PED) to improve the depth training. A novel moving objects mask, computed using the panoptic label, is used to mask the photometric loss.

$$pe(I_t, I_{s\to t}) = \frac{\alpha}{2}(1-\text{SSIM}(I_t, I_{s\to t}))+(1-\alpha)\left\|I_t - I_{s\to t}\right\|_1 \quad (3)$$

In practice, we adopt the two photometric loss masking schemes from [12]. For occlusions we implement the minimum reprojection loss from all source images. To account for the case when the camera is stationary, that can be manifest as 'holes' of infinite depth in the predicted depth map, we filter out pixels where the reprojection error of the synthesized image $I_{s\to t}$ is lower than the original image I_s [12]. The photometric loss is computed as the average of the photometric losses at four scales. The predicted lower resolution depth maps at each scale are first upsampled to scale $1/2$ from the original resolution and then are used for reprojection. Based on the assumption that depth discontinuities occur at image edges, we adopt an edge-aware smoothness loss $\mathcal{L}_{sm}$ [12] that encourages adjacent pixels to have similar depth values unless an image edge is present:

$$\mathcal{L}_{sm} = |\partial_x \bar{d}_t| e^{-|\partial_x I_t|} + |\partial_y \bar{d}_t| e^{-|\partial_y I_t|} \quad (4)$$

where $\bar{d}_t$ is the mean normalized inverse depth.

3.3. Improving Depth with Panoptic Guidance

We propose two main mechanisms to improve the performance of the depth estimation by panoptic guidance. First, we start from the observation that the panoptic segmentation has a strong correlation with the depth map and introduce three panoptic guided losses. Second, we generate motion

masks using consecutive panoptic labels that are applied to the photometric loss. Figure 3 presents visual results of the proposed panoptic-guided mechanisms.

Panoptic-guided Smoothness Loss. We introduce a smoothness loss term [23] that enforces similar depth values for adjacent pixels inside a panoptic segment. This loss is derived from $\mathcal{L}_{sm}$, which assumes depth smoothness in the presence of low image gradient. On the other hand, we observe that there is a stronger alignment between depth edges and panoptic contours. To this end, we introduce the following loss:

$$\mathcal{L}_{pgs} = |\partial_x \bar{d}_t|(1 - \partial_x P_t) + |\partial_y \bar{d}_t|(1 - \partial_y P_t) \quad (5)$$

where P_t represents the panoptic ground truth label, ∂P_t are the panoptic contours and $\bar{d}_t$ is the mean normalized inverse depth. For two adjacent pixels (p_0, p_1), we define the $\partial_x P_t(p_0, p_1)$ as the Iverson bracket:

$$\partial_x P_t(p_0, p_1) = [P(p_0) \neq P(p_1)] \quad (6)$$

Panoptic-guided Edge Discontinuity Loss. Based on the observation that adjacent pixels across the panoptic edges may have large depth discontinuities, we design the following panoptic-guided edge discontinuity term:

$$\mathcal{L}_{ped} = \partial_x P_t e^{-|\partial_x \bar{d}_t|} + \partial_y P_t e^{-|\partial_y \bar{d}_t|} \quad (7)$$

This loss enforces a gradient peak in the disparity map at panoptic edges, when we have different panoptic identifiers for adjacent pixels. It represents an extension of the cross-domain discontinuity loss [33] from the semantic to the panoptic domain and is also similar to the panoptic-guided alignment loss from [23].

Panoptic-guided Triplet Loss. We extend the semantic-guided triplet loss [15] to the panoptic domain. The idea behind this loss is that pixels across the panoptic contours should have a large depth difference. The problem with the original formulation that uses semantic contours [15] is that instances with the same semantic class belong to one segment and the edges between instances are missing. On the other hand, instance edges are present in the panoptic map and panoptic edges are better aligned to the depth edges. The triplet loss is defined as follows. The panoptic segmentation map is divided into 5×5 patches and those that do not intersect the panoptic contours are discarded. For the remaining patches a triplet loss is defined. This loss is applied in the feature representation space on the normalized depth feature maps at four scales before the last $[1 \times 1, 1]$ convolution. Features in each patch are grouped in three classes: anchor, positive P_i^+ and negative P_i^-. The anchor is located at the center of the patch, while positive features are the ones that have the same panoptic class with the anchor and the negative features are the ones with different

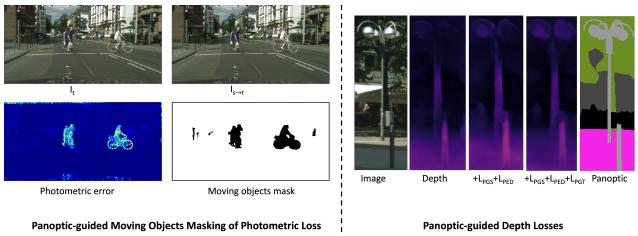

Figure 3: **Moving Objects Masking and Panoptic-Guided Losses.** On the left we illustrate the high photometric loss for moving objects which corrupts the training signal and the moving objects mask. On the right, panoptic-guided depth losses improve the depth prediction.

panoptic class. The triplet loss increases the L2 distance d_i^- between the anchor and the negative features and reduces the L2 distance d_i^+ to the positive features inside a patch. The triplet loss with a margin m is adopted:

$$\mathcal{L}_{pgt} = \max(0, d_i^+ + m - d_i^-) \quad (8)$$

Panoptic-guided Motion Masking. In self-supervised depth estimation the scene is assumed to be static and only the ego motion is modeled. Because the object motion is not taken into consideration, moving objects corrupt the training signal with false high photometric loss. In order to solve this issue, we propose a novel scheme to detect moving objects based on the panoptic labels of consecutive frames, where instance identifiers are temporally consistent. Our goal is to define a moving object mask, which contains 0 where a potentially moving object is present in the target frame I_t or the geometrically warped source frames $I_{s\to t}$, and 1 otherwise. In order to compute the moving object mask we employ the panoptic segmentation pseudo ground truth for the target frame. Since our panoptic pseudo-labels are not temporally consistent, we synthesize panoptic labels for the adjacent source frames from the target panoptic label, in order to ensure that an instance has the same identifier across frames. To achieve this, we employ an external pre-trained optical flow network [34] to warp the target panoptic map P_t to source $\hat{P}_{t\to s}$. The advantage of using optical flow is that it can model both ego and object motion. An occlusion mask $O_{t\to s} = [\exp(-|I_s - \hat{I}_{t\to s}|) > r]$ is designed to remove occluded pixels, where $[\cdot]$ is the Iverson bracket. Then we employ the predicted depth and the geometric projection model from equation 2 to reconstruct the target panoptic map $P_{s\to t}$ using nearest neighbor interpolation. The reconstructed panoptic map has the following formulation:

$$P_{s\to t} = (O_{t\to s}\hat{P}_{t\to s})\langle p' \rangle \quad (9)$$

where p' is the location in the source frame of pixel p in the target frame.

Next, we measure the consistency between the reconstructed $P_{s\to t}$ and the true P_t target panoptic map filtered by the instance masks which correspond to potentially moving object classes. Since the geometric projection model accounts only for ego-motion, we assume a high level of consistency between $P_{s\to t}$ and P_t for a static scene and reduced consistency for moving objects. We measure the consistency as the intersection over union (IoU) between instance masks having the same panoptic identifier in $P_{s\to t}$ and P_t. We define a threshold T for the IoU, such that if the IoU is lower than T, then that instance is considered as a moving object. Pixel locations which correspond to moving objects are excluded from the photometric loss computation. In practice, we obtain the best results with a linear scheduling for threshold T. Instead of a fixed value, we set an initial threshold $T = 0.7$, which linearly decreases with each iteration. The intuition behind this is that, at the beginning we want the network to learn from static pixels, but as the training progresses we allow more noisy samples to account for potential warping errors.

3.4. Panoptic 3D Point Cloud

The depth outputs are up to scale and differ from real-world depth values by a scale factor. Also, each depth map requires a different scale factor, as the depth maps are not inter-frame scale consistent. To recover the true depth, common practice [12] is to perform per-image median scaling: each predicted depth map is scaled with the ratio between the median of the ground truth and the depth prediction. After scaling the depth maps to real-world values, we generate the panoptic 3D point cloud. To obtain the 3D point in the camera coordinate system for each pixel in the image, we backproject the depth map. Since we also have a panoptic output aligned with the depth map, we augment each 3D point with the panoptic identifier of its corresponding pixel, and finally obtain the panoptic 3D point cloud. To completely remove the ground truth dependency, the scale factor can be directly computed from the predicted depth map as the ratio between a known camera height and a computed camera height. The camera height can be computed as the median or average height of all 3D points labeled as road. We adopt the ground truth median scaling for simplicity.

3.5. Implementation Details

During training we optimize nine loss functions. The simple approach of adding up the loss terms results is not optimal, so we balance each loss term with a weighting factor in order to control its importance in the final objective:

$$\begin{aligned}\mathcal{L}_{total} = \gamma_{depth}\mathcal{L}_{depth} + \gamma_{sem}\mathcal{L}_{sem} \\ + \gamma_{instance}\mathcal{L}_{instance} + \gamma_{optical}\mathcal{L}_{optical}\end{aligned} \tag{10}$$

We define the depth loss as a combination of the photometric loss $\mathcal{L}_{photo}$, smoothness loss $\mathcal{L}_{sm}$, panoptic-guided smoothness loss $\mathcal{L}_{pgs}$, panoptic-guided edge discontinuity loss $\mathcal{L}_{ped}$ and panoptic-guided triplet loss $\mathcal{L}_{pgt}$.

$$\begin{aligned}\mathcal{L}_{depth} = \gamma_{photo}\mathcal{L}_{photo} + \gamma_{sm}\mathcal{L}_{sm} + \gamma_{pgs}\mathcal{L}_{pgs} \\ + \gamma_{ped}\mathcal{L}_{ped} + \gamma_{pgt}\mathcal{L}_{pgt}\end{aligned} \tag{11}$$

Following [6], we define $\mathcal{L}_{instance}$ as the weighted sum between mean squared error (MSE) for instance center prediction head and L1 loss for center offset head. Instance weights are the same as in [6]. We set $\gamma_{sem} = 1$, $\gamma_{depth} = 50$, $\gamma_{instance} = 1$, $\gamma_{optical} = 10$, $\gamma_{photo} = 1$, $\gamma_{sm} = 0.001$, $\gamma_{pgs} = 0.01$, $\gamma_{ped} = 0.0001$, $\gamma_{pgt} = 0.1$. The weights have been set such that the main losses have a similar magnitude. Details about network training can be found in the supplementary material in Section B.

4. Experiments

4.1. Experimental Setup

Datasets. Cityscapes-DVPS [22] is an urban driving dataset that extends Cityscapes [7] to video by providing temporally consistent panoptic annotations to every 5th frame in a 30-frame video snippet. The training, validation, test sets have 2,400, 300 and 300 frames. We extend the training set by generating pseudo-labels for every frame in the video sequence that does not have human annotation. The extended training set contains 14,100 images, 11,700 panoptic pseudo-labels and 2,400 panoptic labels. **SemKITTI-DVPS** [22] is based on the odometry split of the KITTI dataset [10, 2] and provides annotations for every frame in the sequence. The sparse panoptic annotations are obtained by projecting 3D point clouds acquired by LiDAR and augmented with semantic and temporally-consistent instance information to the image plane. The dataset contains 19,020 training, 4,071 validation and 4,342 test images. Annotations are provided for 8 *things* classes and 11 *stuff* classes for both datasets.

Evaluation Metrics. We adopt standard evaluation metrics: Panoptic Quality (PQ) for panoptic segmentation [17], Video Panoptic Quality (VPQ) for video panoptic segmentation [16], Depth-aware Video Panoptic Quality (DVPQ) [22] for both depth and video panoptic. For depth estimation we evaluate using absolute relative error (absRel), squared relative error (sqRel) and root mean squared error (RMS) [9]. We measure the inference time of the network with all the post-processing steps, with batch size of one on a NVIDIA Tesla V100 GPU.

4.2. Cityscapes-DVPS Results

In this section, we provide ablation studies and comparison to state-of-the-art on the Cityscapes-DVPS dataset.

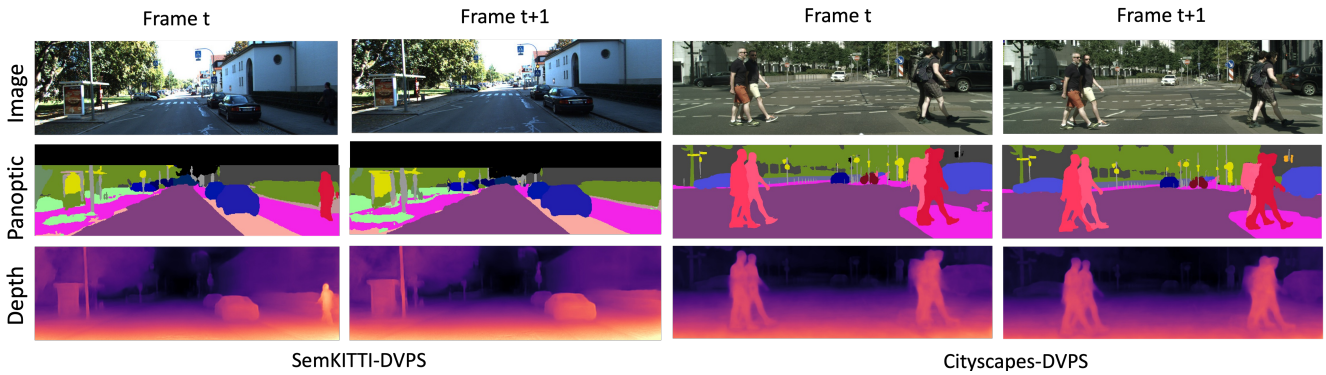

Figure 4: **Qualitative Results.** Video panoptic and depth predictions on SemKITTI-DVPS and Cityscapes-DVPS.

Model	PQ ↑	absRel ↓
Panoptic only	63.5	-
Depth only	-	0.098
MTL Baseline $\gamma_{depth} = 1$	62.9	0.151
+ Loss Balancing $\gamma_{depth} = 50$	63.6	0.102
+ Loss Balancing $\gamma_{depth} = 100$	63.2	0.102
+ Panoptic-guided depth	63.6	0.098
+ Extended train set	**66.5**	**0.082**

Table 1: **Multi-task Learning (MTL).** Comparison between single task and several multi-task training settings.

Model	absRel ↓	sqRel ↓	RMS ↓
Self-Supervised Depth only	0.098	0.731	4.919
MTL Supervised Depth	0.070	0.368	3.675
MTL Self-Supervised Depth	0.106	0.841	5.270
+ Loss Balancing	0.102	0.767	5.034
+ $\mathcal{L}_{PGS}$ + $\mathcal{L}_{PED}$ + $\mathcal{L}_{PGT}$	0.099	0.747	4.988
+ Moving Objects Masking	0.098	0.701	4.864
+ Extended dataset	**0.082**	**0.515**	**4.198**

Table 2: **Panoptic-guided Depth Evaluation in a Multi-task Setting.** Ablation study for loss balancing, panoptic-guided depth losses and moving objects masking.

Multi-task Learning Ablation. In Table 1, we report the results of single-task baselines and our multi-task network. In the multi-task learning baseline setting, with depth loss weight $\gamma_{depth} = 1$, we report a loss in accuracy for both panoptic segmentation and depth compared to single-task baselines. We set $\gamma_{depth} = 50$, as this yields the best PQ and absRel metrics. To further improve the depth prediction, we adopt panoptic-guided losses during training and reach the single-task depth performance. Self-supervised depth estimation does not use depth ground truth during training, therefore can be trained on large-scale unlabeled datasets. By training on the extended train set with panoptic pseudo-labels, we obtain major improvements for both tasks. We present panoptic and depth visualizations on two consecutive frames in Figure 4.

Panoptic-guided Depth Ablation. In Table 2 we evaluate the depth under different settings. First, we compare the depth estimation trained in a supervised vs self-supervised regime on the train set in a multi-task setting. For supervised training we formulate depth estimation as regression and adopt the scale-invariant log loss from [8]. As expected, supervised depth outperforms self-supervised depth in all metrics. To bridge the gap, we propose several improvements in the self-supervised training process. We balance the multi-task loss by increasing the depth weight and introduce panoptic-guided losses $\mathcal{L}_{PGS}, \mathcal{L}_{PED}, \mathcal{L}_{PGT}$ to reduce the depth error. In order to avoid corrupting the training signal in the moving objects region, we design a moving objects masking scheme that further increases the performance. We finally extend the training set (2,400 to 14,410 frames), which significantly reduces the error, showing that a large dataset is a very important element in self-supervised depth training. The supplementary material contains ablation studies for the panoptic-guided losses in Table 7 and for the moving objects masking in Table 6.

Depth-aware Video Panoptic Segmentation. We evaluate DVPQ on Cityscapes-DVPS in Table 3. As expected, DVPQ decreases for larger window size k, as the temporal consistency is reduced, and for lower threshold λ on depth absRel. We observe a larger performance drop in DVPQ-Things than DVPQ-Stuff when decreasing λ for all k, which suggests that depth errors are larger on instances than on *stuff* pixels. We also train MonoDVPS in a fully supervised regime for depth and video panoptic segmentation (MonoDVPS S-MDE) and obtain higher DVPQ than the multi-task network trained in a self-supervised depth regime (MonoDVPS Average). Compared to MonoDVPS S-MDE, depth errors are higher on instances and lower on *stuff* pixels, as

$DVPQ^k_\lambda$ on Cityscapes-DVPS	k = 1			k = 2			k = 3			k = 4			DVPQ Average		
MonoDVPS $\lambda = 0.50$	65.9	55.7	73.3	59.0	43.0	70.6	55.8	36.9	69.5	53.5	32.5	68.8	58.6	42.0	70.6
MonoDVPS $\lambda = 0.25$	59.3	45.4	69.4	53.0	34.2	66.7	50.2	28.9	65.7	48.5	26.1	64.7	52.8	33.7	66.7
MonoDVPS $\lambda = 0.10$	39.0	23.7	50.0	35.1	17.5	47.9	33.4	14.5	47.1	32.5	13.1	46.6	35.0	17.2	48.0
MonoDVPS Average	54.7	41.6	64.2	49.0	31.6	61.7	46.5	26.8	60.8	44.8	23.9	60.0	**48.8**	**31.0**	**61.7**
MonoDVPS S-MDE	57.2	48.4	63.6	51.0	37.0	61.0	47.9	31.0	60.0	45.7	27.0	59.3	**50.4**	**35.9**	**61.0**
ViP-DeepLab (WR-41) [22]	61.9	55.9	66.3	55.6	44.3	63.8	52.4	38.4	62.6	50.4	34.6	61.9	**55.1**	**43.3**	**63.6**
ViP-DeepLab* (ResNet-50) [1]	47.4	38.8	53.7	44.0	28.1	51.6	39.0	23.3	50.5	37.5	20.2	50.0	**42.0**	**27.6**	**51.5**

$DVPQ^k_\lambda$ on SemKITTI-DVPS	k = 1			k = 5			k = 10			k = 20			DVPQ Average		
MonoDVPS $\lambda = 0.50$	48.7	44.7	51.7	43.0	33.3	50.0	41.6	30.7	49.6	40.4	28.4	49.2	43.4	34.2	50.1
MonoDVPS $\lambda = 0.25$	45.3	39.7	49.4	39.8	28.9	47.8	38.5	26.7	47.2	37.6	25.0	46.8	40.3	30.0	47.8
MonoDVPS $\lambda = 0.10$	35.9	28.0	41.6	31.6	20.0	40.0	30.6	18.4	39.4	29.8	17.3	39.0	32.0	21.0	40.0
MonoDVPS Average	43.3	37.5	47.5	38.1	27.4	46.0	36.9	25.2	45.4	36.0	23.6	45.0	**38.6**	**28.4**	**46.0**
ViP-DeepLab [22]	48.9	42.0	53.9	45.8	36.9	52.3	44.4	34.6	51.6	43.4	33.0	51.1	**45.6**	**36.6**	**52.2**

Table 3: **Depth-aware Video Panoptic Segmentation on Cityscapes-DVPS and SemKITTI-DVPS.** Each cell shows $DVPQ^k_\lambda$| $DVPQ^k_\lambda$-Things | $DVPQ^k_\lambda$-Stuff. k is the number of frames and λ is the threshold of relative depth error. MonoDVPS S-MDE is our network trained in a fully supervised regime for both panoptic and depth. Our networks use the ResNet-50 backbone. ViP-DeepLab uses the heavier WR-41 backbone, Mapillary Vistas pretraining and test-time augmentations. ViP-DeepLab* with the ResNet-50 backbone is evaluated using the author's code and pretrained model.

indicated by the lower DVPQ-Things and higher DVPQ-Stuff. Our network surpasses ViP-DeepLab with ResNet-50 [1] and sets a new state-of-the-art. We present more DVPS results in the supplementary material in Table 5.

Video Panoptic Segmentation. In Table 4, we compare our MonoDVPS results with the state-of-the-art for video panoptic segmentation. ViP-DeepLab [22] with WR-41 [4] backbone and Mapillary Vistas [20] pretraining has been designed for accuracy and achieves state-of-the-art results, however it is slow due to the costly test-time augmentations. When using the same ResNet-50 backbone, our network surpasses all other networks, including ViP-DeepLab.

4.3. SemKITTI-DVPS Results

We evaluate our MonoDVPS network on the SemKITTI-DVPS dataset in Table 3. ViP-DeepLab with WR-41 backbone and test-time augmentations surpasses our results, however our network would also benefit from heavier backbone and these costly operations. Compared to our results on Cityscapes-DVPS, we observe that as the absolute relative depth threshold λ decreases, $DVPQ^\lambda_k$ drops are smaller. For example, on Cityscapes-DVPS the difference $DVPQ^{0.5}_1$ - $DVPQ^{0.25}_1$ is 6.6%, while on SemKITTI-DVPS it is 3.4%. The effect is even more pronounced for smaller λ. A reason why Cityscapes-DVPS is more sensitive to λ could be that the dataset is more complex, with a larger number of instances per image and more difficult scenarios, and our depth has higher errors on instances than on background.

Model	Backbone	k = 1	k = 2	k = 3	k = 4	VPQ ↑	Time (s)
VPSNet [16]	ResNet-50	62.7	56.9	53.3	51.3	56.1	0.77
Siam-Track [30]	ResNet-50	64.6	57.6	54.2	52.7	57.3	0.22
VPS-Transformer [21]	ResNet-50	64.8	57.6	54.4	52.2	57.3	0.11
ViP-DeepLab* [1]	ResNet-50	60.6	53.1	49.9	47.7	52.8	-
ViP-DeepLab [22]	WR-41	**70.4**	**63.6**	**60.1**	**58.1**	**63.1**	54*
MonoDVPS (ours)	ResNet-50	66.5	59.6	56.3	54.0	59.1	0.10

Table 4: **Video Panoptic Segmentation.** k is the window size used for evaluation. In this paper $k = \{1, 2, 3, 4\}$ is equivalent to $k = \{1, 5, 10, 15\}$ from [16, 30, 21]. ViP-DeepLab* is evaluated with the author's code.

5. Conclusions

In this work, we have developed a novel multi-task network for depth-aware video panoptic segmentation with mixed training regimes: self-supervised depth estimation and semi-supervised video panoptic segmentation. By leveraging large amounts of unlabeled images, we improve the performance of both tasks. To further reduce the depth error, we introduce panoptic guidance during training with panoptic-guided losses and a novel panoptic motion masking. The final model achieves competitive performance to the state-of-the-art and offers a good trade-off between inference speed and accuracy.

Acknowledgement

This work was supported by the "DeepPerception - Deep Learning Based 3D Perception for Autonomous Driving" grant funded by Romanian Ministry of Education and Research, code PN-III-P4-PCE-2021-1134.

References

[1] Vip-deeplab. https://github.com/google-research/deeplab2.

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019.

[3] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Unsupervised monocular depth and ego-motion learning with structure and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[4] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *European Conference on Computer Vision*, pages 695–714. Springer, 2020.

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2624–2632, 2019.

[6] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12485, 2020.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[8] 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.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*, 2014.

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[12] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019.

[13] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. In *International Conference on Learning Representations*, 2020.

[14] 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.

[15] Hyunyoung Jung, Eunhyeok Park, and Sungjoo Yoo. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12642–12652, 2021.

[16] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9859–9868, 2020.

[17] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

[18] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *European Conference on Computer Vision*, pages 582–600. Springer, 2020.

[19] 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.

[20] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4990–4999, 2017.

[21] Andra Petrovai and Sergiu Nedevschi. Time-space transformers for video panoptic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 925–934, 2022.

[22] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3997–4008, 2021.

[23] Faraz Saeedan and Stefan Roth. Boosting monocular depth with panoptic segmentation maps. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3853–3862, 2021.

[24] Ashutosh Saxena, Sung Chung, and Andrew Ng. Learning depth from single monocular images. *Advances in neural information processing systems*, 18, 2005.

[25] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11166–11175, 2019.

[26] Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Luigi Di Stefano, and Stefano Mattoccia. Distilled semantics for comprehensive scene understanding from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4654–4665, 2020.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[28] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019.

[29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[30] Sanghyun Woo, Dahun Kim, Joon-Young Lee, and In So Kweon. Learning to associate every segment for video panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2705–2714, 2021.

[31] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019.

[32] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019.

[33] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. In *Asian Conference on Computer Vision*, pages 298–313. Springer, 2018.

[34] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I Chang, Yan Xu, et al. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6278–6287, 2020.

[35] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.