Enhancing Self-supervised Monocular Depth Estimation via Piece-Wise Pose Estimation and Geometric Constraints

Pranjay Shyam
Faurecia IRYStec Inc.
Montreal, Canada
pranjay.shyam.psm@forvia.com

Alexandre Okon
Faurecia IRYStec Inc.
Montreal, Canada
alexandre.okon@forvia.com

HyunJin Yoo
Faurecia IRYStec Inc.
Montreal, Canada
hyunjin.yoo@forvia.com

Abstract

Existing single and multi-frame monocular depth estimation (MDE) approaches lack depth estimation consistency around object edges, while single-frame approaches generate scale-ambiguous depth albeit at a lower computational complexity. We revisit the framework design to address these limitations and propose a joint approach that intertwines depth estimation and panoptic segmentation networks. We present an instance-aware patch-based contrastive loss to ensure depth consistency within an object in feature space. This approach disentangles the embedding triplet and independently refines anchor-positive and anchor-negative pairs, providing coherent depth within objects. Leveraging the panoptic information, we propose masking small objects during photometric loss computation while extracting 6-DoF pose estimates for dynamic objects in a piece-wise approach, thus facilitating depth estimation in dynamic scenes. We demonstrate this mechanism to be suited for single and multi-frame MDE. In addition, to ensure scale fidelity in single-frame MDE, we capitalize on the inherent linear relationship between computed depth and ground truth when using self-supervised photometric loss-based MDE. For this, we propose using a multi-frame depth estimation as a teacher network to inject geometric insight into the student MDE via a global scaling factor, thus generating absolute depth. We further improve the teacher network architecture by introducing a multi-scale feature fusion mechanism that benefits scenarios with significant camera motion. We perform a comprehensive evaluation to validate the efficacy of the proposed mechanism and obtain state-of-the-art performance on the KITTI dataset.

1. Introduction

MDE holds paramount significance across various domains, encompassing autonomous vehicles, mobile robotics, and aerial systems. At the forefront of current advancements, convolutional neural networks (CNNs) have propelled the field, operating within the confines of supervised learning paradigms [1, 12, 45]. These networks learn the intricate mapping between input images and corresponding high-density ground truth. However, generating such ground truth for training and evaluation remains unfeasible due to cost constraints, labor-intensive annotations, and scene dynamics that lead to occlusion errors when aggregating LIDAR-derived point clouds. To circumvent these limitations, self-supervised learning emerges as a cost-effective alternative for training MDE networks, harnessing scene geometry as a guiding principle [21]. This approach intertwines the joint estimation of depth and motion during training and then reconstructs original frames using the derived estimates. The photometric loss comes to the fore, compelling the alignment of reconstructed frames with their originals, constituting the supervisory beacon for network training. Despite its scalability to novel scenes, such self-supervised strategies inadvertently introduce scale ambiguity in-depth estimation [16, 21, 59], curtailing their broader application. Additionally, the adoption of photometric loss presupposes a static world, a premise violated by dynamic objects, consequently destabilizing the training process. These multifaceted challenges beckon innovative solutions to enhance the accuracy and applicability of self-

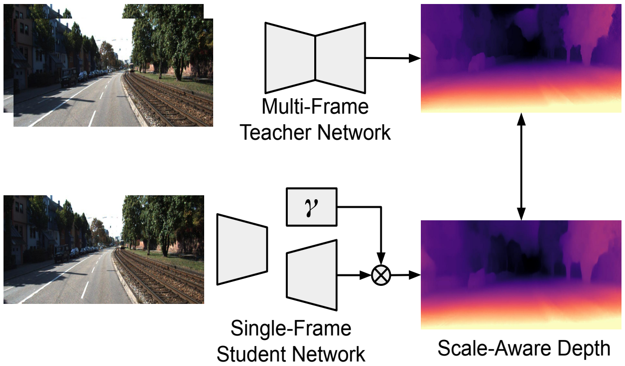

Figure 1. Overview of the proposed mechanism to introduce scale into student depth estimation network using multi-frame teacher network. For simplicity we donot show internal mechanism which is presented in Fig. 4.

supervised MDE frameworks.

Efforts to surmount these inherent limitations have led to many advancements, specifically directed at enhancing the performance of self-supervised MDE algorithms. To ensure the static world assumption holds, several strategies have emerged, encompassing the utilization of semantic segmentation [7,29], auto-masking [16], and optical flow [55] techniques. Moreover, for obtaining absolute depth, a variety of approaches have surfaced, with temporally aligned images being seamlessly integrated into frameworks such as the multi-view geometry paradigm, generating cost-volumes [51], or adopting structure-from-motion (SfM) methodologies [18]. These endeavors have also incorporated supplementary information, including car velocity [18], GPS location [4], and IMU measurements [56]. Despite the collective progress, challenges persist, notably the performance shortfall of MDE in edge-rich regions, coupled with the delivery of scale-ambiguous depth, all within the constraints of a computationally efficient framework. These intricacies underscore the ongoing pursuit of innovative solutions to address these nuanced shortcomings and redefine state-of-the-art MDE techniques.

To ensure edge consistent depth estimation, [29] proposed utilizing semantic segmentation as an auxiliary signal and used patch-based triplet loss to ensure features within each object have similar depth. In contrast, those outside the object have depth differences. However, the efficacy of such a framework faces challenges in scenarios involving occlusion, where the semantic map might need to improve distinguishing between discrete entities, leading to akin depth values for distinct objects. We illustrate instances of such scenarios in the supplementary Appendix-C, underscoring the framework's limitations. Moreover, the direct application of the triplet loss exhibits suboptimal outcomes, as it aims to maintain the distance between anchor-negative (d^-) greater than the distance between anchor-positive (d^+) by a predefined margin (m), i.e., $d^- > d^+ + m$. Notably, this approach overlooks the concurrent objective of minimizing d^+ for consistent depth. To address this, we propose an innovative sampling technique coupled with adjustments to the original triplet loss, enabling independent optimization of the distances. Furthermore, we pivot from semantic segmentation to panoptic for scenarios demanding depth consistency amidst occlusions. This allows us to identify distinct objects, ensuring depth differences in case of occlusion.

When adhering to the static scene assumption, the integrity of pose estimation emerges as a pivotal determinant, crucially impacting the quality of depth estimation outcomes. Notable strides have been made to tackle this issue by prior endeavors such as [3, 34] that diligently sought to rectify this limitation by identifying and estimating the motion of dynamic objects. However, comprehensive motion estimation for all objects proves computationally burdensome and less accurate for smaller objects. While previous approaches often resorted to auxiliary pose estimation networks to determine the pose of objects, we take a novel approach. Leveraging the panoptic segmentation results, we reconfigure the pose estimation network to facilitate piece-wise pose estimation for individual objects within the scene. This involves identifying static objects via Intersection over Union (IoU) comparison between instance masks produced by the panoptic branch. Subsequently, we match temporally adjacent instances based on the highest overlap and perform a piece-wise pose estimation. Finally, these poses are aggregated and combined with depth estimation results to generate an accurate warping while considering dynamic objects. Through this novel methodology, we ensure that dynamic scenes do not impede the training process.

Finally, addressing the critical issue of scale ambiguity, our approach underscores the intrinsic linear relationship between scale-ambiguous estimations in self-supervised MDE and absolute depth. This paves the way for equipping self-supervised MDE with the capability to yield absolute depth predictions. To achieve this, we enact a crucial modification in the depth encoder branch, introducing the prediction of a scale factor (γ). To guide scale prediction, we leverage computationally expensive multi-frame depth estimation network [51] and use this as a distillation mechanism to inject geometric information into the student network. We introduce a multi-scale feature fusion mechanism to improve the quality of depth estimation generated by the teacher network to ensure improved feature representation. This benefits the cost volume and subsequent feature matching, overcoming challenges posed by significant camera motion. We present this distillation pipeline in Fig. 1 and summarize our methodology as,

- To improve the impact of triplet loss, we propose a disentangled version that independently optimizes positive and negative distances.
- To improve pose estimation in dynamic scenes, we leverage the panoptic masks to estimate poses for different objects, which are subsequently assembled for performing the warping.
- To obtain absolute depth using MDE, we propose the integration of a global scale factor estimated via the distillation of depth from a multi-frame teacher network.
- To improve the feature quality for higher feature matching, we propose a multi-scale feature fusion mechanism within the teacher network.

2. Related Works

2.1. Single-Frame Monocular Depth Estimation

MDE is an ill-posed problem as multiple mappings exist from 3D points to pixels within an image. However, su-

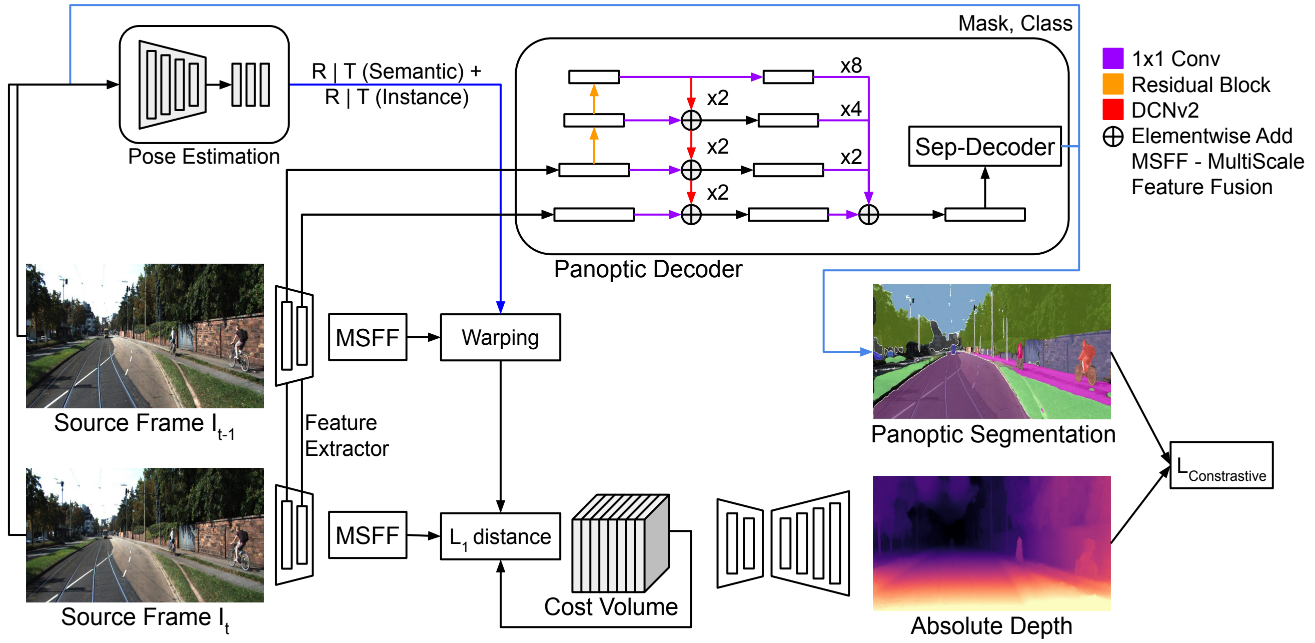

Figure 2. Overview of the proposed framework for performing multi-frame self-supervised monocular depth estimation with panoptic segmentation as an auxiliary task. We further use the panoptic labels to generate piece-wise instance pose estimation.

pervised learning-based approaches [12, 45] overcame this by learning the mapping between an input image and output depth map. However, the requirement of a high-quality depth map for paired training is undesirable for fine-tuning on new domains. To alleviate this, self-supervised approaches are preferred as they combine depth and motion information to be utilized as a guidance signal during optimization.

2.1.1 Architectural Modifications

Initial works [59] proposed an encoder-decoder-based architecture following UNet [46] for extracting depth and ego-motion information from a video. Subsequently, Monodepth2 [17] proposed an encoder using ImageNet [9] pretrained Resnet [22], which was used as a baseline architecture for subsequent works. [39] improved the feature utilization by redesigning the skip connections such that features from multiple scales are fused. To further boost performance, [30] proposed a multi-task framework for joint prediction of depth and semantic segmentation using a shared encoder. To preserve features lost by downsampling and upsampling operations, [18] proposed 3D convolutions-based packing and unpacking operations as their replacement. To further improve performance, attention and transformer-based models [5, 27, 44, 49, 53] are used to correlate features from different regions of an image. However, these methodologies increase the computational cost of the underlying method. To reduce the computational cost, [52] incorporated a CRF-based mechanism to fuse information from multiple scales.

2.1.2 Enforcing Geometric Constraints

While architectural modifications within the underlying depth estimation network demonstrated improved performance, several works focusing on enforcing geometric constraints were also proposed. [25] proposed fusing multi-scale features while [54] utilized virtual normals to estimate 3D scenes robustly. Subsequently, [35] proposed joint prediction of both depth and depth gradients, which are subsequently fused to obtain a refined depth map. In the same line, GeoNet [43] jointly predicts surface normals and depths from a single image to further improve the performance of MDE networks.

2.1.3 Multi-task Architectures

Several works have explored Multi-Task architectures with the motivation of leveraging different scene features to improve performance on the core task. Focusing on improving depth estimation performance, [57] proposed multi-task learning to jointly perform semantic, depth, and surface normal estimation. To ensure feature sharing between these tasks, the authors proposed pattern-affinitive propagation. Sharing the same principle, [26] proposed a geometry-based distillation of semantic features using a pre-trained segmentation network. Recently [29] propose integration of semantic information to improve depth estimation around edged via a semantic segmentation branch and triplet loss.

2.2. Multi-Frame Monocular Depth Estimation

Despite significant advancements made there remains an untapped potential in harnessing information from previous frames during inference, a feature not explored by existing architectures. A notable work, [51], underscored the value of leveraging multiple frames during test-time, casting the problem as a multi-view stereo task. This insight led to the computation of a cost volume using adjacent frames, yielding metric depth estimations. Another notable development, DepthFormer [37], introduced the integration of transformers and grouped self-attention mechanisms to enhance the robustness of outcomes. While our research shares the foundational motivation of [51], our approach surmounts certain limitations associated with dynamic scenes. By capitalizing on panoptic segmentation-based piece-wise pose estimation and its amalgamation, we achieve superior outcomes in the realm of multi-frame self-supervised depth estimation.

3. Methodology

We first elaborate upon the mechanisms that can improve the performance of multi-frame and single-frame MDE approaches, i.e., improving the edge details via panoptic segmentation, improving depth consistency within objects, and handling dynamic objects. Subsequently, we delve into the linear relationship between scale-invariant and absolute depth and how it can be leveraged to ensure scale-aware self-supervised single-frame depth estimation. Finally, we elaborate upon the multi-scale feature fusion mechanism to improve feature aggregation from temporally adjacent frames for computing the cost volume.

3.1. Improving Edge Details

We address the limitations of previous approaches [7,29] that aimed to enhance depth estimation by incorporating semantic segmentation using a shared encoder. Although this strategy can boost depth estimation, it fails to distinguish between occluded objects of the same category. Consequently, depth consistency is maintained across objects, irrespective of their true depth values. We introduce a novel integration of a panoptic segmentation branch to rectify this issue and achieve more accurate edge details while effectively discriminating between occluded objects of identical classes. Incorporating panoptic segmentation in a resource and compute efficient manner, we adopt the innovative You Only Segment Once (YOSO) approach [24], which synergizes panoptic and semantic segmentation by learning a kernel that discriminates unique objects or semantic categories. We present a comprehensive overview of our proposed architecture for multi-frame self-supervised depth estimation, built upon the principles of [51], in Fig. 2. Further elaboration on our framework can be found in Appendix-A whereas we would redirect the readers to [17] and [51] for insights into self-supervised and multi-frame self-supervised depth estimation respectively.

3.2. Revisiting efficacy of Triplet Loss

We reevaluate the motivation behind the triplet loss [29] with the availability of object instances within the scene. Our motivation centers on ensuring the depth estimation network accurately detects edges, which becomes evident through depth discontinuities around object boundaries. Specifically, our observation emphasizes that in occluded scenarios, the inability to distinguish foreground and background pixels effectively obscures boundaries, as the photometric loss equates background pixels with foreground due to shared disparity. We provide a brief overview of the approach presented in [29], which outlines utilizing semantic maps to enforce geometric constraints. This involves partitioning a given semantic label into $K \times K$ patches with a stride of 1. These patches' centers serve as anchors, while same-class features function as positives and others as negatives. The triplet loss is employed to maximize the distance between anchor-positive (d^+) and anchor-negative (d^-) instances, governed by a margin (m). The distances are computed as the mean Euclidean difference of L_2-normalized depth features [29]. Despite its performance improvement, we spotlight two crucial drawbacks: equal weighting of all negative pixels and joint optimization of anchor-positive and anchor-negative samples, leading to sub-optimal results. To overcome these issues, we leverage panoptic masks to introduce a supervised contrastive loss paradigm. Under this, pixels within the mask are classified as positives, while those outside the mask serve as negatives within the same patch. This approach supersedes the triplet loss and employs the supervised contrastive loss [31] using L2 distance $(\cdot)$, denoted as:

$$L_{Constrastive} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp(z_i \cdot z_p / \tau)}{\sum_{n \in N(i)} exp(z_i \cdot z_n / \tau)} \quad (1)$$

Here, $P(i)$ and $N(i)$ refer to indices of positive and negative features, respectively, while z_i, z_p, and z_n represent anchor, positive, and negative features. The temperature parameter τ is introduced to adjust the magnitude of distance computation.

3.3. Piece-wise Pose Estimation

Prior approaches in self-supervised depth estimation have commonly adopted an approach of masking out dynamic objects during training to ensure consistent warping. However, this strategy inadvertently excludes dynamic objects from the optimization process, thereby deviating from the desired goal of accounting for their presence. We undertake a comprehensive rethinking of the global scene pose estimation pipeline to rectify this limitation. Our solution involves a novel proposition: instance-specific pose estimation, which is made feasible by integrating panoptic labels.

By utilizing panoptic segmentation information for two consecutive frames, we initiate a matching process grounded in the mean Intersection over Union (mIoU) metric. Objects with pixel counts below a predefined threshold ($\alpha\%$) of the image resolution) are excluded from consideration, as they typically correspond to distant objects prone to pose estimation errors. The scene's global dynamics encompass both static and dynamic objects. The static elements encapsulate classes categorized as stuff, such as road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, and sky. In contrast, dynamic objects pertain to things categories like person, rider, car, truck, bus, train, motorcycle, and bicycle. As such, the overall scene dynamics can be represented as a fusion of global poses for static objects and instance-wise poses for dynamic elements. This piece-wise approach for capturing global scene translation can be seamlessly integrated into prevailing self-supervised depth estimation models, enabling their application in dynamic scenes without the need for masking. To facilitate this formulation and its self-supervised training, stuff labels are utilized to establish binary masks for objects sharing the same pose. Consequently, in a pair of temporally adjacent frames, a pose estimation network is employed, with the masked static scene as input, to deduce a global pose. Similarly, instance-wise pose estimation is computed employing the masked instance image. With global and instance-wise pose estimations at hand, standard self-supervised practices for forward and backward warping are applied. The comprehensive framework is summarized in Fig. 3.

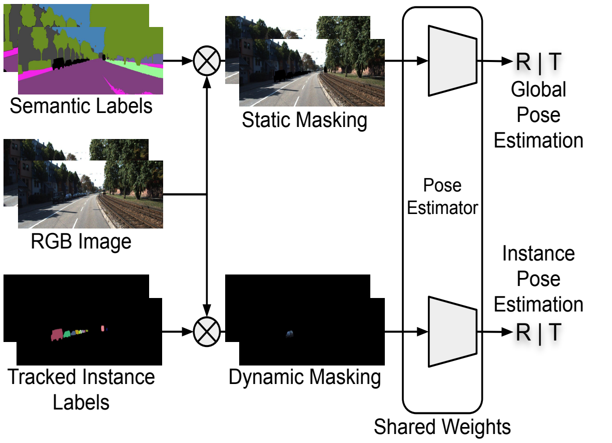

Figure 3. Overview of the proposed piece-wise pose estimation using tracked instance and semantic labels.

3.4. Scale-Aware Single Frame MDE

Single-frame MDE networks offer computational efficiency, yet their prediction of scale-invariant depth poses limitations on their utility. Prior methods, such as [17], attempted to address this limitation by estimating the scale factor through the computation of a median value, aligning the predicted depth with LiDAR-generated ground truth. However, this approach contradicts the essence of self-supervised learning. As an alternative, we propose leveraging the benefits of multi-frame networks to calculate absolute depth. This pseudo-absolute depth can then be harnessed to train a single global scale factor, effectively enabling the conversion of relative depth predictions to absolute depth using a single-frame MDE (MDE) network. This is particularly relevant in the context of monocular videos, where a constant global scale factor can be assumed to provide absolute depth information. In light of this, we embed the computation of depth scaling within the framework of the single-frame MDE architecture. This involves utilizing four 3×3 convolutional layers on encoder-derived features, followed by a global average pooling layer and a sigmoid activation function. A visual overview of the proposed single-frame MDE architecture is illustrated in Figure 4.

We follow the knowledge distillation framework, enforced via L_1 loss, presented in Fig. 1 to leverage absolute depth generated by multi-frame MDE and infuse geometric constraints in form of global scaling factor in single-frame MDE.

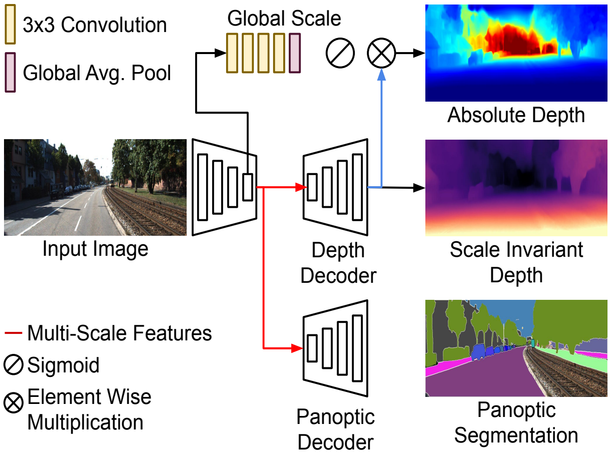

Figure 4. Overview of the proposed framework for performing scale aware single frame self-supervised MDE with panoptic segmentation as an auxiliary task. We use different color maps to highlight scale-invariant (*magma*) and absolute depth (*jet*)

3.5. Multi-Scale Feature Fusion

In the context of utilizing temporally adjacent frames for image projection in feature matching, it becomes evident that the scale of objects can undergo significant changes. This phenomenon arises due to camera motion, introducing variations in object dimensions. Conventional convolutional methods prove inadequate in capturing and representing such intricate scale variations, leading to suboptimal results. To address this challenge, we introduce a novel multi-scale feature fusion mechanism. This mechanism is designed to incorporate a range of multiscale features into the feature representation process, a crucial step utilized for both feature matching and cost volume computation. The proposed fusion approach is depicted in Fig. 5, while more comprehensive insights are provided in Appendix-E of supplementary material.

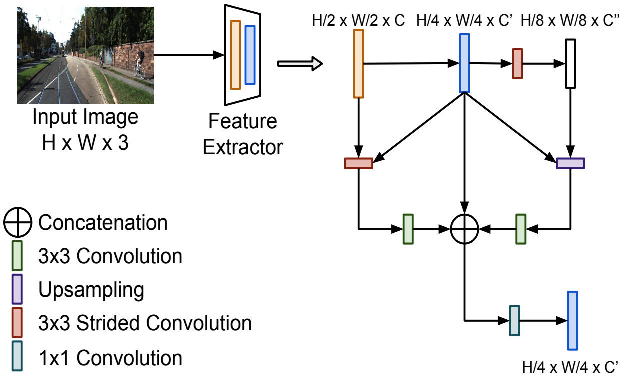

Figure 5. Overview of the proposed multi-scale feature fusion for improving the feature representation quality for performing multi-frame depth estimation.

4. Experimental Evaluation

4.1. Datasets and Evaluation Metrics

We use KITTI [14] dataset and filter the images following [10] as well as removing static frames [59, 60] resulting in 39810 and 4424 training and validation image triplets comprising frames at timestamp $t, t-1, t+1$. To convert predictions by self-supervised methods (except our scale-aware student network) containing scale information, we use median ground truth scaling [16, 59] and evaluate depth up to 80m [10, 13, 15, 16]. For quantitative evaluation we follow prior works to compute mean absolute error (Abs Rel), squared relative error (Sq Rel), root mean squared error (RMSE), log root mean squared error (RMSE log) and accuracy under threshold ($\delta < 1.25^i,\ i = 1, 2, 3$).

4.2. Implementation and Training Details

4.2.1 Pseudo Panoptic Labels

Since we integrate an auxiliary panoptic branch within both single and multi-frame MDE, we require access to training labels.which are not available for KITTI [14] dataset. Hence we use a cityscapes [8] pretrained YOSO [24] network to generate pseudo labels which are then used to train the panoptic segmentation branch in a supervised learning mechanism. Since we jointly predict *stuff* and *things* we use bipartite matching loss [2] for training the panoptic branch with a weight of 0.1 when trained either with multi-frame or single-frame MDE.

For integrating the pose estimation within depth estimation and optimized via a weakly supervised mechanism we use the same loss function as Mask2Former [6] comprising of binary cross entropy loss (L_{ce}), dice loss [40] (L_{dice}), and classification loss (L_{cls}).

$$L_{Pan} = \lambda_{cls} * L_{cls} + \lambda_{dice} * L_{dice} + \lambda_{ce} * L_{ce} \qquad (2)$$

where $\lambda_{cls}, \lambda_{cls}, \lambda_{cls}$ are set to 5.0, 5.0 and 2.0 following [6]. Furthermore when integrated into the depth estimation branch we set the weight of λ_{Pan} to 0.35.

4.2.2 Multi-frame MDE

We first train the multi-frame MDE to ensure generation of absolute depth for distillation to single-frame MDE. We conduct our experiments on a system with 4x 4090 GPU using Pytorch [41] framework and keeping batch size fixed to 12.

Following [17], we use color and flip augmentations and use an input resolution of $640x192$ and use images at instance $t, t-1$ to compute the cost volume. We use ADAM [32] optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) with a learning rate of 0.0001 for 20 epochs and reduce the learning rate by 0.1 for the last 5 epochs. Since our construction of piece-wise pose estimation is consistent with prior global pose estimation we donot modify the aggregate loss function. In our experiments we refer to this network as *ManyDepth+*. We fix the value of α to 1 and include the parameter sweep in Appendix-G of supplementary.

4.2.3 Single-frame MDE

We consider two versions of single-frame MDE based on the final output i.e. scale invariant depth (Ours-Student-I i.e. HRDepth+) and scale-aware (Ours-Student-II i.e. HRDepth++). We follow the approach of monodepth2 [17] to obtain the scaling factor for *Ours-Student-I* model. We train the models for 20 Epochs with an initial learning rate of 0.0001 at a resolution of 640 × 192 consistent with Monodepth2 [17]. We optimize the network using Adam [32] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For enforcing knowledge distillation to learn the global scaling factor, we use L_1 loss with a weight of 0.5. It should be noted that the introduction of knowledge distillation results in increase of training time from 7 to 12 hours on the aforementioned system configuration.

4.3. Comparison with SoTA

We summarize the qualitative results of different SoTA in Tab. 1 and Fig. 6 for an input resolution of 640 × 192. The quantitative results for higher resolution i.e 1024 × 320 and 960 × 320 for single-frame and multi-frame MDE are included in Appendix-H of supplementary. Based on the quantitative results we can conclude our framework to achieve SoTA performance on all metrics at various input resolutions. We emphasize that irrespective of the backbone network (ResNet-18 (R-18) [22], HRNet-18 [50] or Swin-Transformer (SwinL-w7-22k) [38]), our approach surpasses the performance of prior works.

Method	Test Frame	Backbone	Semantic	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SfMLeaner [59]	1	R-18		0.183	1.595	6.709	0.270	0.734	0.902	0.959
Monodepth2 [17]	1	R-18		0.114	0.864	4.817	0.192	0.875	0.959	0.981
Guizilini et al. [19]	1	R-18	✓	0.117	0.854	4.714	0.191	0.873	0.963	0.981
SGDepth [33]	1	R-18	✓	0.113	0.835	4.693	0.191	0.879	0.961	0.981
SAFENet [7]	1	R-18	✓	0.112	0.788	4.582	0.187	0.878	0.963	0.983
Mono-Uncertainty [42]	1	R-18		0.111	0.863	4.756	0.188	0.881	0.961	0.982
PackNet-SfM [18]	1	PackNet		0.111	0.785	4.601	0.189	0.878	0.960	0.982
HRDepth [39]	1	R-18		0.109	0.792	4.632	0.185	0.884	0.962	0.983
FSRE-Depth [29]	1	R-18	✓	0.105	0.722	4.547	0.182	0.886	0.964	0.984
Insta-DM [34]	1	R-18	✓	0.112	0.777	4.772	0.191	0.872	0.959	0.982
DiffNet [58]	1	HRNet-18		0.102	0.764	4.483	0.180	0.896	0.965	0.983
RA-Depth [23]	1	HRNet-18		0.096	0.632	4.216	0.171	0.903	0.968	0.985
Monodepth2 [17]	1	R-50		0.110	0.831	4.642	0.187	0.883	0.962	0.982
FeatDepth [48]	1	R-50		0.104	0.729	4.481	0.179	0.893	0.965	0.984
Guizilini et al. [19]	1	R-50	✓	0.113	0.831	4.663	0.189	0.878	0.971	0.983
Li et al. [36]	1	R-50	✓	0.103	0.709	4.471	0.180	0.892	0.966	0.984
SGDepth [33]	1	R-50	✓	0.112	0.833	4.688	0.190	0.884	0.961	0.981
Johnston et. al. [28]	1	R-101		0.106	0.861	4.699	0.185	0.889	0.962	0.982
ManyDepth [51]	2 (-1, 0)	R-18		0.098	0.770	4.459	0.176	0.900	0.965	0.983
TC-Depth [47]	3 (-1, 0, +1)	R-18		0.103	0.746	4.483	0.180	0.894	0.965	0.983
DepthFormer [37]	2 (-1, 0)	SwinL-w7-22k		0.090	0.661	4.149	0.175	0.905	0.967	0.984
Dynamicdepth [11]	2 (-1, 0)		✓	0.096	0.720	4.458	0.175	0.897	0.964	0.984
DRAFT [20]	2 (-1, 0)	-		0.097	0.647	3.991	0.169	0.899	0.968	0.984
Ours-Student-I	1	HRNet-18	✓	0.093	0.667	4.287	0.172	0.907	0.966	0.985
Ours-Student-II	1	HRNet-18	✓	0.090	0.658	4.221	0.171	0.911	0.967	0.986
ManyDepth+	2 (-1, 0)	HRNet-18	✓	0.086	0.701	4.158	0.168	0.919	0.969	0.986

Table 1. Qualitative results of SoTA on KITTI-2015 Eigen Split trained using monocular videos with and without additional segmentation priors for an input resolution of 640 × 192. For Abs Rel, Sq Rel, RMSE, and RMSE log, lower is better, whereas $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$ highlights metrics where higher is better.

4.3.1 Single-Frame MDE

We first compare the performance with single-frame SoTA algorithms wherein we observe (Ours-Student-I) to improve the performance of baseline DiffNet without making architectural modifications such as RA-Depth. Unlike RA-Depth that utilized a specially designed HRDecoder to better leverage multi-scale features, we proposed integration of a panoptic segmentation head based on class and instance specific kernel prediction. From ablation, we demonstrate this to provide better results in depth estimation. We further highlight a significant performance boost when compared to prior work Insta-DM [34] that leverages similar concept of instance-wise pose estimation using prior access to instance labels. We highlight the prerequisites of Insta-DM to make it a two stage network, whereas the proposed approach predicts panoptic labels simultaneously with depth that enables refinement of both panoptic labels based on depth features and vice-versa using contrastive loss. We highlight this to be the primary reason for significant performance boost. Another direction for performing self supervised depth estimation while overcoming the restrictions imposed by static-scene assumption is to perform optical flow and scene flow [20]. However such approaches are computationally expensive during both training and inference. Furthermore from the qualitative results we demonstrate that we are able to achieve superior performance for both the student and teacher networks.

We also compare the performance of the Student-I network with prior works that leveraged additional semantic information for improving the performance of depth estimation such as Guizilini et al. [19], SGDepth [33], SafeNet [7], FSRE-Depth [29], Insta-DM [34], Li. et al. [36]. Herein we observe a significant performance boost across models which we attribute from ablation to arise from better ability to distinguish different occluded objects and assigning them different depths as should be the case. From qualitative results in Appendix-C we can also conclude that using segmentation branch results in poor occlusion handling wherein the object boundaries are thicker.

4.3.2 Multi-Frame MDE

In case of multi-frame MDE, we highlight the proposed teacher network to achieve SoTA performance. While the original manydepth [51] was performace bound around edges and in dynamic scenes this was subsequently addressed by Dynamicdepth [11] and DRAFT [20]. Dynamicdepth proposed construction of occlusion aware cost-volume to overcome this issue whereas DRAFT performs optical and scene flow to overcome these limitations. However in this work we overcome these limitations via construction of instance wise pose estimation allowing to model different dynamic objects thus aiding in image reconstruction which in turn ensures accurate depth estimation.

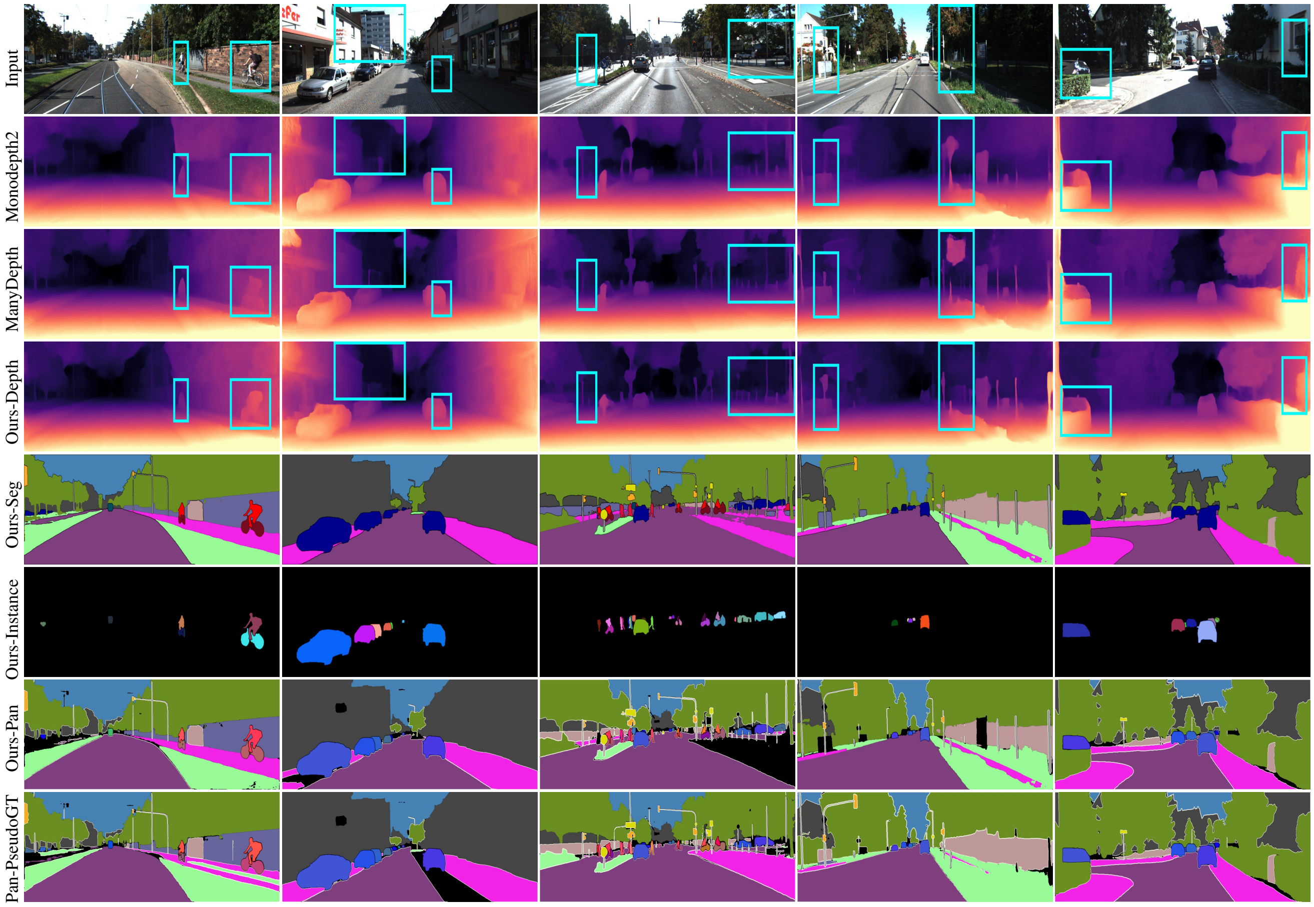

Figure 6. Qualitative comparison of SoTA self-supervised MDE algorithms alongside proposed Ours-Student-I. We highlight performance around object boundaries by cyan colored boxes to demonstrate superior estimation performance. Additional qualitative results are presented in Appendix-F of supplementary.

5. Conclusion

In this work we tackle the inherent challenges associated with self-supervised MDE (MDE) algorithms, offering a viable alternative to supervised methods that rely on high-density ground truth data. Existing single and multi-frame MDE approaches exhibit certain limitations, including imprecise depth estimations around object boundaries and scale-ambiguous depth in single-frame solutions, albeit with lower computational demands. To surmount these challenges, we propose a novel joint framework that intertwines depth estimation and panoptic segmentation networks. Leveraging an instance-aware patch-based contrastive loss, we ensure coherent depth representations within individual objects, effectively disentangling embedding triplets to independently refine anchor-positive and anchor-negative pairs. Capitalizing on panoptic information, we mask small objects during photometric loss computation and facilitate the extraction of 6-DoF pose estimates for dynamic objects in a piece-wise manner, thus enhancing depth estimation in scenes with motion. Notably, our approach caters to both single and multi-frame MDE paradigms. Additionally, to ensure scale consistency in single-frame MDE, we leverage the intrinsic linear relationship between computed depth and ground truth by introducing self-supervised photometric loss-based MDE. We propose using a multi-frame depth estimation network as a teacher to impart geometric understanding to the student MDE via a global scaling factor, ultimately achieving absolute depth prediction. Furthermore, our enhancements extend to the teacher network's architecture through the incorporation of a multi-scale feature fusion mechanism, a solution particularly beneficial in scenarios characterized by substantial camera motion. Rigorous evaluation substantiates the efficacy of our proposed mechanisms, culminating in state-of-the-art performance on the KITTI dataset. This work not only pushes the frontiers of self-supervised MDE but also addresses crucial aspects like scale fidelity and dynamic scene adaptability, thereby contributing significantly to the advancement of computer vision in challenging real-world scenarios.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 6

[3] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8001–8008, 2019. 2

[4] Hemang Chawla, Arnav Varma, Elahe Arani, and Bahram Zonooz. Multimodal scale consistency and awareness for monocular self-supervised depth estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5140–5146. IEEE, 2021. 2

[5] Yuru Chen, Haitao Zhao, Zhengwei Hu, and Jingchao Peng. Attention-based context aggregation network for monocular depth estimation. *International Journal of Machine Learning and Cybernetics*, 12(6):1583–1596, 2021. 3

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 6

[7] Jaehoon Choi, Dongki Jung, Donghwan Lee, and Changick Kim. Safenet: Self-supervised monocular depth estimation with semantic-aware feature extraction. *arXiv preprint arXiv:2010.02893*, 2020. 2, 4, 7

[8] 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. 6

[9] 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, 2009. 3

[10] 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. 6

[11] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In *European Conference on Computer Vision*, pages 228–244. Springer, 2022. 7

[12] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 1, 3

[13] 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. 6

[14] 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. 6

[15] 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. 6

[16] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 1, 2, 6

[17] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019. 3, 4, 5, 6, 7

[18] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020. 2, 3, 7

[19] Vitor Guizilini, Rui Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *arXiv preprint arXiv:2002.12319*, 2020. 7

[20] Vitor Guizilini, Kuan-Hui Lee, Rareş Ambruş, and Adrien Gaidon. Learning optical flow, depth, and scene flow without real-world labels. *IEEE Robotics and Automation Letters*, 7(2):3491–3498, 2022. 7

[21] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1

[22] 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. 3, 6

[23] Mu He, Le Hui, Yikai Bian, Jian Ren, Jin Xie, and Jian Yang. Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *European Conference on Computer Vision*, pages 565–581. Springer, 2022. 7

[24] Jie Hu, Linyan Huang, Tianhe Ren, Shengchuan Zhang, Rongrong Ji, and Liujuan Cao. You only segment once: Towards real-time panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17819–17829, 2023. 4, 6

[25] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE, 2019. 3

[26] Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson WH Lau, and Thomas S Huang. Geometry-aware distillation for indoor semantic segmentation. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2869–2878, 2019. 3

[27] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 4756–4765, 2020. 3

[28] Adrian Johnston and G. Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4755–4764, 2020. 7

[29] 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. 2, 3, 4, 7

[30] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 3

[31] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 4

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[33] 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. 7

[34] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon. Learning monocular depth in dynamic scenes via instance-aware projection consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1863–1872, 2021. 2, 7

[35] Jun Li, Reinhard Klein, and Angela Yao. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3372–3380, 2017. 3

[36] Rui Li, Xiantuo He, Danna Xue, Shaolin Su, Qing Mao, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Learning depth via leveraging semantics: Self-supervised monocular depth estimation with both implicit and explicit semantic guidance. *arXiv preprint arXiv:2102.06685*, 2021. 7

[37] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022. 4, 7

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[39] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. *arXiv preprint arXiv:2012.07356*, 6, 2020. 3, 7

[40] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 6

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[42] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and S. Mattoccia. On the uncertainty of self-supervised monocular depth estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3224–3234, 2020. 7

[43] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 3

[44] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 3

[45] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 1, 3

[46] 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. 3

[47] Patrick Ruhkamp, Daoyi Geo, Hanzhi Chen, Nassir Navab, and Benjamin Busam. Attention meets geometry: Geometry guided spatial-temporal attention for consistent self-supervised monocular depth estimation. In *IEEE International Conference on 3D Vision (3DV)*, December 2021. 7

[48] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, pages 572–588. Springer, 2020. 7

[49] Wen Su, Haifeng Zhang, Quan Zhou, Wenzhen Yang, and Zengfu Wang. Monocular depth estimation using information exchange network. *IEEE Transactions on Intelligent Transportation Systems*, 22(6):3491–3503, 2020. 3

[50] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions*

on pattern analysis and machine intelligence, 43(10):3349–3364, 2020. 6

[51] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1164–1174, 2021. 2, 4, 7

[52] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3917–3925, 2018. 3

[53] Xinchen Ye, Shude Chen, and Rui Xu. Dpnet: Detail-preserving network for high quality monocular depth estimation. *Pattern Recognition*, 109:107578, 2021. 3

[54] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693, 2019. 3

[55] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 2

[56] Sen Zhang, Jing Zhang, and Dacheng Tao. Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating imu motion dynamics. In *European Conference on Computer Vision*, pages 143–160. Springer, 2022. 2

[57] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4106–4115, 2019. 3

[58] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. In *British Machine Vision Conference (BMVC)*, 2021. 7

[59] 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. 1, 3, 6, 7

[60] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 36–53, 2018. 6