Fine-grained Semantics-aware Representation Enhancement for Self-supervised Monocular Depth Estimation

Hyunyoung Jung
Seoul National University
gusdud1500@gmail.com

Eunhyeok Park
POSTECH
eh.park@postech.ac.kr

Sungjoo Yoo
Seoul National University
sungjoo.yoo@gmail.com

Abstract

Self-supervised monocular depth estimation has been widely studied, owing to its practical importance and recent promising improvements. However, most works suffer from limited supervision of photometric consistency, especially in weak texture regions and at object boundaries. To overcome this weakness, we propose novel ideas to improve self-supervised monocular depth estimation by leveraging cross-domain information, especially scene semantics. We focus on incorporating implicit semantic knowledge into geometric representation enhancement and suggest two ideas: a metric learning approach that exploits the semantics-guided local geometry to optimize intermediate depth representations and a novel feature fusion module that judiciously utilizes cross-modality between two heterogeneous feature representations. We comprehensively evaluate our methods on the KITTI dataset and demonstrate that our method outperforms state-of-the-art methods. The source code is available at https://github.com/hyBlue/FSRE-Depth.

1. Introduction

Depth measurement is a critical task in various applications, including robotics, augmented reality, and self-driving vehicles. It measures the distance from all or a part of the pixels in the imaging device to target objects using active/passive sensors. Equipping such devices requires high cost and continuous operation, which makes its use limited. Monocular depth estimation estimates the depth of pixels in a given 2D image without additional measurement. It facilitates the understanding of 3D scene geometry from a captured image, which closes the dimension gap between the physical world and an image.

Because of its importance and cost benefits, there have been lots of studies [53, 35, 33, 48, 67] that have improved depth estimation accuracy, temporal consistency and depth ranges. Owing to the success of the convolutional neural
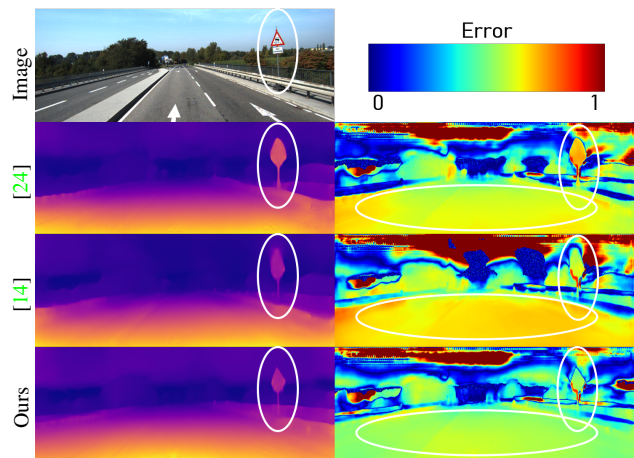

Figure 1: Depth predictions and error maps of recent state-of-the-arts [14, 24]. Apart from ours, both lack accuracy in low-texture regions and object boundaries, owing to weak supervision using contemporary self-supervised training.

network, it has also been adapted to monocular depth estimation and has produced great improvements.

Many existing monocular depth estimation methods train their networks with supervised depth labels computed via synthetic data or estimated from depth sensor [30, 9, 28, 25]. Although such methods have provided significant improvements in depth estimation, they still have multiple concerns related to the high cost of labeling and obtaining the depth labels on pixels, the limited available ground-truth depth data, the restricted depth range of sampled data, and the noticeable noise in the depth values. To avoid these shortcomings, self-supervised training methods have recently been proposed.

Notably, the SfM-Learner [64] method utilizes the ensembles of consecutive frames in video sequences for joint training depth and pose networks. It demonstrates comparable performance to extant supervised methods; however, recent works [69, 39, 12, 2] based on SfM-Learner mostly rely on photometric loss [54] and smoothness constraints; hence, they suffer from limited supervision of weak texture

regions. Furthermore, moving objects and uncertainty in the pose network destabilize training, leading to incorrect depth values, especially on object boundaries (see Fig. 1).

Several recent methods have attempted to overcome this weakness by employing cross-domain knowledge learning, including leveraging scene semantics to improve monocular depth predictions [24, 27, 13, 3]. They remove dynamic objects or explicitly model the object motion from the semantic instances to incorporate them into the scene geometry. In addition, a regularization of the depth smoothness within corresponding semantic objects enforces consistency between depth and semantic predictions [4, 38, 65].

In this study, we aim to improve self-supervised monocular depth estimation via the implicit use of semantic segmentation. We do not explicitly identify moving objects or regularize depth values in accordance with the semantic labels. Instead, we focus on representation enhancement, optimizing the depth network in the representation spaces, to produce semantically consistent intermediate depth representations.

Inspired by the recent use of deep metric learning [52, 44, 23], we suggest a novel semantics-guided triplet loss to refine depth representations according to implicit semantic guidance. Here, our goal is to take advantage of local geometric information from the scene semantics. For example, the adjacent pixels within each object have similar depth values, whereas those that cross semantic boundaries may have large differences. Combined with a simple but effective patch-based sampling strategy, our metric-learning approach exploits the semantics-guided local geometry information to optimize pixel representations near the object boundary, thereby yielding improved depth predictions.

We also design a cross-task attention module for refining depth features more semantically consistent. It computes the similarity between the reference and target features through multiple representation subspaces and effectively utilizes the cross-modal interactions among the heterogeneous representations. As a result, we quantify the semantic awareness of depth features as a form of attention and exploit it to produce better depth predictions.

Our contributions are summarized as follows. First, we present a novel training method that extracts semantics-guided local geometry with patch-based sampling and utilizes it to refine depth features in a metric-learning formulation. Second, we propose a new cross-task feature fusion architecture that fully utilizes the implicit representations of semantics for learning depth features. Finally, we comprehensively evaluate the performances of these two methods using the KITTI Eigen split and demonstrate that our method outperforms recent state-of-the-art self-supervised monocular depth prediction works in every metric.

2. Related Work

2.1. Depth Estimation with Neural Network

The recent success of neural networks has stimulated significant improvements to monocular depth estimation as a supervised regression method [30, 9, 25]. Recently, unsupervised training methods have been actively investigated. [11] used predicted disparity to synthesize a virtual image and minimized its photometric loss for training. [64] trained the depth network jointly with an additional pose network, requiring only monocular sequences. Based on these approaches, they have widely been tackled [12, 2, 57]. Many researchers have made further improvements along multiple lines, such as regularizing consistency with optical flow [47, 69, 39, 60] or functional geometric constraints between feature maps [45, 43].

Several recent works have proposed self-supervised depth prediction with semantics. They have enforced cross-task consistency and smoothness [4, 38, 65] and removed dynamic objects [24] or explicitly modeled object motions [27, 13, 3]. [15] targeted semantics-aware representations for depth predictions, enabling it via knowledge transfer from a fixed teacher segmentation network with pixel-adaptive convolution [46]. In contrast, we design a multi-task network with cross-task multi-embedding attention and semantics-guided triplet loss to successfully produce semantics-aware representation.

2.2. Neural Attention Network

[50, 8] designed a self-attention scheme that captures long-range dependencies to resolve the locality of recurrent operations. They proposed multi-head attention for utilizing information from different representation subspaces. Recently, cross-attention schemes have been utilized to extract features across heterogeneous representations, such as image, speech, and text [17, 55, 61]. Additionally, for self-supervised depth estimation, [22, 62] applied self-attention to capture the global context for estimating depth and combining multi-scale features from dual networks. In this study, inspired by the use of multi-head attention and cross attention, we propose a novel method of judiciously utilizing cross-task features across depth and segmentation.

2.3. Multi-task Architecture

The combination of features from multiple tasks has been widely used in recent multi-task architectures. [56, 49, 20] applied a convolutional layer to extract local information from the reference task feature for multimodal distillation. [63, 59, 21, 5] adopted affinity-guided message passing to propagate the relationship of spatially different features within the reference task to the target one. Instead, we propose a cross-task attention to fully utilize cross-modal interactions between geometry and semantics.

2.4. Deep Metric Learning

Deep metric learning [52, 44, 23] has been widely applied in various fields, such as face recognition [41, 18] and image ranking [51, 34, 1]. Inspired by recent successes, we propose a semantics-guided triplet loss to refine feature representations for improving depth predictions by exploiting implicit geometry from semantic supervision.

3. Methods

Here, we review our baseline approach, Monodepth2 [12], and present our current methodology in the following subsections.

3.1. Depth Estimation and Semantic Segmentation

3.1.1 Self-supervised Monocular Depth Estimation

Given consecutive RGB images, I_{t-1} and I_t, one can predict Z_t, the depth of every pixel on I_t, and compute a six degree-of-freedom relative pose, $T_{t \to t-1}$, using a pose network. With known camera intrinsics, $K \in \mathbb{R}^{3 \times 3}$, we can derive the projected pixel coordinates and use them from I_{t-1} as:

$$p' = K T_{t \to t-1} Z_t K^{-1} p \qquad \hat{I}_t = W_t(I_{t-1}, p'), \quad (1)$$

where p is the homogeneous coordinates of the pixel in I_t, and p' is the transformed coordinates of p by $T_{t \to t-1}$. $W(\cdot)$ is a sub-differentiable bilinear sampler [19] that obtains nearby pixels at p' in I_{t-1} and assigns the linearly interpolated pixel at p in $\hat{I}_t$. Ideally, I_t and $\hat{I}_t$ should be aligned if both depth and pose networks are optimally trained. These two networks are jointly optimized to minimize the discrepancy between I_t and $\hat{I}_t$. We utilize the structural similarity index measure (SSIM) [54] combined with L1 loss as a photometric loss, L_{ph} [11]:

$$L_{ph} = \alpha \frac{1 - \text{SSIM}(I_t, \hat{I}_t)}{2} + (1 - \alpha)|I_t - \hat{I}_t| \quad (2)$$

We compute L_{ph} for the two frame pairs, $[I_{t-1}, I_t]$ and $[I_t, I_{t+1}]$ to deal effectively with occlusions. We apply the minimum reprojection [12], which selects the pixel having a smaller loss between the two reference frames $[I_{t-1}, I_{t+1}]$, and we apply an auto-mask [12]. The following edge-aware smoothness loss [11], L_{sm}, is also added.

$$L_{sm} = |\partial_x d_t| e^{-|\partial_x I_t|} + |\partial_y d_t| e^{-|\partial_y I_t|} \quad (3)$$

The loss function of our baseline is obtained as follows:

$$L_{baseline} = L_{ph} + \beta \cdot L_{sm} \quad (4)$$

where β controls the relative strength of the smoothness factor.

3.1.2 Supervised Semantic Segmentation

A typical network model for semantic segmentation has an encoder-decoder architecture [40] for extracting features and upsampling them for dense predictions. This structure is similar to our baseline depth network [12], wherein basic features are extracted first prior to being fed into the decoder. Therefore, we adopt a shared-encoder architecture to reduce computations and benefit from both tasks.

In our proposed method, we train semantic segmentation with a pseudo-label generated by an off-the-shelf segmentation model [66]. We do not require per-image ground-truth of segmentation in the training dataset; thus, it is more practically applicable. We used the cross-entropy loss, L_{CE}, for training, and the training loss includes γL_{CE} with the baseline loss (Eqn. 4), where γ is a control parameter.

3.2. Semantics-guided Triplet Loss

Based on the local geometric relation from scene semantics, adjacent pixels within each object instance have similar depth values, whereas those across semantic boundaries may have large depth differences. Thus, we apply this intuition through a representation learning problem inspired by the recent usage of deep metric learning [52, 23]. We first separate pixels of the local patch on the semantic label into triplets (i.e., anchor, positive, and negative), and we then divide features from the lth layer of the depth decoder (F_d^l) in accordance with the corresponding location of those triplets. We aim to optimize the distance among these triplets, following the intuition described above. However, we do not directly optimize the depth value itself. Our key idea is that the distance should be defined and optimized in the representation space. Hence, the depth decoder can produce more discriminative features on the boundary regions so that the output depth map becomes more aligned with the semantic boundaries.

3.2.1 Patch-based Candidate Sampling

We first divide the semantic label into the $K \times K$ size of image patches with a stride of one. For each patch, we selected center of each patch as the anchor pixel and those that have the same class as that of the anchor as positive pixels. The negative pixels have different classes from those of the anchor pixels. Subsequently, we define $\mathcal{P}_i{}^+$ and $\mathcal{P}_i{}^-$, the sets of positive and negative pixels in the local patch $\mathcal{P}_i$, of which the spatial location of the anchor is i. We use $\mathcal{P}_i{}^+$ and $\mathcal{P}_i{}^-$ to determine whether $\mathcal{P}_i$ intersects the semantic borders. For example, $|\mathcal{P}_i{}^-| = 0$ means that $\mathcal{P}_i$ is located inside a specific object and does not cross the borders. On the other hand, if $|\mathcal{P}_i{}^+|$ and $|\mathcal{P}_i{}^-|$ are both larger than zero, it indicates that $\mathcal{P}_i$ intersects the boundaries across objects.

Additionally, the semantic labels may not be accurate or consistent because they are predictions of pre-trained seg-
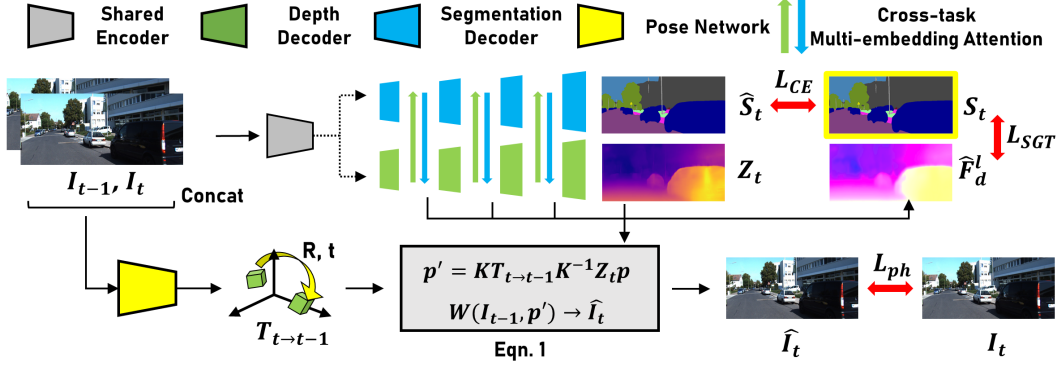
Figure 2: Overview of our proposed architecture. We adopt a shared encoder and two separate decoders, each for depth and segmentation, and we train both tasks jointly. The semantics-guided triplet loss (L_{SGT}) is calculated with the segmentation pseudo-label (S_t), and it optimizes L2 normalized depth feature maps ($\hat{F}_d^l$) with semantic guidance. The cross-task multi-embedding attention module is located between the layers of each decoder, and it enables cross-modal interactions between two task-specific decoders.
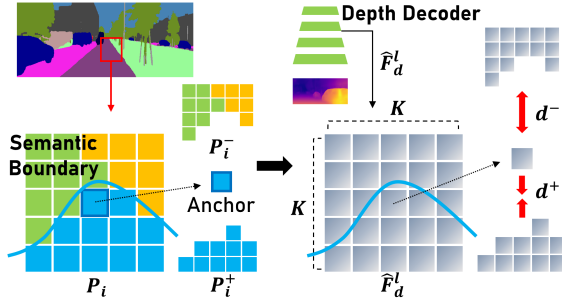

Figure 3: Overview of proposed semantics-guided triplet loss. We first divide $K \times K$ size local patch ($\mathcal{P}^i$) of the semantic label into triplets. Then, we optimize the distance between normalized depth features ($\hat{F}_d^l$) following the corresponding pixel locations in the semantic patch.

mentation networks. To reduce misclassification caused by these imperfect labels, we set a threshold, T, and determine $\mathcal{P}_i$ intersects with the boundaries when $|\mathcal{P}_i^+|$ and $|\mathcal{P}_i^-|$ are both larger than T.

3.2.2 Triplet Margin Loss

We grouped the features in each patch of the depth feature map into three classes (i.e., anchor, positive, and negative) following the corresponding pixel locations in the semantic image patch. We define positive distance d^+ and negative distance d^- as the mean of the Euclidean distance of the L2 normalized depth feature pairs.

$$d^+(i) = \frac{1}{|\mathcal{P}_i^+|} \sum_{j \in \mathcal{P}_i^+} \sqrt{(\hat{F}_d^l(i) - \hat{F}_d^l(j))^2} \qquad (5)$$

$$d^-(i) = \frac{1}{|\mathcal{P}_i^-|} \sum_{j \in \mathcal{P}_i^-} \sqrt{(\hat{F}_d^l(i) - \hat{F}_d^l(j))^2}, \qquad (6)$$

where $\hat{F}_d^l = F_d^l / \|F_d^l\|$.

We aim to reduce the distance between the anchor and positive features, and increase the distance between the anchor and negative features. However, naively maximizing d^- as far as possible does not lead to our desired outcome because the semantic border does not always guarantee that the depth of two separate objects differs by a large amount. Instead, we adopt the triplet loss [42, 51] with a margin so that the distance is no longer optimized when the negative distance exceeds a positive distance more than a specific margin m, as a hyper-parameter.

$$\mathcal{L}_{\mathcal{P}_i} = max(0, d^+(i) + m - d^-(i)) \qquad (7)$$

The semantics-guided triplet loss L_{SGT} is the average of $\mathcal{L}_{\mathcal{P}_i}$, only containing $\mathcal{P}_i$ satisfying the condition described in Sec. 3.2.1.

$$\mathcal{L}_{SGT} = \frac{\sum_i \mathbb{1}[|\mathcal{P}_i^+|, |\mathcal{P}_i^-| > T] \cdot \mathcal{L}_{\mathcal{P}_i}}{\sum_i \mathbb{1}[|\mathcal{P}_i^+|, |\mathcal{P}_i^-| > T]} \qquad (8)$$

We sum over the L_{SGT} of depth features across multiple layers and include into the total loss the sum multiplied by control parameter δ.

3.3. Cross-task Multi-embedding Attention (CMA) Module

We propose a CMA module to produce semantics-aware depth features through the representation subspaces and utilize them to refine depth predictions. As illustrated in Fig. 2, the CMA modules are located in the middle of each decoder layer and utilize the information from the other decoder. A single CMA module has uni-directional data flow, e.g., a CMA module refines the target feature with the reference feature. We use two CMA modules simultaneously to enable bidirectional feature enhancement, where depth

(segmentation) becomes the target (reference) in one CMA module while their roles change in the other. In the following paragraphs, we only describe a single case where the depth feature is the target for ease of explanation.

In our model, each decoder comprises five blocks ($l = 0, 1, 2, 3, 4$) and the spatial resolution of the feature map is doubled for each. The depth (segmentation) decoder generates a feature map, F_d^l (F_s^l), which has a spatial resolution of $(H/2^{4-l}, W/2^{4-l})$, where H and W are the height and width of the input image, respectively. The CMA modules can be attached to any of the five candidates.

The CMA module performs a pixel-wise operation on the two feature maps, F_d^l and F_s^l, through several operations. It first computes the semantic awareness of the depth features as a pixel-wise attention score through cross-task similarity (Sec. 3.3.1). We then extend this computation with multiple linear projections so that the similarity can be computed from different representation subspaces (Sec. 3.3.2). This enables selective extraction of depth features from multiple embeddings upon the corresponding semantic awareness, maximizing the utilization of cross-modality. Subsequently, the fusion function combines the input feature map, F_d^l, with the refined one, $F_d^{l'}$ (Sec. 3.3.3). We explain the details in the following sections.

3.3.1 Cross-task Similarity

We define *cross-task similarity* as $F_d^l(i)^T F_s^l(i)$, where i is the spatial index of each feature map, and $F^l(i)$ is a *C-dimensional* feature vector. This indicates quantitative amounts of semantic representation that each depth feature implicitly refers to. However, direct computation with raw feature vectors is infeasible, owing to the different nature of the tasks. We apply a linear projection, ϕ, that transforms the input feature from the original dimension, C, to C'. This indirectly computes the cross-task similarity through the representation subspace. The refined feature is computed as follows:

$$F_d^l(i)' = \rho(A(i)) \times \phi_v(F_d^l(i)), \qquad (9)$$

$$\text{where} \quad A(i) = \frac{\phi_k(F_d^l(i))^T \phi_q(F_s^l(i))}{\sqrt{C'}} \qquad (10)$$

Here, ρ is a normalization factor scaling the input. We apply three separate linear embeddings, and each acts as query (ϕ_q), key (ϕ_k), and value (ϕ_v) functions. The target feature map, F_d^l, becomes the input for the key and value embeddings, and the reference feature map, F_s^l, becomes the input for the query embedding.

For depth prediction, this imposes large attention scores ($A(i)$) on the specific depth features which are consistent with semantics, so that it can implicitly utilize semantic region information. As mentioned above, this module is bidirectional, and the semantic feature F_s^l acts as the target si-
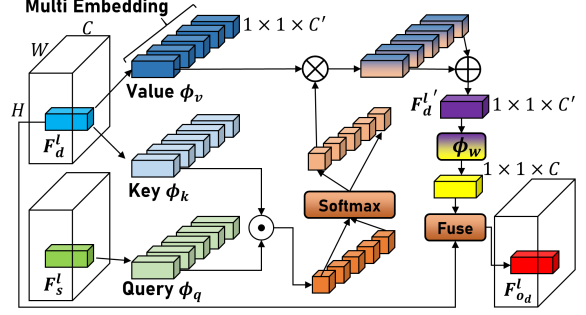

Figure 4: Architecture of the cross-task multi-embedding attention module. The feature map from the lth layer of the depth decoder, F_d^l, acts as a target in this case, and the segmentation feature, F_s^l, acts as a reference. The CMA module refines the target feature with the reference feature, and produces the output feature $F_{o_d}^l$.

multaneously. At this time, the depth feature is used to learn the features for semantic prediction so that backpropagation from the segmentation loss (L_{CE}) optimizes depth layers while offering more semantics-aware representations.

Compared with the affinity matrix for cross-task feature fusion [63, 59, 21, 5], which is computed solely based on features from a single task, the CMA module computes the attention score based on features from both tasks. Hence, it can effectively handle cross-modal interactions for multi-task predictions.

3.3.2 Multi-embedding Attention

Inspired by multi-head attention [50, 8], we adopt multiple linear projections to compute the similarity between feature vectors through different representation subspaces. This refines depth features with implicit semantic representations more effectively, as verified in Sec. 4. We use H distinct projection functions, $\phi^h (h = 1, ..., H)$; hence, the queries, keys, and values are mapped to H independent subspaces. The cross-task similarity in Eqs. 9-10 can be directly extended to a multi-embedding scheme as follows:

$$F_d^{l^h}(i)' = \rho(A^h(i)) \times \phi_v^h(F_d^l(i)), \qquad (11)$$

$$\text{where} \quad \rho(A^h(i)) = \frac{e^{A^h(i)}}{\sum_{h' \in H} e^{A^{h'}(i)}}, \qquad (12)$$

$$\text{and} \quad A^h(i) = \frac{\phi_k^h(F_d^l(i))^T \cdot \phi_q^h(F_s^l(i))}{\sqrt{C'}} \qquad (13)$$

The refined feature, $F_d^l(i)'$, is the summation of the feature maps refined from each embedding function:

$$F_d^l(i)' = \sum_h F_d^{l^h}(i)' \qquad (14)$$

In the above equations, h represents the index of multiple linear embeddings. We adopted $softmax$ as a normalization function, ρ, to compute the importance of each embedding. Thus, we can selectively exploit the outputs from multiple attentions. This process is illustrated in Fig. 4.

In contrast to the original multi-head attention where the results from each embedding head are concatenated and equally handled, we compute the attention score among multiple heads and measure the significance of results from each embedding on the corresponding attention scores.

3.3.3 Fusion Layer

Finally, the refined feature map, $F_d^{l'}$, is projected to the original dimension C (ϕ_w in Fig. 4) and fused with the initial feature map, F_d^l, to produce final output, $F_{o_d}^l$. We apply two convolution layers to concatenated feature maps, $[\phi_w(F_d^{l'}), F_d^l]$, to produce $F_{o_d}^l$. $F_{o_d}^l$ becomes the input of $l + 1^{th}$ layer of the depth decoder.

4. Experiments

4.1. KITTI Dataset

The KITTI dataset [10] has been widely adopted for depth prediction benchmarks. We used the Eigen split [9] for this purpose, and preprocessing was performed to remove static frames, as in [12, 64]; thus, 39,910 and 4,424 images were used for training and validation, respectively, and 697 images were used for evaluation.

For training semantic segmentation, we generated pseudo-labels using an off-the-shelf network [66]. To evaluate the segmentation performance, we used 200 images and labels provided in the training set of the KITTI semantic segmentation benchmark corresponding to KITTI 2015 [32].

4.1.1 Evaluation

To evaluate the capability of depth prediction, we conducted per-image median-scaling with ground-truth following the evaluation protocol in [12]. The maximum depth is 80 m, as in recent studies [12, 36, 24]. We evaluated semantic segmentation in the mean intersection over union (mIoU).

4.2. Implementation Details

4.2.1 Network Architecture

The depth and segmentation network has a standard encoder-decoder architecture [40] with skip connections, as in Monodepth2 [12]. The shared encoder and the pose network encoder are ResNet-18 [16], pre-trained on ImageNet [7]. For the CMA module, we adopt four ($H = 4$) embeddings of the multi-embedding scheme. The dimension ratio between the original feature and the embedded

feature is two, such that $C' = 2 * C$. Thus, the projected vectors have twice the dimensions of the the corresponding input features. The normalization factor, ρ, is the identity function when $H = 1$ (without multi-embedding) and $softmax$ when $H > 1$ (w/ multi-embedding). We apply CMA module to three of decoder layers, $l = 0, 1, 2$.

4.2.2 Training Details

For training, we resized the original image into a resolution of 192×640 and used a batch size of 12. The Adam optimizer was used with an initial learning rate of 1.5e-4, and we trained for 20 epochs while the learning rate was decayed by 0.1 twice, after 10 and 15 epochs of training. We used SSIM [54] with L_1 loss for L_{ph}, with $\alpha = 0.85$ following the previous work [12]. We set the loss parameters as follows: $\beta = 0.001$, $\gamma = 0.3$, and $\delta = 0.1$. The local patch size, K, is set to five, and the margin, m, is set to 0.3 for L_{SGT}. This loss is applied to features from three layers, $l = 1, 2$ and 3. The threshold T is set to $K - 1$.

4.3. Quantitative Results and Ablation Study

Table 1a compares our proposed method with recent works. Ours achieves state-of-the-art results on the KITTI Eigen test split and outperforms previous works in every metric. Our network adopts ResNet-18 as a backbone, but we also use ResNet-50 and compare it with others adopting ResNet-50. Ours (ResNet-50) also achieved the best results. Note that the PackNet versions of [15] adopted a significantly large backbone ($> 10\times$ larger than ResNet-18). Therefore, we compare the ResNet-18 and ResNet-50 versions of [15] and show that our method outperforms it by a large margin. Additionally, in our multi-task network, semantic information is required only for training. In contrast, [15] and [29] require the semantics for both training and testing. [15] requires a teacher segmentation network for feature distillation during inference and [29] requires semantic label or pre-computed segmentation results as the network input. Finally, our network is highly compatible with more advanced networks [31, 14] which have architectural differences from our baseline, Monodepth2. This indicates the potential for further improvement.

In Table 1b, we also evaluate the effectiveness of each proposed method. The addition of semantic segmentation to depth (L_{CE}) via shared encoder shows an improvement. Applying the semantics-guided triplet loss and the CMA module further improves the baseline. This verifies that more semantics-aware representation improvements of depth predictions are produced. Finally, the combination of both methods significantly improves the performance. Both techniques are designed to refine the depth representation via semantic knowledge, and they offer highly synergistic improvements.

Table (a):

Method	Backbone	Sem	Lower is better				Higher is better		
			AbsRel	SqRel	RMS	RMSlog	< 1.25	< 1.25²	< 1.25³
SceneNet [4]	DRN [58]	✓	0.118	0.905	5.096	0.211	0.839	0.945	0.977
Monodepth2 [12] *	ResNet18		0.114	0.864	4.817	0.192	0.875	0.959	0.981
Zou *et al.* [68]	ResNet18		0.115	0.871	4.778	0.191	0.874	0.963	0.984
Guizilini *et al.* [15]	ResNet18	✓	0.117	0.854	4.714	0.191	0.873	0.963	0.981
SGDepth [24]	ResNet18	✓	0.113	0.835	4.693	0.191	0.879	0.961	0.981
Lee *et al.* [27]	ResNet18	✓	0.112	0.777	4.772	0.191	0.872	0.959	0.982
Poggi *et al.* [37]	ResNet18		0.111	0.863	4.756	0.188	0.881	0.961	0.982
Patil *et al.* [36]	ResNet18		0.111	0.821	4.650	0.187	0.883	0.961	0.982
SAFENet [5]	ResNet18	✓	0.112	0.788	4.582	0.187	0.878	0.963	0.983
HRDepth [31]	ResNet18		0.109	0.792	4.632	0.185	0.884	0.962	0.983
Ours	ResNet18	✓	**0.105**	**0.722**	**4.547**	**0.182**	**0.886**	**0.964**	**0.984**
SGDepth [24]	ResNet50	✓	0.112	0.833	4.688	0.190	0.884	0.961	0.981
Monodepth2 [12]	ResNet50		0.110	0.831	4.642	0.187	0.883	0.962	0.982
Guizilini *et al.* [15]	ResNet50	✓	0.113	0.831	4.663	0.189	0.878	**0.971**	0.983
FeatDepth [43]	ResNet50		0.104	0.729	4.481	0.179	**0.893**	0.965	0.984
Li *et al* [29]	ResNet50	✓	0.103	0.709	4.471	0.180	0.892	0.966	0.984
Ours	ResNet50	✓	**0.102**	**0.675**	**4.393**	**0.178**	**0.893**	0.966	**0.984**
Johnston *et al* [22]	ResNet101		0.106	0.861	4.699	0.185	0.889	0.962	0.982
PackNet-SfM [14]	PackNet		0.111	0.785	4.601	0.189	0.878	0.960	0.982
Guizilini *et al.* [15]	PackNet	✓	0.102	0.698	4.381	0.178	0.896	0.964	0.984

* We re-implemented Monodepth2 and the result has slightly improved from the original paper.

(a)

Table (b):

L_{CE}	L_{SGT}	CMA module	HR	AbsRel	SqRel	RMS	RMSlog	< 1.25	< 1.25²	< 1.25³
				0.114	0.864	4.817	0.192	0.875	0.959	0.981
✓				0.112	0.823	4.705	0.189	0.879	0.961	0.982
✓	✓			0.108	0.755	4.618	0.186	0.882	0.962	0.983
✓		✓		0.107	0.741	4.586	0.184	0.884	0.962	0.983
✓	✓	✓		**0.105**	**0.722**	**4.547**	**0.182**	**0.886**	**0.964**	**0.984**
✓	✓	✓	✓	0.102	0.687	4.366	0.178	0.895	0.967	0.984

(b)

Table 1: (a) Comparison of self-supervised monocular depth estimation with recent works on KITTI Eigen split. All methods are trained with monocular images with size of 192×640, except SceneNet. Sem denotes training with semantic information. (b) Ablation of the proposed methods in depth predictions. HR denotes training with high-resolution images (320×1024).

Table 2:

Method	Train	MIoU
SceneNet [4]	CS	37.7
SGDepth [24]	CS	51.6
Ours	K	56.6
Ours (HR)	K	**60.6**
Ours w/o CMA	K	56.1
Ours w/o CMA (HR)	K	59.1

Table 2: Semantic segmentation results on KITTI 2015 training set. CS denotes Cityscapes, and K represents KITTI. HR refers to training using high-resolution image.

In Table 2, we also evaluate the semantic segmentation performance on KITTI 2015 [32]. Though the proposed method outperforms others, it is not fair to compare with the works that trained semantic segmentation with Cityscapes [6] ground-truth (they trained depth on KITTI.). Hence, we focus more on how segmentation benefits from depth estimation via CMA rather than the final performance. As shown in the last two rows, the proposed CMA module also improves the segmentation performance. Thanks to its bi-directional flow, the CMA module also refines semantic features as a target with reference to depth representations. Additionally, it is more effective when the resolution is high.

4.4. Qualitative Evaluation

We qualitatively compare our method with recent methods, SGDepth [24] and PackNet-Sfm [14], as shown in Figs. 1 and 5. In Fig. 1, we compare the depth predictions and error distributions[1], fixing AbsRel between 0 and 1. Similar to ours, [24] also adopted multi-task training with semantic segmentation via a shared encoder. However, only enhancing the encoder in a multi-task setting cannot fully exploit the semantic information. As shown in both figures, our method captures fine-grained detail, leading to

---

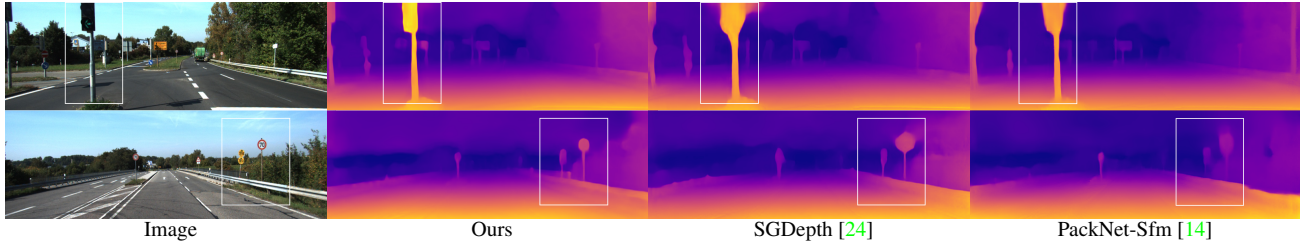[1]Owing to the sparsity of depth ground-truth, we computed the mismatch with top-performing supervised depth network [26].

Figure 5: Qualitative comparison of the depth predictions with recent works.

l	AbsR	SqR	<1.25	K	AbsR	SqR	<1.25
0-3	0.108	0.757	0.882	3	0.109	0.757	0.878
1-3	**0.108**	0.755	**0.882**	5	**0.108**	**0.755**	**0.882**
1-4	0.108	**0.742**	0.879	7	0.110	0.776	0.880
(a)				(b)			

Table 3: Ablations of semantics-guided triplet loss with (a) different layers of depth decoder the loss is applied, (b) the size of local patch K.

more accurate depth predictions compared with others, especially at object borders. This verifies the effectiveness of the proposed fine-grained semantics-aware enhancement of representation.

4.5. Further Analysis

Table 3a shows the results of varying layers ($l = 0, 1, 2, 3, 4$) to which semantics-guided triplet loss is applied. We selected layers $l = 1, 2$ and 3 because it showed the best results. Applying L_{SGT} to $l = 4$ degrades the performance as it has a significantly low channel dimension (16); hence, the distance cannot be properly computed. In Table 3b, we compare the effect of the patch size, K. Because separating a local patch into triplets relies on semantic labels from the off-the-shelf network, there must be noisy labels. When K is small (i.e., $K = 3$), the number of samples $|\mathcal{P}^{+,-}|$ decreases, and each noisy label contributes more to the mean distance, $d^{+,-}$. When K is large (i.e., $K = 7$), each local patch contains more non-boundary pixels and the negative distance can easily exceed the margin. In other words, the loss is computed from more *easy samples*, and the improvements are limited. In our experiments, $K = 5$ was the balanced point, which was the best option. We further compare the results of different margins, m, in the supplement.

Table 4 lists the results of the CMA module with varying parameters. As shown in Table 4a, our bidirectional CMA provides better results than the unidirectional CMA. This confirms that, to benefit from cross-modal representation, it is more beneficial to simultaneously improve both depth and semantics features than to improve just one.

Table 4b shows the effectiveness of the proposed multi-embedding scheme. It can fully utilize cross-modality as the number of embeddings grows. As shown in Fig. 6, the more
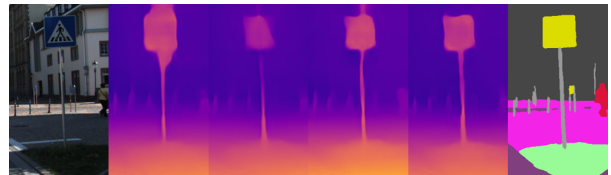

Figure 6: Comparison among baseline ($+L_{CE}$), and CMA module applied ($H = 1, 2, 4$) from left to right.

T	AbsR	SqR	<1.25	H	AbsR	SqR	<1.25
D	0.108	0.759	0.882	1	0.108	0.762	0.881
S	0.111	0.797	0.877	2	0.107	0.749	0.882
D,S	**0.107**	**0.741**	**0.884**	4	**0.107**	**0.741**	**0.884**
(a)				(b)			

Table 4: Ablations of CMA module with (a) different target task features and (b) the number of embedding functions. T denotes the target task, D denotes the depth, and S denotes segmentation.

embeddings used, the more precise the object boundary of the depth network, and the depth prediction becomes more aligned to semantics. This demonstrates that the depth network can have more semantics-aware representations, owing to our proposed multi-embedding scheme.

5. Conclusion

This paper proposed novel methods for accurate monocular depth prediction (i.e., semantics-guided triplet loss and cross-task multi-embedding attention) to make the best use of semantics-geometry cross-modality. Semantics-guided triplet loss offered a new and effective supervisory signal for optimizing depth representations. The CMA module allowed us to utilize rich and spatially fine-grained representations for multi-task training of depth prediction and semantic segmentation. The enhanced representation from these two methods exhibited a highly synergistic performance boost. Our extensive evaluation on the KITTI dataset demonstrates that the proposed methods outperformed extant state-of-the-art methods, including those that use semantic segmentation.

References

[1] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian D. Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Neural Information Processing Systems (NeurIPS)*, 2019.

[3] Vincent Casser, Sören Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. *Advancement of Artificial Intelligence (AAAI)*, 2019.

[4] 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. *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Jaehoon Choi, Dongki Jung, Donghwan Lee, and Changick Kim. Safenet: Self-supervised monocular depth estimation with semantic-aware feature extraction. *Neural Information Processing Systems Workshops (NeurIPSW)*, 2020.

[6] 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. *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition (CVPR)*, 2009.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Neural Information Processing Systems (NeurIPS)*, 2014.

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. *Computer Vision and Pattern Recognition (CVPR)*, 2012.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. *International Conference on Computer Vision (ICCV)*, 2019.

[13] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. *International Conference on Computer Vision (ICCV)*, 2019.

[14] V. Guizilini, Rares Ambrus, S. Pillai, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. *Computer Vision and Pattern Recognition (CVPR)*, pages 2482–2491, 2020.

[15] V. Guizilini, R. Hou, Jie Li, Rares Ambrus, and Adrien Gaidon. Semantically-guided representation learning for self-supervised monocular depth. *International Conference on Learning Representations (ICLR)*, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. *Neural Information Processing Systems (NeurIPS)*, 2019.

[18] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[19] Max Jaderberg, K. Simonyan, Andrew Zisserman, and K. Kavukcuoglu. Spatial transformer networks. *Neural Information Processing Systems (NeurIPS)*, 2015.

[20] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson W. H. Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. *European Conference on Computer Vision (ECCV)*, 2018.

[21] Jianbo Jiao, Yunchao Wei, Zequn Jie, Humphrey Shi, R. Lau, and T. Huang. Geometry-aware distillation for indoor semantic segmentation. *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] A. Johnston and G. Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[23] Mahmut Kaya and Hasan Sakir Bilge. Deep metric learning: A survey. *Symmetry*, 2019.

[24] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and T. Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. *European Conference on Computer Vision (ECCV)*, 2020.

[25] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. *International Conference on 3D Vision (3DV)*, 2016.

[26] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv:1907.10326*, 2019.

[27] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon. Learning monocular depth in dynamic scenes via instance-aware projection consistency. *Advancement of Artificial Intelligence (AAAI)*, 2021.

[28] Bo Li, Yuchao Dai, and Mingyi He. Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. *Pattern Recognition*, 2018.

[29] 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:2102.06685*, 2021.

[30] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian D. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

[31] 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. *Artificial Intelligence (AAAI)*, 2021.

[32] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[33] Theo Moons, Luc Van Gool, Maarten Vergauwen, et al. 3d reconstruction from multiple images part 1: Principles. *Foundations and Trends® in Computer Graphics and Vision*, 2010.

[34] Tony Ng, Vassileios Balntas, Yurun Tian, and Krystian Mikolajczyk. Solar: Second-order loss and attention for image retrieval. *European Conference on Computer Vision (ECCV)*, 2020.

[35] Irina Nurutdinova and Andrew W. Fitzgibbon. Towards pointless structure from motion: 3d reconstruction and camera parameters from general 3d curves. *International Conference on Computer Vision (ICCV)*, 2015.

[36] V. Patil, Wouter Van Gansbeke, Dengxin Dai, and L. Gool. Don't forget the past: Recurrent depth estimation from monocular video. *Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

[38] Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. *Asian Conference on Computer Vision (ACCV)*, 2018.

[39] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[42] Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. *Neural Information Processing Systems (NeurIPS)*, 2003.

[43] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. *European Conference on Computer Vision (ECCV)*, 2020.

[44] Hyun Oh Song, Stefanie Jegelka, Vivek Rathod, and Kevin Murphy. Deep metric learning via facility location. *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[45] J. Spencer, R. Bowden, and S. Hadfield. Defeat-net: General monocular depth via simultaneous unsupervised representation learning. *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[46] Hang Su, V. Jampani, Deqing Sun, O. Gallo, E. Learned-Miller, and J. Kautz. Pixel-adaptive convolutional neural networks. *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[47] Fabio Tosi, Filippo Aleotti, Pierluigi Zama Ramirez, M. Poggi, S. Salti, L. Stefano, and S. Mattoccia. Distilled semantics for comprehensive scene understanding from videos. *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[48] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 1979.

[49] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. *European Conference on Computer Vision (ECCV)*, 2020.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neural Information Processing Systems (NeurIPS)*, 2017.

[51] Jiang Wang, Yang song, Thomas Leung, Chuck Rosenberg, Jinbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[52] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. *International Conference on Computer Vision (ICCV)*, 2017.

[53] Ting-Chun Wang, Manohar Srikanth, and Ravi Ramamoorthi. Depth from semi-calibrated stereo and defocus. *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[54] 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 (TIP)*, 2004.

[55] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[56] D. Xu, Wanli Ouyang, X. Wang, and N. Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[57] H. Xu, J. Zheng, J. Cai, and J. Zhang. Region deformer networks for unsupervised depth estimation from unconstrained monocular videos. *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2019.

[58] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[59] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth,

surface normal and semantic segmentation. *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[60] Wang Zhao, S. Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[61] Yingzhu Zhao, Chongjia Ni, Cheung Chi Leung, Shafiq Joty, Eng Siong Chng, and Bin Ma. Cross attention with monotonic alignment for speech transformer. *Conference of the International Speech Communication Association (Interspeech)*, 2020.

[62] Junsheng Zhou, Yuwang Wang, Kai huai Qin, and Wen-Jun Zeng. Unsupervised high-resolution depth learning from videos with dual networks. *International Conference on Computer Vision (ICCV)*, 2019.

[63] L. Zhou, Z. Cui, Chunyan Xu, Z. Zhang, C. Wang, Tong Zhang, and Jian Yang. Pattern-structure diffusion for multi-task learning. *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[64] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[65] Shengjie Zhu, Garrick Brazil, and Xiaoming Liu. The edge of depth: Explicit constraints between segmentation and depth. *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[66] Yi Zhu, Karan Sapra, Fitsum A. Reda, Kevin J. Shih, Shawn D. Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[67] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rhemann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew W. Fitzgibbon, Charles T. Loop, Christian Theobalt, and Marc Stamminger. Real-time non-rigid reconstruction using an RGB-D camera. *ACM Transactions on Graphics (TOG)*, 2014.

[68] Yuliang Zou, Pan Ji, Quoc-Huy Tran, Jia-Bin Huang, and Manmohan Chandraker. Learning monocular visual odometry via self-supervised long-term modeling. *European Conference on Computer Vision (ECCV)*, 2020.

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