Towards Comprehensive Representation Enhancement in Semantics-guided Self-supervised Monocular Depth Estimation

Jingyuan Ma*, Xiangyu Lei*, Nan Liu, Xian Zhao, and Shiliang Pu**

Hikvision Research Institute
{majingyuan,leixiangyu,liunan6,zhaoxian,pushiliang.hri}@hikvision.com

Abstract. Semantics-guided self-supervised monocular depth estimation has been widely researched, owing to the strong cross-task correlation of depth and semantics. However, since depth estimation and semantic segmentation are fundamentally two types of tasks: one is regression while the other is classification, the distribution of depth feature and semantic feature are naturally different. Previous works that leverage semantic information in depth estimation mostly neglect such representational discrimination, which leads to insufficient representation enhancement of depth feature. In this work, we propose an attention-based module to enhance task-specific feature by addressing their feature uniqueness within instances. Additionally, we propose a metric learning based approach to accomplish comprehensive enhancement on depth feature by creating a separation between instances in feature space. Extensive experiments and analysis demonstrate the effectiveness of our proposed method. In the end, our method achieves the state-of-the-art performance on KITTI dataset.

Keywords: Monocular depth estimation, self-supervised learning, feature metric learning, representation enhancement

1 Introduction

Depth estimation is one of the fundamentals in many computer vision applications such as robotics, augmented reality and autonomous driving. A depth map reflects the distance between image plane and corresponding objects in real world. Such depth map can be acquired from various sensor setups. Owing to the low cost of single camera setup, monocular depth estimation has been actively researched. Although conventional methods, using SfM or SLAM algorithm [11, 36, 39], fail to produce satisfying results, deep-learning based methods [63, 1, 17, 14, 29] have achieved significant improvement. Still, estimating depth from monocular image remains challenging.

* First two authors contribute equally to this work
** Corresponding Author

Deep-learning based monocular depth estimation can be generally categorized into supervised and self/un-supervised learning. Currently, supervised methods [14, 29, 12, 41] have yielded satisfying advancement in monocular depth estimation. However, since acquiring accurate pixel-level annotations is expensive and limited, self-supervised learning has gained attention because of its independence from annotations and better scalability in data. Under self-supervised settings, monocular depth and egomotion are jointly estimated from separate networks [63, 16, 17, 42, 57, 59, 1], whose training process is self-supervised by photometric loss [51]. Recently, since depth and semantics are spatially aligned, some approaches attempt to leverage semantic information in depth estimation via direct feature fusion [27, 20, 18] or representational enhancement [25, 6]. However, depth estimation and semantic segmentation are two different tasks. Thus, their feature distributions are significantly different as shown in Fig 1. Such cross-task feature inconsistency exists between instances and within an instance. Thus, enhancing depth feature solely from cross-task spatial consistency is not sufficient. Task-specific representational uniqueness should be identified.

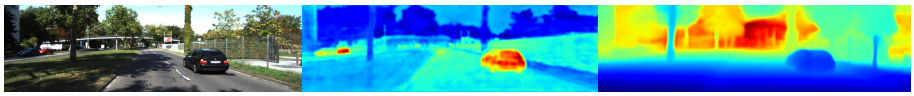

Fig. 1: The visualizations of feature heatmap for depth and semantic segmentation. (a)Colored image. (b)Semantic feature heatmap from Deeplabv3 [3]. (c)Feature heatmap of depth feature heatmap from Monodepth2 [16].

To address task-specific feature uniqueness within an instance, we design a novel and efficient intra/cross-task multi-head attention module (IC-MHA) that adequately fuses task-specific representational uniqueness with cross-task spatial consistency. Inspired by the recent success of vision transformers [33, 21, 9, 58], task-specific representational uniqueness is addressed as window-based self-attention mechanism on task-specific feature. Additionally, the similarity between depth feature and semantic feature is computed using cross-attention mechanism [25], which represents cross-task spatial consistency. A simple fusion layer is implemented to incorporate the generated task-specific self attention and cross attention with input task-specific feature.

To further enhance depth feature by addressing its representational uniqueness between instances, we propose a hardest non-boundary triplet loss whose anchors, positives and negatives are sampled with minimum-distance based candidate mining strategy. Such triplet loss achieves comprehensive enhancement on depth feature over all regions of an image. Extensive experiments and analysis prove the effectiveness of our proposed method, which achieves the state-of-the-art self-supervised monocular depth prediction on KITTI Eigen split [10]. Here, we summarize our contribution in three-fold.

- A novel and efficient intra/cross-task multi-head attention module is proposed to enhance task-specific features by emphasizing their representational uniqueness within instances while preserving their cross-task spatial consistency.
- An effective hardest non-boundary triplet loss using minimum-distance based candidate mining strategy is proposed to further enhance depth feature by addressing its representational uniqueness between instances.
- Our proposed method outperforms previous state-of-the-art self-supervised monocular depth estimation works on KITTI Eigen split.

2 Related Work

2.1 Self-supervised Monocular Depth Estimation

As a pioneering work in self-supervised monocular depth estimation, SfMLearner [63] jointly estimates pose and depth information using two networks, which are self-supervised by photometric loss [51]. Later, under this framework, many approaches are proposed to tackle occlusions [16, 1], dynamic objects [30, 24], low-texture regions [42, 57] and scale-inconsistency [49, 59]. Furthermore, some approaches attempt to utilize consistency between consecutive frames [1, 52, 38] or between various SfM tasks [59, 55, 66]. Also, a couple of better encoders [19, 62] are implemented to improve depth estimation. Considering that depth and semantics of an image are spatially aligned, some recent works propose to improve depth prediction by targeting dynamic objects [27] or exploiting semantics-aware depth feature [20, 6, 32, 25]. In our work, since task-specific feature has its unique distribution, we propose to further refine depth feature by efficiently fusing task-specific representational uniqueness with cross-task spatial consistency.

2.2 Vision Transformer

Inspired by [46], various vision transformers [21] have been proposed and demonstrated superior performance on many tasks such as image recognition [9, 4], object detection [2, 45, 60] and semantic segmentation [50, 61]. Amongst them, some works propose to perform local attention inside of image patches [9] or windows [33]. Inspired by window-based vision transformer [33], we propose to address the uniqueness of task-specific feature as multi-head self-attention within locally partitioned windows, which is then efficiently fused with cross-task spatial consistency.

2.3 Deep Metric Learning

Deep metric learning [28, 48, 44] aims to cluster samples with similar characteristics closer in feature space using proper candidate mining strategy. It has proven its success in various fields like face recognition [23], image retrieval [48, 44], keypoint detection [54, 7, 43] and depth estimation [25]. To further refine

depth feature, we propose a hardest non-boundary triplet loss whose positives and negatives are sampled based on the distance between their anchor sample and semantic boundary.

3 Methods

3.1 Proposed Model

To properly emphasize task-specific representational uniqueness and cross-task spatial consistency, we propose an intra-/cross-task multi-head attention (IC-MHA) module to enhance features for two task: depth estimation and semantic segmentation. Our overall pipeline is shown in Fig 2. Following [25, 16, 63, 42], a 6-DoF $T \in \mathbb{SE}(3)$ is estimated from PoseNet, whose input is a concatenated consecutive image pair. Taking the target image of size $[H, W]$ as input, our DepthSegNet consists of a shared encoder, task-specific decoders and our proposed IC-MHA modules. The IC-MHA module, whose input features' spatial dimension is $[\frac{H}{2^s}, \frac{W}{2^s}]$, is inserted between task-specific decoders at multiple levels s. Inside of IC-MHA module, task-specific representational uniqueness is addressed as intra-task local attention, and cross-task spatial consistency is represented by cross-task attention [25].

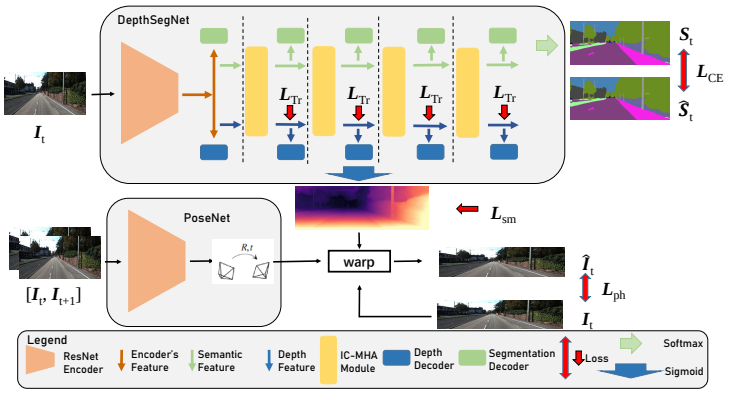

Fig. 2: An overview of our pipeline. DepthSegNet and PoseNet are implemented separately. Proposed IC-MHA module is inserted between task-specific decoders at multiple levels. $\mathcal{L}_{T_r}$ is the metric learning loss, consisting of boundary triplet loss $\mathcal{L}_{BT}$ in [25] and proposed hardest non-boundary triplet loss $\mathcal{L}_{NBT}$ in Eq 8.

Intra-/Cross-task Multi-Head Attention(IC-MHA) Module. The architecture of IC-MHA module is shown in Fig 3. At each level $s < 4$, the

upsampled task-specific features from level $s+1$ are fed into IC-MHA module. Within each IC-MHA module, linear projections with expansion ratio r, $\{\Psi_t^j : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{H\times W\times(r\times C)} | t \in \{D,S\}, j \in \{Q,K,V\}\}$, are applied on task-specific features, $\{F_t^s | t \in \{D,S\}, s < 4\}$, to generate a query-key-value triplet for each task, $\{(Q_t,K_t,V_t) | t \in \{D,S\}\}$. Following [46], we implement these linear projections on multiple heads H. In each head, the intra-task local attention and cross-task attention are computed in parallel. Here, to save computational cost, only one query-key-value triplet is projected for each feature to compute intra-task local attention and cross-task attention. For simplicity, we illustrate such computation on depth feature as an example. Then, the same process is symmetrically identical on semantic feature.

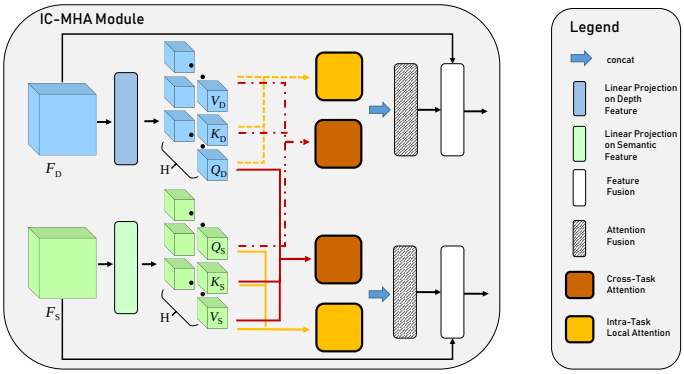

Fig. 3: An overview of our proposed IC-MHA module. Query-key-value triplet of depth and semantic feature firstly projected with Ψ_t^j. Then, intra-task local attention and cross-task attention are computed.

The overall process is shown in Fig 4. Inspired by [33], the self-attention mechanism is applied inside locally partitioned windows to properly compute intra-task local attention. In contrast to [33], in each head, query-key-value triplet, (Q_D^h, K_D^h, V_D^h), of depth feature is partitioned by a square window of size w_h, instead of depth feature itself. Additionally, we apply windows of different sizes on different heads instead of uniform window size on all heads in [33], such that our proposed self-attention mechanism can incorporate information from various local region efficiently and effectively. Denoting partitioned query, key, value as $(\hat{Q}_D^h, \hat{K}_D^h, \hat{V}_D^h)$, the intra-task local attention is computed as:

$$F_{S_D}^h(i) = \frac{e^{(\hat{Q}_D^h(i)(\hat{K}_D^h(i))^{\mathrm{T}}/\sqrt{C'})}}{\sum_{i'<w_h^2} e^{(\hat{Q}_D^h(i')(\hat{K}_D^h(i'))^{\mathrm{T}}/\sqrt{C'})}} \cdot \hat{V}_D^h(i) \tag{1}$$

where $i, i^{'} \in \mathbb{N}$ is the local index of feature map within one window and $C^{'} = r \cdot C$. Then, the local attention of each head is reversed back to the spatial dimension

of the inputs and concatenated along channel dimension. The concatenated local attention map is projected back to the feature dimension of the inputs, C, to form the final intra-task local attention map F_{S_D}.

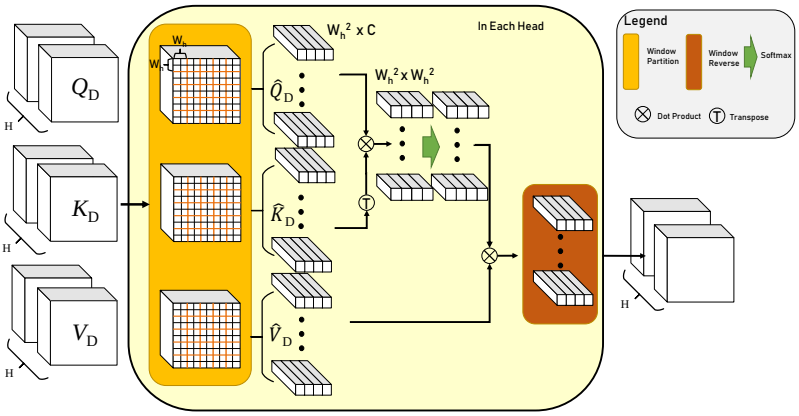

Fig. 4: The computation process of our proposed intra-task local attention. Here, we draw one head as an example. Query-key-value triplet of depth feature is firstly partitioned by a window of size w_h. Then, intra-task local attention is computed as in Eq 1. In the end, local attention of each head is reversed back to the spatial resolution of inputs, $[H^{'}, W^{'}]$.

To represent cross-task spatial consistency, we compute cross-task attention [25] from the key-value pair of depth feature and the query of semantic feature. Here, we do not apply window partition on the input query, key and value because the purpose of addressing such consistency is to align depth boundaries with semantic boundaries. Thus, computing cross-task attention directly from query-key-value triplets of depth feature and semantic feature is more optimal. Then, for each head, such attention is computed as:

$$F^h_{C_D}(j) = \frac{e^{(Q^h_S(j)(K^h_D(j))^{\mathrm{T}}/\sqrt{C'})}}{\sum_{h'<H} e^{(Q^{h'}_S(j)(K^{h'}_D(j))^{\mathrm{T}}/\sqrt{C'})}} \cdot V^h_D(j) \tag{2}$$

where j is the spatial index of feature map and $h \in \mathbb{N}$ and $C^{'} = r \cdot C$. Then, the cross attention is summed over head h and projected back to the feature dimension of the inputs, C. The process for computing cross-task attention is visualized is in Fig 2 of Supplementary Material.

Later, a linear projection is applied on the concatenated feature $[F_{S_D}, F_{C_D}]$ to generate the final attention map, F_{A_D}. In the end, a fusion layer, consisting of two convolution layers, is implemented to incorporate attention map, F_{A_D}, with input depth feature of IC-MHA module. The output of IC-MHA module is fed

into depth decoder to generate depth estimation of this level. Detailed schematic of IC-MHA module is shown in Sec 3 of Supplementary Material.

3.2 Photometric Loss and Edge-aware Smoothness Loss.

Given a pair of consecutive color images, I_s and I_t, estimated pose $T \in \mathbb{SE}(3)$ and estimated dense depth map D_t, the reconstructed target image $\hat{I}_t$ can be generated from source image I_t via:

$$\hat{I}_t(p) = I_s(\hat{p}), \quad \hat{p} = KTD_tK^{-1}p \tag{3}$$

where p is pixel's homogeneous coordinate in target image I_t, $\hat{p}$ is transformed coordinate of p, and $K \in \mathbb{R}^{3\times3}$ is a known camera intrinsic. Then, the photometric loss [16, 42, 25, 59] is the weighted sum of structural similarity index measure(SSIM) [51] and L1-loss [17]:

$$L_{ph} = \sum_{p\in I_t}(\alpha\frac{1-SSIM(I_s(\hat{p}), I_t(p))}{2} + (1-\alpha)|I_s(\hat{p}) - I_t(p)|) \cdot M(p) \tag{4}$$

where $\alpha = 0.85$. Following [16], two pairs of consecutive images, $[I_{t_0}, I_{t_{-1}}]$ and $[I_{t_0}, I_{t_1}]$, are used, and minimum reprojection with auto-masking is applied, which is M in Eq 4. To further encourage depth prediction aligned with edges of objects in an image, an edge-aware smoothness loss [63, 16] is computed as:

$$L_{sm} = \sum_{p\in I_t}\sum_{i\in\{x,y\}} |\partial_i d_t^*| e^{-|\partial_i I_t|} \tag{5}$$

where $d_t^* = d_t/\bar{d}_t$ is the mean-normalized inverse depth from [47].

3.3 Hardest non-boundary Triplet Loss with Minimum-distance Based Candidate Mining Strategy

Although IC-MHA module identifies task-specific representational uniqueness and cross-task spatial consistency, it can only identifies task-specific representational uniqueness within instances because of its local windowed attention mechanism. Thus, further enhancement on depth feature can be achieved by creating separation in feature space between various instances using deep metric learning techniques. In [25], Jung *et al.* propose semantics-guided triplet loss using pseudo semantic labels. Here, we name such triplet loss as boundary triplet loss, $\mathcal{L}_{BT}$. Such boundary triplet loss is effective but not sufficient since the depth feature at non-boundary region remains unrefined. This leads to higher prediction error for pixels that are away from the boundary, shown in Fig 8. Therefore, we propose a triplet loss that aims to fine-grain the depth feature at non-boundary region. However, where to sample an anchor in a non-boundary triplet requires careful design. Within an image, objects from the same semantic class might have various depth value, and one object can be separated or occluded, *e.g*, a turning

car is sliced into parts by traffic signs. Moreover, how to sample a non-boundary triplet's positives and negatives should be properly handled. Considering that instances in a scene are either static or rigidly moving, depth value or feature of distant pixels in the same objects could be different or dissimilar, *e.g.*, centers and edges of road. Thus, incorrect sampling strategy could result in overly similarity between depth feature within the same object. To overcome these two issues, we propose a *minimum-distance based candidate mining strategy* to properly sample anchors, positives and negatives for non-boundary-triplet loss.

Minimum-distance based Candidate Mining Strategy. To correctly mine non-boundary triplet, anchors and their positives should be sampled from the same instance, which is defined as a group of connected pixels with the same semantic labels. Such instances, denoted as $\mathcal{I}$, can be generated by applying labeling algorithm [13, 53] on pixels with the same semantic label, shown as (a) in Fig 5, and then over all semantic classes in an image. Concurrently, a boundary mask B is generated using patch-based sampling strategy[1] [25] to identify non-boundary region, shown as (b) in Fig 5, along with boundary anchors $\mathcal{P}_B$. For each boundary anchor, $b \in \mathcal{P}_B$, we denote its positives as $\mathcal{P}_b^+$ and its negatives as $\mathcal{P}_b^-$. With these information, our sampling strategy within each instance $\mathcal{I}_i \in \mathcal{I}$ is described as follow.

1. Mask out non-boundary pixels in each instance $\mathcal{I}_i$ and randomly sample non-boundary anchors $\mathcal{P}_{NB}^i$ with $|\mathcal{P}_{NB}^i| = N_i^s$ from them .
2. Mask out boundary pixels in the same instance $\mathcal{I}_i$, denoted as $\mathcal{P}_B^i$. Here, $\mathcal{P}_B^i$ is a subset of boundary anchors $\mathcal{P}_B$.
3. For every sampled non-boundary anchor $j \in \mathcal{P}_{NB}^i$ in Step 1, find its spatially nearest boundary anchor $b_j \in \mathcal{P}_B^i$ from Step 2.
4. Then, for each non-boundary anchor $j \in \mathcal{P}_{NB}^i$, its set of positives $\mathcal{P}_j^+$ and its set of negatives $\mathcal{P}_j^-$ are sampled from the positives, $\mathcal{P}_{b_j}^+$, and negatives, $\mathcal{P}_{b_j}^-$, of its nearest boundary anchor b_j from Step 3.

Visualizations of above process are shown as (c) and (d) in Fig 5.

In above mining process, we do not randomly sample positives from the same instance and negatives from other semantic classes because depth value within an object might change greatly. Such sampling process will lead to overly similarity between depth features within the same object. Therefore, it is more optimal to sample positives and negatives of each non-boundary anchor from its closet boundary region. The intuition behind this is that we would like to encourage non-boundary anchors' features more similar to spatially nearest boundary pixels' features and to decouple the features of non-boundary anchors from that of boundary anchors concurrently. For computational efficiency, the non-boundary anchors is sampled randomly. For each instance $\mathcal{I}_i$ at each level s, the total number of non-boundary anchors sampled is N_i^s, with $N_i^s = \frac{N_0}{4^s}\frac{|\mathcal{I}_i|}{H'W'}$. Here, N_0 is

[1] Detailed patch-based sampling process and $\mathcal{L}_{BT}$ [25] is in Sec 1 of Supplementary Material.

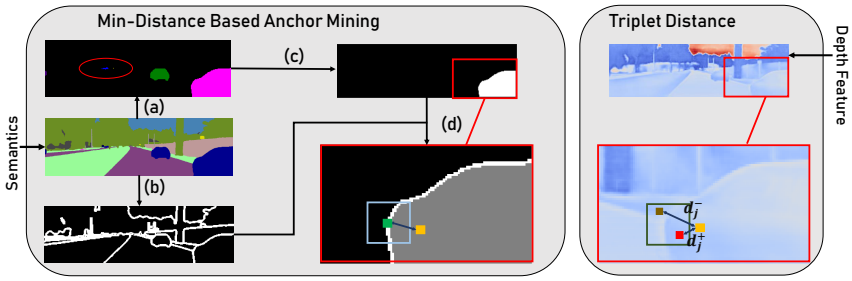

Fig. 5: An Overview of Mining strategy and triplet distance. Inside the left gray box is the minimum-distance based candidate sampling process. Here we visualize such process on one non-boundary anchor as an example: (a)generate instances; (b)generate boundary mask B; (c)Mask out one instance $\mathcal{I}_i$; (d)Randomly sample one non-boundary anchor $j \in \mathcal{P}^i_{NB}$(yellow dot), and find its closest boundary anchor b_j (green dot). The black region in blue box are the set of negatives $\mathcal{P}^-_j$, and the dark gray region in blue box are the set of positives set $\mathcal{P}^+_j$. The white curve inside of the red box is the boundary region of this instance. Inside the right gray box is the positive and negative distance in Eq 6: hardest positive(red dot) and hardest negative(brown dot) for the non-boundary anchor(yellow dot)

the total number of non-boundary anchors sampled at level $s = 0$ and $[H^{'}, W^{'}]$ is the spatial resolution of depth feature at level s.

Hardest Non-Boundary Triplet Loss. For each non-boundary anchor $j \in \mathcal{P}^i_{NB}$ in instance $\mathcal{I}_i$, a set of positives $\mathcal{P}^+_j$ and a set of negatives $\mathcal{P}^-_j$ are sampled with the process described above. Inspired by [7], we select the hardest positives and negatives to compute positive distance d^+_j and negative distance d^-_j, *i.e.* the most dissimilar positive feature in $\mathcal{P}^+_j$ and most similar negative feature in $\mathcal{P}^-_j$, shown in the right gray box in Fig 5:

$$d^+_j = \max_{j^+ \in P^+_j} (\|F_D(j) - F_D(j^+)\|_2), \quad d^-_j = \min_{j^- \in P^-_j} (\|F_D(j) - F_D(j^-)\|_2) \tag{6}$$

where F_D is the normalized depth feature, j^+ and j^- are the positives or negatives of non-boundary anchor i. Thus, the triplet margin loss for all non-boundary anchors $j \in \mathcal{P}^i_{NB}$ in one instance $\mathcal{I}_i$ is

$$\mathcal{L}(\mathcal{P}^i_{NB}) = \frac{\sum_{j \in \mathcal{P}^i_{NB}} \max(0, d^+_j + m - d^-_j)}{|\mathcal{P}^i_{NB}|} \tag{7}$$

where $m = 0.3$ is the margin for feature separation. In practice, instances generated by labeling algorithm could be false because of misclassification in pseudo-labels, shown as red circle in instance mask after process (a) in Fig 5. Thus, our

final hardest non-boundary triplet loss is the mean of $L(\mathcal{P}^i_{NB})$ over all instances, $\mathcal{I}$, whose number of non-boundary pixels is larger than a threshold δ.

$$\mathcal{L}_{NBT} = \frac{\sum_{\mathcal{I}_i \in \mathcal{I}} \mathbb{I}\{|\mathcal{I}_i| > \delta\} \cdot L(\mathcal{P}^i_{NB})}{\sum_{\mathcal{I}_i \in \mathcal{I}} \mathbb{I}\{|\mathcal{I}_i| > \delta\}} \tag{8}$$

where $\mathbb{I}$ is the indicator function. Our final training loss is the weighted sum of photometric loss $\mathcal{L}_{ph}$, edge-aware smoothness loss $\mathcal{L}_{sm}$, boundary triplet loss $\mathcal{L}_{BT}$, hardest non-boundary triplet loss $\mathcal{L}_{NBT}$ and semantic cross-entropy loss $\mathcal{L}_{CE}$:

$$\mathcal{L} = \sum_{s \in S}(\mathcal{L}_{ph} + \beta \cdot \mathcal{L}_{sm} + \gamma \cdot \mathcal{L}_{CE}) + \sum_{s \in S_{BT}} \eta \cdot \mathcal{L}_{BT} + \sum_{s \in S_{NBT}} \kappa \cdot \mathcal{L}_{NBT} \tag{9}$$

where β, γ, η, κ are control parameters and s represents the output level.

4 Experiments

4.1 Datasets.

To ensure fair comparison with previous state-of-the-art works, we conduct experiments on widely-used KITTI dataset [15]. Following [63, 1, 16], we use the Eigen split [10] for depth training and evaluation, which consists of 39,810 images for training, 4,424 images for validation and 697 images for evaluation.

For the supervision of semantic segmentation, following [25], pseudo-labels for the training and validation set of Eigen split are generated using a well-trained segmentation network [65]. To evaluate semantic segmentation, the training set of the KITTI 2015 [35] is used, which contains 200 images with fine-annotated semantic labels.

4.2 Implementation Details

The encoders of DepthSegNet and PoseNet are implemented with ResNet-18 [22] with pretrained weight from ImageNet [8] loaded at initialization. For both PoseNet and DepthSegNet, input image is resized to 192×640. In addition, for fair comparison with previous state-of-the-art works, we also implement our DepthSegNet with ResNet-50 [22] whose input image is of various resolution: 192×640 and 320×1024.

For our IC-MHA module, the number of heads is set to be four($H = 4$) at each level with size of the window for each head $w_h = [2, 2, 4, 4]$. Within each head, the expansion ratio is set as: $r = 2$. We implement of our IC-MHA module at four levels, *i.e.* $s = [0, 1, 2, 3]$.

For hyperparameters of our final training loss $\mathcal{L}$ in Eq 9, we set them as $\beta = 0.001$, $\gamma = 0.3$, $\eta = 0.1$, $\kappa = 0.1$. S and S_{BT} is set to be $\{3, 2, 1, 0\}$, and S_{NBT} is $\{1, 0\}$. Additionally, we set the threshold for pixel number in Eq 8 as: $\delta = 80$. And the total number of non-boundary anchors at $s = 0$, *i.e.* N_0, is

8000. The non-boundary mining process is only employed only during training process, not at inference time.

During the training of our network, the data preprocessing in [16] is applied. We implement our proposed method on PyTorch [37], and the Adam optimizer [26] is used with initial learning rate as 1.5×10^{-4} for 20 epoches. At epoch 10 and 15, the learning rate is decayed to 1.5×10^{-5} and 1.5×10^{-6} respectively. The batch size for training is 12.

4.3 Evaluation Metrics.

For depth prediction, we firstly set its maximum to be 80m and conduct median-scaling using ground-truth as that in [16]. Then, the depth is evaluated by seven standard metrics [16, 63, 42, 25], which are AbsRel, SqRel, RMSE, RMSElog, δ_1, δ_2, δ_3. For semantic segmentation, we evaluate it with the mean intersection over union(mIoU), which is the standard evaluation metric for this task.

Table 1: Comparison with recent state-of-the-art works in self-supervised monocular depth estimation. All methods are trained with monocular video sequences. Methods with (*) utilize semantic information

Methods	Input Res.	BackBone	lower is better				higher is better		
			AbsRel	SqRel	RMSE	RMSElog	δ_1	δ_2	δ_3
SfMLearner [63]	128 × 416	R18	0.208	1.768	6.958	0.283	0.678	0.885	0.957
SC-SfMLearner [1]	128 × 416	R18	0.137	1.089	5.439	0.217	0.830	0.942	0.975
(*)SceneNet [5]	256 × 512	DRN [56]	0.118	0.905	5.096	0.211	0.839	0.945	0.977
MonoDepth2 [16]	192 × 640	R18	0.115	0.903	4.863	0.193	0.877	0.959	0.981
(*)Guizilini *et al.* [20]	192 × 640	R18	0.117	0.854	4.714	0.191	0.873	0.963	0.981
(*)SGDepth [27]	192 × 640	R18	0.113	0.835	4.693	0.191	0.873	0.963	0.981
R-MSFM [64]	192 × 640	R18	0.112	0.806	4.704	0.191	0.878	0.960	0.981
(*)Lee *et al.* [30]	256 × 832	R18	0.112	0.777	4.772	0.191	0.872	0.959	0.982
Poggi *et al.* [40]	192 × 640	R18	0.111	0.863	4.756	0.188	0.881	0.961	0.982
Patil *et al.* [38]	192 × 640	R18	0.111	0.821	4.650	0.187	0.883	0.961	0.982
(*)SAFENet [6]	192 × 640	R18	0.112	0.788	4.582	0.187	0.878	0.963	0.983
Zhao *et al.* [59]	256 × 832	R18	0.113	0.704	4.581	0.184	0.871	0.961	0.984
HRDepth [34]	192 × 640	R18	0.109	0.792	4.632	0.185	0.884	0.962	0.983
Wang *et al.* [49]	192 × 640	R18	0.109	0.779	4.641	0.186	0.883	0.962	0.982
(*)FSRE [25]†	192 × 640	R18	0.107	0.730	4.530	0.182	**0.886**	0.964	**0.984**
(*)FSRE [25]	192 × 640	R18	0.105	0.722	4.547	0.182	**0.886**	0.964	**0.984**
(*)**Ours**	192 × 640	R18	**0.104**	**0.690**	**4.473**	**0.179**	**0.886**	**0.965**	**0.984**
(*)SGDepth [27]	192 × 640	R50	0.112	0.833	4.688	0.190	0.884	0.961	0.981
(*)Guizilini *et al.* [20]	192 × 640	R50	0.113	0.831	4.663	0.189	0.878	**0.971**	0.983
MonoDepth2 [16]	192 × 640	R50	0.110	0.831	4.642	0.187	0.883	0.962	0.982
(*)Li *et al.* [31]	192 × 640	R50	0.103	0.709	4.471	0.180	0.892	0.966	0.984
(*)FSRE [25]	192 × 640	R50	**0.102**	0.675	4.393	0.178	**0.893**	0.966	0.984
(*)**Ours**	192 × 640	R50	**0.102**	**0.656**	**4.339**	**0.175**	0.892	0.967	**0.985**
PackNet [19]	375 × 1224	PackNet	0.104	0.758	4.386	0.182	0.895	0.964	0.982
FeatDepth [42]	320 × 1024	R50	0.104	0.729	4.481	0.179	0.893	0.965	0.984
(*)Guizilini *et al.* [20]	375 × 1224	PackNet	0.100	0.761	4.270	0.175	**0.902**	0.965	0.982
(*)**Ours**	320 × 1024	R50	**0.099**	**0.624**	**4.165**	**0.171**	**0.902**	**0.969**	**0.986**

† We re-trained [25] with its official implementation, since no pretrained model is available.

4.4 Experiment Results and Ablation Study

Comparison with previous state-of-the-art methods. Comparison with recent state-of-the-art results is shown in Table 1. The table shows that our method achieves the state-of-the-art performance on KITTI Eigen test split. Specifically, our proposed method outperforms previous state-of-the-art work significantly on SqRel, RMSE, RMSElog. For AbsRel, δ_1, δ_2, δ_3, our proposed method yields comparable or better performance than previous state-of-the-art methods. In addition, our method with low input resolution and lighter backbone outperforms some previous state-of-the-art approaches with higher resolution [59, 30] or heavier backbone [27, 20, 19]. See Table 4 for detailed timing and parameter number of ours and previous methods. The testing device is NVidia V100 GPU. Furthermore, our proposed method with high resolution input and deeper network achieves significant improvement over previous state-of-the-art methods, which indicates that our proposed approach gains performance boost with better backbone network(ResNet-50).

Table 2: Ablations on proposed IC-MHA module and non-boundary triplet loss $\mathcal{L}_{NBT}$. IC-MHA* represents IC-MHA module with intra-task attention only.

CMA	$\mathcal{L}_{BT}$	IC-MHA*	IC-MHA(All)	$\mathcal{L}_{NBT}$	AbsRel	SqRel	RMSE	RMSElog	δ_1	δ_2	δ_3
✓	✓				0.107	0.730	4.530	0.182	0.886	0.964	**0.984**
	✓	✓			0.106	0.731	4.527	0.181	0.886	0.964	**0.984**
	✓		✓		0.105	0.734	4.516	0.180	**0.887**	**0.965**	**0.984**
	✓		✓	✓	**0.104**	**0.690**	**4.473**	**0.179**	0.886	**0.965**	**0.984**

Table 3: The segmentation result of proposed methods against baseline [25]

Methods	mIoU
FSRE† [25]	55.8
Ours	56.3

Table 4: Inference time and parameter number of ours against previous methods

	Time(ms)	Param.#(M)
FSRE[25](R18)	10	28.6M
Ours(R18)	12	30.3M
Ours(R50)	31	45.5M
[20](PackNet)	60	70M

Table 5: Ablations on Different Loss term on proposed IC-MHA module.

$\mathcal{L}_{ph} + \mathcal{L}_{sm}$	$\mathcal{L}_{CE}$	$\mathcal{L}_{BT}$	$\mathcal{L}_{NBT}$	AbsRel	SqRel	RMSE	RMSElog	δ_1	δ_2	δ_3
✓	✓			0.110	0.794	4.610	0.187	0.879	0.962	0.982
✓	✓	✓		0.105	0.734	4.516	0.180	**0.887**	**0.965**	**0.984**
✓	✓		✓	0.105	0.727	4.498	0.180	0.885	**0.965**	**0.984**
✓	✓	✓	✓	**0.104**	**0.690**	**4.473**	**0.179**	0.886	**0.965**	**0.984**

Ablation Study. Ablations on our proposed IC-MHA module and our non-boundary triplet loss $\mathcal{L}_{NBT}$ is shown in Table 2. We compare our proposed methods with our baseline [25], which consists of CMA module and boundary triplet loss $\mathcal{L}_{BT}$ to emphasize cross-task correlation. Since no pretrained model is available on the official implementation of [25], we re-train the model using its official implementation for multiple times and take the best result. The experiment result verifies the effectiveness of our proposed IC-MHA module and non-boundary triplet loss $\mathcal{L}_{NBT}$. Also, the result in Table 2 shows effectiveness of our proposed intra-task local attention. To further demonstrate the effectiveness of our proposed method, the visualization of feature heatmap of IC-MHA module and CMA module is shown in Fig 6. Such visualization is generated via PCA decomposition by normalizing the summation of top principle channels, who contributes 90% totally to feature map. It shows that our refined depth feature is much more smoothing within instances, and it is more aligned with the actual depth distribution. The heatmap of cross-task attention and intra-task attention is shown in Fig 7. It suggests that cross-task attention is consistent with semantic feature, while intra-task attention is consistent with depth feature. The ablation study on different loss term on IC-MHA module is shown in Table 5. It shows that IC-MHA module gains performance boost from $\mathcal{L}_{BT}$ or $\mathcal{L}_{NBT}$. These two loss terms together achieve best result.

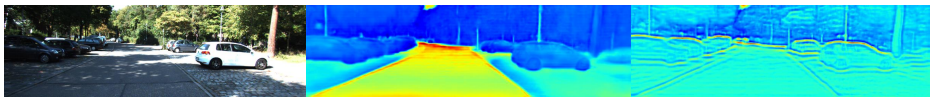

Fig. 6: Depth feature heatmap visualization: (a)Colored image. (b)Depth feature heatmap of IC-MHA module. (c)Depth feature heatmap of CMA module [25].

Additionally, in Table 3, we compare the segmentation result of our methods with that of baseline. The mean intersection-of-union (mIoU) of proposed method is better than that of baseline by 0.5. This verifies that addressing representational uniqueness of task-specific feature in IC-MHA module can improve prediction of both tasks: depth and semantic segmentation. Extra ablation study on hyperparameters of IC-MHA module and $\mathcal{L}_{NBT}$ is included in Sec 5 of Supplementary Material.

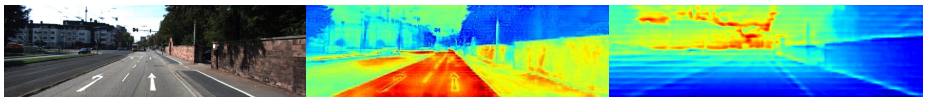

Fig. 7: Heatmap of cross-task attention feature(middle) and intra-task attention feature(right)

4.5 Qualitative Results.

Qualitative comparison between our proposed method and previous state-of-the-art work, FSRE [25] is shown in Fig 8. Error distribution[2] uses absolute error between reference depth and predicted depth. The maximum of the absolute error map is set to be 10, and then it is rescaled to $[0, 1]$. The figure proves that our proposed method not only preserves object boundaries but improves estimation at non-boundary region of each object as well. More qualitative examples are shown in Supplementary Material.

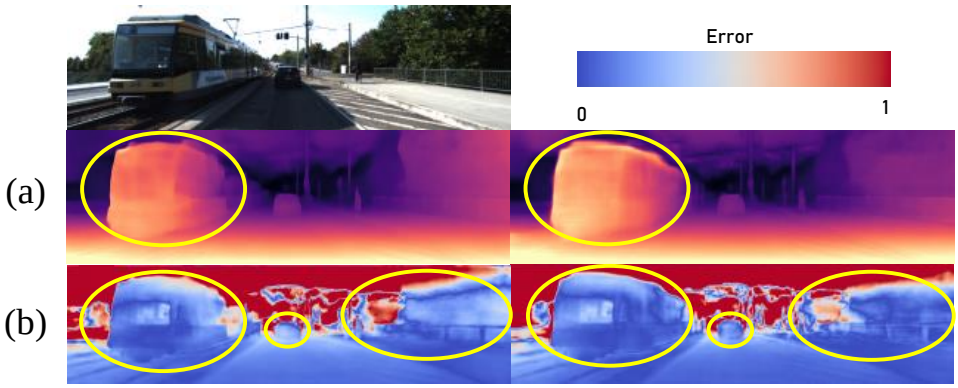

Fig. 8: Qualitative results of depth estimation. (a)Depth output of [25](left) and ours(right). (b)Error map of [25](left) and ours(right).

5 Conclusions

In this work, we propose a novel method for self-supervised monocular depth estimation by emphasizing task-specific uniqueness in feature space. Specifically, our proposed IC-MHA module exploits more fine-grained features by addressing representational uniqueness of task-specific feature within instances in parallel with cross-task spatial consistency. Additionally, the proposed hardest non-boundary triplet loss further enhances depth feature by addressing its uniqueness between instances, which is a full refinement on depth feature of all regions in an image. Our whole method is end-to-end trainable and achieves state-of-the-art performance on KITTI Eigen test split.

Acknowledgements This work is supported by National Key R&D Program of China(Grant No.2020AAA010400X).

[2] Considering ground-truth depth is sparse, we use estimation of top-performance supervised depth network [29] as reference

References

1. Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.M., Reid, I.: Unsupervised scale-consistent depth and ego-motion learning from monocular video. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
4. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International Conference on Machine Learning. pp. 1691–1703. PMLR (2020)
5. Chen, P.Y., Liu, A.H., Liu, Y.C., Wang, Y.C.F.: Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2619–2627 (2019)
6. Choi, J., Jung, D., Lee, D., Kim, C.: Safenet: Self-supervised monocular depth estimation with semantic-aware feature extraction. In: Thirty-fourth Conference on Neural Information Processing Systems, NIPS 2020. NeurIPS (2020)
7. Choy, C., Park, J., Koltun, V.: Fully convolutional geometric features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8958–8966 (2019)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
10. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014)
11. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE transactions on pattern analysis and machine intelligence **40**(3), 611–625 (2017)
12. Farooq Bhat, S., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4008–4017 (2021)
13. Fiorio, C., Gustedt, J.: Two linear time union-find strategies for image processing. Theoretical Computer Science **154**(2), 165–181 (1996)
14. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep Ordinal Regression Network for Monocular Depth Estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
15. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
16. Godard, C., Aodha, O.M., Firman, M., Brostow, G.: Digging into self-supervised monocular depth estimation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3827–3837 (2019)

17. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
18. Goel, K., Srinivasan, P., Tariq, S., Philbin, J.: Quadronet: Multi-task learning for real-time semantic depth aware instance segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 315–324 (January 2021)
19. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
20. Guizilini, V., Hou, R., Li, J., Ambrus, R., Gaidon, A.: Semantically-guided representation learning for self-supervised monocular depth. In: International Conference on Learning Representations (ICLR) (April 2020)
21. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D.: A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2022)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
23. Hu, J., Lu, J., Tan, Y.P.: Discriminative deep metric learning for face verification in the wild. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1875–1882 (2014)
24. Jiang, H., Ding, L., Sun, Z., Huang, R.: Unsupervised monocular depth perception: Focusing on moving objects. IEEE Sensors Journal **21**(24), 27225–27237 (2021)
25. Jung, H., Park, E., Yoo, S.: Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12642–12652 (October 2021)
26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
27. Klingner, M., Termöhlen, J.A., Mikolajczyk, J., Fingscheidt, T.: Self-Supervised Monocular Depth Estimation: Solving the Dynamic Object Problem by Semantic Guidance. In: European Conference on Computer Vision (ECCV) (2020)
28. Kulis, B.: (2013)
29. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019)
30. Lee, S., Im, S., Lin, S., Kweon, I.S.: Learning monocular depth in dynamic scenes via instance-aware projection consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2021)
31. Li, R., He, X., Xue, D., Su, S., Mao, Q., Zhu, Y., Sun, J., Zhang, Y.: Learning depth via leveraging semantics: Self-supervised monocular depth estimation with both implicit and explicit semantic guidance (2021)
32. Li, R., Mao, Q., Wang, P., He, X., Zhu, Y., Sun, J., Zhang, Y.: Semantic-guided representation enhancement for self-supervised monocular trained depth estimation (2020)
33. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021)
34. Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., Chen, X., Yuan, Y.: Hr-depth: High resolution self-supervised monocular depth estimation. arXiv preprint arXiv:2012.07356 **6** (2020)

35. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
36. Mur-Artal, R., Montiel, J.M.M., Tardós, J.D.: Orb-slam: A versatile and accurate monocular slam system. IEEE Transactions on Robotics **31**(5), 1147–1163 (2015)
37. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
38. Patil, V., Van Gansbeke, W., Dai, D., Van Gool, L.: Don't forget the past: Recurrent depth estimation from monocular video. IEEE Robotics and Automation Letters **5**(4), 6813–6820 (2020)
39. Pire, T., Fischer, T., Castro, G., De Cristóforis, P., Civera, J., Jacobo Berlles, J.: S-PTAM: Stereo Parallel Tracking and Mapping. Robotics and Autonomous Systems (RAS) **93**, 27 – 42 (2017)
40. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3224–3234 (2020)
41. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)
42. Shu, C., Yu, K., Duan, Z., Yang, K.: Feature-metric loss for self-supervised learning of depth and egomotion. In: ECCV (2020)
43. Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., Moreno-Noguer, F.: Discriminative learning of deep convolutional feature point descriptors. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 118–126 (2015)
44. Song, H.O., Jegelka, S., Rathod, V., Murphy, K.: Deep metric learning via facility location. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2206–2214 (2017)
45. Sun, Z., Cao, S., Yang, Y., Kitani, K.M.: Rethinking transformer-based set prediction for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3611–3620 (2021)
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
47. Wang, C., Buenaposada, J.M., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2022–2030 (2018)
48. Wang, J., Zhou, F., Wen, S., Liu, X., Lin, Y.: Deep metric learning with angular loss. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2612–2620 (2017)
49. Wang, L., Wang, Y., Wang, L., Zhan, Y., Wang, Y., Lu, H.: Can scale-consistent monocular depth be learned in a self-supervised scale-invariant manner? In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12727–12736 (October 2021)
50. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8741–8750 (2021)
51. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)

52. Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: Self-supervised multi-frame monocular depth. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1164–1174 (2021)
53. Wu, K., Otoo, E., Shoshani, A.: Optimizing connected component labeling algorithms. In: Medical Imaging 2005: Image Processing. vol. 5747, pp. 1965–1976. SPIE (2005)
54. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: European conference on computer vision. pp. 467–483. Springer (2016)
55. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: CVPR (2018)
56. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 636–644 (2017)
57. Zhan, H., Garg, R., Weerasekera, C.S., Li, K., Agarwal, H., Reid, I.M.: Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 340–349 (2018)
58. Zhao, H., Jia, J., Koltun, V.: Exploring self-attention for image recognition. In: CVPR (2020)
59. Zhao, W., Liu, S., Shu, Y., Liu, Y.J.: Towards better generalization: Joint depth-pose learning without posenet. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
60. Zheng, M., Gao, P., Zhang, R., Li, K., Wang, X., Li, H., Dong, H.: End-to-end object detection with adaptive clustering transformer. arXiv preprint arXiv:2011.09315 (2020)
61. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021)
62. Zhou, H., Greenwood, D., Taylor, S.: Self-supervised monocular depth estimation with internal feature fusion. In: British Machine Vision Conference (BMVC) (2021)
63. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6612–6619 (2017)
64. Zhou, Z., Fan, X., Shi, P., Xin, Y.: R-msfm: Recurrent multi-scale feature modulation for monocular depth estimating. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12777–12786 (October 2021)
65. Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8848–8857 (2019)
66. Zou, Y., Luo, Z., Huang, J.B.: Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In: European Conference on Computer Vision (2018)