Self-supervised Monocular Depth Estimation with Semantic-aware Depth Features

Jaehoon Choi¹*, Dongki Jung¹,², Donghwan Lee¹, Changick Kim²

¹NAVER LABS
²Korea Advanced Institute of Science and Technology
{jaehoon.c, donghwan.lee}@naverlabs.com, {jdk9405, changick}@kaist.ac.kr

Abstract. Self-supervised monocular depth estimation has emerged as a promising method because it does not require ground truth depth during training. As an alternative for ground truth depth, a photometric loss enables to provide self-supervision on depth prediction by matching the input image frames. However, the photometric loss has various problems, resulting in less accurate depth values compared to supervised approaches. In this paper, we propose to leverage semantic information to overcome the limitations of the photometric loss. Our key idea is to exploit semantic-aware depth features which integrate the semantic and geometric knowledge. We introduce a multi-task approach to incorporate semantic-awareness into the depth feature representations. Our proposed modules for multi-task learning can be widely adopted to self-supervised models based on both stereo images and monocular video sequences. Experiments on the KITTI dataset demonstrate that our methods compete or outperform the state-of-the-art algorithms. Furthermore, extensive experiments show that semantic-aware depth features are robust to a wide array of conditions, such as low-light or adverse weather.

Keywords: Self-supervised learning, Monocular depth estimation, Semantic segmentation, Multi-task learning

1 Introduction

Monocular Depth Estimation, aiming at producing dense depth estimates from a single image, is an important task for autonomous driving, augmented reality, and robotics. Most supervised methods [12, 28, 13] show that Convolutional Neural Networks (CNNs) are powerful tools to produce dense depth images. Nevertheless, collecting large-scale dense depth maps for groundtruth is very difficult due to data sparsity and expensive depth sensing devices [16], such as LiDAR. In light of this, self-supervised monocular depth estimation [15, 17, 54, 18] has gained attention in recent years because it does not require image and ground truth pairs. Self-supervised depth learning is a training method to regress the depth values via the error function, named photometric loss. This function computes errors between the reference image and the geometrically reprojected image from other

* First two authors contributed equally.
Fig. 1. Example of monocular depth estimation based on self-supervision from either stereo images (top row) or monocular video sequences (bottom row). The third column illustrates the results of Monodepth2 [18] (top) and Monodepth [17] (bottom).

viewpoints. The reference and the image of other viewpoints can be either a calibrated pair of left and right images in stereo [15,17] or adjacent frames with the relative camera pose in a video sequence [54,18]. However, previous studies [26,53,18,15] show that the brightness change of pixels, low texture regions, repeated patterns, and occlusions can cause differences in the photometric loss distribution and hinder the training. To address such limitations of the photometric loss, we propose a novel method, which fuses the feature level semantic information to geometric representations. Depth features guided by semantic supervision can involve the spatial context of the input image. This information serves as complementary knowledge to interpret the 3D Euclidean space and improves the depth estimation performance. For example, Fig. 1 shows that our method has a consistent depth range for each instance. In the first row, the distorted car shape of the baseline prediction is recovered with ours. Despite these advantages, a general method to learn semantic-aware depth features has not been explored widely in the current self-supervised monocular depth estimation approaches.

To learn semantic-aware depth features, we investigate a multi-task learning approach that imposes semantic supervision from supervised segmentation training to self-supervised depth training. However, multi-task learning (MTL) often suffers from task interference in that features learned to perform one task may not be suitable for others [27]. Thus, it is essential to distinguish the features between the task-specific and task-shared properties, which represent whether or not to share information for the different tasks. We present modules to obtain semantic-aware depth features by taking only the portions of the semantic features that are helpful for accurate depth estimation. In the encoder stage, we exploit the Residual Adapter [43] and the Squeeze and Excitation module [20] to learn adaptive features for each task. We demonstrate that these simple modules improve the performance of depth estimation. Besides, in the decoder stage, we introduce cross propagation units inspired by [37,22] and affinity propagation units to integrate the intermediate representation from both tasks. With these two auxiliary modules, the depth decoder can take into account the intermediate representation of the semantic-awareness in both spatial and channel dimensions.
Our proposed strategy can be easily extended to both types of self-supervised approaches; video sequences and stereo images. Furthermore, we experimentally validate the superiority of semantic-aware depth features under low light and adverse weather conditions. In summary, the contributions of this paper are shown as follows:

– We propose a multi-task approach to obtain semantic-aware depth features in self-supervised monocular depth estimation networks.
– We demonstrate that the obtained semantic-aware depth features can overcome drawbacks of the photometric loss and allow our network to improve monocular depth estimation performance.
– Our method achieves state-of-the-art results on the KITTI dataset \cite{16}, and extensive experiments show that our method is more robust to various adverse conditions than current algorithms.

2 Related Work

2.1 Self-supervised Training with Stereo Vision

Depth estimation from a single image is an ill-posed problem since one 2D image can be created from countless 3D scenes. Supervised monocular depth estimation models \cite{12,35,28,24,46,7} require a large-scale groundtruth dataset, which is expensive to collect and has different characteristics depending on the sensors. To mitigate this issue, Garg et al. \cite{15} and Godard et al. \cite{17} propose self-supervised training methods for monocular depth estimation. These approaches exploit the warping function to transfer the coordinates of the left image to the right image plane. In particular, \cite{17} design a photometric loss combining SSIM \cite{49} with $L_1$ term and geometric warping using the spatial transformer network \cite{21}. These ideas are extended to the trinocular assumption \cite{39} or the generative adversarial loss function \cite{138}.

2.2 Self-supervised Training with Monocular Video Sequences

Zhou et al. \cite{54} propose a method to perform depth estimation through camera ego-motion from video sequence images. Instead of left-right consistency, this method computes the photometric loss by reprojecting adjacent frames to a current frame with predicted depth and relative camera pose. Monodepth2 \cite{18} enhance performance using techniques such as minimizing the minimum of reprojection error and auto-masking. Multiple studies depend on one assumption that image frames consist of rigid scenes, i.e., appearance change among context is caused by the camera motion. For this reason, \cite{54} applies network predicted masks to moving objects, and \cite{18} compares the per-pixel loss to ignore regions where this assumption is violated. Besides, many studies have been conducted using additional cues to improve the quality of regression, such as surface normal \cite{51}, optical flow \cite{32,41}, and edges \cite{50}. Recently, the methods in \cite{28} apply geometric constraints as well as the photometric loss to achieve state-of-the-art performance.
2.3 Multi-task Learning

MTL has been developed for a single CNN model to handle a multitude of tasks and yield better results in all of them. Previous MTL methods based on CNNs commonly utilize parameter sharing, which share some layers across all tasks and add task-specific layers on the top of the shared networks \cite{37,31,27}. These naive approaches have two limitations. First, since these methods combine all the task-specific losses without considering optimal weight parameters, the model cannot learn multiple objectives properly. Thus, some papers \cite{9,23,45} propose ways to assign the weights to balance each task. Second, task-specific features may discourage the network from performing other tasks. Alternative studies are presented to learn task-shared features and task-specific features, respectively. In \cite{30}, task-specific attention modules allow the shared network to achieve this goal. Maninis et al. \cite{34} also apply the attention mechanisms, such as Squeeze and Excitation blocks \cite{20} and Residual Adapters \cite{42,43} to calibrate intermediate features. These approaches enable the separate learning of task-specific and task-shared features.

2.4 Self-supervised Training with Semantic Segmentation

Although semantic supervision is helpful for self-supervised monocular depth estimation, as far as we know, there are only a few works that handle this aspect. For self-supervision from stereo pairs, Ramirez et al. \cite{40} utilize the shared encoder and separate decoders to train both tasks jointly. Chen et al. \cite{6} design a left-right semantic consistency and semantics-guided smoothness regularization showing that semantic understanding strengthens the depth prediction accuracy. For video sequence models, some previous works \cite{3,36} also utilize information from either semantic or instance segmentation masks for moving objects in the frames. In contrast to these works, we aim to learn feature level semantic-awareness for monocular depth estimation. To the best of our knowledge, this is the first attempt to utilize the semantic-aware depth features to overcome the problems of self-supervised monocular depth estimation.

3 Proposed Approach

3.1 Motivation

In this section, we discuss the mechanism of photometric loss and its limitations. Besides, we explain the reason why we exploit semantic supervision to overcome the problems.

Photometric Loss for Self-supervision. Self-supervised monocular depth estimation relies on the photometric loss through warping between associated frames, $I_m$ and $I_n$. These two images are sampled from the left-right pair in stereo vision or the adjacent time frames in the monocular video sequence. The
photometric loss with SSIM \cite{49} is formulated as follows:

\[
L_{photo} = \frac{1}{N} \sum_{p \in N} \left( \alpha \frac{1 - \text{SSIM}_{mn}(p)}{2} + (1 - \alpha) \| I_m(p) - I'_m(p) \| \right),
\]

where \( I'_m \) is the arranged image by warping \( I_n \) with the predicted depth, \( N \) is the number of valid points that are successfully projected, and \( \alpha \) is 0.85. In the case of video sequence model, camera pose and intrinsic parameters are included in the warping process. For more details, please refer to the supplementary material. However, this loss has a severe drawback that depth regression from RGB images is vulnerable to environmental changes. We hypothesize that depth features jointly trained by semantic segmentation, termed semantic-aware depth features, are capable of leveraging semantic knowledge to guide the depth estimation. Therefore, we propose semantic supervision to solve the issues of the photometric loss through multi-task learning.

**Semantic Supervision.** Semantic-awareness can give prior knowledge that if certain 3D points are projected to adjacent pixels with the same semantic class, those points should locate in similar positions in the 3D space. Besides, even where the RGB values are indistinguishable, understanding the spatial context from the semantic information can lead to the individual characteristics of the pixels in that region.

In order to guide the geometric reconstruction by the feature level of semantics, we design a method to learn two tasks through joint training rather than simply using segmentation masks as input. For the supervised framework in the semantic segmentation task, pre-trained DeepLabv3+ \cite{5} is used to prepare pseudo labels of semantic masks, and the loss function is cross-entropy.

### 3.2 Network Architecture

Without a direct association between tasks, task interference can occur, which can corrupt each task-specific feature. We propose a network with the parameter sharing that two tasks share an encoder and have each decoder branch. Therefore, the task-specific schemes are designed to prevent corruption in single encoder, and each subnetwork for the decoders has task-sharing modules to make synergy between tasks.

**Encoder.** To avoid interference between the tasks of depth estimation and segmentation, we build the encoder using three techniques of \cite{34}, shown in Fig. 2. First, the Squeeze and Excitation (SE) block \cite{20} inserts global average pooled features into a fully connected layer and generates activated vectors for each channel via a sigmoid function. The vectors that pass through SE modules are multiplied with the features and give attention to each channel. We allocate different task-dependent parameters so that SE blocks can possess distinct characteristics. Second, Residual Adapters (RA) \cite{43}, ensuring a small number of extra parameters that can have task-specific attribute and rectify the shared
features, are added to existing residual layers:
\[ L_T(x) = x + L(x) + RA_T(x), \]
where \( x \) is processed features and \( T \in \{ \text{Depth, Seg} \} \). \( L(\cdot) \) and \( RA_T(\cdot) \) denote a residual layer and a task-specific Residual Adapter of task \( T \), respectively. Third, we obtain task-invariant features through batch normalization per individual tasks as it exploits calculated statistics which have task-dependent properties.

**Decoder.** As illustrated in Fig. 3, we design two separate decoders for each task. The separate decoders are allowed to learn task-specific features, but making it difficult to exploit other task’s features. We have experimented with two information propagation approaches to handling this issue. The first approach is inspired by the success of the sharing units between two task networks in [37,22]. Instead of weighted parameters suggested by previous works, we utilize 1×1 convolutions \( H_{1x1}^1(\cdot) \), \( B_{1x1}^1(\cdot) \) to share intermediate representations from the other task. It is worth mentioning that the 1×1 convolutions with stride 1 only perform feature modulations across channel dimensions. Before upsampling layers, we add \( H_{1x1}^1(\cdot) \), \( B_{1x1}^1(\cdot) \) that enable both the decoders to share intermediate features automatically, rather than tuning parameters for every features manually. Also, we adopt 1×1 convolutional shortcut \( H_{2x1}^1(\cdot) \), \( B_{2x1}^1(\cdot) \) to reduce the negative effect of propagation interruption [22], meaning that features propagated from one task interfere with performing each other task. Given a segmentation feature \( s_t \) and depth feature \( d_t \), task-shared features \( s_{t+1} \) and \( d_{t+1} \) can be obtained as:
\[ d_{t+1} = d_t + H_{1x1}^1(s_t) + H_{2x1}^1(d_t), \quad s_{t+1} = s_t + B_{1x1}^1(d_t) + B_{2x1}^1(s_t). \]
We refer to this module as the cross propagation unit (CPU).

The second approach is to propagate affinity information from segmentation to depth estimation. Since all the above mentioned sharing units are composed of 1×1 convolutions, the depth decoder is not able to fuse the features
Fig. 3. Overview of the proposed framework. In top part, our network consists of one shared encoder and two separate decoders for each task. This network can take either monocular video sequences or stereo images for self-supervised training. The bottom part shows the proposed modules to propagate information between two different tasks in order to learn semantic-aware depth features. See the detailed architecture in the supplementary material.

at different spatial locations or learn semantic affinity captured by the segmentation decoder. Thanks to the feature extraction capability of CNN, the high-dimension features from the segmentation decoder are used to compute the semantic affinity information. To learn non-local affinity matrix, we first feed segmentation feature $s_t$ into two $1 \times 1$ convolution layers $K^{1 \times 1}(\cdot)$ and $F^{1 \times 1}(\cdot)$, where $K^{1 \times 1}(s_t), F^{1 \times 1}(s_t) \in \mathbb{R}^{C \times H \times W}$. Here, H, W, and C denote height, width, and the number of channels of the feature. After reshaping them to $\mathbb{R}^{C \times H \times W}$, we perform a matrix multiplication between transpose of $F^{1 \times 1}(s_t)$ and $K^{1 \times 1}(s_t)$. By applying the softmax operation, the affinity matrix $A \in \mathbb{R}^{H \times W \times H \times W}$ can be formulated as:

$$a_{j,i} = \frac{\exp(F^{1 \times 1}(s_t)^T \cdot K^{1 \times 1}(s_t))}{\sum_{i=1}^{H \times W} \exp(F^{1 \times 1}(s_t)^T \cdot K^{1 \times 1}(s_t))},$$

where $a_{j,i}$ is the affinity propagation value at location $j$ from the $i$-th region, and $T$ is the transpose operation. Different from a non-local block [48], the
obtained semantic affinity matrix is propagated to the depth features to transfer a semantic correlation of pixel-wise features. We conduct a matrix multiplication between depth features from $G^{1 \times 1}()$ and semantic affinity matrix $A$. Then we can obtain depth features guided by the semantic affinity matrix. To mitigate the propagation interruption [22], we add the original depth feature to the result of affinity propagation. The affinity propagation process can be expressed as

$$d_{t+1} = BN(P^{1 \times 1}(AG^{1 \times 1}(d_t))) + d_t,$$

(5)

where $P^{1 \times 1}$ and $BN$ are a $1 \times 1$ convolution layer and the batch normalization layer. This module is named as the affinity propagation unit (APU). This spatial correlation of semantic features is significant to estimate depth accurately in the self-supervised regime.

### 3.3 Loss Functions

Our loss function consists of supervised and self-supervised loss terms. For semantic supervision, either pseudo labels or groundtruth annotations are available. We define the semantic segmentation loss $L_{\text{seg}}$ using cross entropy. As described above, we use photometric loss $L_{\text{photo}}$ in 3.1 for self-supervised training. In addition, to regularize the depth in low texture or homogeneous region of the scene, we adopt the edge-aware depth smoothness loss $L_{\text{smooth}}$ in [17]. The overall loss function is formulated as follows,

$$L_{\text{tot}} = L_{\text{photo}} + \lambda_{\text{smooth}} L_{\text{smooth}} + \lambda_{\text{seg}} L_{\text{seg}},$$

(6)

where $\lambda_{\text{seg}}$ and $\lambda_{\text{smooth}}$ are the weighting terms selected through grid search. Our network can be trained in an end-to-end manner. All the parameters in task-shared modules of the encoder, APU and CPU are trained by back-propagation of $L_{\text{tot}}$, while the parameters in task-specific modules of the encoder and decoders are learned by the gradient of the task-specific loss, namely either $L_{\text{seg}}$ or $L_{\text{photo}} + L_{\text{smooth}}$. For instance, all the specific layers for the segmentation task in both the encoder and the decoder are not trained with $L_{\text{photo}}$ and $L_{\text{smooth}}$, and vice versa.

Furthermore, for self-supervised training with the monocular video sequence, we train an additional pose network and the proposed encoder-decoder model simultaneously. The pose network follows the same training protocols described in Monodepth2 [18]. We also incorporate techniques in [18], including auto-masking, applying per-pixel minimum reprojection loss, and depth map upsampling to obtain improved results.

### 4 Experiments

In this section, we evaluate the proposed approach on self-supervised monocular depth estimation that includes both stereo and sequence scenes, and compare with other state-of-the-art methods.
Table 1. Quantitative results on the KITTI 2015 [16] by the split of Eigen. * indicates updated results from Github. D is supervised training with depth labels, and M is the self-supervised method with video sequence input. We additionally show better performance on high resolution 1024×320. This table does not include online refinement performance for a fair comparison.

| Method      | Train | Lower is better. | Higher is better. |
|-------------|-------|------------------|-------------------|
|             |       | Abs Rel Sq Rel RMSE RMSE log δ < 1.25 δ < 1.25² δ < 1.25³ |
| Eigen [12]  | D     | 0.203 1.548 6.307 0.282 | 0.702 0.890 0.957 |
| Liu [29]    | D     | 0.201 1.584 6.471 0.273 | 0.680 0.898 0.967 |
| DORN [13]   | D     | **0.072 0.307 2.727 0.120** | 0.932 0.984 0.994 |
| Zhou [54]*  | M     | 0.183 1.595 6.709 0.270 | 0.734 0.902 0.959 |
| Yang [51]   | M     | 0.182 1.481 6.501 0.267 | 0.725 0.906 0.963 |
| LEGO [50]   | M     | 0.162 1.352 6.276 0.252 | - - - |
| Mahjourian [33] | M     | 0.163 1.240 6.220 0.250 | 0.762 0.916 0.968 |
| GeoNet [52]* | M     | 0.149 1.060 5.567 0.226 | 0.796 0.935 0.975 |
| DDVO [37]   | M     | 0.151 1.257 5.583 0.228 | 0.810 0.936 0.974 |
| DF-Net [50] | M     | 0.150 1.124 5.507 0.223 | 0.806 0.933 0.973 |
| EPC++ [35]  | M     | 0.141 1.029 5.350 0.216 | 0.816 0.941 0.976 |
| Struct2depth [3] | M     | 0.141 1.026 5.291 0.215 | 0.816 0.945 0.979 |
| SC-SM Learner [2] | M     | 0.137 1.089 5.439 0.217 | 0.830 0.942 0.975 |
| CC [41]     | M     | 0.140 1.070 5.326 0.217 | 0.826 0.941 0.975 |
| SIGNet [36] | M     | 0.133 0.905 5.181 0.208 | 0.825 0.947 0.981 |
| GLNet [3]   | M     | 0.135 1.070 5.230 0.210 | 0.841 0.948 0.980 |
| Monodepth2 [18] | M     | 0.115 0.903 4.863 0.193 | **0.877** 0.959 0.981 |
| Ours        | M     | **0.110 0.743 4.489 0.183** | 0.879 0.964 0.984 |
| Ours (1024×320) | M     | **0.110 0.743 4.489 0.183** | 0.879 0.964 0.984 |

4.1 Experimental Settings

**Dataset.** We used the KITTI dataset [16] as in Zhou et al. [54], which consists of 39,810 triple frames for training and 4,424 images for validation in the sequence model. In stereo model, we used Eigen [12]’s splits of 22,600 left-right pairs for training and 888 pairs for validation. The test split is composed of 697 images in both models. These images have no segmentation labels, so we prepared semantic masks of 19 categories from DeepLabv3+ pre-trained on Cityscapes [10]. The pre-trained model attains the semantic segmentation performance of mIoU 75% on the KITTI validation set. To show that our method has robust performance in the adverse weather, we experimented with Virtual KITTI (vKITTI) [14], which is synthetic data composed of various weather conditions in five video sequences and 11 classes of semantic labels. We divided vKITTI into six weather conditions as given in [14]. The training set has relatively clean 8464 sequence triplets that belong to morning, sunset, overcast, and clone. The 4252 fog and clone images, which are challenging because of very different environments to the training set, were tested to show each performance. The predicted depth range of KITTI and vKITTI is clipped to 80m to match the Eigen following [18].

**Implementation Details.** We implemented the proposed deep model using PyTorch. We built our encoder based on the ResNet-18 [19] backbone with SE modules, and bridged to the decoder with skip connections based on the general U-Net architecture [44]. Each layer of the encoder was pre-trained on ImageNet.
Fig. 4. Qualitative results on the KITTI Eigen split. Our models in the last row produce better visual outputs, especially the sharpest boundaries of the objects. In the second row, Semantic denotes the segmentation results from DeepLabv3+ on the test set.

[11], while parameters in the task-specific modules of the encoder, two decoders, CPU and APU were randomly initialized. In terms of training with monocular video sequence, we used a pose network based on ResNet-18 and pre-trained it using ImageNet. Architectural details of the pose network follow Monodepth2 [18]. We trained our model in a batch size of 8 using Adam optimizer [25]. We used the learning rate of $10^{-4}$ and the weight decay $\beta = (0.9, 0.999)$. The training is done end-to-end with images and precomputed segmentation masks resized to $640 \times 192$ ($512 \times 256$ for stereo). We set $\lambda_{seg} = 1$ and $\lambda_{smooth} = 10^{-3}$ to balance the loss function. The remaining details follow [17] for the stereo or [18] for the sequence, which is the base network of our method.

4.2 Experimental Results

Comparison with State-of-the-art. The quantitative results of self-supervised monocular depth estimation on KITTI are shown in Table [1]. Our method out-
Table 2. Ablation for sequence model. Ours indicates our reimplementation of [18], and Seg is multi-task learning with segmentation. R and N denote the task-specific Residual Adapter and batch normalization per each task.

| Model          | Seg R/N CPU APU | Lower is better | Higher is better | \(\delta < 1\) | \(\delta < 1.25\) | \(\delta < 1.25^2\) | \(\delta < 1.25^3\) |
|----------------|-----------------|-----------------|------------------|----------------|----------------|----------------|----------------|
| Monodepth2     |                 | 0.115 0.903 4.863 0.193 | 0.874 0.959 0.981 |
| Ours with SE   | ✓               | 0.116 0.918 4.842 0.193 | 0.874 0.959 0.981 |
| Ours with SE   | ✓ ✓             | 0.116 0.883 4.703 0.189 | 0.877 0.961 0.982 |
| Ours with SE   | ✓ ✓ ✓           | 0.111 0.815 4.665 0.187 | 0.881 0.962 0.982 |
| Ours with SE   | ✓ ✓ ✓ ✓         | 0.114 0.775 4.589 0.186 | 0.872 0.962 0.984 |

Table 3. Ablation for stereo model. Ours indicates our reimplementation of [17] with ResNet-18 backbone, and pp means the post-processing method [17].

| Model          | Seg R/N CPU APU | Lower is better | Higher is better | \(\delta < 1\) | \(\delta < 1.25\) | \(\delta < 1.25^2\) | \(\delta < 1.25^3\) |
|----------------|-----------------|-----------------|------------------|----------------|----------------|----------------|----------------|
| Garg et al. [15]*|                 | 0.152 1.226 5.849 0.246 | 0.784 0.921 0.967 |
| Monodepth [17]  |                 | 0.133 1.142 5.533 0.230 | 0.830 0.936 0.970 |
| 3Net [5]        |                 | 0.129 0.996 5.281 0.223 | 0.831 0.939 0.974 |
| Chen et al. [6] + pp ✓ |         | 0.118 0.905 5.096 0.211 | 0.839 0.945 0.977 |
| Ours            |                 | 0.150 1.304 5.884 0.247 | 0.789 0.919 0.964 |
| Ours with SE    | ✓               | 0.128 1.242 5.348 0.225 | 0.847 0.941 0.971 |
| Ours with SE    | ✓ ✓             | 0.118 0.972 5.107 0.213 | 0.850 0.947 0.975 |
| Ours with SE    | ✓ ✓ ✓           | 0.120 0.940 5.006 0.213 | 0.851 0.946 0.975 |

performs not only Monodepth2 but also other networks for most of the metrics. We also show a further increase in performance through high-resolution images.

The qualitative results in Fig. 4 show that our approach reduces the problem that training with photometric losses is inappropriate to where ambiguous boundaries or complicate shapes exist. For example, road signs in the first and last columns are the hard objects to describe, so all the other methods except ours fail to estimate the depth accurately. As our method with semantic-aware depth features perceives the representation of the target objects, the outlines of instances become clear. In other words, the limitation of the photometric loss, which compares individual errors at the pixel level, can be improved by supervision from the feature level semantic information.

Ablation Study. We conduct experiments to explore the effects of the proposed methods while removing each module in Table 2. When semantic knowledge is delivered through multi-task learning with segmentation, the performance is enhanced. Furthermore, the more improvement occurs in almost all the metrics when semantic-aware depth features are created by our techniques that divide task-specific and task-shared parameters. CPU and APU process the features in the channel and spatial dimensions, respectively, and show better results when both of them are included in the networks.

In order to demonstrate the scalability of our method in self-supervised monocular depth estimation, the proposed modules are applied to Monodepth, which train the networks from stereo cues. Table 3 shows that semantic-aware depth features in the stereo model also increase the performance comparable to state-of-the-art Chen et al. [6], which only focus on self-supervised training.
Table 4. Adverse weather experiments on vKITTI [14]. For a fair comparison, we test after adding SE modules into the base architecture of Monodepth2.

| Method          | Weather | Lower is better. | Higher is better. |
|-----------------|---------|------------------|-------------------|
|                 |         | Abs Rel Sq Rel  | RMSE              |
|                 |         | RMSE log $\delta < 1$ | $\delta < 1.25\delta < 1.25^2\delta < 1.25^3$ |
| Monodepth2 [18] (SE) fog | 0.218  | 2.823 | 10.392 | 0.370 | 0.686 | 0.871 | 0.919 |
| Ours            | fog     | **0.213** | **2.478** | **9.018** | **0.317** | **0.690** | **0.872** | **0.936** |
| Monodepth2 [18] (SE) rain | 0.200  | 1.907 | 6.965 | 0.263 | 0.734 | 0.901 | 0.961 |
| Ours            | rain    | **0.145** | **1.114** | **6.349** | **0.222** | **0.800** | **0.937** | **0.977** |

(a) Results according to Light Conditions

(b) Qualitative results.

Fig. 5. Robustness for light intensity changes. (b) Top to bottom: Input RGB images, predicted depth map of GeoNet [52], SIGNet [36], Monodepth2 [18], and ours.

with stereo vision. On the other hand, our method can be globally adjusted to self-supervised networks regardless of stereo or sequence input. Hence, we expect better performance if loss functions proposed by [6] is combined with ours.

**Low Light Conditions.** Assuming low light situations, we measure the performance of networks multiplying the input images by a scale between zero and one. Figure 5 shows that our proposed method has shown consistent results regardless of illuminance. When the value of darkness becomes 0.9, our approach produces a smaller increase than others in the square relative error. This proves that our strategy complements the depth estimation by identifying semantics rather than simply regressing depth values from RGB information. In the case of zero intensity, only SIGNet [36] shows some valuable performance, because it exploits segmentation masks as input to the network during the test.

**Weather Conditions.** In addition to low light experiments, we experiment with vKITTI to show that the proposed method is robust to the adverse weather. We test the case of rain and fog that are challenging for depth estimation, after training with the other condition data, to prove the effectiveness of our methods. Table 4 demonstrates that the performance increase when the depth estimation is performed using semantic-aware depth features. Correspondingly, Fig. 6 shows the depth hole (1st column) or infinite depth on moving objects (4th column) problems are reduced, and the shape of the objects is predicted better.
Reflective Material Problems. Figure 7 shows that our approach has better qualitative results in the regions where the Lambertian assumption is violated. Without semantic-awareness, Monodepth2 [18] often fails to learn proper depths for distorted, reflective, or color-saturated regions like windows of vehicles. However, our model is aware of semantic information which can tell whether a group of neighboring pixels belongs to the same object category or not. Therefore, the distances of the windows are similar to those of their vehicles compared to [18].

Further Discussion about Semantic Supervision. Since our network training of the segmentation layers relies on pseudo labels generated by DeepLabv3+ [5], this training scheme may have problems when DeepLabv3+ does not work well. The performance of DeepLabv3+ is good enough, but there are several hard cases on the test set, as shown in Fig. 8. Likewise, the segmentation masks from our semantic decoder are coarse and lose some details in those cases. However, our segmentation results are reasonable because they are derived from not only semantic supervision but also geometric features through joint learning. Besides, our approach exploits not a single segmentation mask as input but feature level semantic knowledge across the entire data, so our coarse semantic learning is sufficient to make the depth features semantic-aware.

To demonstrate the strength of semantic-aware depth features directly, performance evaluation for each class is shown in Fig. 9. We exploit the pseudo labels as the masks per each class to evaluate the class-specific depth estimation performance. With semantic information, our method shows that absolute relative difference is reduced in all classes except for the sky class. In particular, people (0.150 to 0.137) and poles (0.223 to 0.215) have significant performance
Fig. 8. Segmentation and depth estimation results for the test set. Segmentation masks in the second row are never considered by our network during training, but we present these results for a fair comparison with our results.

Fig. 9. Comparison of depth estimation error in distinct classes. Our method increases the performance in all classes except for sky which has infinite depth.

improvement. Accurate depth values of these categories are difficult to learn by photometric loss because of the exquisite shape, but the semantic-aware features delineate the contour of objects better. Besides, semantic-awareness shows that it is also helpful for estimating the distances of the moving classes such as riders (0.197 to 0.180) and trains (0.125 to 0.109) that violate the assumption of rigid motions in self-supervised monocular depth training.

5 Conclusions

This paper points out the problems of the photometric loss and introduces how to mediate those issues with semantic information. Through the designed multi-task approach, our self-supervised depth estimation network can learn semantic-aware features to improve the performance of depth prediction. We also demonstrate that our modules can be applied to universal self-supervision depth networks, regardless of whether the type of training images is either stereo or video sequence. Furthermore, to prove our method is robust to environmental changes, various experiments are conducted under different conditions. The experimental results show that our framework is more effective than other state-of-the-art networks. In future work, we will investigate the still existing limitation of photometric loss in semi-supervised depth training, which uses small amounts of groundtruth depth and explore the way to apply semantic information.
References

1. Aleotti, F., Tosi, F., Poggi, M., Mattoccia, S.: Generative adversarial networks for unsupervised monocular depth prediction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
2. 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. pp. 35–45 (2019)
3. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8001–8008 (2019)
4. Chang, W.G., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7354–7362 (2019)
5. 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). pp. 801–818 (2018)
6. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2624–2632 (2019)
7. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: Advances in neural information processing systems. pp. 730–738 (2016)
8. Chen, Y., Schmid, C., Sminchisescu, C.: Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7063–7072 (2019)
9. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 80, pp. 794–803. PMLR, Stockholmsmässan, Stockholm Sweeden (10–15 Jul 2018)
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
11. 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. Iee (2009)
12. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. pp. 2366–2374 (2014)
13. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2002–2011 (2018)
14. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4340–4349 (2016)
15. Garg, R., BG, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European Conference on Computer Vision. pp. 740–756. Springer (2016)
16. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361. IEEE (2012)
17. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 270–279 (2017)
18. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3828–3838 (2019)
19. 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)
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
21. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025 (2015)
22. Jiao, J., Cao, Y., Song, Y., Lau, R.: Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 53–69 (2018)
23. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018)
24. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 66–75 (2017)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
26. Klodt, M., Vedaldi, A.: Supervising the new with the old: learning sfm from sfm. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 698–713 (2018)
27. Kokkinos, I.: Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6129–6138 (2017)
28. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). pp. 239–248. IEEE (2016)
29. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. IEEE transactions on pattern analysis and machine intelligence 38(10), 2024–2039 (2015)
30. Liu, S., Johns, E., Davison, A.J.: End-to-end multi-task learning with attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1871–1880 (2019)
31. Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., Feris, R.: Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5334–5343 (2017)
32. Luo, C., Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R., Yuille, A.: Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. arXiv preprint arXiv:1810.06125 (2018)
33. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5667–5675 (2018)
34. Maninis, K.K., Radosavovic, I., Kokkinos, I.: Attentive single-tasking of multiple tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1851–1860 (2019)
35. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4040–4048 (2016)
36. Meng, Y., Lu, Y., Raj, A., Sunarjo, S., Guo, R., Javidi, T., Bansal, G., Bharadia, D.: Signet: Semantic instance aided unsupervised 3d geometry perception. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9810–9820 (2019)
37. Misra, I., Shrivastava, A., Gupta, A., Hebert, M.: Cross-stitch networks for multi-task learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3994–4003 (2016)
38. Pilzer, A., Xu, D., Puscas, M., Ricci, E., Sebe, N.: Unsupervised adversarial depth estimation using cycled generative networks. In: 2018 International Conference on 3D Vision (3DV). pp. 587–595. IEEE (2018)
39. Poggi, M., Tosi, F., Mattoccia, S.: Learning monocular depth estimation with unsupervised trinocular assumptions. In: 2018 International Conference on 3D Vision (3DV). pp. 324–333. IEEE (2018)
40. Ramirez, P.Z., Poggi, M., Tosi, F., Mattoccia, S., Di Stefano, L.: Geometry meets semantics for semi-supervised monocular depth estimation. In: Asian Conference on Computer Vision. pp. 298–313. Springer (2018)
41. Ranjan, A., Jampani, V., Balles, L., Kim, K., Sun, D., Wulf, J., Black, M.J.: Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12240–12249 (2019)
42. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. In: Advances in Neural Information Processing Systems. pp. 506–516 (2017)
43. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Efficient parametrization of multi-domain deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8119–8127 (2018)
44. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
45. Sener, O., Koltun, V.: Multi-task learning as multi-objective optimization. In: Advances in Neural Information Processing Systems. pp. 527–538 (2018)
46. Ummenhofer, B., Zhou, H., Uhrig, J., Mayer, N., Ilg, E., Dosovitskiy, A., Brox, T.: Demon: Depth and motion network for learning monocular stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5038–5047 (2017)
47. Wang, C., Miguel Buenaposada, J., 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, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
49. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13(4), 600–612 (2004)
50. Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R.: Lego: Learning edge with geometry all at once by watching videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 225–234 (2018)
51. Yang, Z., Wang, P., Xu, W., Zhao, L., Nevatia, R.: Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
52. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1983–1992 (2018)
53. Zhang, Y., Khamis, S., Riemenschneider, C., Valentin, J., Kowdle, A., Tancik, V., Schoenbrun, M., Izadi, S., Funkhouser, T., Fanello, S.: Activestereonet: End-to-end self-supervised learning for active stereo systems. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 784–801 (2018)
54. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1851–1858 (2017)
55. Zou, Y., Luo, Z., Huang, J.B.: Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 36–53 (2018)