Multi-stage Information Diffusion for Joint Depth and Surface Normal Estimation[⋆,⋆⋆]

Zhiheng Fu[a], Siyu Hong[b], Mengyi Liu[c], Hamid Laga[e], Mohammed Bennamoun[a], Farid Boussaid[a] and Yulan Guo[b,*]

[a]Department of Computer Science and Software Engineering, The University of Western Australia, 6009 Perth, Australia.
[b]School of Electronics and Communication Engineering, The Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, 518107, China.
[c]Department of Content Security, Kuaishou, Beijing, 100085, China.
[e]School of Information Technology, and Harry Butler Institute and Centre for Healthy Aging, Murdoch University.

ARTICLE INFO

Keywords:

Depth Estimation

Surface Normal Estimation

Multitask Learning

Attention Map

Multi-Stage Information Fusion

ABSTRACT

Depth and surface normal estimations are important for 3D geometric perception, which has numerous applications including autonomous vehicles and robots. In this paper, we propose a lightweight Multi-stage Information Diffusion Network (MIDNet) for the simultaneous prediction of depth and surface normals from a single RGB image. To obtain semantic and detail-preserving features, we adopt a high-resolution network as our backbone to learn multi-scale features, which are then fused into shared features for the two tasks. To mutually boost each task, a Cross-Correlation Attention Module (CCAM) is proposed to adaptively integrate information for the prediction of the two tasks in multiple stages, including feature-level information interaction and task-level information interaction. Ablation studies show that the proposed multi-stage information diffusion strategy can boost the performance gain for the two tasks at different levels. Compared to current state-of-the-art methods on the NYU Depth V2, Stanford 2D-3D-Semantic and KITTI datasets, our method achieves superior performance for both monocular depth and surface normal estimation.

1. Introduction

Monocular depth estimation aims to predict distances between objects in a scene and the camera using a single image. Monocular surface normal estimation aims to predict the relative directions to the camera viewpoint for each pixel from a single image. Monocular depth estimation is an ill-posed problem due to the geometric ambiguity caused by the projection from a 3D scene to a 2D plane. Both depth and surface normal estimation tasks become extremely challenging in the presence of complex backgrounds and occlusions when no prior information of the scene is available [1–3].

Driven by deep learning techniques, several monocular depth estimation [4–12] and surface normal estimation [13] [14] [15–17] methods have achieved impressive results using a large number of labelled data. Most of these methods perform depth estimation and surface normal estimation separately without sufficient consideration on their closely underlying geometric relationship. This commonly causes inconsistent predictions, e.g., the predicted depth maps could be distorted in planar regions [18]. This can be inferred that joint monocular depth and surface normal estimation has the potential to boost the performance of each individual prediction.

In the past few decades, multi-task learning has been extensively studied, e.g., for depth estimation and image decomposition [19], depth estimation and semantic segmentation [3, 20–25] and depth and surface normal estimation

*Corresponding author at: School of Electronics and Communication Engineering, The Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, 518107, China. e-mail: guoyulan@sysu.edu.cn.
22907304@student.uwa.edu.au (Z. Fu); guoyulan@sysu.edu.cn (Y. Guo)
http://yulanguo.me/ (Y. Guo)
ORCID(s): 0000-0002-3818-5659 (Z. Fu)

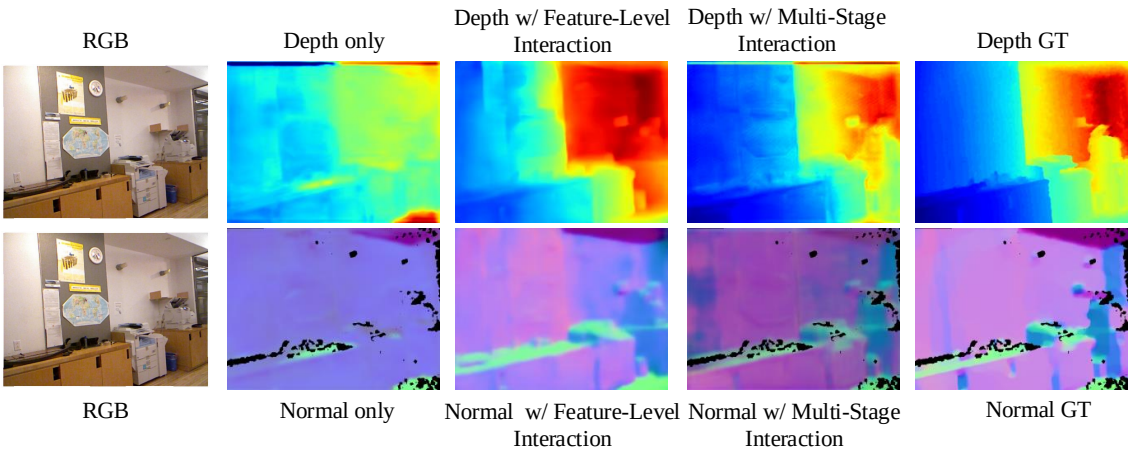


Figure 1: Comparison between the results of depth and surface normal estimation with feature-level interaction and with multi-level interaction.

[18, 26–30]. However, most existing methods first learn common features using shared layers before the learning of task-specific features using separate branches [31–33]. Although different tasks may be correlated to some extent, these previously reported methods just perform information interaction through a combination of task-specific losses on the same backbone.

Recently, more effective multi-task learning approaches have been proposed to focus on task-level interaction [27–29, 34] or feature-level interaction [35]. However, these existing methods normally only focus on only one-stage interaction. As a result, they cannot fully learn the interactive information between depth and surface normal estimations.

In this paper, we propose a joint prediction approach of depth and surface normal using a multi-stage interaction strategy with a Cross-Correlation Attention (CCAM). As shown in Fig. 1, both depth and surface normal estimations with feature-level interaction achieve a significantly better performance compared to the single tasks. Moreover, further interaction at the task level can boost the performance to a certain extent. To this end, we adopt a high-resolution network (HRnet) [36] as our shared backbone to perform depth prediction and surface normal estimations. This choice is motivated by the fact that high-resolution features are beneficial to the pixel-level tasks. From Fig. 1, it can be seen that HRnet can preserve detailed information for both depth and surface normal estimation, e.g., the picture on the wall. Then, we design a multi-scale feature fusion module to fuse low-resolution features with semantic information and high-resolution features with detailed information. We also designed a Cross-Correlation Attention Module (CCAM) to perform fusion of task-specfic features for the first stage interaction. Once the initial depth and surface normal estimations are obtained, the proposed CCAM is used to conduct task-level information fusion for the second stage interaction. Experimental results on the NYU Depth V2 [37], Stanford 2D-3D-Semantic [38] and KITTI [39] datasets demonstrate that our network achieves superior performance for both monocular depth and normal estimation.

The major contributions of our work can be summarized as follows:

- We introduce a multi-stage information diffusion strategy to achieve interactions for joint-task learning through the proposed end-to-end deep neural network (MIDNet).
- We design a Cross-Correlation Attention Module to effectively capture interactive information at different levels between depth and surface normal estimations.
- Our MIDNet achieves state-of-the-art performance for joint monocular depth estimation and surface normal estimation on the NYU Depth V2 [37] and Stanford 2D-3D-Semantic [38] datasets, and achieves similar performance to prior art on the KITTI [39] dataset.

The rest of this paper is organized as follows. In Section 2, we briefly review related works. In Section 3, we describe the proposed network in details. In Section 4, experimental results are presented. Finally, we conclude this paper in Section 5.

2. Related Works

In this section, we briefly review related works of the relevant approaches for monocular depth estimation, surface normal estimation, and multi-task learning.

2.1. Monocular Depth Estimation

Monocular depth estimation has long been tackled using Markov Random Fields (MRFs). Recently, learning-based methods have reached promising performance in monocular depth estimation. Eigen et al. [4] proposed a deep neural network with two stacks to predict a coarse depth map from an entire image, and then refine the original prediction locally. Liu et al. [40] integrated continuous Conditional Random Field (CRF) into a CNN framework to capture consistency information of neighbouring pixels. Roy et al. [41] combined neural regression forests with convolutional neural networks to predict depths in the continuous domain via regression. Xu et al. [42] proposed a deep model to predict depth in an end-to-end manner by integrating a cascade of multiple CRFs as a unified graphical model. Xu et al. [43] proposed continuous CRFs to fuse information using a structured attention model, which automatically regulates the amount of information transferred between corresponding features at different scales. Fu et al. [5] adopted dilated convolution to capture multi-scale semantic features and designed an ordinal regression loss for depth estimation. Chen et al. [44] considered monocular depth estimation as a multi-class dense labeling problem. They proposed an attention-based context aggregation network to adaptively learn task-specific similarities between pixels. Zhang et al. [45] proposed a hard-mining network and a hard-mining objective function to achieve monocular depth estimation. Besides, intra-scale and inter-scale refinement sub-networks were designed to accurately localize and refine those hard regions. Kim et al. [46] proposed a deep variational model to effectively integrate heterogeneous predictions from global and local convolutional neural networks for depth prediction from a single image. More recently, Su et al. [47] proposed a general information exchange convolutional neural network and a mutual channel attention mechanism to perform information exchange between the high-resolution and low-resolution features for monocular depth estimation. Chen et al. [48] and Lee et al. [49] proposed to decode multi-scale features to multi-resolution depth predictions to take advantage of the multi-scale loss functions and explicitly constraint multi-scale features. Multi-scale features and multi-resolution depth predictions can be matched directly in this manner, but computational cost increases significantly. Huynh et al. [50] implicitly embedded the coplanarity constraint into the monocular depth network by designing a Depth-Attention Volume (DAV). Ye et al. [51] proposed a non-local spatial attention module by introducing a non-local filtering strategy to explicitly exploit the non-local correlation in the spatial domain to facilitate depth details inference. In contrast to these previous works, our method exploits multi-task learning across depth and surface normal predictions at multiple information interaction stages.

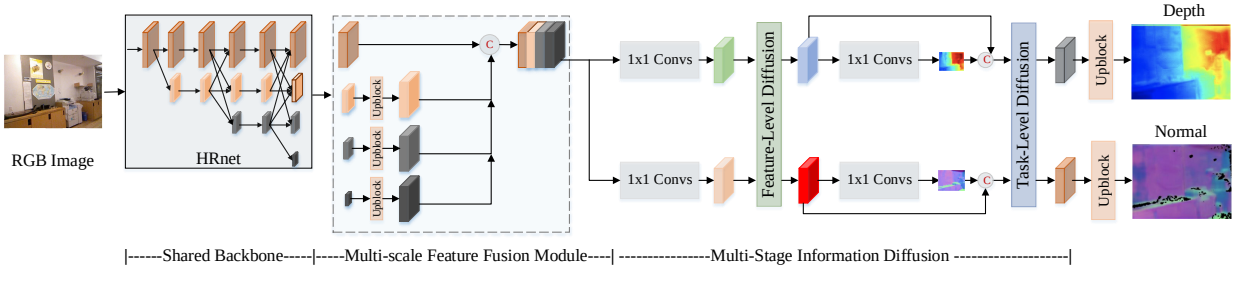


Figure 2: Overview of the proposed MIDNet.

2.2. Surface Normal Estimation

Most surface normal estimation methods exploit the strong feature representation capability of deep neural networks. Ladicky et al. [15] combined contextual and segment-based cues to build a regressor in a boosting framework. They transformed surface normal estimation into the regression of coefficients of a local encoding. Eigen et al. [31] adopted a hierarchical network for depth/normal prediction in a coarse-to-fine manner. Wang et al. [17] proposed a deep neural network to learn both global and local features for coarse prediction. They then used a fusion module to perform the final estimation. Li et al. [26] proposed a deep convolutional neural network model to learn the mapping from multi-scale image patches to depth or surface normal values at the super-pixel level. Then, the estimated

super-pixel depth or surface normal is refined to the pixel level by exploiting various potentials in the depth or surface normal map. Qi et al. [16] converted depth and surface normal estimation into spherical regression to obtain the final prediction. More recently, Wang et al. [52] proposed a method to combine traditional line and vanishing point analyses with a deep learning approach for single-view surface normal estimation.

2.3. Multi-Task Learning

Misra et al. [53] proposed a cross-stitch network for multi-task learning. Although this method outperforms the baselines, it suffers from propagation interruption if the combination of weights degenerate into 0. Besides, the network design with two parallel sub-networks also increases the number of parameters and learning complexity. Kokkinos et al. [54] proposed an Ubernet to implement various tasks on diverse datasets. Huang et al. [55] applied dense connections in each layer of a network for recognition tasks. With fully-dense connections, all the information can be shared. Meanwhile, memory consumption is also increased. Xu et al. [43] proposed the PADNet to distill knowledge from different tasks for the target tasks using a structured attention module. Qi et al. [27] proposed GeoNet to perform task-level interaction between depth and surface normal estimations using depth-to-normal and normal-to-depth networks.

Jiao et al. [35] proposed a synergy network and an attention-driven loss to automatically learn the information sharing strategies between the two tasks. However, a unified weight is directly assigned to a feature map without assigning different weights to individual values. Liu et al. [32] designed a multi-task attention network for the prediction of depth, surface normal and semantic labels. They used a soft-attention module to extract task-specific features from a single shared backbone. However, this method can only learn common features from the shared backbone. Zhang et al. [29] proposed a pattern-affinitive propagation method to learn an affinity matrix for each task of depth estimation, normal estimation and semantic segmentation. They then adaptively combined these three affinity matrices to boost each task by extracting shared information from the other two tasks. Furthermore, Zhou et al. [28] proposed intra-task Pattern-Structure Diffusion (PSD) and inter-task PSD structures to transmit intra-task pattern-structures at the feature-level among different tasks.

In contrast to these previous works, we explore a multi-stage information fusion strategy for joint depth and surface normal estimation. We also design a unified Cross-Correlation Attention Module to measure the relevance between multiple tasks at the task-specific feature level and the task level.

3. Multi-Stage Information Diffusion

In this section, we first describe the proposed MIDNet, and then introduce the multi-scale feature fusion module, and the Cross-Correlation Attention Module for multi-stage interactions, and finally provide the loss function integrating two different pixel-level prediction tasks. The architecture of the proposed MIDNet is shown in Fig. 2.

3.1. Overview of MIDNet

In this work, we use HRnet [36] as our architecture backbone to learn the high-resolution features of a scene. HRnet uses a high-resolution sub-network at the first stage, followed by high-to-low resolution sub-networks one by one to form more stages, and connects the multi-resolution sub-networks in parallel. For HRnet [36], the high-resolution sub-network is designed to preserve detailed features, which are potentially beneficial to pixel-wise prediction. In [56], the pixel-wise task of semantic segmentation has been explored using HRnet [36]. In this paper, we focus on the monocular depth and surface normal estimations.

As shown in Fig. 2, once the multi-scale features are extracted by the shared backbone (i.e., HRnet), a multi-scale feature fusion module is introduced to fuse these features. These multi-scale features are then concatenated and further processed using several 1×1 convolutional layers to produce raw task-specific features. Next, CCAM (shown in Fig. 3 (a)) is used to capture interactive information at the feature level between multiple tasks to generate fused features, which are further processed by 1×1 convolutional layers to aggregate the different channel features, resulting in the initial estimates of depth and surface normals. The initial estimates are then refined at the task-level interaction by the CCAM module. Finally, the output of the last CCAM is up-sampled to the same resolution as the input of MIDNet using an up-block module (Fig. 3 (b)) to obtain the final prediction.

3.2. Multi-scale Feature Fusion

As shown in Fig. 2, HRnet produces feature maps at four scales, with resolutions of $H \times W \times 32$, $\frac{H}{2} \times \frac{W}{2} \times 64$, $\frac{H}{4} \times \frac{W}{4} \times 128$ and $\frac{H}{8} \times \frac{W}{8} \times 256$. The feature map with higher resolution contains detailed image information, while the

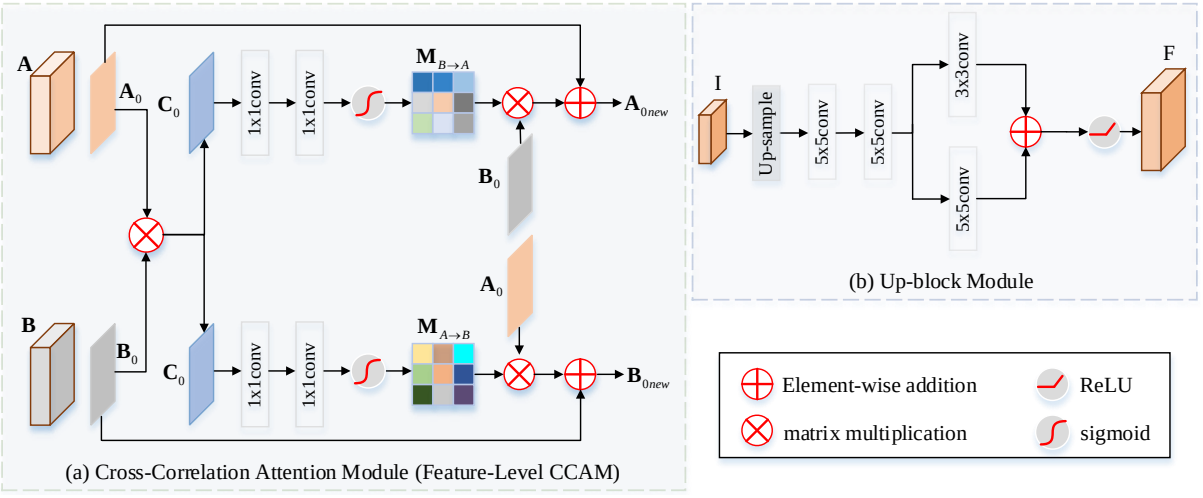


Figure 3: Architecture of the CCAM and up-block modules. $\mathbf{A}_0$ and $\mathbf{B}_0$ are the 0-th channels of feature maps $\mathbf{A}$ and $\mathbf{B}$, respectively. $\mathbf{C}_0$ is the cross-correlation matrix of $\mathbf{A}_0$ and $\mathbf{B}_0$. $\mathbf{I}$ and $\mathbf{F}$ represent the initial and final estimations for one of the two tasks, respectively.

feature maps with lower resolutions contain high-level semantic information. To fully use multi-level information, we propose an efficient multi-scale feature fusion module to integrate information from multi-scale features. Specifically, we use convolutional kernels of different sizes for the low-resolution branches to ensure their number of channels is the same as the high-resolution branch. Then, bilinear interpolation is used to resize these transformed feature maps to produce features with the same spatial resolution as the high-resolution branch. Finally, these resized features are concatenated to produce raw features, which are further exploited in the subsequent steps.

An advantage of this feature fusion module is the reduction of the number of parameters. This is because the feature map produced by the multi-scale feature fusion module has only 128 channels. In contrast, direct concatenation of feature maps with different resolutions produces a feature map with 480 channels (as in [36]).

3.3. Multi-Stage Information Diffusion

In this work, we propose CCAM (shown in Fig. 3 (a)) to capture the interactive information between the different tasks at different stages for the prediction of depth and surface normals.

3.3.1. Cross-Correlation Attention Module

A major challenge of joint-task learning is to share useful information across different tasks. Although the shared backbone can learn the common features of the two tasks, the feature-level interaction and the task-level interaction are more important to improve the final performance. Therefore, we propose CCAM to handle the multi-stage information diffusion. In this section, CCAM is described at the feature-level for simplification.

Given two feature maps $\mathbf{A}$ and $\mathbf{B} \in \mathbb{R}^{H\times W\times C}$ learned from the two branches (Fig. 3), they are fed to the cross-correlation module to generate a cross-correlation matrix $\mathbf{C} \in \mathbb{R}^{H\times W\times C}$, which is used as a guide to learn the relevance between the two tasks in Eq. (1).

$$\mathbf{C} = \mathbf{A} \otimes \mathbf{B}, \tag{1}$$

where $\otimes$ denotes the element-wise matrix multiplication. $\mathbf{C}$ is fed to two 1×1 convolutional layers in each branch, resulting in two feature maps $\mathbf{Q}_A \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{Q}_B \in \mathbb{R}^{H\times W\times C}$. Next, $\mathbf{Q}_A$ and $\mathbf{Q}_B$ are respectively encoded into two attention maps $\mathbf{M}_{B\to A}$ and $\mathbf{M}_{A\to B}$ using two sigmoid layers. Therefore, the contribution of each value of a feature map from task$_B$ to task$_A$ can be represented by the attention map $\mathbf{M}_{B\to A}$. Similarly, the contribution of each value of a feature map from task$_A$ to task$_B$ can be represented by the attention map $\mathbf{M}_{A\to B}$. Consequently, each value of the feature map of task$_A$ is assigned a relevance probability to the corresponding feature map of task$_B$.

3.3.2. Feature-Level Information Diffusion

In this stage, we use two attention maps ($\mathrm{M}_{B\to A}$ and $\mathrm{M}_{A\to B}$) to fuse features $\mathbf{A}$ and $\mathbf{B}$ for feature-level information diffusion. Specifically, the fused feature maps $\mathbf{A}_{new}$ and $\mathbf{B}_{new}$ for task_A and task_B are respectively represented as:

$$\begin{aligned}\mathbf{A}_{new} &= \mathbf{A} + (\mathbf{M}_{B\to A} \otimes \mathbf{B}) \\ \mathbf{B}_{new} &= \mathbf{B} + (\mathbf{M}_{A\to B} \otimes \mathbf{A})\end{aligned}, \tag{2}$$

where $\mathbf{A}_{new}$ absorbs the interactive information from task_B and $\mathbf{B}_{new}$ absorbs the interactive information from task_A. Here, $\mathbf{A}_{new}$ and $\mathbf{B}_{new}$ are fine task-specific features produced by CCAM. Subsequently, the initial predictions are estimated from $\mathbf{A}_{new}$ and $\mathbf{B}_{new}$ with the interactive information between two tasks.

3.3.3. Task-Level Information Diffusion

Once the initial predictions for task_A and task_B are obtained from the first round of the feature-level interaction, we repeat CCAM to conduct task-level information diffusion. In this stage, CCAM is explicitly used to measure the relevance between the original depth and the surface normal estimation, and perform the task-level interaction. Specifically, the initial depth map is concatenated with the output feature maps of the first CCAM to form the new input for the second round of information interaction. The same process is also applied to the surface normal estimation task. The rationale behind the task-level information diffusion is that initial estimates of task_A and task_B contain more task-orientated information. Therefore, the second CCAM can explicitly mine the relationship between these two tasks. Besides, the concatenated feature-level information and initial estimates can further be integrated by the second CCAM. Finally, we iteratively implement this operation to conduct deep interaction between these two tasks. The experimental results in Section 4.2.2 demonstrate that iterative CCAM can further improve the performance of these individual two tasks.

3.4. Loss for Joint-task Learning

We use BerHu loss [7] for depth estimation. That is:

$$\mathcal{L}_d = \begin{cases} \left|d_{pre} - d_{gt}\right|, & \text{if } \left|d_{pre} - d_{gt}\right| \le \tau \\ \frac{(d_{pre}-d_{gt})^2+\tau^2}{2\tau}, & \text{if } \left|d_{pre} - d_{gt}\right| > \tau \end{cases}, \tag{3}$$

where d_{pre} and d_{gt} are the predicted and ground truth depth maps. Berhu loss [7] provides a good trade-off between $\mathcal{L}_1$ and $\mathcal{L}_2$ norms. It assigns large weights to pixels with high gradient residuals due to the use of its $\mathcal{L}_2$ term. Meanwhile, its $\mathcal{L}_1$ term gives more consideration to the pixels with smaller gradient residuals than $\mathcal{L}_2$ norm. In Eq. 3, τ is a threshold, which is set to 20% of the maximum error between predictions and ground truth.

We then use $\mathcal{L}_2$ loss for the surface normal estimation task:

$$\mathcal{L}_n = \left|\left|\boldsymbol{n}_{pre} - \boldsymbol{n}_{gt}\right|\right|_2, \tag{4}$$

where $\boldsymbol{n}_{pre}$ and $\boldsymbol{n}_{gt}$ are the predicted and ground truth normals of a 3D point, respectively. The $\mathcal{L}_2$ loss can effectively learn accurate normal predictions since it focuses more on the three axes of normals than the cosine loss [57].

To balance these two tasks, we simply assign the same weight for both depth and surface normal estimations. Therefore, the total loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_d + \mathcal{L}_n. \tag{5}$$

In addition, the total loss (Eq. 5) is applied to the multi-stage information diffusion, i.e., the total loss is used after each CCAM for the supervision of each estimation stage.

4. Experiments

In this section, we first introduce our experimental settings and then conduct ablation studies to test the proposed MIDNet. In addition, we further compare our method with several relevant state-of-the-art single-task and joint-task methods.

4.1. Experimental Settings

4.1.1. Datasets

We tested our method on three publicly available datasets: the NYU Depth V2 [37], Stanford 2D-3D-Semantic [38], and KITTI [39] datasets.

The NYU Depth v2 dataset consists of around 400k RGB-D images of 464 indoor scenes. For training, we randomly selected 12k images from the official training scenes following Eigen split [4]. These 12k images have ground truth depth maps but no ground truth for the surface normals. Therefore, we used a toolbox provided by the Nathan et al. [37] to generate the ground truth surface normals for these images. Note that, we follow the Stanford 2D-3D-Semantic dataset [38] protocol to set the direction of surface normal towards the camera as the ground truth's direction.

The Stanford 2D-3D-Semantic dataset [38] is a large-scale indoor dataset collected in 6 areas of over $600m^2$. This dataset contains over 70k RGB images and their corresponding depth and surface normal maps.

The KITTI dataset [39] contains over 93k depth maps with their corresponding raw LiDAR scans and RGB images. Following the split of [4], training was performed on 23488 uninpainted images of 32 scenes and test was conducted on 697 images of 29 scenes. During training, images were randomly cropped to a resolution of 352×512.

4.1.2. Implementation Details

We implemented our network using Pytorch [58] on a server with 8 Nvidia GTX 2080Ti GPUs. We built our model based on HRnet-W32, which is pre-trained on the ImageNet classification task [59]. During training, the NYU Depth V2 images (with a resolution of 640×480) were center-cropped to a resolution of 544×416 to remove invalid points around the boundaries of images. For images in the Stanford 2D-3D-Semantic and KITTI datasets, two data augmentation strategies were used to increase the diversity of the data, including cropping and flipping.

For the NYU Depth V2 and Stanford 2D-3D-Semantic datasets, all models were first optimized using SGD with a batch size of 6. The initial learning rate was set to 1×10^{-4} and reduced to half after every 20 epochs. The training process was stopped after 200 epochs. For the KITTI dataset, since there is no ground truth for surface normals, we first trained our model on NYU Depth V2 for the surface normal estimation, and then froze the surface normal estimation branch to train the depth estimation branch on KITTI for 100 epochs.

4.1.3. Evaluation Metrics

To achieve fair comparison with previous works, the same evaluation metrics are used to test the depth estimation performance, including the average relative error (REL), root mean squared error ($RMSE$), root mean squared error in logarithmic space ($RMSE(log)$), mean absolute error in logarithmic space (Log10) and accuracy with a threshold δ, where $\delta \in (1.25, 1.25^2, 1.25^3)$. These metrics are defined as:

$$REL = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|d_{pre}(i) - d_{gt}(i)\right|}{d_{gt}(i)}, \tag{6}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_{pre}(i) - d_{gt}(i))^2}, \tag{7}$$

$$RMSE(log) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log d_{pre}(i) - \log d_{gt}(i))^2}, \tag{8}$$

$$\begin{aligned} percent &= \frac{\sum_{i=1}^{N}\|d_{pre}(i)\|_0}{N} \times 100 \\ s.t. \max &\left(\frac{d_{pre}(i)}{d_{gt}(i)}, \frac{d_{gt}(i)}{d_{pre}(i)}\right) \leq \delta \end{aligned} \tag{9}$$

where $d_{pre}(i)$ and $d_{gt}(i)$ are the estimated and ground truth depth values at position i, respectively. N is the number of valid pixels and $\delta = 1.25, 1.25^2, 1.25^3$.

For the evaluation of the surface normal prediction, several commonly used metrics [17, 31, 60] are adopted,

Table 1
Ablation study of CCAM on NYU Depth V2.

Models	$RMSE$	$rmse-n$
Depth only	0.544	-
Normal only	-	28.3
Depth & Normal w/o CCAM	0.510	27.0
Feature-level interaction with CCAM	0.440	25.9
Feature-level and task-level interactions with CCAMs	**0.427**	**24.7**

including mean of angle error (*mean*), median of angle error ($median$), root mean square error of normal ($rmse-n$), and pixel accuracy. Here, pixel accuracy is defined as the percentage of pixels with angle errors smaller than a threshold θ, where $\theta \in (11.25°, 22.50°, 30°)$.

4.2. Ablation Study

In this section, we present several experiments on the NYU Depth V2 dataset [37] to justify our design choices. Specifically, we used 12k images to train all the models and 654 images for test. Due to the limit of GPU memory, we downsampled the input images from a resolution 544×416 to 136×104 for the high-resolution branch. During test, we cropped the 654 images of resolution of 640×480 to the same size as the training samples.

4.2.1. Multi-scale Feature Fusion

We test the effectiveness of our multi-scale feature fusion module (Fig. 2) only for monocular depth prediction. More specifically, we introduce a depth estimation baseline by directly concatenating multi-scale features and then regressing the depth map. Compared to the baseline, we adopt the designed multi-scale feature fusion module for depth estimation. The method with our module achieves an $RMSE$ of 0.544, which is lower than the baseline (i.e., 0.556). That is because direct concatenation of multi-scale features in the baseline can weaken the discriminativeness of the high-resolution features. In contrast, we first transform the multi-scale features to the same number of channels as the high-resolution branch, and then concatenate these multi-scale features at the same resolution. This operation can aggregate the semantic information of the low-resolution features and simultaneously improve the utilization rate of the high-resolution feature maps in the subsequent parts of the network. This is because the ratio of high-resolution features is increased by compressing the number of feature channels with low resolutions. Compared to ResNet, HRnet achieves a similar performance, while it preserves the finer details. Specifically, the HRnet-w32 model trained with 12k images achieves an $RMSE$ of 0.556. In contrast, the ResNet18 and ResNet50 models trained with 795 images achieve $RMSE$ values of 0.572 and 0.510 [28], respectively.

4.2.2. Multi-Stage Information Diffusion With CCAM

In this section, we analyze the benefits introduced by our multi-stage information diffusion strategy for the two tasks using CCAM. In Figure 4, we show the visualized Cross-Correlation Attention Map (CCAM). The depth-to-normal attention map (a) illustrates that pixels with similar depths have similar attention values and pixels that are farther away have higher attention values, which helps in the normal estimation process. The normal-to-depth attention map (b) shows that different surfaces have different attention values, indicating that different surfaces contribute to their corresponding local depth estimations.

It is noted that the model of *Depth & Normal w/o CCAM* is implemented to predict depth and surface normal estimates following the method in [25]. It can be seen from Table 1 that joint-task models (for both without or with interactions) achieve better performance than single task models. That is because, the joint-task model without CCAM can still learn shared information between multiple tasks from the shared backbone, while the models with only feature-level interaction, 2 feature-level interactions or with both feature-level and task-level interactions can further conduct interactions using the designed CCAM. Besides, joint-task models with feature-level interaction outperform joint-task models without CCAM by 15.7% and 4.6% in terms of $RMSE$ and $rmse-n$, respectively. Compared to the module of feature-level and task-level interactions with CCAMs, the module of 2 feature-level interactions with CCAMs was trained without the supervision of ground truth of depth and surface normal for the initial estimations, leading to slightly inferior performance. This result also demonstrates the effectiveness of the proposed multi-stage information diffusion strategy.

input (a) D2N attention (b) N2D attention

Figure 4: Visualization of CCAM mechanism. The D2N attention and N2D attention represent depth-to-normal attention map and normal-to-depth attention map, respectively.

Table 2
Comparison to the state-of-the-art monocular depth estimation methods on the NYU Depth V2 dataset. The log and Data represent $RMSE(log)$ and the training size for each method, respectively. The single task algorithms [4, 5, 7, 40–42, 61, 62] and the MTL-based algorithms [18, 21, 26, 27, 29, 34, 43, 65] are separated into two parts.

Method	Data	$RMSE$	REL	log	δ_1	δ_2	δ_3
		Lower the better			Higher the better		
DCNF [40]	795	0.824	0.230	-	0.614	0.883	0.971
NR forest [41]	795	0.744	0.187	-	-	-	-
Xu et al. [61]	795	0.593	0.125	-	0.806	0.952	0.986
Eigen [4]	120k	0.877	0.214	0.285	0.611	0.887	0.971
MS-CRF [42]	95k	0.586	0.121	-	0.811	0.954	0.987
FCRN [7]	12k	0.573	0.127	0.194	0.811	0.953	0.988
AdaD-S [62]	100k	0.506	0.114	-	0.856	0.966	0.991
DORN [5]	120k	0.509	0.115	-	0.828	0.965	0.992
PAD-Net [43]	795	0.582	0.120	-	0.817	0.954	0.987
Wang et al. [21]	795	0.745	0.220	0.262	0.605	0.890	0.970
HCRF [26]	795	0.821	0.232	-	0.621	0.886	0.968
SURGE [18]	795	0.643	0.155	0.214	0.768	0.951	0.989
GradNorm [65]	81k	0.629	-	-	-	-	-
GeoNet [27]	16k	0.569	0.128	-	0.834	0.960	0.990
TRL [34]	12k	0.501	0.144	0.181	0.815	0.962	0.992
PAPNet [29]	12k	0.497	0.121	0.175	0.846	0.968	0.994
SharpNet [64]	-	0.495	0.139	0.157	0.888	**0.979**	**0.995**
MIDNet (d+n)	12k	**0.427**	**0.099**	**0.105**	**0.892**	0.973	0.990

The $RMSE$ performance of monocular depth estimation is improved from 0.440 to 0.427 with both feature-level and task-level iterations, as shown in Table 1. However, the improvement is almost saturated after the two task-level iterations. That is because, useful interactive information between different tasks has been fully absorbed by each single task after the second task-level iteration, further task-level iterations cannot improve their performance. To reach a compromise between accuracy and computational complexity, feature-level and task-level iterations are both adopted once in our work on the NYU Depth V2, Stanford 2D-3D-Semantic and KITTI datasets. These results demonstrate that the proposed CCAM module can effectively fuse features from the two tasks.

4.3. Comparison to State-of-the-Art

4.3.1. The NYU Depth V2 Dataset

We compare our MIDNet with a number of monocular depth estimation methods [4, 5, 7, 21, 26, 27, 29, 34, 40–43, 61–63] and surface normal estimation methods [13] [14] [15, 17, 18, 27, 29, 60, 64] on this dataset. The depth estimation task was first trained using the selected 12k images. Then, the depth estimation and surface normal estimation tasks were jointly trained using the CCAM module. Finally, the joint-task learning network was tested on 654 images.

Table 3
Comparison to the state-of-the-art surface normal estimation methods on the NYU Depth V2 Dataset. The *mea*, *med* and $rm-n$ represent *mean*, *median* and $rmse-n$ metrics, respectively. The single task algorithms [13–15, 17, 60] and the MTL-based algorithms [18, 27, 29] are separated into two parts.

Method	*mea*	*med*	$rm-n$	θ_1	θ_2	θ_3
	Lower the better			Higher the better		
3DP [13]	36.3	19.2	-	16.4	36.6	48.2
UNFOLD [14]	35.2	17.9	-	40.5	54.1	58.9
Discr. [15]	33.5	23.1	-	27.7	49.0	58.7
Deep3D [17]	26.9	14.8	-	42.0	61.2	68.2
SkipNet [60]	19.8	12.0	28.2	47.9	70.0	77.8
SURGE [18]	20.6	12.2	-	47.3	68.9	76.6
GeoNet [27]	19.0	11.8	26.9	48.4	71.5	79.5
PAPNet [29]	18.6	**11.7**	25.5	**48.8**	72.2	79.8
MIDNet (d+n)	**17.8**	12.1	**24.8**	48.6	**72.3**	**82.1**

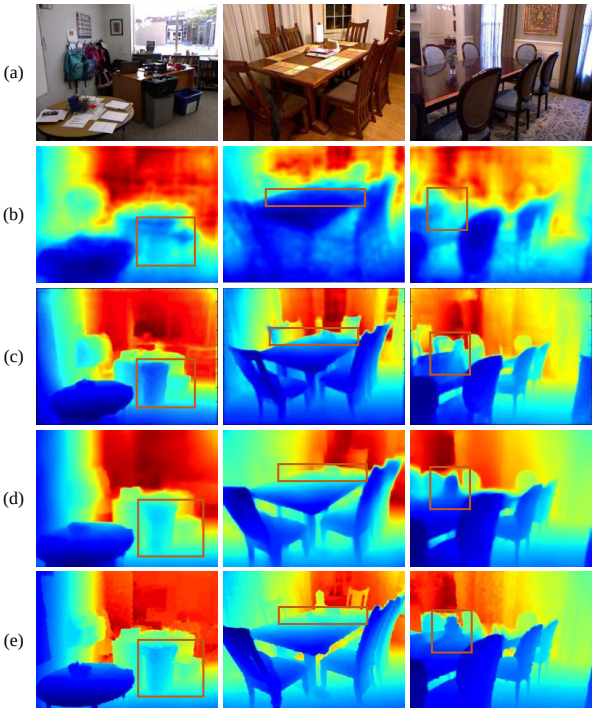
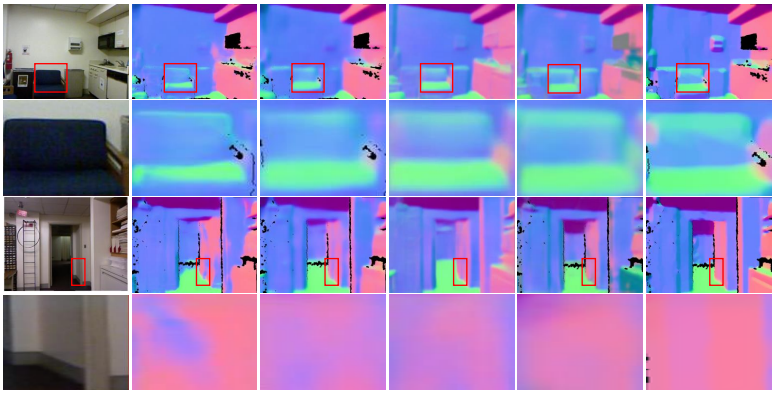


Figure 5: A comparison of depth maps estimated by different methods on the NYU Depth V2 dataset. The images used (with a resolution of 640×480) have been center-cropped to a resolution of 544×416 to remove any invalid points around the boundaries. (a) original RGB image; (b) depth maps generated by the method of [61]; (c) PAPNet [29]; (d) our proposed MIDNet; (e) ground truth.

Quantitative Comparisons: Monocular depth estimation results are shown in Table 2. It can be observed that our MIDNet achieves the best performance on the NYU Depth V2 dataset. Our MIDNet outperforms the second best method (i.e., PAPNet [63]) by 14% in terms of $RMSE$. The REL, $RMSE(log)$ and δ_1 results achieved by our MIDNet are better than PAPNet [63] by a large margin although the performance of δ_3 is slightly lower than PAPNet [63]. This is because the proposed method performs joint prediction of depth and normal estimations with a multi-stage information diffusion strategy, while PAPNet [63] only implements interaction at the task level. Through multi-stage



(a) Image (b) SkipNet (c) GeoNet (d) PAPNet (e) Our CPNet (f) GT

Figure 6: Visualization of surface normal maps estimated by different methods on the NYU Depth V2 dataset and images (with a resolution of 640×480) were center-cropped to a resolution of 544×416 to remove invalid points around the boundaries of images. (a) RGB image; (b) SkipNet [60]; (c) GeoNet [27]; (d) PAPNet [29]; (e) our MIDNet; (f) ground truth. Note that, invalid points are shown in black. The areas within the red rectangles in the first row are magnified and shown in the second row.

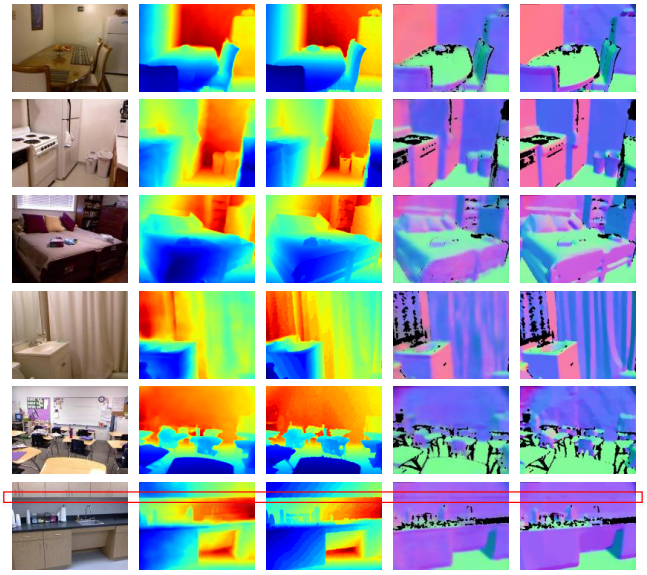


Figure 7: Visualization of depth maps and surface normal maps estimated by our MIDNet on the NYU Depth V2 dataset and images (with a resolution of 640×480) were center-cropped to a resolution of 544×416 to remove invalid points around the boundaries of images.

information interaction, the task-specific features from the depth estimation task can be adequately absorbed by the features for surface normal estimations, and vice versa. our MIDNet outperforms TRL [34] by a large margin in terms of $RMSE$, REL, $RMSE(log)$ and δ_1 although TRL conducts feature-level interaction in a hierarchical manner. This is because, our CCAM obtains the relevance of the two tasks with the guidance of the cross-relation matrix of the two task-specific feature maps, while TRL [34] directly concatenates the two task-specific feature maps to obtain the relevance. Besides, the task-level interaction proposed in our model can further boost the final performance.

The surface normal estimation results are shown in Table 3. Our MIDNet outperforms GeoNet [27] and PAPNet [29] in terms of *mean*, *median*, θ_2 and θ_3. That is mainly because our CCAM can extract more useful information from the depth estimation task for the surface normal estimation. In addition, our CCAM can boost the interaction

Table 4
Comparison to the state-of-the-art monocular depth estimation methods on the Stanford 2D-3D-Semantic Dataset.

Method	$RMSE$	REL	$log10$
	Lower the better		
Ron et al. [66]	0.825	0.219	0.094
Our MIDNet (d+n)	**0.424**	**0.126**	**0.055**

between depth maps and surface normal maps at multiple stages.

Qualitative Comparisons: Visualization results are shown in Figs. 5 and 6. It can be seen from Fig. 5 that our predicted depth maps are closer to ground truth depth maps than the other methods, especially in challenging areas (as shown by the red rectangles in Fig. 5). Besides, our MIDNet clearly produces finer details. For example, our MIDNet can infer that different parts of a table should have different depth values (as shown in the selected regions of the second column of Fig. 5). In contrast, other methods cannot achieve such detailed reasoning. In addition, our MIDNet can accurately predict the depth maps of small objects, such as the vase on the table in the third column of Fig. 5. In addition, the proposed algorithm obtains slightly worse depth estimation than that of PAPNet [29] above the red box in the second column of Fig. 5. However, we achieve the more accurate depth estimation in the remaining parts that contains more pixels, leading to better performance in average.

It can also be seen from the outlined regions in Fig. 6 that, our MIDNet has a stronger reasoning ability than other methods. Specifically, our MIDNet can accurately predict the surface normals of a sofa, especially on the backrest part, as shown in the second row of Fig. 6. Another example on the third and fourth rows also demonstrates similar results. Although interaction is performed between depth and surface normal estimation in GeoNet [21] and PAPNet [23], only the task-level interaction is used in these methods, which cannot extract sufficient information from the other task.

In Fig. 7, we provide more visualization results on this dataset to further demonstrate our accurate predictions. It can be observed from the bottom row of Fig. 7 that, although the split line between the cupboard and the wall is not labelled in the ground truth of surface normal map, our MIDNet can adaptively absorb useful information from the depth estimation branch to predict a clear split line in the normal map. These examples, in Fig. 7, also demonstrate that our MIDNet can predict accurate depth and surface normal estimations at the same time.

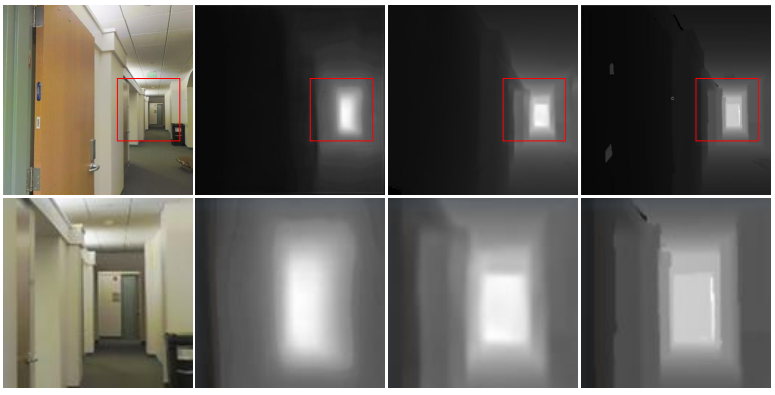
4.3.2. The Stanford 2D-3D-Semantic Dataset

We compare our proposed MIDNet with the state-of-the-art method [66] on this dataset. We used the same network structure as the NYU Depth V2 dataset to jointly perform depth and surface normal estimations. For a fair comparison, we followed the same testing protocol as [66]. Specifically, we used a subset of areas 2, 4 and 5 for training and a subset of areas 1, 3 and 6 for test. The Stanford 2D-3D-Semantic dataset also provides a binary mask for invalid raw depth pixels. The loss was only calculated on valid pixels during training, and evaluation was only performed on valid pixels during test.

Quantitative Comparisons: Quantitative depth estimation results are shown in Table 4. Our network outperforms the state-of-the-art method [66] by a very large margin. For example, our $RMSE$ (i.e., 0.424) is almost reduced to half as compared to [66] (i.e., 0.825). This clearly demonstrates the superiority of our network on the Stanford 2D-3D-Semantic dataset. That is because, our CCAM fully employs interactive information between the two tasks (i.e., depth estimation and surface normal estimation) to achieve optimal prediction on all pixels. In contrast, the previous method [66] only focuses on user-selected parts by introducing ordinal constraints. For surface normal estimation, our MIDNet also achieves impressive results (i.e., $mean$: 24.5° and $median$: 18.1°).

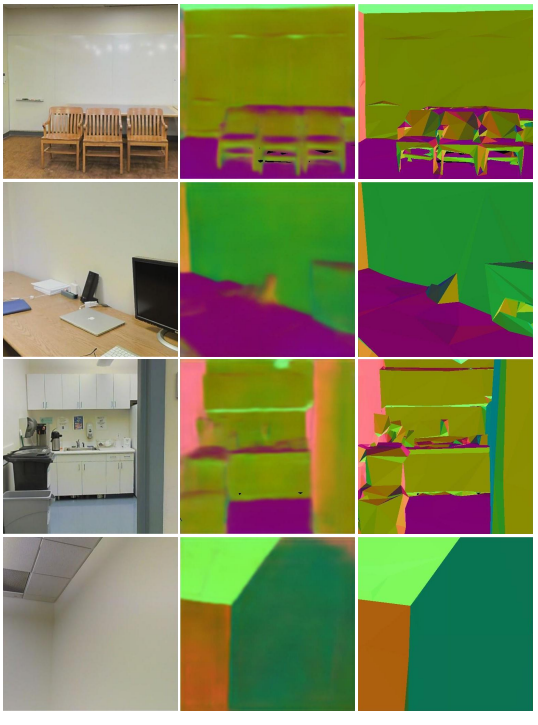
Qualitative Comparisons: Visualization results of depth estimation are shown in Fig. 8. It is clear that our predicted depth maps are closer to the ground truth. More specifically, the door in the outlined region in the first row of Fig. 8 is accurately predicted by our MIDNet, while only a blurry outline of the door is predicted by [66]. Besides, our MIDNet can predict the depth of the wall with detailed boundaries. However, almost the same depth is predicted by [66] for the whole wall, as shown in the second row of Fig. 8.

In Figs. 9 and 10, more qualitative results are shown for a better illustration. It can be observed that our MIDNet can simultaneously achieve accurate predictions of depth and surface normals. Besides, features can be automatically fused by the normal maps from their corresponding depth maps to improve normal prediction performance, such as the chairs in the third row of Fig. 10.



(a) (b) (c) (d)

Figure 8: Visualization of depth maps estimated by different methods on the Stanford 2D-3D-Semantic dataset and images (with a resolution of 1080×1080) were down-sampled to a resolution of 640×640 due to the limit of GPU memory. (a) RGB image; (b) Method of [66]; (c) our MIDNet; (d) ground truth. The areas within the red rectangles in the first row are magnified and shown in the second row.



(a) (b) (c)

Figure 9: Visualization of surface normal maps estimated by our MIDNet on the Stanford 2D-3D-Semantic dataset and images (with a resolution of 1080×1080) were down-sampled to a resolution of 640×640 due to the limit of GPU memory. (a) RGB image; (b) predicted normal map; (c) ground truth normal map.

Note that, only the task of monocular depth estimation is performed in [66]. In contrast, our MIDNet jointly predicts depth and surface normal maps. As shown in Fig. 9, our predicted normal maps are very close to the ground truth maps and contain accurate semantic information, e.g., the chairs. Besides, the detailed structures of objects are shown in predicted results.

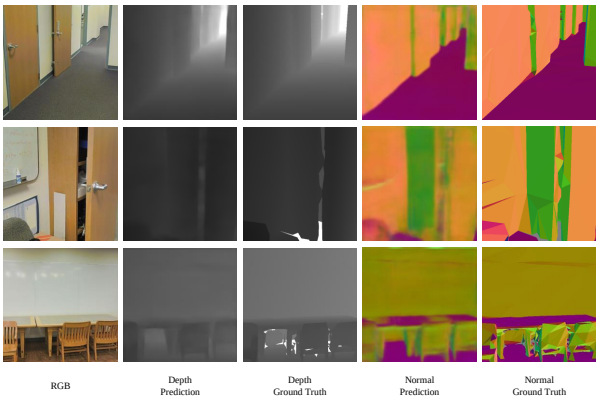


Figure 10: Visualization of depth maps and surface normal maps estimated by our MIDNet on the Stanford 2D-3D-Semantic dataset and images (with a resolution of 1080×1080) were down-sampled to a resolution of 640×640 due to the limit of GPU memory.

Table 5
Comparison to the state-of-the-art monocular depth estimation methods on the KITTI dataset.

Method	$RMSE$	REL	δ_1	δ_2	δ_3
	Lower the better		Higher the better		
DORN [5]	**2.727**	**0.072**	0.932	0.984	0.994
VNL [69]	3.258	**0.072**	**0.938**	**0.990**	**0.998**
Make3D [67]	8.734	0.280	0.601	0.820	0.920
Eigen et al. [4]	7.156	0.190	0.692	0.899	0.967
Liu et al. [40]	4.935	0.114	0.647	0.882	0.961
Semi.[68]	4.621	0.113	0.862	0.960	0.986
Guo et al.[6]	3.258	0.090	0.902	0.969	0.986
MIDNet (d)	3.462	0.085	0.916	0.975	0.986
MIDNet (d+n)	3.298	0.077	0.926	0.987	0.996

Table 6
Comparison to DORN and VNL in terms of FLOPs and number of parameters. Note that, $1GFLOPs = 10^9$ FLOPs.

Method	Input setting	FLOPs	#Para.
DORN [5]	257×353	120.62G	110.28M
VNL [63]	512×512	317.54G	90.44M
MIDNet (d+n)	512×512	**69.74G**	**30.02M**

4.3.3. The KITTI Dataset

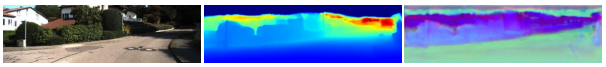
We compare our MIDNet with a number of monocular depth estimation methods [4–6, 40, 67–69] on this dataset. Two models of MIDNet were trained for fair comparison, including MIDNet(d) and MIDNet(d+n). The MIDNet(d) model was obtained by training our MIDNet on the KITTI dataset. In contrast, the MIDNet(d+n) model was obtained by first training our MIDNet on NYU Depth V2 for the surface normal estimation, and then training the depth estimation branch on the KITTI dataset (with the surface normal estimation branch being frozen).

Quantitative Comparisons: Depth estimation results are shown in Table 5. Although our MIDNet is slightly inferior to DORN [5] and VNL [69], it outperforms other existing methods [4, 6, 40, 67, 68] in almost all metrics and achieves better performance than the model without surface normal priors.

Two reasons can be concluded for this observation. **First**, our MIDNet is more lightweight than DORN [5] and VNL [63] (as shown in Table 6). Specifically, our MIDNet has only **30.02M** parameters for the whole model. In

Table 7
The complexity of each part of the proposed network. Note that, the input setting is set to 512×512 and the $1GFLOPs = 10^9$ FLOPs.

Modules	Backbone	Feature-level CCAM	Task-level CCAM
FLOPs	30.9G	12.68G	13.07G



(a) (b) (c)

Figure 11: Visualization of predictions on the KITTI dataset and images were randomly cropped to a resolution of 352×512. (a) RGB image; (b) predicted depth map; (c) predicted normal map.

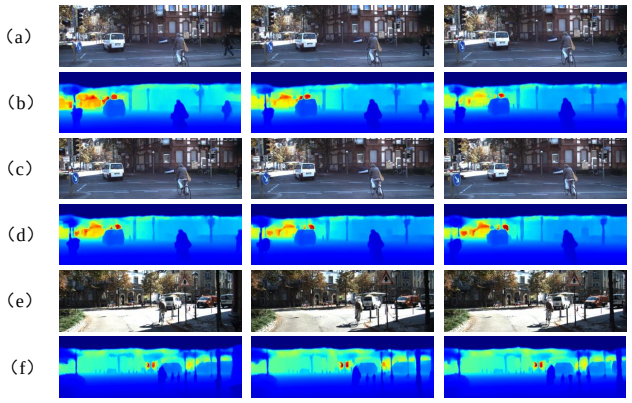


Figure 12: Visualization of predictions on the KITTI dataset and images were randomly cropped to a resolution of 352×512. (a), (c) and (e) RGB image; (b), (d) and (f) predicted depth map.

contrast, the encoder part of DORN has **51M** parameters, while VNL has **90.44M** parameters in total. In addition, the Floating Point Operations (FLOPs) and number of parameters (#Para.) of our MIDNet, DORN [5] and VNL [63] achieved on the KITTI dataset are listed in Table 7. It is clear that our MIDNet uses a significantly smaller number of parameters and consumes fewer calculations than DORN [5] and VNL [63]. Note that, the FLOPs and parameter numbers of DORN [5] and VNL [63] are calculated using their publicly available models [1,2].

Second, a different neural network structure is used in our MIDNet. Specifically, the encoder-decoder structures used in DORN [5] and VNL [63] can extract richer information. In contrast, the HRnet [36] used in our MIDNet can preserve detailed information but MIDNet cannot efficiently process the encoded features at our decoder stage for a large-scale outdoor environment.

Qualitative Comparisons: Visualization results are shown in the first row of Fig. 11. It can be observed that our MIDNet can predict both accurate depth and surface normal maps, even though there is no surface normal ground truth for network training on the KITTI dataset. Besides, our MIDNet can capture rich details of a scene, such as the boundaries of the road in Fig. 11 (b-c). More qualitative results are shown In Fig. 12. It can be observed that our MIDNet achieves reliable results on the validation split.

5. Conclusions

In this paper, we have presented a novel approach to jointly predict depth and surface normal maps from a single RGB image using a multi-stage information diffusion strategy. Our method employed a high-resolution, lightly shared backbone for extracting common features for both tasks and a multi-scale feature fusion module for integrating features at different scales. Additionally, we have introduced a cross-correlation attention map module to effectively mine

[1] https://github.com/dontLoveBugs/DORN_pytorch
[2] https://github.com/YvanYin/VNL_Monocular_Depth_Prediction

useful information between the two tasks at multiple stages for interactive learning. The results of our experiments on the NYU Depth V2, Stanford 2D-3D-Semantic and KITTI datasets demonstrate the effectiveness of our proposed information diffusion strategy. In future work, we will investigate the use of a more efficient encoder-decoder structure to improve performance in large-scale outdoor environments.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China (Nos. U20A20185, 61972435), the Natural Science Foundation of Guangdong Province (2022B1515020103), the Shenzhen Science and Technology Program (JCYJ2019080715220 9394), and the Australian Research Council (Grants DP150100294 and DP150104251).

References

[1] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[2] Hamid Laga, Yulan Guo, Hedi Tabia, Robert B Fisher, and Mohammed Bennamoun. *3D Shape analysis: fundamentals, theory, and applications*. John Wiley & Sons, 2018.

[3] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3D point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[4] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pages 2366–2374, 2014.

[5] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[6] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *European Conference on Computer Vision*, pages 484–500, 2018.

[7] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *International Conference on 3D vision*, pages 239–248. IEEE, 2016.

[8] Yuanzhouhan Cao, Tianqi Zhao, Ke Xian, Chunhua Shen, Zhiguo Cao, and Shugong Xu. Monocular depth estimation with augmented ordinal depth relationships. *IEEE Transactions on Image Processing*, 2018.

[9] Wenhui Zhou, Enci Zhou, Gaomin Liu, Lili Lin, and Andrew Lumsdaine. Unsupervised monocular depth estimation from light field image. *IEEE Transactions on Image Processing*, 29:1606–1617, 2020.

[10] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *Ieee transactions on pattern analysis and machine intelligence*, 2020.

[11] Shuai Li, Jiaying Shi, Wenfeng Song, Aimin Hao, and Hong Qin. Hierarchical object relationship constrained monocular depth estimation. *Pattern Recognition*, 120:108116, 2021.

[12] Feng Xue, Junfeng Cao, Yu Zhou, Fei Sheng, Yankai Wang, and Anlong Ming. Boundary-induced and scene-aggregated network for monocular depth prediction. *Pattern Recognition*, 115:107901, 2021.

[13] David F Fouhey, Abhinav Gupta, and Martial Hebert. Data-driven 3D primitives for single image understanding. In *IEEE International Conference on Computer Vision*, pages 3392–3399, 2013.

[14] David Ford Fouhey, Abhinav Gupta, and Martial Hebert. Unfolding an indoor origami world. In *European Conference on Computer Vision*, pages 687–702. Springer, 2014.

[15] L Ladickỳ, Bernhard Zeisl, and Marc Pollefeys. Discriminatively trained dense surface normal estimation. In *European Conference on Computer Vision*, volume 2, page 4, 2014.

[16] Shuai Liao, Efstratios Gavves, and Cees GM Snoek. Spherical regression: Learning viewpoints, surface normals and 3D rotations on n-spheres. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9759–9767, 2019.

[17] Xiaolong Wang, David Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 539–547, 2015.

[18] Peng Wang, Xiaohui Shen, Bryan Russell, Scott Cohen, Brian Price, and Alan L Yuille. Surge: Surface regularized geometry estimation from a single image. In *Advances in Neural Information Processing Systems*, pages 172–180, 2016.

[19] Seungryong Kim, Kihong Park, Kwanghoon Sohn, and Stephen Lin. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *European conference on computer vision*, pages 143–159. Springer, 2016.

[20] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014.

[21] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Towards unified depth and semantic prediction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2809, 2015.

[22] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.

[23] Liang Zhang, Le Wang, Xiangdong Zhang, Peiyi Shen, Mohammed Bennamoun, Guangming Zhu, Syed Afaq Ali Shah, and Juan Song. Semantic scene completion with dense crf from a single depth image. *Neurocomputing*, 318:182–195, 2018.

[24] Zhengfa Liang, Yulan Guo, Yiliu Feng, Wei Chen, Linbo Qiao, Li Zhou, Jianfeng Zhang, and Hengzhu Liu. Stereo matching using multi-level cost volume and multi-scale feature constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[25] Shuaifeng Zhi, Michael Bloesch, Stefan Leutenegger, and Andrew J Davison. Scenecode: Monocular dense semantic reconstruction using learned encoded scene representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11776–11785, 2019.

[26] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015.

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

[28] Ling Zhou, Zhen Cui, Chunyan Xu, Zhenyu Zhang, Chaoqun Wang, Tong Zhang, and Jian Yang. Pattern-structure diffusion for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4514–4523, 2020.

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

[30] Hao Xing, Yifan Cao, Maximilian Biber, Mingchuan Zhou, and Darius Burschka. Joint prediction of monocular depth and structure using planar and parallax geometry. *Pattern Recognition*, page 108806, 2022.

[31] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[32] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1871–1880, 2019.

[33] Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63(11):1–11, 2020.

[34] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *European Conference on Computer Vision*, pages 235–251, 2018.

[35] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *European Conference on Computer Vision*, pages 53–69, 2018.

[36] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[37] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[38] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2D-3D-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[39] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *IEEE International Conference on 3D Vision*, pages 11–20. IEEE, 2017.

[40] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):2024–2039, 2015.

[41] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5506–5514, 2016.

[42] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5354–5362, 2017.

[43] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. PAD-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018.

[44] Richard Chen, Faisal Mahmood, Alan Yuille, and Nicholas J Durr. Rethinking monocular depth estimation with adversarial training. *arXiv preprint arXiv:1808.07528*, 2018.

[45] Zhenyu Zhang, Chunyan Xu, Jian Yang, Junbin Gao, and Zhen Cui. Progressive hard-mining network for monocular depth estimation. *IEEE Transactions on Image Processing*, 27(8):3691–3702, 2018.

[46] Youngjung Kim, Hyungjoo Jung, Dongbo Min, and Kwanghoon Sohn. Deep monocular depth estimation via integration of global and local predictions. *IEEE Transactions on Image Processing*, 27(8):4131–4144, 2018.

[47] Wen Su, Haifeng Zhang, Quan Zhou, Wenzhen Yang, and Zengfu Wang. Monocular depth estimation using information exchange network. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[48] Xiaotian Chen, Xuejin Chen, and Zheng-Jun Zha. Structure-aware residual pyramid network for monocular depth estimation. *arXiv preprint arXiv:1907.06023*, 2019.

[49] Jae-Han Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2019.

[50] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä. Guiding monocular depth estimation using depth-attention volume. In *European Conference on Computer Vision*, pages 581–597. Springer, 2020.

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

[52] Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, and Jan-Michael Frahm. Vplnet: Deep single view normal estimation with vanishing points and lines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 689–698, 2020.

[53] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016.

[54] Iasonas Kokkinos. 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*, pages 6129–6138, 2017.

[55] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[56] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

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

[58] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[59] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[60] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2D-3D alignment via surface normal prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5965–5974, 2016.

[61] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3917–3925, 2018.

[62] Jogendra Nath Kundu, Phani Krishna Uppala, Anuj Pahuja, and R Venkatesh Babu. Adadepth: Unsupervised content congruent adaptation for depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2656–2665, 2018.

[63] Yin Wei, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *IEEE International Conference on Computer Vision*, 2019.

[64] Michael Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[65] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pages 794–803. PMLR, 2018.

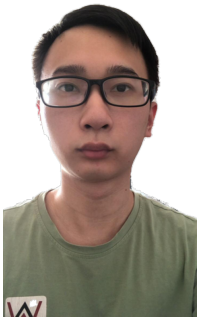
[66] Daniel Ron, Kun Duan, Chongyang Ma, Ning Xu, Shenlong Wang, Sumant Hanumante, and Dhritiman Sagar. Monocular depth estimation via deep structured models with ordinal constraints. In *IEEE International Conference on 3D Vision*, pages 570–577. IEEE, 2018.

[67] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3D: Learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2008.

[68] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6647–6655, 2017.

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

Zhiheng Fu received the B.E. degree in electric engineering from Northestern University (NEU), Shenyang, China, in 2015, and the M.E. degree in information and communication engineering from National University of Defense Technology (NUDT), Changsha, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Software Engineering (CSSE), UWA. His current research interests include 3D vision and deep learning.



Siyu Hong received the B.E. degree in computer science from Sun Yat-Sen University (SYSU), Guangzhou, China, in 2018, and the M.E. degree in electric and communication engineering from SYSU, Guangzhou, China, in 2020. His current research interests include 3D vision and deep learning.

Mengyi Liu received the B.S. degree in computer science and technology from Wuhan University, Hubei, China, in 2012, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2017. She was a Research Associate with the Center for Machine Vision Research, University of Oulu, Finland, in 2014. She was a Research Associate with the Language Technologies Institute, Carnegie Mellon University, USA, from 2015 to 2016. She was a Senior Algorithm Engineer in Alibaba Group from 2018 to 2021. She joined Kuaishou technology in 2021, where she is currently an Algorithm Engineering Manager. Her research interests include computer vision, pattern recognition, and machine learning.

Hamid Laga is currently a Professor at Murdoch University (Australia). His research interests span various fields of machine learning, computer vision, computer graphics, and pattern recognition, with a special focus on the 3D reconstruction, modeling and analysis of static and deformable 3D objects, and on machine learning for agriculture and health. He is the recipient of the Best Paper Awards at SGP2017, DICTA2012, and SMI2006.

Mohammed Bennamoun is Winthrop Professor in the Department of Computer Science and Software Engineering at UWA and is a researcher in computer vision, machine/deep learning, robotics, and signal/speech processing. He has published 4 books (available on Amazon), 1 edited book, 1 Encyclopedia article, 14 book chapters, 160+ journal papers, 250+ conference publications, 16 invited & keynote publications. His h-index is 57 and his number of citations is 13,000+ (Google Scholar). He was awarded 65+ competitive research grants, from the Australian Research Council, and numerous other Government, UWA and industry Research Grants. He successfully supervised 26+ PhD students to completion. He won the Best Supervisor of the Year Award at QUT (1998), and received award for research supervision at UWA (2008 & 2016) and Vice-Chancellor Award for mentorship (2016). He delivered conference tutorials at major conferences, including: IEEE Computer Vision and Pattern Recognition (CVPR 2016), Interspeech 2014, IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP) and European Conference on Computer Vision (ECCV). He was also invited to give a Tutorial at an International Summer School on Deep Learning (DeepLearn 2017).

Farid Boussaid received the M.S. and Ph.D. degrees in microelectronics from the National Institute of Applied Science, Toulouse, France, in 1996 and 1999, respectively. He joined Edith Cowan University, Perth, Australia, as a Post-Doctoral Research Fellow, and the Visual Information Processing Research Group as a member in 2000. He joined The University of Western Australia, Crawley, Australia, in 2005, where he is currently a Professor. His current research interests include smart CMOS vision sensors and image processing.

Yulan Guo received the B.E. and Ph.D. degrees from National University of Defense Technology (NUDT) in 2008 and 2015, respectively. He has authored over 150 articles at highly referred journals and conferences. His research interests lie in 3D vision, low-level vision, and machine learning. He served as an associate editor for IEEE Transactions on Image Processing, IET Computer Vision, IET Image Processing, and Computers & Graphics. He also served as an area chair for CVPR 2023/2021, ICCV 2021, and ACM Multimedia 2021. He is a Senior Member of IEEE and ACM.