

# Structure-Attentioned Memory Network for Monocular Depth Estimation

Jing Zhu<sup>1,2,3</sup>Yunxiao Shi<sup>1,2</sup>Mengwei Ren<sup>1,2</sup>Yi Fang<sup>1,2,3\*</sup>Kuo-Chin Lien<sup>4</sup>Junli Gu<sup>4</sup><sup>1</sup>NYU Multimedia and Visual Computing Lab, USA<sup>2</sup>New York University, USA<sup>3</sup>New York University Abu Dhabi, UAE<sup>4</sup>XMotors.ai

{jingzhu, yunxiao.shi, mengwei.ren, yfang}@nyu.edu {kuochin, junli}@xmotors.ai

## Abstract

Monocular depth estimation is a challenging task that aims to predict a corresponding depth map from a given single RGB image. Recent deep learning models have been proposed to predict the depth from the image by learning the alignment of deep features between the RGB image and the depth domains. In this paper, we present a novel approach, named Structure-Attentioned Memory Network, to more effectively transfer domain features for monocular depth estimation by taking into account the common structure regularities (e.g., repetitive structure patterns, planar surfaces, symmetries) in domain adaptation. To this end, we introduce a new Structure-Oriented Memory (SOM) module to learn and memorize the structure-specific information between RGB image domain and the depth domain. More specifically, in the SOM module, we develop a Memorable Bank of Filters (MBF) unit to learn a set of filters that memorize the structure-aware image-depth residual pattern, and also an Attention Guided Controller (AGC) unit to control the filter selection in the MBF given image features queries. Given the query image feature, the trained SOM module is able to adaptively select the best customized filters for cross-domain feature transferring with an optimal structural disparity between image and depth. In summary, we focus on addressing this structure-specific domain adaption challenge by proposing a novel end-to-end multi-scale memorable network for monocular depth estimation. The experiments show that our proposed model demonstrates the superior performance compared to the existing supervised monocular depth estimation approaches on the challenging KITTI and NYU Depth V2 benchmarks.

## Introduction

Depth estimation is an important component in many 3D computer vision tasks like visual Simultaneous Localization and Mapping (visual SLAM). Traditional approaches have made significant progress in binocular or multi-view depth estimation by taking advantage of geometry constraints of either spatial (i.e. stereo camera) or temporal (i.e. video sequence) pairs. With the prevalence of deep convolutional neural networks, researchers have been trying to relax the constraints by tackling monocular depth estimation. Recent works (Wang et al. (2015); Roy and Todorovic (2016);

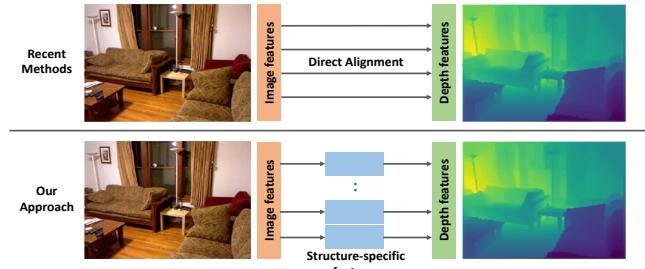


Figure 1: Different from the recent methods that directly align the features from different domains, we focus on the structure-specific domain adaption.

Kuznetsov, Stckler, and Leibe (2017); Kim et al. (2016); Fu et al. (2018); Eigen and Fergus (2015)) have demonstrated promising results using regression-based deep learning models. Their models are trained by minimizing image-level losses with supervised signal on predicted results. Nevertheless, the cross-modality variance between the RGB image and the depth map still makes monocular depth prediction an ill-posed problem. Based on this observation, some researchers have considered solving the problem with additional feature-level structural constraints by minimizing the cross-modality residual complexity between image features and depth features. Most existing methods either consider the pixel-wise or structure-wise alignment in this regard. For instance, several architectures utilize the micro discrepancy loss as similarity measures such like sum of squared differences, correlation coefficients (Myronenko and Song (2010)) and maximum mean discrepancy (Ghifary et al. (2015); Long et al. (2015)) to align the RGB images features with depth features from pixel to pixel independently without considering the spatial dependencies. Another line of work has tried to apply the adversarial adaptation methods (Kundu et al. (2018); Tzeng et al. (2017); Hoffman et al. (2015)) in conjunction with task-specific losses that concentrate on macro spatial distribution similarity between the image features and depth ones. In this paper, we seek a way to address this domain adaption challenge on both pixel-wise discrepancies and the structure dependencies by extracting the structure-specific information between the two domains (as shown in Figure 1).

\*indicates corresponding author

In order to explore the pixel-wise discrepancies as well as the structure dependencies between the image features and depth features, we propose a memorable domain adaptation network, with an image-encoder-depth-decoder regression network backbone, and a specifically designed Structure-Oriented Memory (SOM) module coupled with a cross-modality residual complexity loss to minimize the gap between latent distribution of the image and depth map from both the pixel-level and structure-level. Given the observation that similar type of scenes (e.g. roadside scenes) often share common structural regularities (e.g. repetitive structure patterns, planar surfaces, symmetries), a set of filters could be trained to learn a specific structural image-depth residual patterns. Therefore, in our SOM module, we build a Memorable Bank of Filters (MBF) to store and learn the structure-aware filters, then we construct an Attention Guided Controller (AGC) to learn to automatically select the appropriate filters (from the MBF) to capture the significant information from the given image features (generated by the image encoder) for the further depth estimation. Finally, the customized image features are fed into the depth decoder network to output the corresponding depth maps. Importantly, comparing to the direct alignment between the two domains features (e.g. direct applying  $L_1$  loss between  $Z_i$  and  $Z_d$ ), our introduced SOM module not only improves the fitting ability, but also reduces the training burden of the image encoder simultaneously. The experiments conducted on two well-known large scale benchmarks KITTI and NYU Depth V2, demonstrate that our proposed model obtains the state-of-the-art performance on monocular depth estimation tasks. Moreover, the performance margin between model trained with SOM and the one trained with direct alignment, validate the effectiveness of our proposed SOM module. In summary, our contributions in this paper are as follows:

- We introduce memory strategies to address monocular depth estimation by designing a novel Structure-Oriented Memory (SOM) module with a Memorable Bank of Filters (MBF) and an Attention Guided Controller (AGC) for feature-level cross-modality domain adaptation.
- We propose a novel end-to-end deep learning Structure-Attentioned Memory Network, which seamlessly integrates a front-end regression network with the SOM module that operates at feature-level to substantially improve the depth prediction performance.
- We achieve state-of-the-art performance on two large scale benchmarks: KITTI and NYU Depth V2, which validates the effectiveness of the proposed method.

The remainder of our paper is organized as follows. We present a brief review of the related literature in Section *Related Works*, after which we introduce the proposed method in details in Section *Proposed Method*. In Section *Experiments*, we provide the qualitative and quantitative experimental results, as well as ablation studies that demonstrate the effectiveness of the proposed method. Finally, we conclude the paper in Section *Conclusion*.

## Related Works

Monocular depth estimation is a fundamental problem in computer vision which has widespread application in graphics, robotics and AR/VR. While previous works mainly tackle this using hand-crafted image features or probabilistic models such as Markov Random Fields (MRFs) (Saxena, Sun, and Ng (2009)), recent success of deep learning based methods (Wang et al. (2015); Roy and Todorovic (2016); Kuznetsov, Stckler, and Leibe (2017); Kim et al. (2016); Fu et al. (2018); Eigen and Fergus (2015)) have inspired researchers to use deep learning techniques to address the challenging depth estimation problem. The learning based monocular depth estimation approaches can be mainly summarized into two categories, the supervised and the unsupervised/semi-supervised methods.

**Supervised Methods** A majority of works focus on supervised learning to use the learned features from CNNs to do accurate depth prediction. Eigen, Puhrsich, and Fergus (2014) first brought CNNs to depth regression task by integrating coarse and refined features with a two-stage network. The multi-task learning strategies were also applied in depth estimation to boost the performance. Liu, Gould, and Koller (2010) utilized the semantic segmentation as objectness cues for depth estimation. Furthermore, Shi and Pollefeys (2014) and Xu et al. (2018) performed joint prediction of the pixel-level semantic labels as well as the depth. Surface normal information was also adopted in many recent works (Eigen and Fergus (2015); Zhou et al. (2017); Wang et al. (2015); Qi et al. (2018)). Besides, some research works also demonstrated the robustness of multi-scale feature fusion in pixel-level prediction tasks (e.g. semantic segmentation, depth estimation). Fu et al. (2018) adopted the dilated convolution to enlarge the perceptive field without decreasing spatial resolution of the feature maps. In Buysens, Elmoataz, and Lzoray (2012)'s work, inputs at different resolutions are utilized to build a multi-stream architecture. Instead of regression, there are also methods that discretize the depth range and transfer the regression problem to a classification problem. In the work of Fu et al. (2018), the space-increasing discretization is proposed to reduce the over-strengthened loss for the large depth values.

**Unsupervised/Semi-supervised Methods** Another line of methods on monocular image depth prediction goes along the unsupervised/semi-supervised direction which mostly takes advantage of geometry constraints (e.g. epipolar geometry) on either spatial (between left-right pairs) or temporal (forward-backward) relationship. Garg et al. (2016) proposed to estimate the depth map from a pair of stereo images by imposing the left-right consistency loss. Zhan et al. (2018) jointly learned a single view depth estimator and monocular odometry estimator using stereo video sequences, which enables the use of both spatial and temporal photometric warp constraints. Moreover, following the trend of adversarial learning, the generative adversarial networks (GANs) have been utilized in the depth estimation problem. Kundu et al. (2018) proposed an unsupervised domain adaptation strategy for adapting depth predictions from synthetic RGB-D pairs to natural scenes in the depth estimation task.

**Cross-Modality Domain Adaptation** In addition to

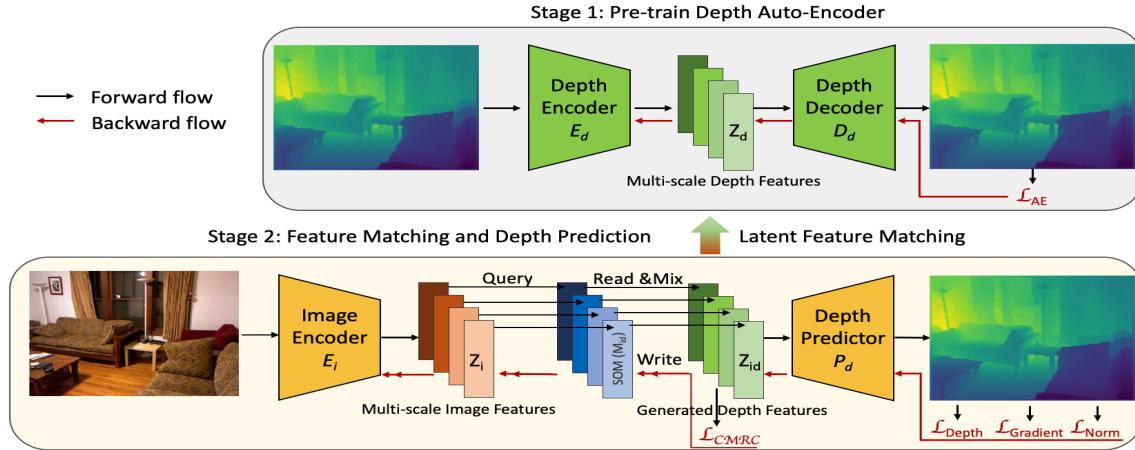


Figure 2: The pipeline of our proposed Structure-Attentioned Memory Network.

the recent depth estimation methods, research works focused on the cross-modality domain adaption are also highly relevant to ours. The existence of cross modality, or domain shift, is commonly seen in real-world application, which is the consequence of data captured by different sensors (e.g. optical camera, LiDAR or stereo camera), or varying conditions (i.e. background). In different domains, semantic labels are shared whereas the data distributions are usually different to a large extent. For example, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) in biomedical image analysis (Dou et al. (2018)); RGB images, depth maps and point clouds in 2D, 2.5D and 3D computer vision tasks. Numerous approaches have been proposed to address the domain adaptation needs in different visual tasks. Here we briefly review some domain adaption methods using deep learning techniques.

Most deep domain adaptation methods utilize a siamese architecture with two streams for source and target models respectively, and the network is trained with a discrepancy loss to minimize the pixel-wise shift between domains. Long et al. (2015) used maximum mean discrepancy together with a task-specific loss to adapt the source and target, while Sun and Saenko (2016) proposed the deep correlation alignment algorithm to match the mean and covariance. Bloesch et al. (2018) proposed to learn a dense representation using an auto-encoder. Mandikal et al. (2018) trained the network with  $L_1$  constrain in latent space to transfer feature from 2D to 3D in order to directly predict 3D point cloud from a single image. In our work, we aim to design a domain adaptive (SOM) module using memory mechanism, so that the image features can be automatically customized to obtain a better depth prediction.

## Proposed Method

The monocular depth estimation problem can be defined as a nonlinear mapping  $f : I \rightarrow Y$  from the RGB image  $I$  to the geometric depth map  $Y$ , which can be learned in a supervised fashion given a training set  $X = \{I^t, Y^t\}_{t=1}^N$ . To learn the mapping function, we propose Structure-Attentioned Memory Network as shown in Figure 2, which is composed of a (pre-trained) depth auto-encoder, an image encoder and

a depth predictor equipped with SOM module. All the components are trained into two stages. In the first stage, a series of ‘target’ depth features  $\{Z_d^t\}_{t=1}^k \in R^k$  are learned by training a depth map auto-encoder  $(E_d, D_d)$ . In the second stage, we train an image encoder  $E_i$ , SOM modules  $M_{id}$  and a depth predictor  $P_d$  to map the 2D image to the depth map in an end-to-end manner. Particularly,  $E_i$  encodes the RGB image to the ‘source’ image features  $\{Z_i^t\}_{t=1}^k \in R^k$ , which act as queries to obtain image-depth residual patterns from SOM module. The residual is then concatenated to the source feature to form a newly transferred feature set  $\{Z_{id}^t\}_{t=1}^k \in R^k$  (which is expected to be aligned with the target feature  $\{Z_d^t\}_{t=1}^k$  with supervision) is fed to the predictor  $P_d$  to estimate the output depth map. We will elaborate the network structures from two stages separately.

### Stage 1: Depth Auto-Encoder

In order to learn a strong and robust prior over the depth map as a reference in the latent matching process, we train a depth auto-encoder  $(E_d, D_d)$  which takes a ground truth depth map  $Y_d \in R^{M \times N}$  as input, and outputs a reconstructed depth map  $\hat{Y}_d \in R^{M \times N}$ . As shown in Figure 3 (Stage 1), we use the DenseNet based encoder-decoder structure. Specifically, DenseNet-121 is utilized for constructing the depth encoder (Figure 3 (a)), in which four feature maps with cascading resolutions are extracted from different blocks (shallow to deep) for depth decoding. In order to make sure that the object contours as well as details are well preserved, we use a Feature Pyramid Network (FPN) to build the depth decoder, fusing multi-scale features in a pyramid structure. Specifically, as shown in Figure 3 (b), four features with sizes  $1/4, 1/8, 1/16$  and  $1/32$  of the input are derived. Starting from the deepest feature, each feature map is first upsampled by a factor of 2, and element-wisely added to its following feature map. After the fusion process of the multi-scale feature maps, each of the newly generated feature maps is upsampled to size of  $1/4$  the original input (or the size of the shallowest feature map), and concatenated together to form a feature volume. Finally, the output depth map is predicted via extra CNN layers on the concatenated feature volume. The FPN decoder is able to preserve details

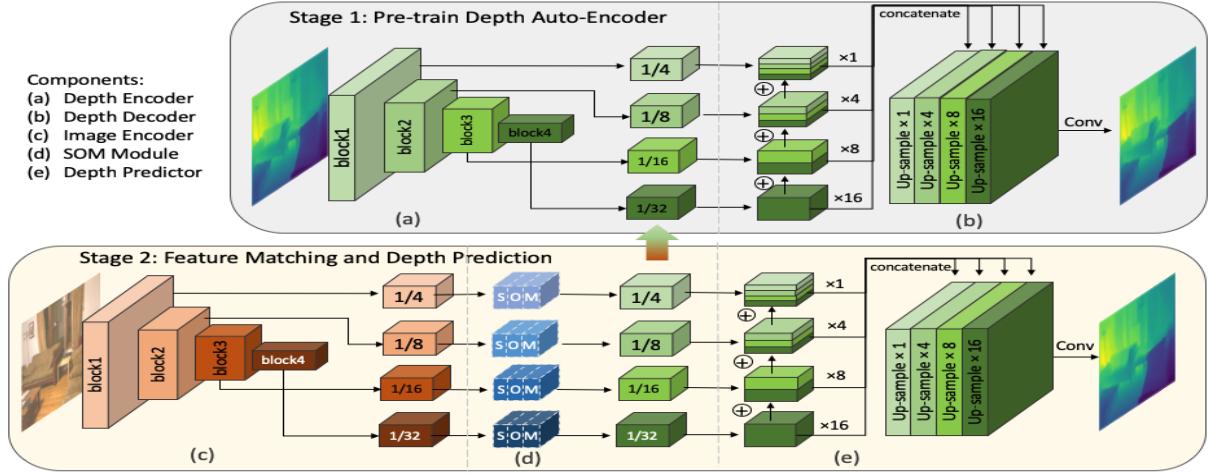


Figure 3: The network structure of Structure-Attentioned Memory Network.

in the depth map decoding process. We will show more experimental comparison between different decoder structure to demonstrate its effectiveness in the *Experiments* section.

## Stage 2: Depth Prediction with SOM Module for Latent Space Adaptation

In the second stage, we aim to train the network in an end-to-end manner to effectively transfer the features derived from image encoder  $E_i$  from image domain to depth domain, as a strong prior over the ground truth depth, so as to better deduce the depth from the transferred prior. To this end, this stage contains three major components as shown in Figure 3 (c), (d) and (e): the image encoder, the SOM module for latent space adaptation, and the depth predictor ( $E_i, M_{id}, P_d$ ). Each component of the network will be explained below.

**Image Encoder and Depth Predictor as Regression Backbone** In order to make sure that the network derive both depth features and image features at the same scale, we design the encoder-decoder based backbone ((c) and (d) in Figure 3) for stage 2 exactly the same as those of stage 1 but without weight sharing. Specifically, the structure of image encoder  $E_i$  ((c) in Figure 3) is identical to that of depth encoder  $E_d$  ((e) in Figure 3), and similarly for  $D_d$  ((b)) and  $P_d$  ((e)).

**SOM Module for Latent Space Adaptation** In the latent space, we propose an additional structure oriented memory module consisting of two collaborative units: a Memorable Bank of Filters (MBF) that stores a bank of learned filters to detect the cross-modality residual complexity between the depth feature and the image feature, and an Attention Guided Controller (AGC) which controls the interaction between the image feature with the MBF. The image feature as a specific query feature selects filters from MBF with an attention guided read controller, and the MBF is updated through a write controller that is naturally integrated into the back propagation to make the network can be trained end-to-end. The proposed SOM reading and writing process are as follows.

**SOM Reading** Different from reading by ‘addressing’ in general memory concept, the proposed SOM module

is reading by ‘attention’, which means each memory slot is assigned with a weight, and the whole memory is merged per weights as reading output. As demonstrated in Figure 4, given the query feature  $Z_i$ , in order to obtain weights for each memory slot, we build a LSTM-based read controller to learn the weights. Specifically, each filter from the memory slot  $\{M_t\}_{t=1}^n$  is firstly convolved on the feature, and the intermediate outputs are denoted as  $\{x_t\}_{t=1}^n$ , where  $n$  is the memory size, and  $x_t$  is formulated as:  $x_t = W_t * Z_i + b_t$ ,  $M_t = (W_t, b_t)$ ,  $W_t$  is the kernel,  $b_t$  is the bias, and  $*$  is the convolution operation. The intermediate outputs  $\{x_t\}_{t=1}^n$  could be thought of as the ‘unweighted/unbiased’ output that takes each filter/memory slot equally. Then in order to further add weighted attention on the result pool, a Bi-Directional Convolutional Long Short Term Memory is applied as the read controller on  $\{x_t\}_{t=1}^n$  to explore the correlation within the pool, so as to aggregate the memory slots with strong attention. Particularly, read controller processes  $\{x_t\}_{t=1}^n$  from two directions and computes the forward hidden sequence  $h_f$  by iterating the input from  $t = 1$  to  $n$ , and the backward hidden sequence  $h_b$  by iterating the input from  $t = n$  to 1. The forward/backward flow of the LSTM cell is formulated as below:

$$\begin{aligned} i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \\ f_t &= \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \\ o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \circ c_t + b_o) \\ h_t &= o_t \circ \tanh(c_t) \end{aligned}$$

where  $h$  is the hidden sequence,  $\sigma$  is the logistic sigmoid function,  $*$  is the convolution operator and  $\circ$  denotes the Hadamard product.  $i_t, f_t, o_t, c_t$  represent input gate, forget gate, output gate, and cell activation vector respectively, and  $W_{hi}$  is the hidden-input gate matrix, while  $W_{xo}$  is the input-output gate matrix. The final attention sequence  $\alpha$  is computed with regard to both  $h_f$  and  $h_b$  as follows:  $\alpha_t = \text{softmax}(W_{hfy}h_{f(t)} + W_{hby}h_{b(t)} + b_y)$ , where  $t = 1$  to  $n$ , and each  $y$  after softmax operation in the output sequence is associated with the weight for each memory slot (refer to  $\alpha$  value in Figure 4, the redder the color, the higher

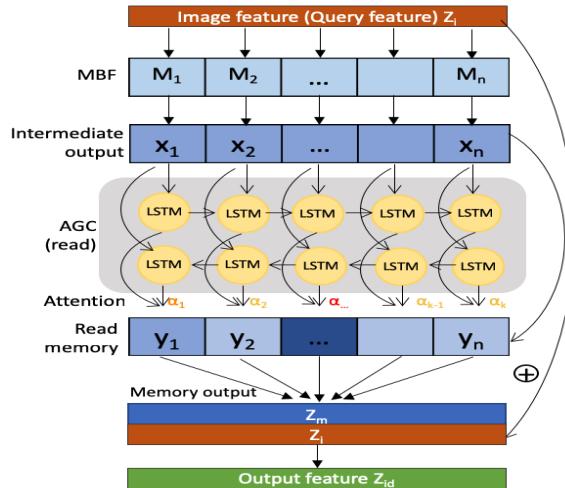


Figure 4: The SOM reading process (of a single SOM module).

the attention), therefore  $\sum_{i=1}^k \alpha_i = 1$ . The memory output  $Z_m$  is a combination of the output sequence that focuses more on the slot with higher attention, while less on lower attention value:  $Z_m = \sum_{t=1}^n y_t, y_t = \alpha_t x_t$ . Finally,  $Z_m$  is concatenated with the query feature itself to reproduce a transferred feature  $Z_{id}$  that is supposed to match the distribution of the depth feature  $Z_d$ .

**SOM Writing** The proposed memory writer can be seamlessly integrated to network back propagation. The attention learned from the read controller will also operate in the memory writing process, and specifically, the slot with higher attention will be updated to a larger extent and vice versa. As shown in Figure 2, there are two backward flows that affect the writing of the memory (red arrows in Figure 2): one comes from the output branch, and the other comes from the latent matching branch. The update rule could be formulated (in a simplified form) as  $W_t \leftarrow W_t + \alpha_t \eta \Delta_{W_t}$ , where  $\alpha_t$  is the attention for each slot,  $\eta$  is the learning rate, and  $\Delta_{W_t}$  is the total gradient from both branches.

## Learning objectives

We design multiple objectives to constrain the joint training of the network with details as follows.

**Depth Estimation Objective** The depth estimation objective poses constraints on the front-end pipeline of the single image depth estimation. A common way for supervising regression tasks is to adopt  $L_1$  or  $L_2$  loss between the prediction and the ground truth, which means that larger values have much heavier influence on the loss. However, in depth estimation task, the larger the depth value is, the farther the object is to the camera, which means that the information is less rich for the estimator, leading to unnecessarily large loss (Fu et al. (2018)). Therefore, in order to reduce the over-emphasized error on large depth values, we use the logarithm mean squared error ( $\text{RMSE}_{\log}$ ) loss to make the predictor focus more on closer objects which makes up the main portion in a depth map. The objective is formulated as  $\mathcal{L}_{depth} = \sqrt{\frac{1}{N} \sum_{i \in N} \|\log(d_i) - \log(d_i^*)\|^2}$ , where  $d$  is the ground truth depth map, while  $d^*$  is the predicted depth map.

**Auto-Encoder Objective** The objective for the depth auto-encoder is utilized in the first training stage. To make sure that the depth features and the image features are in the same scale with same constraints, we also applied the  $\text{RMSE}_{\log}$  on the auto-encoder as  $\mathcal{L}_{AE} = \sqrt{\frac{1}{N} \sum_{i \in N} \|\log(d_i) - \log(\hat{d}_i)\|^2}$ , where  $d$  is the ground truth depth map, while  $\hat{d}$  is the reconstructed depth map.

**Cross-Modality Residual Complexity Objective** The latent adaptation objective is applied to constrain the SOM module to minimize feature distribution discrepancies. We use  $L_1$  loss between the ‘target’ depth features (pretrained from stage 1) and the SOM transferred image features. The objective is a sum of feature alignment losses at different levels as  $\mathcal{L}_{CMRC} = \sum_k \|Z_{id}^k - Z_d^k\|_1$ , where  $k$  is the number of features involved in latent matching.

**Gradient and Surface Normal Constraints** To further strengthen the network by pulling out the model from local minima, we added extra constraints on the predicted depth map including the gradient loss and the surface normal loss to finetune the training following commonly used techniques. The gradient loss is defined as  $\mathcal{L}_{gradient} = \frac{1}{N} \sum_{i=1}^N \|\nabla d_i - \nabla d_i^*\|_1$ , and specifically, we adopt Sobel filter to calculate the gradient both vertically and horizontally;  $\nabla d$  is the image gradient of the ground truth depth map, while  $\nabla d^*$  is the image gradient of the predicted depth map. The surface normal loss is defined as the similarity between the surface normal of the ground truth depth map with the predicted depth map as  $\mathcal{L}_{normal} = \frac{1}{N} \sum_{i=1}^N (1 - \frac{\langle \nabla d_i, \nabla d_i^* \rangle}{\|\nabla d_i\|_2 \|\nabla d_i^*\|_2})$ , formulated with the corresponding gradient.

In total, the training objectives are summarized as follows: (1) In training stage 1, the total loss is:  $\mathcal{L}_{S_1} = \mathcal{L}_{AE}$ ; (2) In training stage 2, the total loss is a weighted sum of  $\mathcal{L}_{depth}$ ,  $\mathcal{L}_{CMRC}$ ,  $\mathcal{L}_{gradient}$  and  $\mathcal{L}_{normal}$ , which is formulated as:  $\mathcal{L}_{S_2} = \lambda_{depth} \mathcal{L}_{depth} + \lambda_{CMRC} \mathcal{L}_{CMRC} + \lambda_{gradient} \mathcal{L}_{gradient} + \lambda_{normal} \mathcal{L}_{normal}$ , where  $\lambda$  is the weight for each objective.

## Experiments

In this section, we present our experiments on two large-scale datasets by introducing the implementation details, benchmark performance, and ablation studies validating the effectiveness of the proposed approach.

**Implementation Details** The proposed method is implemented using the TensorFlow 1.10 framework and runs on a single NVIDIA TITAN X GPU with 12 GB memory. The encoder-decoder structure from both stage 1 and stage 2 are identical but without weight sharing. The depth auto-encoder is trained from scratch, while the image encoder is initialized with ImageNet (Russakovsky et al. (2015)) pre-trained parameters. For multi-scale feature fusion, we consider four levels of feature maps which are derived from different blocks of the DenseNet-121 backbone with the feature map sizes 1/4, 1/8, 1/16 and 1/32 of the input images. For instance, in NYU Depth V2 dataset, with the input resolution  $480 \times 640$ , four feature maps with cascading sizes  $120 \times 160, 60 \times 80, 30 \times 40, 15 \times 20$  are extracted. The

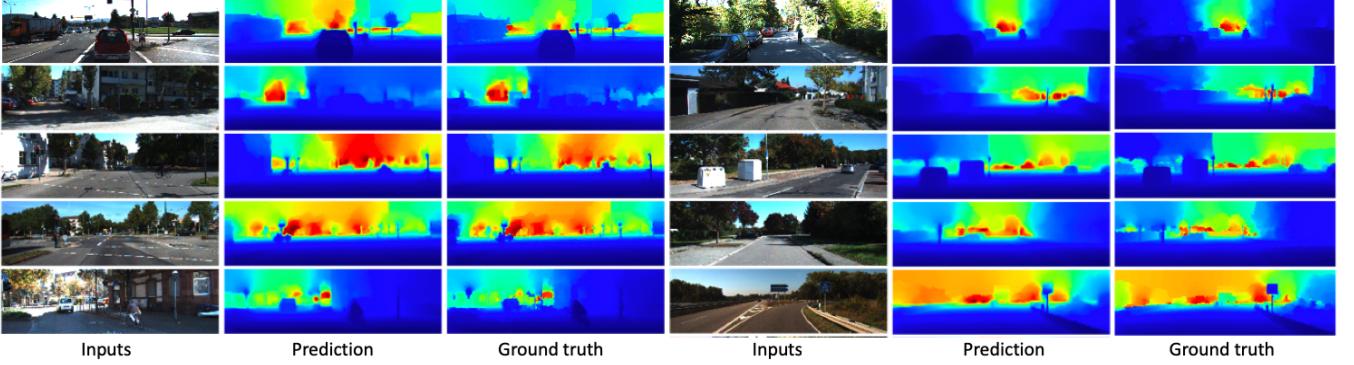


Figure 5: Results on KITTI validation set.

Table 1: Performance on KITTI validation set. All scores are evaluated on Eigen split (Eigen and Fergus (2015)).

Method	Error (lower is better)				Accuracy (higher is better)		
	Abs Rel	Sq Rel	RMSE	RMSE <sub>log</sub>	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Saxena, Sun, and Ng (2009)	0.280	3.012	8.734	0.361	0.601	0.820	0.926
Liu et al. (2016)	0.217	1.841	6.986	0.289	0.647	0.882	0.961
Zhou et al. (2017)	0.208	1.768	6.858	-	0.678	0.885	0.957
Eigen, Puhrsch, and Fergus (2014)	0.190	1.515	7.156	0.270	0.692	0.899	0.967
Garg et al. (2016)	0.177	1.169	5.285	-	0.727	0.896	0.962
Kundu et al. (2018)	0.167	1.257	5.578	0.237	0.771	0.922	0.971
Zhan et al. (2018)	0.135	1.132	5.585	0.229	0.820	0.933	0.971
Godard, Aodha, and Brostow (2017)	0.114	0.898	4.935	0.206	0.861	0.949	0.976
Kuznetsov, Stckler, and Leibe (2017)	0.113	0.741	4.621	0.189	0.862	0.960	0.986
<b>Ours</b>	<b>0.097</b>	<b>0.398</b>	<b>3.007</b>	<b>0.133</b>	<b>0.913</b>	<b>0.985</b>	<b>0.997</b>

network is trained with initial learning rate 0.001, and decreased every 10 epochs. The weight decay and momentum set to  $10^{-6}$  and 0.9 respectively. We used the Adam optimizer and batch normalization during training, with normalization decay 0.97. We set the weights for each objective as  $\lambda_{depth} = 1$ ,  $\lambda_{gradient} = 1$ ,  $\lambda_{norm} = 1$ , and  $\lambda_{CMRC} = 2$ . The gradient loss is added after 4k steps of training, and the surface normal loss is added after 8k steps of training.

**Data Augmentation** We employ several data augmentation techniques on NYU Depth V2 dataset to prevent overfitting from limited amount of data, including: (i) *Random Cropping* by 0–10% of the image height/ width; (ii) *Scaling* the original image by the factor interval of [0.75, 1.25]; (iii) *Random Flipping* 50% of the images horizontally; (iv) *Rotating* the images randomly with the degree of  $[-10^\circ, 10^\circ]$ ; (v) *Color jitter* of brightness (by -10 to 10 of original value), contrast (by a factor of 0.5 to 2.0), saturation and hue (by -20 to 20 of original value).

**Evaluation Metrics** Below is a list of evaluation metrics the quantitative evaluation is performed: (1) the absolute mean relative error (Abs Rel):  $\frac{1}{N} \sum_{i \in N} \frac{|d_i - d_i^*|}{d_i^*}$ , (2) the squared relative error (Sq Rel):  $\frac{1}{N} \sum_{i \in N} \frac{\|d_i - d_i^*\|^2}{d_i^*}$ , (3) the root mean squared error (RMSE):  $\sqrt{\frac{1}{N} \sum_{i \in N} \|d_i - d_i^*\|^2}$ , (4) log mean squared error (RMSE<sub>log</sub>):  $\sqrt{\frac{1}{N} \sum_{i \in N} \|\log(d_i) - \log(d_i^*)\|^2}$ , (5) average log 10 error (Avg log<sub>10</sub>):  $\frac{1}{N} \sum_{i \in N} |\log_{10}(d_i) - \log_{10}(d_i^*)|$ , and (6) accuracy with threshold  $t$  ( $t=1.25, 1.25^2, 1.25^3$ ):

$$\frac{1}{N} \sum_{i \in N} \mathbb{1}_{\{\delta = \max(\frac{d_i^*}{d_i}, \frac{d_i}{d_i^*}) < t\}}.$$

**Results on KITTI Dataset (Eigen split)** The KITTI dataset is a large scale dataset for autonomous driving, which contains depth images captured with LiDAR sensor mounted on a driving vehicle. In our experiment, to compare the results at the same level, we follow the experimental protocol proposed by Eigen and Fergus (2015), in which around 22600 images (resolution  $384 \times 1280$ ) from 32 scenes are utilized as training data, and around 800 images from 29 scenes are used for validation. Following the previous works, the depth value of the RGB image is scaled to 0-80m. During training, the depth maps are down-scaled to resolution  $192 \times 640$ , and up-sampled to the original size in evaluation process. Table 1 shows the comparison with the state-of-the-art methods on KITTI dataset. We compared with state-of-the-art methods (Saxena, Sun, and Ng (2009); Liu, Shen, and Lin (2015); Zhou et al. (2017); Eigen, Puhrsch, and Fergus (2014); Garg et al. (2016); Kundu et al. (2018); Zhan et al. (2018); Godard, Aodha, and Brostow (2017); Kuznetsov, Stckler, and Leibe (2017)). Particularly, the methods proposed by Saxena, Sun, and Ng (2009); Liu, Shen, and Lin (2015); Zhou et al. (2017); Eigen, Puhrsch, and Fergus (2014); Kundu et al. (2018) only employ monocular images in both training and testing, while approaches in Zhan et al. (2018); Garg et al. (2016); Kuznetsov, Stckler, and Leibe (2017); Godard, Aodha, and Brostow (2017) are unsupervised methods that use stereo images in training and apply single image during testing. The proposed method outperforms all these methods by a large margin, and Figure 5 displays a few visualized prediction results on examples ran-

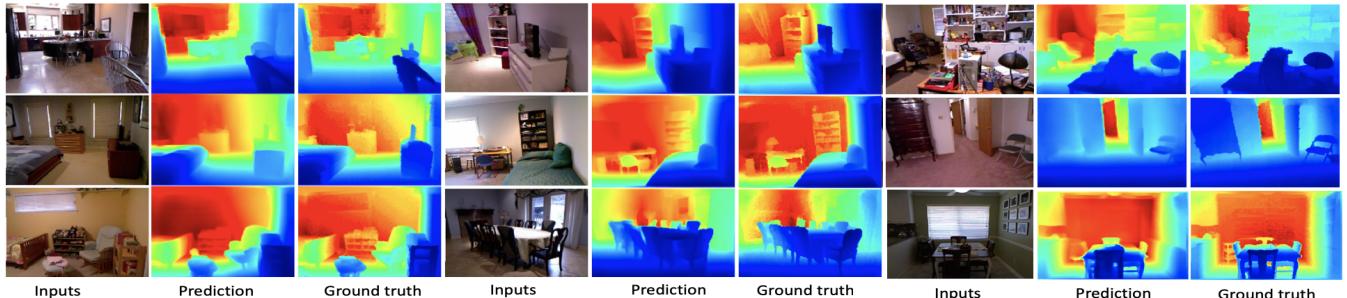


Figure 6: Examples of predicted depth maps on NYU V2 Depth dataset.

Table 2: Performance on NYU Depth V2.  $\delta_1 : \sigma < 1.25$ ,  $\delta_2 : \sigma < 1.25^2$ ,  $\delta_3 : \sigma < 1.25^3$ .

Method	Error			Accuracy		
	Rel	RMSE	$\log_{10}$	$\delta_1$	$\delta_2$	$\delta_3$
Saxena, Sun, and Ng (2009)	0.349	1.214	-	0.447	0.745	0.897
Karsch, Liu, and Kang (2012)	0.35	1.2	0.131	-	-	-
Liu, Salzmann, and He (2014)	0.335	1.06	0.127	-	-	-
Shi and Pollefeys (2014)	-	-	-	0.542	0.829	0.941
Zhuo et al. (2015)	0.305	1.04	-	0.525	0.838	0.962
Li et al. (2015)	0.232	0.821	0.094	0.621	0.886	0.968
Wang et al. (2015)	0.220	0.745	-	0.605	0.890	0.970
Xu et al. (2018)	0.214	0.792	0.091	0.643	0.902	0.977
Liu et al. (2016)	0.213	0.759	0.087	0.650	0.906	0.976
Roy and Todorovic (2016)	0.187	0.744	-	-	-	-
<b>Ours (<math>E_i + D_{pure}</math>)</b>	<b>0.231</b>	<b>0.828</b>	<b>0.095</b>	<b>0.631</b>	<b>0.889</b>	<b>0.968</b>
<b>Ours (<math>E_i + D_{FPN}</math>)</b>	<b>0.229</b>	<b>0.803</b>	<b>0.092</b>	<b>0.633</b>	<b>0.891</b>	<b>0.969</b>
<b>Ours (<math>E_i + D_{FPN} + align</math>)</b>	<b>0.148</b>	<b>0.627</b>	<b>0.075</b>	<b>0.802</b>	<b>0.944</b>	<b>0.986</b>
<b>Ours (<math>E_i + D_{FPN} + SOM</math>)</b>	<b>0.136</b>	<b>0.604</b>	<b>0.067</b>	<b>0.814</b>	<b>0.959</b>	<b>0.990</b>

domly chosen from the validation dataset.

**Results on NYU Depth V2 Dataset** The NYU Depth V2 dataset contains 120K pairs of RGB-D (resolution  $480 \times 640$ ) captured by Kinect. The dataset is manually selected and annotated into 1449 RGB-D pairs, in which 795 images are used for training, and the rest for validation. The depth value ranges from 0 to 10m. In the training process, the depth maps are down-scaled to resolution  $120 \times 160$ , and in testing/evaluation, the predicted depth map is upsampled to the original resolution. Table 2 shows the comparison of the proposed method with state-of-the-art methods (official test split). We compare with both hand-crafted feature based approaches (Saxena, Sun, and Ng (2009); Karsch, Liu, and Kang (2012); Shi and Pollefeys (2014)) and deep learning based ones (Liu, Salzmann, and He (2014); Zhuo et al. (2015); Li et al. (2015); Wang et al. (2015); Xu et al. (2018); Liu et al. (2016); Roy and Todorovic (2016)). Figure 6 shows examples of predicted depth maps on the NYU Depth V2 dataset.

**Ablation Studies** To further demonstrate the effectiveness of the proposed method, we conduct ablation studies from two aspects on NYU Depth V2 dataset. Firstly, we compare the performance of the depth estimation pipeline with different decoder structures: (1) The decoder that simply uses symmetric structure with the encoder that cascadingly upsample the feature map until the output size. (2) The decoder that takes four different feature maps from the encoder and fuses them in a pyramid fashion (as described in Section ). The qualitative comparison are shown in Table 2 ( $E_i + D_{pure}$  and  $E_i + D_{FPN}$ ). As can be seen from the evaluation results, the decoder structure with pyramid multi-

sacle feature fusion out-performs the one that only takes the latent feature as input by a large margin, especially in the  $\delta_1 < 1.25$  metric. Therefore, it is obvious that the mixture of features from different levels are beneficial for the details compensation (i.e. contour, edges).

To validate the effectiveness of the proposed SOM module, we compare the performance of the proposed method with SOM settings against direct alignment and analyze the results. Firstly, we add the feature alignment loss for latent feature maps based on the  $E_i + D_{FPN}$  structure to test the performance of direct feature alignment ( $E_i + D_{FPN} + align$ ). The quantitative results of direct alignment rarely improved compared with the one that is trained without feature alignment loss, reflecting the limited capability of the encoder for feature adaptation. Then, we add the SOM module at feature level ( $E_i + D_{FPN} + SOM$ ) and compare the results with the baseline structure that goes without memory. The large margin quantitative improvement in Table 2 implies that structure-specific feature alignment with memory mechanism (SOM) is superior to other approaches such as direct alignment.

## Conclusion

In this paper, we developed a novel memory guided network named Structure-Attentioned Memory Network for monocular depth estimation, consisting of the encoder-decoder based structure, as well as the external SOM module which is trained to learn and memorize the structure attentioned image-depth-residual pattern in cross-modality latent alignment. The proposed method achieves state-of-the-art performance on challenging large-scale benchmarks, and each component is validated to be effective in the ablation study.

## References

- Bloesch, M.; Czarnowski, J.; Clark, R.; Leutenegger, S.; and Davison, A. J. 2018. Codeslam learning a compact, optimisable representation for dense visual slam. In *CVPR*.
- Buyssens, P.; Elmoataz, A.; and Lzoray, O. 2012. Multiscale convolutional neural networks for visionbased classification of cells. In *Asian Conference on Computer Vision*, 342–352.
- Dou, Q.; Ouyang, C.; Chen, C.; Chen, H.; and Heng, P. A. 2018. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss.

- Eigen, D., and Fergus, R. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *IEEE International Conference on Computer Vision*, 2650–2658.
- Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. In *International Conference on Neural Information Processing Systems*, 2366–2374.
- Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; and Tao, D. 2018. Deep ordinal regression network for monocular depth estimation.
- Garg, R.; Vijay, K. B. G.; Carneiro, G.; and Reid, I. 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. 740–756.
- Ghifary, M.; Kleijn, W. B.; Zhang, M.; and Balduzzi, D. 2015. Domain generalization for object recognition with multi-task autoencoders. 2551–2559.
- Godard, C.; Aodha, O. M.; and Brostow, G. J. 2017. Unsupervised monocular depth estimation with left-right consistency. In *Computer Vision and Pattern Recognition*, 6602–6611.
- Hoffman, J.; Tzeng, E.; Darrell, T.; and Saenko, K. 2015. Simultaneous deep transfer across domains and tasks. 30(31):4068–4076.
- Karsch, K.; Liu, C.; and Kang, S. B. 2012. Depth extraction from video using non-parametric sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(11):2144.
- Kim, S.; Park, K.; Sohn, K.; and Lin, S. 2016. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *European Conference on Computer Vision*, 143–159.
- Kundu, J. N.; Uppala, P. K.; Pahuja, A.; and Babu, R. V. 2018. Adadepth: Unsupervised content congruent adaptation for depth estimation.
- Kuznetsov, Y.; Stckler, J.; and Leibe, B. 2017. Semi-supervised deep learning for monocular depth map prediction. 2215–2223.
- Li, B.; Shen, C.; Dai, Y.; Hengel, A. V. D.; and He, M. 2015. 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*, 1119–1127.
- Liu, F.; Shen, C.; Lin, G.; and Reid, I. 2016. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(10):2024–2039.
- Liu, B.; Gould, S.; and Koller, D. 2010. Single image depth estimation from predicted semantic labels. In *Computer Vision and Pattern Recognition*, 1253–1260.
- Liu, M.; Salzmann, M.; and He, X. 2014. Discrete-continuous depth estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 716–723.
- Liu, F.; Shen, C.; and Lin, G. 2015. Deep convolutional neural fields for depth estimation from a single image. In *Computer Vision and Pattern Recognition*, 5162–5170.
- Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning transferable features with deep adaptation networks. 97–105.
- Mandikal, P.; Murthy, N.; Agarwal, M.; and Babu, R. V. 2018. 3d-Imnet: Latent embedding matching for accurate and diverse 3d point cloud reconstruction from a single image.
- Myronenko, A., and Song, X. 2010. Intensity-based image registration by minimizing residual complexity. *IEEE Transactions on Medical Imaging* 29(11):1882.
- Qi, X.; Liao, R.; Liu, Z.; Urtasun, R.; and Jia, J. 2018. Geonet : Geometric neural network for joint depth and surface normal estimation.
- Roy, A., and Todorovic, S. 2016. Monocular depth estimation using neural regression forest. In *Proceedings of the IEEE CVPR*, 5506–5514.
- Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; and Bernstein, M. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.
- Saxena, A.; Sun, M.; and Ng, A. Y. 2009. *Make3D: Learning 3D Scene Structure from a Single Still Image*. IEEE Computer Society.
- Shi, J., and Pollefeys, M. 2014. Pulling things out of perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, 89–96.
- Sun, B., and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. 443–450.
- Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation.
- Wang, P.; Shen, X.; Lin, Z.; and Cohen, S. 2015. Towards unified depth and semantic prediction from a single image. In *Computer Vision and Pattern Recognition*, 2800–2809.
- Xu, D.; Ouyang, W.; Wang, X.; and Sebe, N. 2018. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing.
- Zhan, H.; Garg, R.; Weerasekera, C. S.; Li, K.; Agarwal, H.; and Reid, I. 2018. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction.
- Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised learning of depth and ego-motion from video. 6612–6619.
- Zhuo, W.; Salzmann, M.; He, X.; and Liu, M. 2015. Indoor scene structure analysis for single image depth estimation. In *Computer Vision and Pattern Recognition*, 614–622.