SUW-Learn: Joint Supervised, Unsupervised, Weakly Supervised Deep Learning for Monocular Depth Estimation

Haoyu Ren*, Aman Raj**, Mostafa El-Khamy* and Jungwon Lee*

*SOC R&D, Samsung Semiconductor, Inc., California, USA
**University of San Diego, California, USA
{haoyu.ren, mostafa.e, jungwon2.lee}@samsung.com amraj@ucsd.edu

Abstract

We introduce SUW-Learn: A framework for deep-learning with joint supervised learning (S), unsupervised learning (U), and weakly-supervised learning (W). We deploy SUW-Learn for deep learning of the monocular depth from images and video sequences. The supervised learning module optimizes a depth estimation network by knowledge of the ground-truth depth. In contrast, the unsupervised learning module has no knowledge of the ground-truth depth, but optimizes the depth estimation network by predicting the current frame from the estimated 3D geometry. The weakly supervised module optimizes the depth estimation by evaluating the consistency between the estimated depth and weak labels derived from other information, such as the semantic information. SUW-Learn trains the deep-learning networks end-to-end with joint optimization of the desired SUW objectives. We benchmark SUW-Learn on the commonly-used KITTI driving-scene and achieve the state-of-the-art performance. To demonstrate the capacity of SUW-Learn in learning the depth of scenes with people from different sources with different domain knowledge, we construct the M&M dataset from the Megadepth and Mannequin Challenge datasets.

1. Introduction

Monocular depth estimation (MDE) is a key feature of understanding the geometric structure of the scene. In particular, the depth map can be used to infer the 3D structure, which is the basic element of many topics in 3D vision, such as image reconstruction, image rendering, and shallow depth of the field. MDE is an ill-posed problem, since a single 2D image may be produced from an infinite number of distinct 3D scenes. To overcome this ambiguity, typical methods focus on exploiting statistically meaningful features, such as perspective and texture information, object locations, and occlusions.

With the prosperity of deep convolutional neural networks (CNNs), many deep learning-based methods have achieved significant advances in MDE. Most CNNs were trained using supervised learning, by minimizing a loss between the estimated depth and the ground-truth depth. Some recent methods adopted unsupervised learning, when video sequences were available [25][9]. They utilized 3D geometry information to project 2D frames back to the 3D space. These projected frames are aligned in the 3D space by the estimated depth and the camera's relative pose. Without ground-truth depth supervision, unsupervised methods can estimate the depth which minimizes a photometric loss after the projection of the estimated 3D frames again to 2D. Recent approaches adopted weakly-supervised training [1][17], where the network is optimized based on the ordinal (farther, closer) relationship between pixel pairs. Unfortunately, their accuracy is still limited, since the pixel-pair based label is very sensitive to noise.

In this paper, we investigate the concurrent training of deep neural networks with 3 different learning strategies, supervised learning (S), unsupervised learning (U), and weakly supervised learning (W), which we call 'SUW-Learn'. We demonstrate that SUW learning of monocular depth estimation, improves the depth estimation accuracy and the generalization capability of deep networks.

Since MDE is an under-determined problem, MDE trained networks tend to work well only on the kind of scenes they have been trained on. There has been some works on robust depth estimation [21]. To improve the robustness of MDE in diverse scenes with people, we construct the M&M dataset by combining the MegaDepth (MD) dataset [17] and the Mannequin Challenge (MC) dataset [16]. The MD dataset consists of single-frame outdoor images for landmarks surrounded by people, and the MC dataset consists of video sequences of persons in mostly

indoor scenarios. The contributions of this paper are highlighted below:

Table 1. RMSE/REL of our SUW-Learn framework with different learning strategies on KITTI Eigen's split with official ground-truth depth. 'S' stands for 'supervised learning', 'U' stands for 'unsupervised learning', and 'W' stands for 'weakly-supervised learning'.

Learning Strategy	REL (%)	RMSE (in meter)
S	6.61	3.058
SU	5.17	2.298
SW	5.17	2.478
SUW	**3.98**	**1.915**

- We show that joint training of MDE networks using both supervised and unsupervised (SU) learning achieves better accuracy compared to training with supervised (S) learning only. We adopted unsupervised learning to learn the 3D geometry by training a pose estimation network and the depth estimation network.
- We show that joint training with both supervised and weakly-supervised (SW) learning together performs better compared to training with supervised learning only. Weakly-supervised learning aims to provide a consistency between the estimated depth and the weak labels. We investigated different methods for generating weakly-supervised labels. We propose a patch-based weak depth label generated by the semantic knowledge only, which captures the similarity and differences of the depth statistics in different semantic regions, which is more discriminative and robust compared to generating pixel-based weak labels.
- By joint supervised, unsupervised, and weakly-supervised (SUW) learning, our SUW-Learn framework achieves the best accuracy, without increasing any of the complexity during the prediction since only the supervised learning module is utilized. In Table 1, we give the accuracy of our SUW-Learn framework with different learning strategies on KITTI Eigen's split. It can be observed that the more learning strategies we use, the better depth accuracy we may get. The visualization results can be found in our supplementary video[1].
- We demonstrate the improved generalization capability when using SUW-Learn to learn the depth from the M&M dataset. Since M&M is constructed from different sources and each source provides different types of labels, only a subset of the different learning strategies are possible on each source. We observe that the proposed SUW-Learn strategies results in better accuracy and generalization compared to the S, SU, and SW models trained on the same M&M dataset. Some visualization of our SUW-Learn results on M&M test set are given in Fig. 1.

[1] https://www.youtube.com/watch?v=fWhvc6OClVg

2. Related work

2.1. Supervised depth estimation

Previous approaches for supervised depth estimation from monocular image can be categorized into two main groups, methods operating on hand-crafted features [10][11][13][3][23], and methods adopting deep neural networks [6][7][14][15]. Most of the recent approaches are based on convolutional neural networks. In the pioneering work [6], Eigen et al. introduced a coarse-to-fine network, and utilized the scale-invariant loss to improve the accuracy of the estimated depth map. This work is further extended in [5], where the depth estimation, surface normal estimation, and semantic segmentation are integrated into one unified network. Fu et al. [7] introduced a spacing-increasing discretization strategy to discretize depth and re-casted depth network learning as an ordinal regression problem. Most of above works use backbones with fully connected layers, which limits the deployment flexibility. To solve this problem, a fully convolutional network was proposed by Laina et al. [14]. A revised version of this work is introduced in [18], where randomly sampled sparse depth is adopted together with the RGB image to predict a dense depth map. Cheng et al. [2] proposed a spatial propagation network (CSPN) to learn the affinity matrix and showed its effectiveness to improve the performance of existing SIDE networks. Lee et al. [15] introduced local planar guidance layers located at multiple stages into the encoder-decoder architecture, which achieved the state-of-the-art accuracy.

2.2. Unsupervised depth estimation

Various approaches for unsupervised learning of depth from multiple frames have been considered. Garg et al. [8] introduced the learning of depth and ego-motion at the same time by generating an inverse warping of the target image using the predicted depth and the known inter-view displacement. Zhou et al. [29] demonstrated a fully differentiable approach where depth and ego-motion are predicted jointly by deep neural networks. Wang et al. [25] learned the depth CNN predictor using a differentiable implementation of direct visual odometry. Ariel et al. [9] addressed occlusions geometrically and differentiably. They also introduced randomized layer normalization during the training. Some other papers improve the depth quality at inference by training on stereo frames. Khamis et al. [12] utilized a Siamese network to extract features from the left and right stereo pairs. Yao et al. [26] extracted deep visual image features from stereo-pairs and built the 3D cost volume upon the reference camera frustum. These unsupervised methods

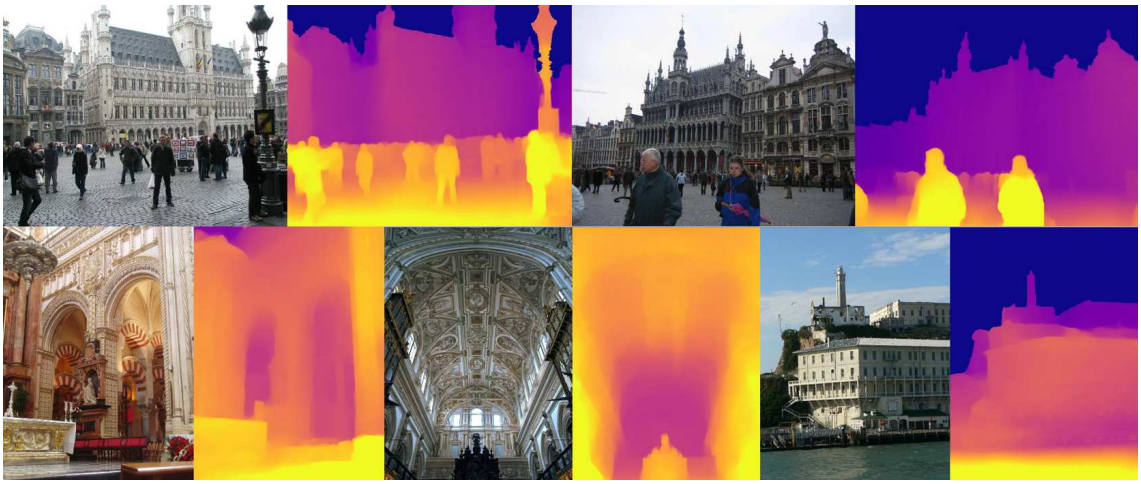

Figure 1. Some qualitative results on M&M test set from our SUW-Learn model.

can obtain reasonable depth estimation results without the ground-truth depth supervision, but their accuracy is lower compared to the supervised methods.

2.3. Weakly-supervised depth estimation

Some other works address the depth estimation problem using weakly labeled pixel pairs. One method is to label the ordinal (farther or closer) relationship. In [1], Chen et al. manually labeled such pixel pairs from web-collected images and constructed the Depth in the wild (DIW) dataset. They trained an hourglass network to generate relative depth maps based on these pixel pair labels. Li et al. [17] used the SfM (structure-from-motion) algorithm to generate the ground-truth depth for landmarks, and further extracted the weakly labeled pixel pairs from foregrounds and landmarks respectively. The accuracy of these weakly supervised methods are still limited since learning with pixel-based weak depth is sensitive to noise.

3. SUW-Learn

The proposed SUW-Learn framework is displayed in Figure 2. SUW-Learn consists of a supervised learning module S, an unsupervised learning module U, and a weakly supervised learning module W. The supervised learning module S estimates the depth D_t of input frame I_t by a depth estimation network. The supervised loss $\mathcal{L}_s$ is calculated between the estimated depth D_t and ground-truth depth D_t^*, as per the Eq. 1. The unsupervised learning module U takes input a sequence of three frames I_{t-1}, I_t, I_{t+1} and estimates relative camera pose (ego-motion) and the objects' rigid motion. We will denote M the set of all motion vectors estimated from frame $t \rightarrow t-1$ and $t \rightarrow t+1$, e.g, $M = \{M_{t\rightarrow t-1}, M_{t\rightarrow t+1}\}$. These motion vectors, together with the knowledge of the intrinsic parameters K, along with estimated depth of target frame D_t and input sequence are utilized to create warped version of target frame I_t namely I'_{t-1} and I'_{t+1}. The unsupervised loss $\mathcal{L}_u$ is calculated between the warped images and target frame by Eq. 2. Weakly supervised learning module W generates a weak depth label $\hat{D}_t^*$ from input frame I_t without using ground-truth depth knowledge. A weakly supervised loss $\mathcal{L}_w$ is calculated between the weak depth label $\hat{D}_t^*$ and the estimated depth map D_t, as given in Eq. 3.

Our networks are trained end-to-end by jointly minimizing $\mathcal{L}_s, \mathcal{L}_u, \mathcal{L}_w$. The supervised loss $\mathcal{L}_s$ and weakly supervised loss $\mathcal{L}_w$ are calculated for the center frame I_t, only. For simplicity, we ignore the subscript t when discussing the supervised learning and the weakly supervised learning in the following sections. During the prediction, only the supervised learning module is utilized. So the complexity will not be increased at all.

$$S(I_t, D_t^*) \rightarrow \{D_t, \mathcal{L}_s(D_t^*, D_t)\} \tag{1}$$

$$U(I_{t-1}, I_t, I_{t+1}, D_t, K) \rightarrow \left\{I'_{t-1}, I'_{t+1}, \mathcal{L}_u(I_t, I'_{t-1}, I'_{t+1})\right\} \tag{2}$$

$$W(I_t, D_t) \rightarrow \{\hat{D}_t^*, \mathcal{L}_w(\hat{D}_t^*, D_t)\} \tag{3}$$

3.1. Supervised learning

The supervised learning module S takes the center frame I and its ground-truth depth map D^*as input. It calculates the supervision loss function $\mathcal{L}_s(D^*, D)$, which is back-propagated to train a network to output the estimated depth map D of the input frame I. We utilized two different methods for supervised depth estimation. The first is soft classification, where the continuous depth is quantized into discrete bins, and a classification loss is minimized [7][21]. The second supervised loss is regression, where the output feature map is regressed to the ground-truth depth in training, and directly utilized as the estimated depth map [1][6].

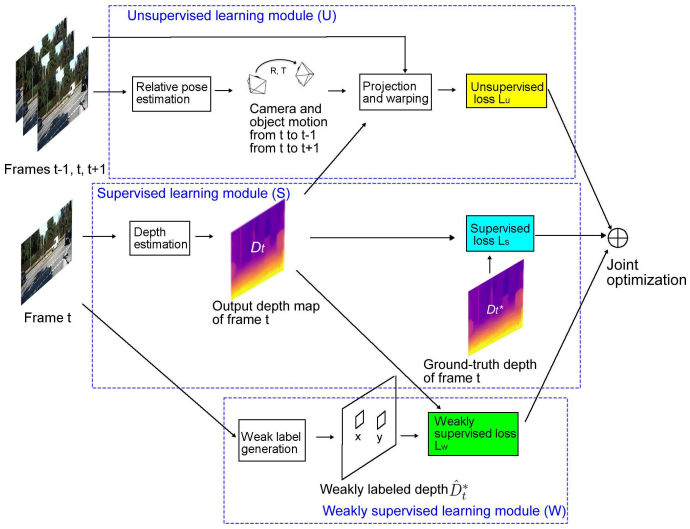

Figure 2. Overall framework of our proposed SUW-Learn framework. The training takes three frames I_{t-1}, I_t, I_{t+1} as input. During the prediction, only single frame is utilized since only the supervised learning module is used for depth estimation. The supervised loss, unsupervised loss, and weakly supervised loss are jointly optimized.

3.1.1 Classification based supervised learning

The soft classification loss $\mathcal{L}_{s,cls}$ used in our SUW-Learn framework calculates the predicted depth at each pixel location as the expected depth. During the training, the desired depth-range $[\alpha, \beta]$ is uniformly quantized into B bins in the log scale. Let p_i^j represents the estimated probability that the depth at pixel i is classified to belong to the bin of the quantized depth j, and B be the maximum quantized depth, then the expected depth D_i of pixel i is calculated as $D_i = \sum_{j=1}^{B} j \times p_i^j$. Please notice that D_i can be a floating point number. A Huber loss is utilized to measure the difference between the predicted depth D_i, and the ground truth quantized depth D_i^*, as given in Eq. 4, where n is the number of valid labeled pixels in D^*.

$$\mathcal{L}_{cls} = \begin{cases} \frac{1}{n}\sum_i 0.5(D_i - D_i^*)^2 & if |D_i - D_i^*| < 1 \\ \frac{1}{n}\sum_i |D_i - D_i^*| - 0.5 & \text{otherwise} \end{cases} \quad (4)$$

In the experiments on the KITTI dataset, we trained a encoder-decoder network. Its supervised module uses the soft classification loss. The encoder-decoder network has a similar architecture to that of [21]. The encoder contains several depth-wise separable convolutional layers to extract the discriminative features from the input image. The decoder consists of several 3×3 convolutional layers and a stacked atrous multi-scale (SAM) module [4]. Inspired by the U-Net architecture [22], we concatenate the feature maps between the encoding blocks and the decoding blocks.

3.1.2 Regression based supervised learning

SUW-Learn deploys a regression loss consisting of three terms, a scale-invariant loss $\mathcal{L}_{mse}$, a gradient loss $\mathcal{L}_g$, and a smoothness loss $\mathcal{L}_{sm}$. The scale-invariant loss computes the mean square error (MSE) between pairs of log-depth as

$$\mathcal{L}_{mse} = \frac{1}{n}\sum_{i=1}^{n} R_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} R_i\right)^2, \quad (5)$$

where $R_i = \log D_i - \log D_i^*$, n is the number of valid pixels in D^*. Since the difference in log values corresponds to ratios of the pairs of depths, they are preserved under scaling. Hence, it is advantageous to use the scale-invariant loss for supervision with datasets whose ground-truth is only known up to a scale, such as that generated by the SfM method, as is the case with our combined M&M dataset.

Regression based depth estimation network prefers generating an over-smooth depth map. To solve this problem, we use a multiple-scale gradient matching loss $\mathcal{L}_g$ on the scale-invariant log-depth map, which encourages sharper depth discontinuities in the predicted depth map, defined as

$$\mathcal{L}_g = \sum_{k=0}^{s} \frac{1}{n_k}\sum_i (|\nabla_x R_i^k| + |\nabla_y R_i^k|), \quad (6)$$

where R_i^k is the difference map of the log depth at position i and scale k (depth map downsampled with scale 2^k), n_k is the number of valid pixels at scale k, and $s+1$ is the number of different scales. We further use an edge-aware smoothness loss $\mathcal{L}_{sm}$ [27][20] weighted by image gradients

$$\mathcal{L}_{sm} = \sum_{k=0}^{s} \frac{1}{n_k 2^k}\sum_i |\nabla D_i^k| \cdot (e^{-|\nabla I_i^k|})^T \quad (7)$$

where $|\cdot|$ denotes element-wise multiplication, ∇ is the vector differential operator applied on pixel i.

For supervised learning on the constructed M&M dataset, we train an hourglass network with the above regression losses. Our hourglass network follows the same architecture as [1]. It consists of a series of the inception [24] modules organized in a hourglass shape. The output feature map is directly utilized as the estimated depth map.

3.2. Unsupervised learning

The inputs to our unsupervised learning module consists of three consecutive frames I_{t-1}, I_t, I_{t+1}, the estimated depth of frame t D_t, and the camera intrinsic matrix K. Similar to [29], we first estimate the global camera motion (ego-motion) between adjacent frames. There will be two ego-motions for three input frames, corresponding to frame t

to frame $t-1$ as $M_{t\to t-1}$, and frame t to frame $t+1$ as $M_{t\to t+1}$.

Let p_t denote the homogeneous coordinates of a pixel p in the frame t (target view), we can obtain p_t's projected coordinates $p_s, s = t-1/t+1$ onto the frame $t-1$ and $t+1$ (source view) by Eq. 8, where $D_t(p_t)$ is the estimated depth of pixel p in frame t.

$$p_s = KM_{t\to s}D_t(p_t)K^{-1}p_t \tag{8}$$

The unsupervised loss is then established using a photometric loss comparing the projected frames $I^{'}_{t-1}$ (warping from $p_s \to p_t, s = t-1$ in Eq. 8), and $I^{'}_{t+1}$ (warping from $p_s \to p_t, s = t+1$ in Eq. 8) onto the actual target frame I_t in RGB space

$$\mathcal{L}_{u,ego} = \sum_p (||I_t(p) - I^{'}_{t-1}(p)|| + ||I_t(p) - I^{'}_{t+1}(p)||). \tag{9}$$

Eq. 9 assumes the scene is static without any moving objects. This is not realistic in most of the scenarios. Inspired by recent work [19], we estimate the object motion as well. Suppose there are k objects aligned in 3 consecutive frames, we extract the object motion $M_{t\to t-1,1}, ..., M_{t\to t-1,k}$ and $M_{t\to t+1,1}, ..., M_{t\to t+1,k}$, where $M_{t\to t-1,m}, 1 \le m \le k$ is the rigid motion from frame t to frame $t-1$ of the mth object. The unsupervised loss based on object motion is defined as

$$\begin{aligned}\mathcal{L}_{u,obj} = \sum_{m=1}^{k}\sum_p (&||I_{t,m}(p) - I^{'}_{t-1,m}(p)|| \\ &+||I_{t,m}(p) - I^{'}_{t+1,m}(p)||).\end{aligned} \tag{10}$$

where $I_{t,m}$ is the masked RGB image of frame t by object m (all the pixels not belonging to object m are set to 0). In our experiments, we show that our SUW-Learn works well with using $\mathcal{L}_{u,ego}$ only, as well as combining $\mathcal{L}_{u,ego}$ and $\mathcal{L}_{u,obj}$ together. Considering object motion can give us better accuracy.

3.3. Weakly supervised learning

Existing weakly supervised depth estimation methods [1][17] adopt weakly labeled depth based on pixel pairs. Each weakly labeled depth $\hat{D}^*$ consists of two pixels (i, j) belonging to the foreground and the background respectively. The loss function $\mathcal{L}_{w,pix}$ calculated based on estimated depth D is given by

$$\mathcal{L}_{w,pix} = \begin{cases} \log(1+e^{P_{ij}}) & if P_{ij} \le 0.25 \\ \log(1+e^{\sqrt{P_{ij}}}) & if P_{ij} > 0.25 \end{cases} \tag{11}$$

where $P_{ij} = -r_{ij}(\log(D_i) - \log(D_j))$, D_i is the estimated depth of pixel i, and r_{ij} is an ordinal depth indicator ($r_{ij} = 1$ if pixel i is farther than pixel j, otherwise $r_{ij} = -1$). $\mathcal{L}_{w,pix}$ encourages the depth difference to be large and ordered between a pair of points, but it is sensitive to noise.

In this paper, we propose the patch-based weak depth label $\hat{D}^* = \{x, y, r_{xy}\}$, where $\{x, y\}$ are rectangles, and r_{xy} is an ordinal indicator. Our weak depth label $\hat{D}^*$ is generated by the semantic knowledge from a pre-calculated semantic segmentation map s. x and y have exact same size $W \times H$. All the pixels in x or y have the same semantic label, e.g., $\forall i, j \in x, s(i) = s(j)$, and $\forall i, j \in y, s(i) = s(j)$. $r_{xy} = 0$ if x and y are from same semantic class, e.g., $r_{ij} = 0 \; iff \; \forall i \in x, \forall j \in y, s(i) = s(j)$ otherwise $r_{xy} = 1/-1$ depending on whether x is farther than y, which is similar to the pixel-based weak label. This farther/closer relationship between x and y is decided by the semantic knowledge instead of using ground-truth depth, more details are given in Section 4.1. We define our weakly supervised loss function $\mathcal{L}_{w,patch}$ based on patch labels as in Eq. 12

$$\mathcal{L}_{w,patch} = \begin{cases} 1 - HI(h(D_x), h(D_y)) + V_{xy} & r_{xy} = 0 \\ \log(1+e^{P_{xy}}) + V_{xy} & r_{xy} \neq 0, P_{xy} \le 0.25 \\ \log(1+e^{\sqrt{P_{xy}}}) + V_{xy} & r_{xy} \neq 0, P_{xy} > 0.25 \end{cases} \tag{12}$$

where $h(D_x)$ is a normalized depth histogram of D_x on the depth range $[min(D_x), max(D_x)]$ with O bins, HI stands for the histogram intersection operation which calculates the distance between two histograms, $V_{xy} = \lambda(\mathrm{var}(D_x) + \mathrm{var}(D_y))$ is a regulator based on the variance of the depth in patch x and y, and λ is a constant weight set as 0.05 empirically. $P_{xy} = -r_{xy}(\mathrm{mean}(D_x) - \mathrm{mean}(D_y))$, which is similar to Eq. 11.

Our weakly supervised loss function in Eq. 12 can be explained as follows. In the first row, if the two patches $\{x, y\}$ come from the same semantic class (e.g., they are all pedestrians), their depth distribution should be similar. This makes sense since the depth distribution of a person should be similar to another person, no matter whether the person is close or far from the camera. The second row and third row are similar to the pixel-based weakly supervised loss in Eq. 11. If $\{x, y\}$ come from different semantic classes, their average depth should be consistent to the farther/closer relationship generated from the semantic knowledge (e.g., sky is farther than trees). We add a regulator V_{xy} into the above loss function to suppress the effect of depth outliers. In the experiment, we show that our SUW-Learn framework works well with both pixel-based weakly supervised loss $\mathcal{L}_{w,pix}$, as well as our proposed patch-based loss $\mathcal{L}_{w,patch}$. Training with $\mathcal{L}_{w,patch}$ achieves significantly better accuracy compared to $\mathcal{L}_{w,pix}$.

Our proposed weakly supervised depth estimation based on patch labels has clear advantages compared to existing solutions based on pixel labels [17][1]. First, our patch-based label is more flexible since it consists of more patterns with different patch size. Second, the pixel-based label is sensitive to noise. In contrast, our patch-based label is evaluated by the depth statistics information (such as mean, variance, and histogram) of patches, which is more robust. Third, existing pixel-based label focuses on the depth ordinal relationship between different object categories (e.g., person vs. landmark), while our patch-based label also considers the depth similarity within same object categories (e.g., person vs. person). By end-to-end training of our networks using SUW-Learn with patch-based weak label, we may obtain more accurate depth map compared to the networks using pixel-based weak label.

4. Experiments on the KITTI dataset

4.1. Training on the KITTI Dataset

For fair comparison with existing methods [9][29][25] we use the Eigen's split of KITTI introduced in [6]. Consecutive 3-frame video clips are cropped as the training example. Since the size of KITTI images varies, we zero pad them to a uniform size 1248×384 for training. During evaluation we use depth cap at 80m and evaluate both on KITTI's official ground-truth and the raw LiDAR data. The commonly-used evaluation metric for KITTI, includes REL (mean relative absolute error), sqREL (mean relative squared error), RMSE (root mean square error), and δ thresholds.

In our SUW-Learn framework we jointly optimize the following losses:

$$\mathcal{L} = \mathcal{L}_s + \alpha\mathcal{L}_u + \beta\mathcal{L}_w. \tag{13}$$

where $\mathcal{L}_s$ is supervised loss term, $\mathcal{L}_u$ is unsupervised loss term, and $\mathcal{L}_w$ is weakly supervised loss term. The weights α, β are constants decided by evaluation.

We train our classification based encoder-decoder network on the KITTI dataset. The ground-truth depth from $[0m, 80m]$ is quantized to 200 bins. The depth higher than $80m$ is truncated to $80m$ during the training. Since KITTI provides video sequences and has consistent ordinal depth relationship from semantic knowledge (e.g., sky is farther than persons), we may calculate all supervised loss, unsupervised loss, and weakly supervised loss for all training images. The supervised loss is the soft classification loss $\mathcal{L}_s = \mathcal{L}_{cls}$ given in Eq. 4. We use both the ego-motion and object-motion to calculate the unsupervised loss $\mathcal{L}_u = \mathcal{L}_{u,ego} + \mathcal{L}_{u,obj}$. All the images are pre-processed to generate object masks using Mask-RCNN. These masks are used in estimating the object motion and computing the photometric error for object-mask images. Since KITTI doesn't provide extrinsic camera parameters for all the frames, we use a pose estimation CNN to predict relative poses between center and adjacent frames.

Figure 3. Using semantic knowledge to generate weakly labeled depth. A: the sky is farther than other objects. B: the depth distribution of two road patches should be similar.

We use the patch-based weak depth label with the weakly supervised loss $\mathcal{L}_w = \mathcal{L}_{w,patch}$ defined in Eq. 12. We generate these weak depth labels with semantic masks obtained from PSPNet [28], as given in Fig. 3, where the 'A' and 'B' are the patch pairs generated from the semantic masks. We set $\alpha = 0.5$ and $\beta = 0.125$ in the Equation 13 when joint training with multiple loss functions. More detailed discussions are given in the supplementary material.

4.2. Results on the KITTI dataset

In Table 2, we observe that our our SUW-Learn framework achieves significantly better results on KITTI's eigen split compared to existing unsupervised depth estimation methods [29][25], supervised depth estimation methods [7][21][15], as well as the methods based on weakly-labeled depth [17]. Compared to the the state-of-the-art supervised methods [7][15], our SUW-Learn benefits from the unsupervised loss which constraints 3D consistency in depth estimation, and the weakly supervised loss which captures ordinal depth relations between objects in the scene. It also performs better than existing unsupervised methods [29][25]. Some visualizations of the output depths are given in Figure 4. We observed that our SUW-Learn achieves lower error and better perceptive quality compared to current state-of-the-art methods DORN [7] and BTS [15].

4.3. Ablation study

In Table 3, we give a detailed comparison of different learning strategies on KITTI. The baseline model utilizes only supervised learning strategy and achieves 3.058m RMSE (row 1). It can been seen that in addition to supervised loss, using the unsupervised loss with ego-motion, the RMSE is reduced from 3.058m to 2.362m (row 2), and further to 2.298m (row 3) by taking object motion into consideration. Next, we introduce the weakly supervised loss and observe that RMSE further reduces to 2.116m (row 4). We also did ablation study on which type of weak labeling method aids the most in the depth estimation task using our SUW-Learn framework. By using a patch-based weakly supervised loss instead of pixel-based version, the RMSE of our framework is further reduced from 2.116m to 1.915m (row 7).

Table 2. Depth estimation accuracy and efficiency of different methods on KITTI Eigen's split. The depth cap is 80m. 'S' stands for 'supervised learning', 'U' stands for 'unsupervised learning', and 'W' stands for 'weakly supervised learning'. * denotes that the method is evaluated using the official ground truth, otherwise, evaluated with raw LiDAR scan data.

Method	Learning	REL (%)	sqREL	RMSE (in meter)	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Speed (sec)
Eigen et al. [5]	S	20.34	1.548	6.307	0.702	0.898	0.967	0.864
Ren et al. [21]	S	9.39	0.595	3.871	0.900	0.965	0.986	0.348
Fu et al. [7]	S	9.90	0.593	3.714	0.897	0.966	0.986	0.748
Lee et al. [15]	S	9.10	0.555	4.033	0.904	0.967	0.984	0.443
Zhou et al. [29]	U	20.81	1.768	6.856	0.678	0.885	0.957	1.367
Wang et al. [25]	U	14.82	1.187	5.469	0.812	0.938	0.975	1.478
Ours	SUW	**8.69**	**0.513**	**3.441**	**0.911**	**0.973**	**0.991**	0.313
Fu et al. * [7]	S	8.10	0.337	2.930	0.936	0.986	0.995	0.748
Ren et al. * [21]	S	6.64	0.271	3.065	0.945	0.989	0.996	0.348
Lee et al. * [15]	S	6.40	0.254	2.815	0.950	0.993	0.999	0.443
Wang et al. * [25]	U	11.37	0.872	4.821	0.913	0.955	0.982	1.478
Ours *	SUW	**3.98**	**0.221**	**1.915**	**0.957**	**0.995**	**0.999**	0.313

Table 3. Depth estimation accuracy of our encoder-decoder network with different learning strategies and loss functions on KITTI Eigen's split. The depth cap is 80m. The official ground-truth is utilized for evaluation. The unit of RMSE is meter.

Learning	Motion	Weak label	REL (%)	RMSE
S	-	-	6.61	3.058
SU	ego	-	5.32	2.362
SU	object	-	5.17	2.298
SUW	ego+object	pixel	4.55	2.116
SW	-	pixel	5.67	2.802
SW	-	patch	5.17	2.478
SUW	ego+object	patch	**3.98**	**1.915**

This validates the effectiveness of our proposed SUW-Learn framework and patch-based weakly supervised loss. This is further substantiated by comparison between SW (pixel-based) and SW (patch-based) in row 5 and row 6 respectively, where patch-based weak labels achieve lower error. This shows the advantages of evaluating the depth distribution of patches compared to single pixels. We also compare different ways to generate the weak labels in the supplementary material.

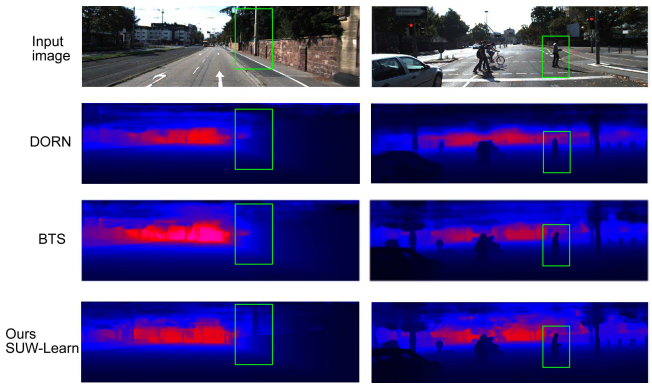

Figure 4. Some qualitative results on KITTI test set and compared with existing methods.

5. Experiments on the M&M dataset

5.1. Training on the M&M dataset

For the purpose of improving the depth estimation in the diverse scenes with people, we propose M&M dataset by combining the MegaDepth dataset [17] and the Mannequin challenge dataset [16] . The MegaDepth (MD) dataset was introduced in [17], that contains downloaded internet photos of well photographed landmarks from Flickr. It provides 100K images with depth data generated from SfM+MVS, and 30K images with pixel based weak depth labels. Mannequin challenge (MC) dataset [16] contains a wide variety of YouTube videos with people of different ages, naturally posing in different group configurations in various types of frozen scenes. The ground-truth depth is generated by the same methods SfM+MVS algorithms as in [17]. Since, the depth generated from SfM+MVS is known upto a scale only, so such a combination is possible. By combining these two datasets together, we have a large number of real-world images with persons and landmarks. The MC [16] only provides YouTube video urls along with timestamp of frames and their respective camera parameters, we were able to download only 1,508 videos out of 2,663 in train set and 703 out of 750 in test set, since some videos are not available on YouTube anymore. We generate 3-frame sequences set for training on MC dataset. Our M&M dataset training set contains the official train/test split of MD images and the downloaded MC train/test sequences. We will call official MC training set [16] as MC+ here which has more 1.7 times of images than ours. We use the scale-invariant RMSE(si-RMSE) metric to evaluate our SUW-Learn framework on M&M dataset where si-RMSE is equivalent to $\sqrt{\mathcal{L}_{mse}}$. We scale all the images in our M&M dataset to a fixed size and correspondingly change it's intrinsic camera parameters, depth maps and the coordinates of weak depth labels. We train our regression based hourglass network with SUW-Learn framework on this dataset. The Megadepth

(MD) dataset only provides single frame images, so it is not possible to extract the camera and object motion for unsupervised learning. Most of the Mannequin Challenge (MC) images are indoor scenarios, which lacks of the consistency of ordinal depth relationship from the semantic knowledge. While training the network on M&M dataset, supervised losses are coming from MC and MD, unsupervised loss from MC images only and weakly supervised loss from MD images only.

In M&M training, the supervised loss is the depth regression loss $\mathcal{L}_s = \mathcal{L}_{mse} + w_1\mathcal{L}_g + w_2\mathcal{L}_{sm}$ discussed in Section 3.1.2, where $w_1 = 0.5, w_2 = 0.10$, are hypertuned weights. The unsupervised loss term is calculated by the ego-motion only $\mathcal{L}_u = \mathcal{L}_{u,ego}$ since all the scenarios in the MC dataset are frozen without any object motion. We directly use the camera parameters (intrinsic and extrinsic) provided by the MC dataset for relative camera pose estimation rather than predicting it using a CNN. We use the weakly supervised loss based on the official pixel-pair weak label provided by MD dataset $\mathcal{L}_w = \mathcal{L}_{w,pix}$. We set $\alpha = 0.25$ and $\beta = 0.005$ in the Eq. 13 when training the networks with SUW-Learn. More details can be found in supplementary material.

5.2. Results on the M&M dataset

We show the evaluation results of our SUW-Learn framework on M&M test set in Table 4 and compare it with existing methods. We used official checkpoints of Megadepth [17] and Mannequin Challenge [16] to evaluate on M&M. Our best method trained jointly with supervised, unsupervised and weakly supervised losses, i.e. SUW (row 8) beats the previous methods [17] and [16] with a large margin having least error. In order to show that our improved performance on M&M is coming from improvement over both MC and MD, we evaluate previous methods [17][16] and our best method (SUW) individually on MC and MD. We show these results in Table 5, where our method performance beats previous state of the art on MegaDepth (MD) and is close in performance to Mannequin Challenge (MC), but overall on M&M we are better. This shows our model is more generalized to real world situations compared to previous methods.

Similarly, we perform ablation study of our SUW-Learn framework on M&M dataset in Table 4. The results show consistent improvement as we incorporate more learning strategies. When training on M&M dataset, our SU and SW models (row 7 and 8) perform consistently better than the model using supervised only (row 6). The accuracy can be further improved by the SUW (row 9). When training on MC dataset only, the observation is similar, where the SU model (row 5) performs clearly better than supervised learning only (row 4). These results support our proposed SUW-Learning strategy to leverage use of supervised, unsupervised and weakly supervised labels.

Table 4. si-RMSE of our SUW-Learn framework on M&M test set. 'MC' stands for Mannequin challenge dataset, and 'MD' stands for MegaDepth dataset. The official model [16] of MC dataset uses more images than us during the training (MC+).

Method	Learning	Train set	M&M (si-RMSE)
Li et al. [16]	S	MC+	0.259
Li et al. [17]	SW	MD	0.252
Ours	S	MC	0.280
Ours	SU	MC	0.269
Ours	S	M&M	0.221
Ours	SU	M&M	0.212
Ours	SW	M&M	0.214
Ours	SUW	M&M	**0.208**

Table 5. si-RMSE of our SUW-Learn framework on M&M test set. 'MC' stands form Mannequin challenge, 'MD' stands for MegaDepth.

Train set \ Test set	MC	MD	M&M
MC+ [16]	**0.295**	0.224	0.259
MD [17]	0.401	0.103	0.255
M&M (SUW)	0.317	**0.100**	**0.208**

6. Conclusion

In this paper, we introduce SUW-Learn as a framework for joint training of deep neural networks using multiple learning strategies, supervised, unsupervised, as well as weakly-supervised learning. We deployed SUW-Learn to address the monocular depth estimation (MDE) problem. We proposed to train MDE networks by the joint optimization of supervised losses between the estimated depth and the ground truth, an unsupervised loss based on object-aware photometric error from multi-frame reconstruction, and a weakly supervised loss using ordinal depths labels generated from semantic information. We propose a new dataset M&M which combines two recent datasets, the Mannequin Challenge dataset and the MegaDepth dataset, to have good coverage of people subjects in diverse scenes, such as outdoor and indoor scenes. Our proposed SUW-Learn framework improves the generalization capacity of the MDE network. Our experimental results on both the KITTI and the M&M datasets demonstrate the verify the effectiveness of our proposed framework. Our results show that by combining different learning strategies, our deep MDE networks are robust and can achieve the state of art accuracies. For future work, we plan extending SUW-Learn to other challenging computer vision tasks with multi-frame inputs, such as video super resolution or video semantic segmentation.

References

[1] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *NIPS*, pages 730–738,

2016.

[2] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *ECCV*, pages 103–119, 2018.

[3] Erick Delage, Honglak Lee, and Andrew Y Ng. A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In *CVPR*, volume 2, pages 2418–2428, 2006.

[4] Xianzhi Du, Mostafa El-Khamy, and Jungwon Lee. Amnet: Deep atrous multiscale stereo disparity estimation networks. *IEEE International Conference on Consumer Electronics*, 2019.

[5] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pages 2650–2658, 2015.

[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, pages 2366–2374, 2014.

[7] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018.

[8] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, pages 740–756, 2016.

[9] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *arXiv 1904.04998*, 2019.

[10] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. *ACM transactions on graphics (TOG)*, 24(3):577–584, 2005.

[11] Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2144–2158, 2014.

[12] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *ECCV*, pages 573–590, 2018.

[13] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *CVPR*, pages 89–96, 2014.

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

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

[16] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, pages 4521–4530, 2019.

[17] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018.

[18] Fangchang Ma and Sertac Karaman. Sparse-to-dense: depth prediction from sparse depth samples and a single image. *arXiv preprint arXiv:1709.07492*, 2017.

[19] Wei-Chiu Ma, Shenlong Wang, Rui Hu, Yuwen Xiong, and Raquel Urtasun. Deep rigid instance scene flow. In *CVPR*, pages 3614–3622, 2019.

[20] Yue Meng, Yongxi Lu, Aman Raj, Samuel Sunarjo, Rui Guo, Tara Javidi, Gaurav Bansal, and Dinesh Bharadia. Signet: Semantic instance aided unsupervised 3d geometry perception. In *CVPR*, pages 9810–9820, 2019.

[21] Haoyu Ren, Mostafa El-khamy, and Jungwon Lee. Deep robust single image depth estimation neural network using scene understanding. *CVPR Workshops*, 2019.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[23] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. 3-d depth reconstruction from a single still image. *International journal of computer vision*, 76(1):53–69, 2008.

[24] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.

[25] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, pages 2022–2030, 2018.

[26] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, pages 767–783, 2018.

[27] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, pages 1983–1992, 2018.

[28] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017.

[29] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 1851–1858, 2017.