SRC-Disp: Synthetic-Realistic Collaborative Disparity Learning for Stereo Matching

Guorun Yang, Zhidong Deng, Hongchao Lu, and Zeping Li

Department of Computer Science, Tsinghua University, Beijing 100084, China[†]
{ygr13,luhc15,li-zp16}@mails.tsinghua.edu.cn,
michael@mail.tsinghua.edu.cn

Abstract. Stereo matching task has been greatly improved by convolutional neural networks, especially the fully-convolutional network. However, existing deep learning methods always overfit to specific domains. In this paper, focus on domain adaptation problem of disparity estimation, we present a novel training strategy to conduct synthetic-realistic collaborative learning. At first, we design a compact model that consists of shallow feature extractor, correlation feature aggregator and disparity encoder-decoder. Our model enables end-to-end disparity regression with fast speed and high accuracy. To perform collaborative learning, we then propose two distinct training schemes, including guided label distillation and semi-supervised regularization, both of which are used to alleviate the lack of disparity labels in realistic datasets. Finally, we evaluate the trained models on different datasets that belong to various domains. Comparative results demonstrate the capability of our designed model and the effectiveness of collaborative training strategy.

Keywords: Stereo matching · Collaborative learning · Disparity · Guided label distillation · Semi-supervised regularization.

1 Introduction

Disparity estimation aims to find corresponding pixels between rectified stereo images [18]. It is a fundamental low-level task in computer vision, which has a wide range of applications such as depth prediction [32], scene understanding [12] and robotics navigation [37]. In recent years, deep learning methods [43, 29, 24, 31, 6, 26, 41] continuously improve the performance on specific scenes, while the domain adaptation for stereo matching gains more attention.

A popular pipeline for disparity estimation gets involved in matching cost computation, cost aggregation, disparity calculation, and disparity refinement [36]. Previous methods often manually design reliable features to describe image patches to localize matching correspondings [15, 5, 34]. These methods are easily

[†] State Key Laboratory of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology.

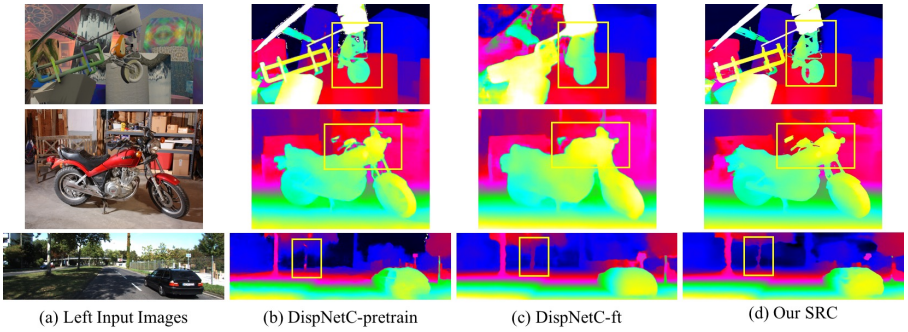

Fig. 1. Examples of predicted results of DispNetC [29] and our SRC model. From top to down, we test the models on Scene Flow dataset [29], Middleburry Stereo dataset [35] and KITTI Stereo 2015 dataset [30]. From left to right, we give left input images and predicted disparity maps by DispNetC-pretrained, DispNetC-finetuned, and our SRC model. All of the disparity maps are colorized by the devkit [30]. With the SRC training, our model can adapt to various domains

affected by textureless areas, shadow regions, and repetitive patterns. Since the convolutional neural network (CNN) exhibits great representative capacity on image classification [25], several approaches replace above-mentioned hand-craft features with CNN features [43, 28], which significantly increases the accuracy of disparity estimation. Inspired by the progress in semantic image segmentation task [27, 7], modern stereo methods adopt fully-convolutional network (FCN) to learn disparity map [29, 24]. These methods utilize siamese structure to process binocular inputs and design a correlation part to automatically encode matching costs. This structure enables an end-to-end disparity regression and further improves estimated accuracy and processing speed. In order to predict reasonable disparities on target scenes, these models are always pretrained on synthetic datasets and finely-tuned on realistic datasets. However, the resulted models are easily overfitted on specific domains. For example, in Fig. 1, the DispNetC model [29] pretrained on scene flow dataset is able to predict considerable results on synthetic scenes, but leading to mediocre outputs on realistic scenes. Meanwhile, the model finely-tuned on KITTI dataset [30] behaves well on road scenes, while suffering on indoor and virtual environments. From the above observation, existing deep methods cannot well address domain adaptation problem.

In this paper, our goal is to train better adaptable model for stereo matching. Instead of successively training on the synthetic and realistic dataset, we propose a novel synthetic-realistic collaborative (SRC) learning strategy, where virtual and real images are fused to train our network synchronously. We hope that the model maintains its properties on various domains through SRC training. Specifically, for virtual images in the synthetic dataset, high-quality disparity labels can be directly fetched to train the model in a supervised learning mode. For real images, since there are not enough disparity labels, we present two different schemes: 1) Guided label distillation. Here, an existing method is employed to

generate disparity maps that are used as guided labels for the realistic dataset, and then the supervised training can be seamlessly migrated to the realistic dataset. 2) Semi-supervised regularization. Unlike the supervised loss computed between predicted disparities and labels, a photometric distance is measured between the source image and the reconstructed image at the referenced view. In our experiments, the reconstructed image is warped from source image at the other view based on the current predicted disparity map. Along with photometric distance, we also add smoothness constraints to penalize disparity incoherence. Thus the semi-supervised regularization means that either the supervised loss or the unsupervised photometric loss is selected depending on whether labels are provided or not. The experimental results in Sec. 4 illustrate that both of guided label distillation and semi-supervised regularization make sense.

To take full advantage of the capacity of SRC learning, we design an end-to-end disparity regression model. The encoder-decoder architecture is also adopted in our model, embedded with a shallow feature extractor and a correlation feature aggregator. We use the extractor to obtain image features and the aggregator to combine matching cost between stereo features. The following encoder is a ResNet-like model to learn disparity information. The decoder is composed of several deconvolutional blocks to regress the full-size disparity map. We evaluate the SRC models on different datasets across various domains, including indoor [35], outdoor [30] and virtual [29] scenes. Compared to the baseline models which are only trained on an individual dataset, the SRC model shows significant superiorities, especially on unseen domains. Besides, we set different ratios between synthetic dataset and realistic dataset to exploit data properties for SRC learning. Our main contributions are summarized below:

- We develop a compact model that integrates shallow feature extractor, correlation aggregator and encoder-decoder to regress disparity map in an end-to-end manner.
- We propose a novel synthetic-realistic collaborative learning strategy. Two schemes as guided label distillation and semi-supervised regularization, are presented to conduct SRC model training.
- Comparative results are evaluated on different stereo datasets across various domains, which demonstrates the effectiveness of our SRC strategy for domain adaption problem.

2 Related Work

Disparity estimation from stereo images has been studied for several decades. Scharstein *et al.* [36] provide a taxonomy of stereo algorithms and analyze the typical four-step pipeline. In this section, we would not track back to early stereo methods but focus on recent deep learning approaches.

Zbontar and LeCun [43] first introduce CNN to describe image patches and compute matching cost. Luo *et al.* [28] design a siamese structure embedded with a product layer to calculate marginal distributions over all possible disparities. A multi-scale deep model presented by Chen *et al.* [8] leverages appearance data

to learn disparity from a rich embedding space. Shaked and Wolf *et al.* [38] propose a highway network along with a hybrid loss to measure the similarity between image patches on multi-levels. Compared to traditional approaches, the above methods adopt CNN features to conduct matching cost computation and improve the accuracy of disparity prediction with a considerable margin. However, these methods are still time-consuming due to the post-processing steps or complex optimization framework.

Inspired by FCN used in semantic segmentation task [27, 7], Mayer *et al.* [29] raise an encoder-decoder architecture called DispNet to enable end-to-end disparity regression. DispNet adopts a correlation operation as FlowNet [11] where the matching cost can be directly integrated to encoder volumes. Pang *et al.* [31] provide a cascade structure to optimize residues between predicted results and ground-truth values. Liang *et al.* [26] propose two-stage pipeline where the second sub-network is used to refine initial estimated disparity by measuring feature constancy. A few methods adopt three-dimensional convolutions to learn disparity. For example, Kendall *et al.* [24] integrate contextual information by 3D convolutions over a cost volume. A two-stream network proposed by Yu *et al.* [42] realizes cost aggregation and proposal selection respectively. Chang *et al.* [6] combine spatial pyramid network with 3D convolutional layers to incorporate global context information. Although these methods achieve state-of-the-art results on several stereo benchmarks by successively training on synthetic dataset [29] and realistic dataset [14, 30], there remains the domain adaptation problem because their models always overfit to specific domains belonging to current training datasets. Unlike the common training schedule, we propose SRC training strategy, which makes our model more reliable to domain shifts.

Another class of approaches attempts to exploit other information to improve stereo matching. Guney and Geiger [17] introduce object-aware knowledge into MRF formulation to resolve possible stereo ambiguities. Yang *et al.* [41] combine the high-level semantic information to optimize disparity prediction. Song *et al.* [39] utilize the cues of edge detection to recover disparity details. In addition, several approaches [1, 33, 4, 9] tackle semantic-level or instance-level information to improve the accuracy of optical flow which is a similar scene-matching task as disparity estimation. We argue that the introduction of other information may not be effective because domain adaptation is a wide-spread problem in vision tasks. Moreover, increased information also brings extra computations.

Recently, some unsupervised learning methods are proposed for depth prediction and scene matching. Garg *et al.* [13] estimate single-view depth by minimizing projection errors in stereo environment. Godard *et al.* [16] conduct left-right consistency check in a fully-differentiable structure. Yu *et al.* [22] fuse photometric loss and smoothness constancy to predict optical flow. Tonioti *et al.* [40] leverage on the confidence measures to finetune a standard stereo model. A guided flow method presented by Zhu *et al.* [45] employs FlowFields [2] to generate flow labels for flow learning. Our idea of SRC learning is inspired from these unsupervised methods. Concretely for the semi-supervised regularization,

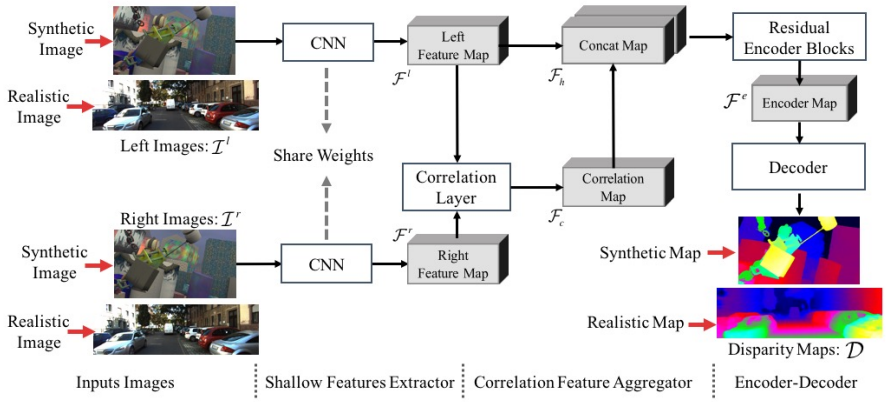

Fig. 2. Model architecture. The model can be divided into three parts: shallow feature extractor, feature aggregator and encoder-decoder. For SRC learning, the input data are fused with synthetic and realistic images

we combine the supervised loss and unsupervised loss together to train our model on the fusion set of synthetic and realistic data.

Our method follows the encoder-decoder architecture to regress disparity map. We utilize ResNet model [19] as the backbone and integrate the correlation part [11, 29] to compute cost volumes between stereo pairs. Focus on the domain adaptation problem, SRC learning strategy is proposed to train the model. Specifically, for the lack of disparity labels in the realistic dataset, we provide two schemes as guided label distillation and semi-supervised regularization. Finally, our experimental results evaluated across different datasets demonstrate the effectiveness of the SRC-learning.

3 Method

In this section, we first describe the model architecture for disparity regression in Sec. 3.1 and then explain the SRC learning strategy, including the guided label distillation and semi-supervised regularization in Sec. 3.2.

3.1 Model Architecture

Our model is shown in Fig. 2 and layer structural definition is detailed in Tab. 1. Given a pair of images $\mathcal{I}^l$ and $\mathcal{I}^r$, the goal is to estimate the dense disparity map $\mathcal{D}$. We use the ResNet-50 [19] as the backbone of our model. According to data flow, the model can be divided into three parts: shallow feature extractor, correlation feature aggregator and encoder-decoder. For the inputs of network at training time, we fuse synthetic and realistic images to conduct SRC learning. Our model enables accurate prediction of disparity map.

Table 1. Layer-by-layer structure of model. The "conv_block" denotes the convolutional block, where a convolutional layer is followed by batch normalization and ReLU activation. The "res_block" denotes the residual block designed by [19]. The "corr_1d" denotes the single-directional correlation [29]. The "deconv_block" denotes the deconvolutional block that is composed of deconvolutional layer, batch normalization and ReLU layer.

Layer	Attributes	Channels I/O	Scaling	Inputs
	1. Shallow Feature Extractor			
conv_block1_1	kernel size = 3, stride = 2	3 / 64	1/2	input stereo images
conv_block1_2	kernel size = 3, stride = 1	64 / 64	1/2	conv_block1_1
conv_block1_3	kernel size = 3, stride = 1	64 / 128	1/2	conv_block1_2
max_pooling	kernel size = 3, stride = 2	128 / 128	1/4	conv_block1_3
res_block2_1	kernel size = 3, stride = 1	128 / 256	1/4	max_pool_block1
res_block2_2	kernel size = 3, stride = 1	256 / 256	1/4	res_block2_1
res_block2_3	kernel size = 3, stride = 1	256 / 256	1/4	res_block2_2
res_block3_1	kernel size = 3, stride = 1	512 / 512	1/8	res_block2_3
	2. Feature Aggregator			
conv_block_pre	kernel size = 3, stride = 1	512 / 256	1/8	res_block3_1
corr_1d	max displacement = 32, single direction [29]	256 / 33	1/8	conv_block_pre
conv_trans	kernel size = 3, stride = 1	256 / 256	1/8	conv_block_pre
concat	aggregate corr_1d and conv_trans	(256 + 33) / 289	1/8	corr_1d, conv_trans
	3-1. Disparity Encoder-Decoder			
res_block3_2	kernel size = 3, stride = 1	409 / 512	1/8	concat
res_block3_3	kernel size = 3, stride = 1	512 / 512	1/8	res_block3_2
res_block3_4	kernel size = 3, stride = 1	512 / 512	1/8	res_block3_3
res_block4_1	kernel size = 3, stride = 1, dilated pattern	512 / 1024	1/8	res_block3_3
res_block4_2	kernel size = 3, stride = 1, dilated pattern	1024 / 1024	1/8	res_block4_1
res_block4_3	kernel size = 3, stride = 1, dilated pattern	1024 / 1024	1/8	res_block4_2
res_block4_4	kernel size = 3, stride = 1, dilated pattern	1024 / 1024	1/8	res_block4_3
res_block4_5	kernel size = 3, stride = 1, dilated pattern	1024 / 1024	1/8	res_block4_4
res_block4_6	kernel size = 3, stride = 1, dilated pattern	1024 / 1024	1/8	res_block4_5
res_block5_1	kernel size = 3, stride = 1, dilated pattern	1024 / 2048	1/8	res_block4_6
res_block5_2	kernel size = 3, stride = 1, dilated pattern	2048 / 2048	1/8	res_block5_1
res_block5_3	kernel size = 3, stride = 1, dilated pattern	2048 / 2048	1/8	res_block5_2
conv_block5_4	kernel size = 3, stride = 1	2048 / 512	1/8	res_block5_3
deconv_block1	kernel size = 3, stride = 2	512 / 256	1/4	conv_block5_4
deconv_block2	kernel size = 3, stride = 2	256 / 128	1/2	deconv_block1
deconv_block3	kernel size = 3, stride = 2	128 / 64	1	deconv_block2
disp_conv	kernel size = 3, stride = 1	64 / 1	1	deconv_block3

Shallow feature extractor We use the shallow part of ResNet-50 model to extract image features $\mathcal{F}^l$ and $\mathcal{F}^r$. This part contains three convolutional blocks, a max-pooling layer and four residual blocks. It subsamples the input images in two stages: "conv_block_1_1" and "max_pool1", which results in 1/8 scaling to raw images. Compared with original images, the features obtained from shallow extractor are more robust to local context.

Correlation Feature Aggregator A correlation layer [29] is adopted to compute cost volumes $\mathcal{F}^c$ between $\mathcal{F}^l$ and $\mathcal{F}^r$. We only perform single-direction search due to epipolar property. Both max displacement and padding size are set to 32 so that channels of $\mathcal{F}^c$ are 33. Besides, left features $\mathcal{F}^l$ are preserved

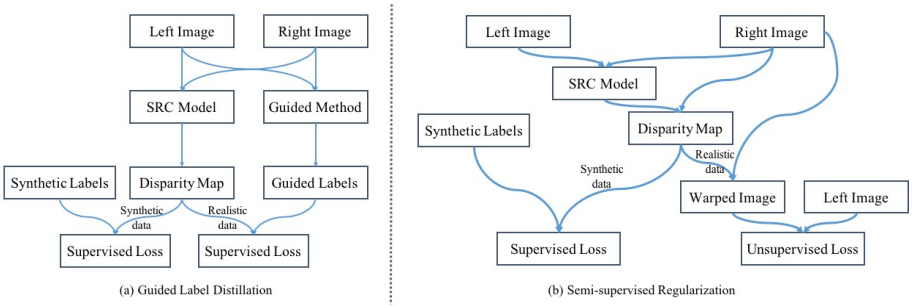

Fig. 3. Diagram of guided label distillation and semi-supervised regularization

for the detailed information on reference view. The cost volumes $\mathcal{F}^c$ and left features $\mathcal{F}^l$ are aggregated to form hybrid map $\mathcal{F}^h$. The feature aggregator integrates image features and matching information together for posterior disparity learning.

Encoder-Decoder After feature aggregation, we feed hybrid map $\mathcal{F}^h$ into encoder-decoder to regress full-size disparity map $\mathcal{D}$. As depicted in Tab. 1, the encoder consists of 12 residual blocks. Several convolutional operations in residual blocks use dilation pattern [7] for larger receptive fields. In the decoder, we place three deconvolutional blocks to gradully upsample the spatial size of feature maps. An extra one-channel convolutional layer is appended at the end to regress full-size disparity map. Our model is also fully-convolutional so that it enables end-to-end disparity learning.

3.2 SRC Learning

As introduced in Sec. 1, the idea of SRC learning is to find an available training solution on fused datasets, especially on the realistic datasets. To enable training on real images, we explain two schemes below:

Guided label distillation A direct way for collaborative learning on realistic dataset is that we adopt an existing method to generate disparity maps. Although the labels predicted by guided method are not exactly accurate as synthetic labels, we find they can help our model converge to a certain level. This scheme has two advantages: 1) The guided method can produce large amounts of disparity labels so that we do not need manual annotations or extra equipments, such as Lidar and depth camera. 2) It enables seamless supervised training on realistic datasets, which makes our model better adaptable to target domains.

In our experiments, we employ SGM algorithm [20] as the guided method. Since the generated disparity maps are not dense, we only measure loss on valid pixels. The loss function is expressed as:

$$\mathcal{L}_{sup} = \frac{1}{N_{\mathcal{V}}} \sum_{i,j \in \mathcal{V}} \|\mathcal{D}_{i,j} - \hat{\mathcal{D}}_{i,j}\|_1, \tag{1}$$

where $\mathcal{V}$ is the set of valid disparity pixels, $N_{\mathcal{V}}$ is the number of valid pixels, $\mathcal{D}$ is the predicted disparity map and $\hat{\mathcal{D}}$ is the disparity label map. Here we adopt the l_1 norm to measure distance between predictions and labels. The experimental results in Sec. 4.3 show that the SRC model trained with guided label distillation outperforms the models trained on the individual datasets.

Semi-supervised regularization We introduce unsupervised training based on spatial transformation to constitute semi-supervised regularization for SRC learning. As shown in Fig. 3(b), stereo images are fed to SRC model and we obtain predicted disparity map. Based on the disparity map, we warp the right image to left view and get the reconstructed image. Our image reconstruction adopts bilinear sampling where the output pixel is the weighted sum of nearest two input pixels [21]. Such sampling operation is differentiable and enables loss propagation. Compared with guided label distillation, semi-supervised regularization further gets rid of the dependence on guided method. Here, we measure the photometric loss [22] between source left image and reconstructed image on all pixels:

$$\mathcal{L}_p = \frac{1}{N} \sum_{i,j} \|\tilde{\mathcal{I}}^l_{i,j} - \mathcal{I}^l_{i,j}\|_1, \tag{2}$$

where $\mathcal{I}^l$ denotes source left image and $\tilde{\mathcal{I}}^l$ denotes reconstructed left image. In addition, we define smoothness loss to penalize discontinuity on disparity maps:

$$\mathcal{L}_s = \frac{1}{N} \sum_{i,j} [\rho_s(\mathcal{D}_{i,j} - \mathcal{D}_{i+1,j}) + \rho_s(\mathcal{D}_{i,j} - \mathcal{D}_{i,j+1})], \tag{3}$$

where $\rho_s(\cdot)$ is the spatial smoothness penalty implemented as generalized Charbonnier function [3]. The photometric loss and the smoothness loss are made as the unsupervised loss for realistic datasets. The overall semi-supervised loss is regularized as:

$$\mathcal{L}_{semi} = \delta \mathcal{L}_{sup} + (1 - \delta)(\lambda_p \mathcal{L}_p + \lambda_s \mathcal{L}_s), \tag{4}$$

where $\delta \in \{0, 1\}$, λ_p denotes the weight of photometric loss and λ_s denotes the weight of smoothness loss. When training SRC model with semi-supervised regularization, δ is set to 1 for synthetic data and 0 for realistic data.

4 Experimental Results

In this section, we fuse Scene Flow dataset [29] and Cityscapes dataset [10] to train the model, where the former is the synthetic dataset and the latter is the realistic dataset. The well-trained models are evaluated on Scene Flow test set [29], KITTI stereo 2015 [30] and Middleburry stereo 2014 [35] which represent virtual, outdoor and indoor domains, respectively. Related datasets and evaluation metrics are introduced in 4.1. Implementation details are described in 4.2. Ablation studies on guided label distillation and semi-supervised regularization are provided in 4.3 and 4.4, respectively. Finally, we compare our method with other methods in 4.5.

4.1 Datasets and Evaluation Metrics

The Scene Flow dataset [29] is a synthetic dataset for scene matching including disparity estimation and optical flow prediction. This dataset is rendered by computer graphics techniques with background scenes and 3D foreground models. It contains 22,390 images for training and 4,370 images for testing. Image size is $H = 540$ and $W = 960$.

The Cityscapes dataset [10] is a realistic dataset that is released for urban scene understanding. It provides stereo images and corresponding disparity maps which are pre-computed by SGM algorithm [20] so that we directly use these disparity maps as guided labels. Gathering the stereo pairs from different subsets, we can fetch over 20,000 images with the size of $H = 1024$ and $W = 2048$. These subsets contain "train", "validate", "test" and "extra train" sets.

The KITTI Stereo 2015 dataset is also released for real-world autopilot scenes. It contains 200 training and 200 testing image pairs. Since the disparity labels for testing set are not released, we evaluate our model on the training set. The average image size is $H = 376$ and $W = 1240$.

The Middleburry Stereo 2015 dataset [35] provides 30 pairs for indoor scenes, where 15 each for training and testing. This dataset offers different resolutions and we select quarter resolution for model evaluation.

We use the Scene Flow training split [29] and Cityscapes dataset [10] to train our model. To keep balance, we respectively choose $22,000$ images from Scene flow dataset and Cityscapes dataset so that a maximum number of $44,000$ images can be used for SRC training. The Scene Flow testing set [29], KITTI Stereo datasets [14, 30] and Middleburry Stereo 2014 [35] are selected for model evaluation. We apply the end-point-error (EPE) and the bad pixel error (D1) as evaluation metrics, where the threshold in D1 is set to 3. For KITTI datasets, the errors in both non-occluded regions (Noc) and all pixels (All) are calculated. In addition, we depict colorized disparity maps and error maps for better visualization. In the error maps such as Fig. 4, blue areas represent correct predictions and red regions indicate mistaken estimates.

4.2 Implementation Details

Our model shown in Tab. 1 is implemented on a customized Caffe [23]. We use the "poly" learning rate policy where current learning rate equals to the base one multiplying $(1 - \frac{iter}{max\_iter})^{power}$ [5, 44]. At training time, we set base learning rate to 0.01, power to 0.9, momentum to 0.9 and weight decay to 0.0001 respectively. The maximum iterations and batch size are set to $200K$ and 16 for ablation studies in Sec. 4.3 and Sec. 4.4. We select the GPU of NVIDIA Titan Xp for model training and testing.

For data augmentation, we adopt random resize, color shift and contrast brightness adjustment. The random factor is between 0.5 to 2.0. The maximum color shift along RGB axes is set to 10 and maximum brightness shift is set to 5. The contrast multiplier is between 0.8 and 1.2. The "cropsize" is set to 513×513 and batch size is set to 16.

For parameters in semi-supervised regularization Eq. 4, the loss weights λ_p and λ_s for photometric term and smoothness term are set to 1.0 and 0.1. The Charbonnier terms α, β and ϵ in smoothness loss term are 0.21, 5.0 and 0.001 as described in [22].

4.3 Ablation Study for Guided Label Distillation

We conduct four groups of experiments on guided label distillation. As described in 4.1, Scene Flow dataset [29] and Cityscapes dataset [10] are selected as synthetic and realistic dataset. Here, the guided labels for realistic data are precomputed by SGM method [20]. The first column in Tab. 2 indicates the current dataset settings for training, where the values of "Synth." and "Real." denote the used ratios of synthetic and realistic images. For example, the values in first line of Group 4 are 1/8 and 1, which means 2,750 synthetic images and 22,000 realistic images are used for training.

The first group of experiments are performed to compare the SRC-trained model with synthetic-trained model and realistic-trained model. On scene flow validation set [29], the error rate of SRC-trained model is flat to synthetic-trained model and much lower than realistic-trained model. On KITTI Stereo 2015 dataset [30], the SRC-trained model performs much better than synthetic-trained model and also achieves higher accuracy than realistic-trained model. On Middleburry dataset [35], the SRC-trained model outperforms the other two models with a large margin, where the EPE is reduced from 2.42 to 1.82 compared to synthetic-trained model, and the D1 error is improved by 8% compared to realistic-trained model. This group of experiments proves the effectiveness of guided label distillation.

Table 2. Results of guided label distillation.

Settings		Scene Flow [29]		KITTI Stereo 2015 [30]				Middleburry [35]	
Synth.	Real.	EPE	D1	Noc EPE	All EPE	Noc D1	All D1	EPE	D1
Group 1: Compare SRC-trained model with individual-trained models.									
1	0	**2.89**	10.69	3.63	3.65	16.90	17.22	2.42	15.08
0	1	6.50	19.05	1.26	1.29	6.21	6.42	3.36	20.93
1	1	2.92	**10.33**	**1.21**	**1.23**	**5.91**	**6.15**	**1.82**	**12.10**
Group 2: SRC models trained with different amounts of data.									
1/8	1/8	3.00	10.94	1.23	1.25	6.12	6.34	1.87	12.37
1/4	1/4	2.96	10.60	**1.20**	**1.22**	**5.83**	**6.09**	1.88	12.39
1/2	1/2	2.94	10.50	1.21	1.24	5.95	6.19	1.94	13.13
1	1	**2.92**	**10.33**	1.21	1.23	5.91	6.15	**1.82**	**12.10**
Group 3: SRC models trained with different amounts of realistic data.									
1	1/8	2.99	12.00	1.24	1.26	6.32	6.59	1.88	12.16
1	1/4	**2.88**	**10.14**	1.24	1.26	6.18	6.44	1.84	**12.09**
1	1/2	2.96	11.02	**1.20**	**1.23**	5.97	6.21	1.88	13.01
1	1	2.92	10.33	1.21	**1.23**	**5.91**	**6.15**	**1.82**	12.10
Group 4: SRC models trained with different amounts of synthetic data.									
1/8	1	3.18	12.25	1.23	1.25	5.99	6.18	2.11	14.16
1/4	1	3.05	11.48	**1.20**	**1.22**	5.91	6.13	2.08	13.33
1/2	1	3.00	10.75	1.22	1.24	**5.86**	**6.08**	1.88	12.90
1	1	**2.92**	**10.33**	1.21	1.23	5.91	6.15	**1.82**	**12.10**

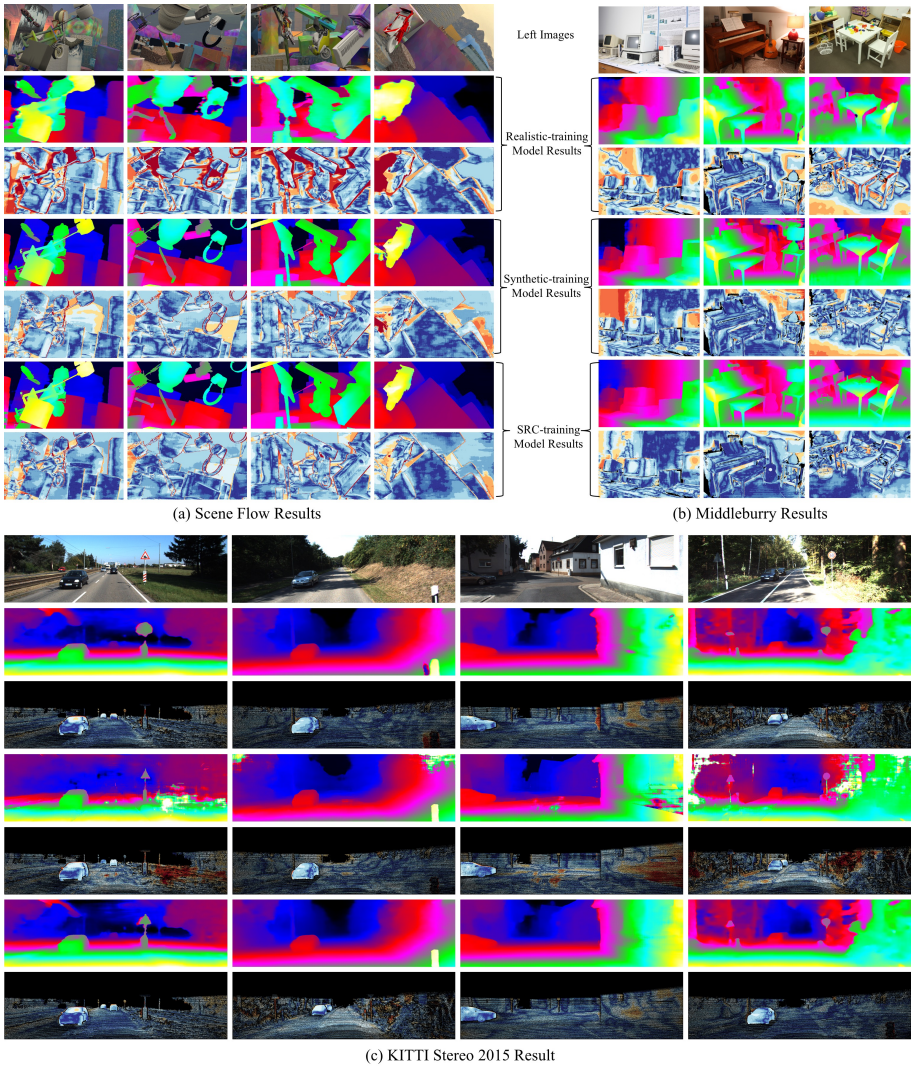

Fig. 4. Qualitative results of SRC-trained models with guided label distillation. These results are tested on Scene Flow validation set [29], Middleburry Stereo 2014 dataset [35] and KITTI Stereo 2015 dataset [30] respectively

In the second group of experiments, we hold the balance between synthetic and realistic datasets and reduce training data. We observe that the SRC models trained by 1/4 and 1/2 ratios yield similar accuracy to the full-data training model. When the ratio decreases to 1/8, the error rates increase on Scene Flow [15] and KITTI Stereo [30] datasets. Based on this group of experiments, we suggest 5k or more images to train SRC models for guaranteed quality.

In the third group, we fix the synthetic data ratio and increase the ratio of realistic data, and no significant improvement is gained from incremental realistic images. In contrast, we keep the quantity of realistic dataset and raise the ratio of synthetic images in fourth group of experiments. The results on Scene Flow [29] and Middleburry dataets [35] are gradually improved. Here, the key difference between synthetic and realistic datasets is the quality of disparity labels. We analyze that, when the model converges to a certain level during training, the potential errors in guided labels may hinder the further boosts. Nevertheless, a certain amount of guided labels for realistic dataset is still a prerequisite of SRC learning.

In Fig. 4, we show several qualitative examples of SRC-trained models on different scenes. From the colorized disparity maps and error maps, we find that the synthetic-trained model is keen to object edges, while the realistic-trained model behaves better consistency on big objects, such as road and car. The SRC-trained model combines the advantages to provide more reasonable predictions on various domains.

4.4 Ablation Study for Semi-supervised Regularization

For SRC learning by semi-supervised regularization, we calculate supervised loss for synthetic data and unsupervised loss for realistic data as described in Sec. 3.2. We conduct two groups of experiments to illustrate the effects of such regularization. In the first group of experiments, the SRC-trained model is compared to synthetic-trained and realistic-trained models. With synthetic labels, the SRC-trained model performs much better than realistic-trained model on all benchmarks. When fed with realistic data and constraints of photometric consistency, the SRC-trained model achieves higher accuracy on KITTI Stereo [30] and Middleburry [35] datasets than synthetic-trained model. These three experiments validate the semi-supervised regularization.

In the second group of experiments, we explore the impacts of different amounts of data. Similar to guided label distillation, we adopt four ratios of training data. With the increase of ratio, no remarkable improvement can be found on Scene Flow dataset [29] and Middleburry dataset [35], and a little

Table 3. Results of semi-supervised regularization.

Settings		Scene Flow [29]		KITTI Stereo 2015 [30]				Middleburry [35]	
Synth.	Real.	EPE	D1	Noc EPE	All EPE	Noc D1	All D1	EPE	D1
Group 1: Compare SRC-trained model with individual-trained models.									
1	0	**2.89**	10.69	3.63	3.65	16.90	17.22	2.42	15.08
0	1	8.81	24.22	1.88	2.11	8.85	9.55	5.69	29.28
1	1	**2.89**	**10.56**	**1.39**	**1.50**	**7.14**	**7.74**	**2.33**	**15.01**
Group 2: SRC models trained with different amounts of data.									
1/8	1/8	2.96	10.62	1.46	1.57	7.31	7.89	**2.19**	**13.86**
1/4	1/4	2.95	10.90	1.52	1.63	7.21	7.82	2.46	15.15
1/2	1/2	2.90	**10.49**	1.48	1.63	7.22	7.93	2.55	15.88
1	1	**2.89**	10.56	**1.39**	**1.50**	**7.14**	**7.74**	2.33	15.01

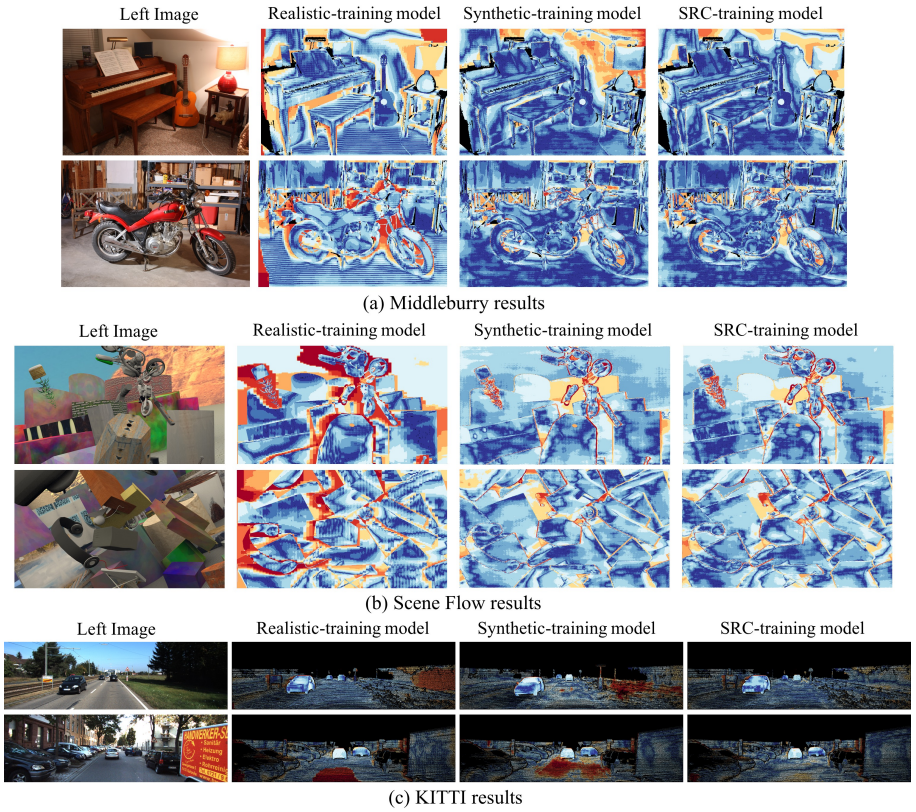

Fig. 5. Error maps of SRC-trained models with semi-supervised regularization. From top to down, we provide Scene Flow [29], Middleburry Stereo 2014 [35] and KITTI Stereo 2015 [30] examples, respectively. From left to right, we provide left input images and error maps by realistic-trained, synthetic-trained and SRC-trained models

progress emerges on KITTI dataset [30]. We argue that the quantity of training data is not the core factor for semi-supervised regularization. Besides, the overall results of semi-supervised regularization are worse than guided label distillation. We analyze that the photometric loss measured between potential corresponding image patches only provides weak guidance. Compare with supervised loss computed on SGM labels [20], the unsupervised loss is easily influenced by ambiguities and illuminations. Even so, the semi-supervised regularization enables SRC learning without ground-truth labels or guided labels on realistic datasets.

In Fig. 5, we show several predictive examples. In contrast to the realistic-trained model that is the purely unsupervised model, the SRC-trained model predicts more accurate disparities on edges, sharp positions, and small objects. Compared to synthetic-trained model, the SRC-trained model further reduces the errors on local ambiguity areas.

Table 4. Comparative results to other methods.

Methods	Scene Flow [29]		KITTI Stereo 2015 [30]		Middleburry [35]		Running time (s)
	EPE	D1	All EPE	All D1	EPE	D1	
SGM [20]	7.29	16.18	5.02	14.79	8.29	25.35	1.47
DispNetC [29]	2.33	10.04	1.61	10.84	3.09	18.85	**0.05**
CRL [31]	1.67	6.70	1.40	8.18	1.77	13.47	0.16
iResNet [26]	**1.27**	**4.90**	**0.70**	**2.38**	1.74	11.06	0.13
EdgeStereo [39]	1.33	5.26	1.48	8.64	**1.57**	11.38	0.21
SRC (Ours)	2.72	8.45	1.12	5.64	1.67	**10.96**	0.29

4.5 Compare with other methods

To exploit the potential of SRC-learning, we adopt guided label distillation, and we increase the training iterations from $200K$ to $500K$, and the batch size from 16 to 32. More training epoches further improve the performance of our model. We compare our SRC method to other classical or deep learning-based methods, including SGM [20], DispNetC [29], CRL [31], EdgeStereo [39] and iResNet [26]. The deep learning-based models [29, 31, 39, 26] are pretrained on Scene Flow dataset without finely tuning on specific datasets.

We list the results of different methods in Tab. 4. We find all of the deep learning methods outperform classical SGM algorithm on three benchmarks. Although our SRC model ranks at penultimate on Scene Flow dataset [29], our method ranks second on both KITTI Stereo 2015 benchmark [30] and Middleburry dataset [35]. It is remarkable that KITTI Stereo 2015 dataset and Middleburry dataset are unseen in the training period so that it illustrates that our SRC-learning can help model better adapt to various domains. We believe that the SRC learning can be used as a general strategy for other deep learning-based stereo matching models.

5 Conclusion

As a core problem in low-level vision, disparity estimation is required to have the properties of fast speed, high accuracy, and adaptability to various domains. In this paper, we develop a compact model that is composed of shallow feature extractor, matching feature aggregator and encoder-decoder. We also present SRC learning strategy for joint training on synthetic and realistic datasets. Two schemes, *i.e.* guided label distillation and semi-supervised regularization, are provided to mitigate for the lack of labels in realistic datasets. Our experimental results evaluated on different datasets demonstrate the effectiveness of our deep learning model and SRC strategy.

Acknowledgment

This work was supported in part by the National Key R&D Program of China under Grant No. 2017YFB1302200 and by Joint Fund of NORINCO Group of China for Advanced Research under Grant No. 6141B010318.

References

1. Bai, M., Luo, W., Kundu, K., Urtasun, R.: Exploiting semantic information and deep matching for optical flow. In: ECCV (2016)
2. Bailer, C., Taetz, B., Stricker, D.: Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In: ICCV (2015)
3. Barron, J.T.: A more general robust loss function. arXiv preprint arXiv:1701.03077 (2017)
4. Behl, A., Jafari, O.H., Mustikovela, S.K., Alhaija, H.A., Rother, C., Geiger, A.: Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In: ICCV (2017)
5. Brown, M., Hua, G., Winder, S.: Discriminative learning of local image descriptors. TPAMI (2011)
6. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: CVPR (2018)
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI (2016)
8. Chen, Z., Sun, X., Wang, L., Yu, Y., Huang, C.: A deep visual correspondence embedding model for stereo matching costs. In: ICCV (2015)
9. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: Joint learning for video object segmentation and optical flow. In: ICCV (2017)
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
11. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: ICCV (2015)
12. Franke, U., Joos, A.: Real-time stereo vision for urban traffic scene understanding. In: IV (2000)
13. Garg, R., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: ECCV (2016)
14. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012)
15. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale stereo matching. In: ACCV (2010)
16. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
17. Guney, F., Geiger, A.: Displets: Resolving stereo ambiguities using object knowledge. In: CVPR (2015)
18. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
20. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. TPAMI (2008)
21. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: NIPS (2015)
22. Jason, J.Y., Harley, A.W., Derpanis, K.G.: Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In: ECCV Workshop (2016)

23. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R.B., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM MM (2014)
24. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: ICCV (2017)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
26. Liang, Z., Feng, Y., Guo, Y., Liu, H., Chen, W., Qiao, L., Zhou, L., Zhang, J.: Learning for disparity estimation through feature constancy. In: CVPR (2018)
27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
28. Luo, W., Schwing, A.G., Urtasun, R.: Efficient deep learning for stereo matching. In: CVPR (2016)
29. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: CVPR (2016)
30. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR (2015)
31. Pang, J., Sun, W., Ren, J., Yang, C., Yan, Q.: Cascade residual learning: A two-stage convolutional neural network for stereo matching. In: ICCV Workshop (2017)
32. Rajagopalan, A., Chaudhuri, S., Mudenagudi, U.: Depth estimation and image restoration using defocused stereo pairs. TPAMI (2004)
33. Ren, Z., Sun, D., Kautz, J., Sudderth, E.B.: Cascaded scene flow prediction using semantic segmentation. In: 3DV (2017)
34. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Deepmatching: Hierarchical deformable dense matching. IJCV (2016)
35. Scharstein, D., Hirschmller, H., Kitajima, Y., Krathwohl, G., Nei, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: GCPR (2014)
36. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV (2002)
37. Schmid, K., Tomic, T., Ruess, F., Hirschmuller, H.: Stereo vision based indoor/outdoor navigation for flying robots. In: IROS (2013)
38. Shaked, A., Wolf, L.: Improved stereo matching with constant highway networks and reflective confidence learning. In: CVPR (2017)
39. Song, X., Zhao, X., Hu, H., Fang, L.: Edgestereo: A context integrated residual pyramid network for stereo matching. arXiv preprint arXiv:1803.05196 (2018)
40. Tonioni, A., Poggi, M., Mattoccia, S., Di Stefano, L.: Unsupervised adaptation for deep stereo. In: ICCV (2017)
41. Yang, G., Zhao, H., Shi, J., Deng, Z., Jia, J.: Segstereo: Exploiting semantic information for disparity estimation. In: ECCV (2018)
42. Yu, L., Wang, Y., Wu, Y., Jia, Y.: Deep stereo matching with explicit cost aggregation sub-architecture. In: AAAI (2018)
43. Zbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. JMLR (2016)
44. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
45. Zhu, Y., Lan, Z., Newsam, S., Hauptmann, A.G.: Guided Optical Flow Learning. In: CVPR Workshop (2017)