PHYSICS-REGULARIZED STEREO MATCHING FOR DEPTH ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Depth estimation from stereo or multi-view images is an essential technology for a wide range of vision and robotics applications. In recent years, many deep learning based methods have been proposed for this purpose. However, training the stereo matching network is challenging and requires a large amount of data, especially for the 3D convolution networks. Existing stereo matching approaches are mostly data-driven, which often converge to a local minimum biased to the training data. In this paper, we propose a novel self-supervised physics regularization framework to improve the training of the networks using physical knowledge or constraints. More specifically, we explore the use of low-level structures as physical constraints for the regularization of the stereo-matching network via multi-task learning. Moreover, a disparity aggregation module is proposed to aggregate the disparity output with image features to consider the association in between. We also find that the canny edge can be used as a pseudo ground truth to train the network with performance comparable to the ideal ground truth edge in the Scene Flow dataset. We combine the proposed physics regularization with four existing stereo matching algorithms. The experimental results in three public datasets, including Scene Flow, KITTI 2012, and KITTI 2015, show the effectiveness and generality of the proposed framework.

1 INTRODUCTION

Depth estimation from stereo or multi-view images is of substantial interest due to a wide range of applications including robotics (Mancini et al., 2016; Saxena et al., 2007), augmented reality (Zhou et al., 2017; Nguyen et al., 2018; Alhaija et al., 2018), medical image analysis (Ong et al., 2020), surgery (Ye et al., 2017) and more (Chen et al., 2015; Zhang et al., 2015). The basic principle for the stereo matching is based on corresponding pixels between the left and right camera images. Given a pixel (x, y) in the left image which is found to be matching with a pixel at $(x - d, y)$ in the right image, the depth z of this pixel is linear to $1/d$: $z = \frac{f \cdot B}{d}$, where f refers to the focal length of the camera and B denotes the distance between centers of the two cameras. Typical stereo matching methods involve finding the corresponding points based on matching cost and post-processing to compute the depth (Scharstein & Szeliski, 2002). With the development of deep learning, convolutional neural networks (CNN) have been adopted to compute the matching cost between two image patches via similarity computation (Shaked & Wolf, 2017; Seki & Pollefeys, 2017). Later, more complex network architectures have been proposed to compute the matching cost. These methods usually first extract image features from the left and right images via two backbones with shared weights. Then the extracted features are used to construct a cost volume. Finally, the cost volume is processed by cost aggregation to estimate the disparity or the depth.

Although the recent deep learning-based approaches have shown to be promising to stereo depth estimation, they are mostly data-driven. A challenge is that the training of such data-driven models especially the 3D convolutions is not easy and requires large amount of training data. Based on recent progress in deep learning, it is expected that the backbone network is able to extract representative features for the left and right images. However, the purely data-driven model optimized for stereo matching may lead to a representation that is locally optimal for the stereo matching task alone. To avoid that, a spontaneous idea is to use additional tasks as constraints to regularize the training. Recent works (Ramirez et al., 2018; Zhang et al., 2019b) in semantic stereo matching have

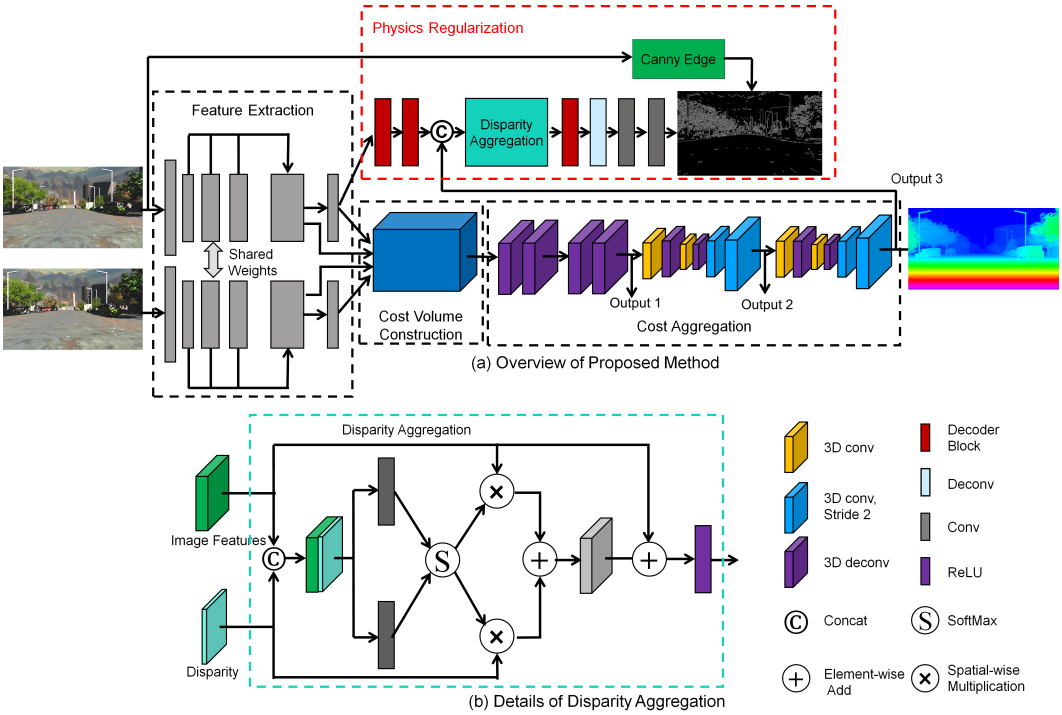

Figure 1: (a) The pipeline of the proposed physics-regularized network. On top of an existing stereo matching network, we further design a physics-regularization network which introduces additional constraints to the existing network. (b) A disparity aggregation module is introduced to integrate the disparity estimation output into the physics-regularization.

shown that the visual clues from semantic labels can be used to improve the performance of stereo matching. These methods can also be considered as using semantic labels as extra constraints to regularize the training. However, it is time-consuming and expensive to annotate semantic labels in large scale, especially for indoor scenarios where both the foreground and backgrounds are complex.

Instead of using semantic labels, we propose to use low-level structures to provide the visual clues. We consider the low-level structures as a type of physical constraints as they indicate the local change of intensities and provide physical clues for disparity or depth. For example, the depths of the neighbouring pixels from the same side of a structure are more likely to closer than those from different sides. Low-level structure information has shown to be useful in many other applications. Zhou et al. (2021) proposed to use the low-level structure as a complementary to semantic structure in anomaly detection. The annotation cost of low-level structures is cheaper than the semantic labels. The advantages of using low-level structure is two-fold. First, it improves the representation capability of the feature extraction module. As low-level structures detection is explainable with physical meanings, it helps to drive the trained model from purely data-driven toward physics-driven. Second, it is easier to obtain the low-level structures than the semantic labels. The low-level structures can be estimated by pretrained deep learning models or traditional handcrafted methods such as Canny edge detector (Canny, 1986). In this paper, we propose to use the Canny edge as a pseudo ground truth of the low-level structure. As the Canny edge can be obtained automatically without manual annotation, we call our framework as self-supervised physics regularization framework.

A straight forward way to use the low-level structure information is to use it via multi-task learning. However, this may ignore the relationship between the low-level structure and the disparity changes. In practice, we often observe that a large change in disparity or depth often leads to a substantial change in intensities. This is reasonable as the objects with difference depths are often under different viewing angles and lightning conditions, which make them appear differently even if they have the same original colors. Motivated from this, we further propose a disparity aggregation module

to enhance the association in between. We name this module as disparity aggregation module as it aggregates the disparity estimation with feature maps from RGB.

Figure 1 illustrates the proposed self-supervised physics regularization framework. As enclosed by dashed lines in black, current stereo matching network such as ACVNet (Xu et al., 2022) usually contains a feature extraction module, a cost volume construction module and a cost aggregation module. We introduce a physics regularization network, enclosed by the dashed lines in red. The proposed physics regularization is essentially a low-level structure detection module via multi-task learning. Moreover, a disparity aggregation module is proposed to aggregate the estimated disparity from the original stereo matching with RGB features for low-level structure detection. It shall be noted that we are not proposing a new disparity estimation method. Instead, we propose a general framework that can be applied to regularize current stereo matching methods and improve their performances. The proposed physics regularization network takes features from one image (left image is used in this work) and the estimated disparity as inputs to compute low-level structures.

We conduct comprehensive experiments to validate the effectiveness of the proposed framework on Scene Flow(Mayer et al., 2016), KITTI 2012 (Geiger et al., 2012), and KITTI 2015 (Menze et al., 2015) datasets. The major contributions of this paper are:

- We propose a novel framework to regularize the stereo matching using physical constraints from low-level structures.
- We propose a disparity aggregation module to leverage on the association between the disparity and RGB information in the physics regularization.
- We propose to use the Canny edge as self-supervised labels for regularization and find that the Canny edge works comparably to ground truth structure in Scene Flow dataset.
- We integrate the proposed physics regularization with four different stereo matching networks, experimental results show that it is generic for different networks.

2 Related works

2.1 Stereo Matching

Many deep learning based algorithms have been proposed for stereo matching. In GC-Net, Kendall et al. (2017) proposed end-to-end learning to estimate disparity using 3D CNN to filter the cost volume. In PSMNet, Chang & Chen (2018) proposed to use spatial pyramid pooling and a stacked hourglass 3D CNN for regularizing cost volume. In GA-Net, Zhang et al. (2019a) integrated semi-global matching (Hirschmuller, 2005) into 3D CNN for cost filtering. In AANet, Xu & Zhang (2020) replaced the time consuming 3D CNN with cost aggregation algorithms. In DeepPruner, Duggal et al. (2019) developed a differentiable patch match module to reduce disparity searching space. In GwcNet, Guo et al. (2019) proposed group-wise correlation to obtain efficient representations for measuring feature similarities. Recently, Xu et al. (2022) proposed to compute attention weights to suppress redundant information and enhance the concatenation volume. Besides the performance, computational cost is another major factor. PSMNet requires more than 6 seconds on an NVIDIA Jetson TX2 module, which is too slow for practical applications on edge side devices such as robots. In RTNet, Chang et al. (2020), proposed to learn adaptive fusion of multi-scale features in a similar way and achieved 12-33 frames per second with a trade-off in performance. In AnyNet, Wang et al. (2019) proposed successive update in multi-scale resolutions for trade-off between computation and accuracy. In STTR, Li et al. (2021) revisited the stereo matching from a transformer perspective and replace cost volume construction with dense pixel matching. These end-to-end approaches demonstrated the state-of-the-art performance on stereo matching. However, these models are mainly data-driven.

2.2 Semantic Information Extraction

Depth and semantic information are often needed for high-level tasks such as reasoning, planning, collision prevention and etc. To reduce the computational cost, joint optimization of semantic segmentation and disparity estimation yields mutual benefit to both tasks while saving the computational cost in backbone feature extraction. The depth estimation in the challenging portions of the

images corresponding to reflective surfaces can be improved by knowing that they belong to an object with defined 3D properties. On the other hand, depth information can be used to reduce ambiguity in the segmentation of vegetation and terrain. Ramirez et al. (2018) proposed to leverage on semantics and geometry by enforcing spatial proximity between depth discontinuities and semantic for monocular depth estimation. Zhang et al. (2019b) proposed a structure and smoothness loss to fuse semantic segmentation with disparity estimation. Wu et al. (2019) proposed pyramid cost volumes to capture semantic and multi-scale spatial information for semantic stereo matching. Dovesi et al. (2020) proposed real-time semantic stereo network for jointly solving the depth estimation task and semantic segmentation task. Although the above approaches improve the performance, semantic labels are required in the training, which is expensive to obtain. For outdoor self-driving scenario which is highly structured, at least 10 to 20 major objects need to be detected in 3D, including cars, trucks, pedestrians, motorcycles, traffic signs, etc. For indoor scenario which is less structured and more complex, more objects need to be detected. Besides semantic segmentation, low-level structure detection such as edge detection is another option. Song et al. (2020) proposed EdgeStereo to incorporate edge information into the disparity branch by the edge feature embedding and edge-aware smoothness loss. Similarly, Yang et al. (2022) proposed to use edge supervision via multi-task learning. These methods still need manual annotation of edges.

Compared with semantics, low-level structure information is easier to be computed. For example, traditional methods such as Canny operator (Canny, 1986) and Sobel operator (Sobel & Feldman, 1973), can be used to compute the low-level structures. Pretrained deep learning models can also be used. In some of recent work, synthetic images are used to train the deep learning models, where the edges can be easily obtained. In SketchGAN, Zhang et al. (2019c) proposed a method composed of sketch generation followed by image painting. Therefore, obtaining images with corresponding edges is much easier and faster.

2.3 AGGREGATION

Attention has been widely utilized in computer vision to aggregate the most important features. Hu et al. (2018) proposed SENet to recalibrate feature responses by modelling interdependencies among different channels. Li et al. (2019) proposed SKNet to select kernel sizes and adjust the receptive field size based on on multiple scale input. Wang et al. (2018) introduced a non-local operation to explore the similarity of paired points in space. Attention has shown to be able to strengthen context information in segmentation. Yu et al. (2018) introduced channel attention to select more discriminative features for semantic segmentation. Fu et al. (2019) proposed dual attention modules to capture the semantic inter-dependencies in both spatial and channel dimensions. Chen et al. (2020b) proposed SA gate for cross-modality feature aggregation and noise suppression from RGBD data. Although we have depth information from disparity estimation which makes our data similar to RGBD, the depth is computed from the RGB images and is not independently obtained.

3 METHOD

As we discussed, joint semantic and depth network requires semantic labels which are costly. In this paper, we propose to use low-level structure information for physics regularization when training the deep stereo matching model. This is also inline with the inspiration that the visual cognitive mechanism of the ventral stream simulates shallow neurons to extract low-level biologically inspired features. In deep learning for stereo matching, the feature extraction module is expected to obtain a good representation of the data. Therefore, it shall also provide information for other tasks such as low-level structure detection.

3.1 PHYSICS REGULARIZATION

We propose to integrate low-level structure with stereo matching via multi-task learning and disparity aggregation. The physics-regularization network (PRNet) includes three decoder blocks, a disparity aggregation module, a deconvolution layer and two convolutional layers, as enclosed by the dashlines in red in Figure 1. Each decoder block includes a 1×1 convolution, a 3×3 transposed convolution and a 1×1 convolution consecutively. The disparity aggregation module is inspired from our observation that large change in disparity or depth often leads to change in RGB inten-

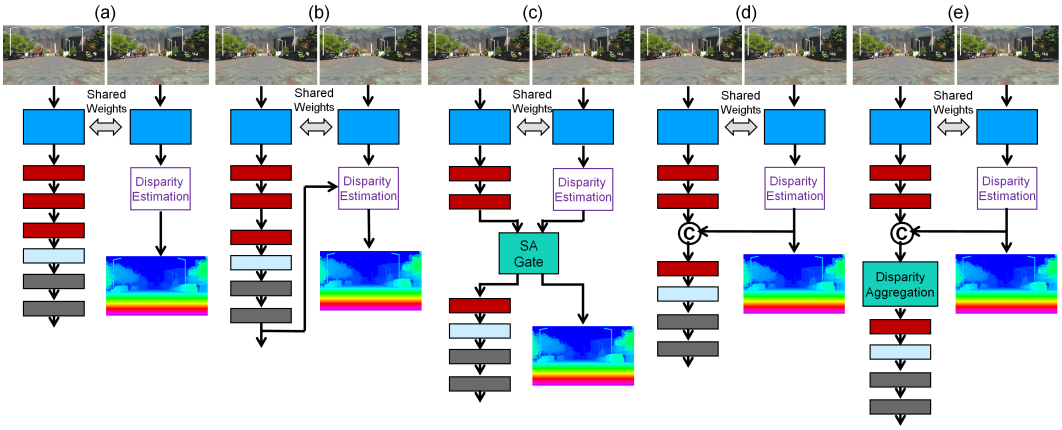

Figure 2: Comparison between proposed aggregation and other options. (a) Multi-task without aggregation (b) Edge aggregation (c) Mutual aggregation via SA gate Chen et al. (2020b) (d) Concat aggregation (e) Proposed disparity aggregation

sities. Therefore, it can be expected that the disparity output shall be helpful for low-level feature extraction. The structure of the disparity aggregation is introduced in details in Section 3.2.

The benefit of the integration is four-fold. First, the joint training of multiple tasks reduce the chance of over-fitting and often benefit each other. Second, low-level structure and depth map complement each other in nature as the edges in the depth map are often edges in RGB images. Third, the low-level structure detection module can be trained using pseudo labels from conventional edge detection without any manual annotation. Finally, it improves the model explainability and pushes the data-driven model toward physics-driven.

3.2 Disparity Aggregation

The diagram of the proposed disparity aggregation module is illustrated in Figure 1b. Given RGB features $\mathbf{f}$ and disparity $\mathbf{d}$, we first concatenate them to get $\mathbf{c} = Concat(\mathbf{f}, \mathbf{d})$. Then we define two functions to map $\mathbf{c}$ to two different spatial-wise gates $G_\mathbf{f}$ and $G_\mathbf{d}$ via two 1×1 convolution. We obtain

$$A_\mathbf{f} = G_\mathbf{f}(\mathbf{c}), A_\mathbf{d} = G_\mathbf{d}(\mathbf{c}). \tag{1}$$

A softmax function is further applied as:

$$S_\mathbf{f} = \frac{e^{A_\mathbf{f}}}{e^{A_\mathbf{f}} + e^{A_\mathbf{d}}}, S_\mathbf{d} = \frac{e^{A_\mathbf{d}}}{e^{A_\mathbf{f}} + e^{A_\mathbf{d}}}. \tag{2}$$

A merged feature m can be computed by weighted sum of the two maps:

$$\mathbf{m} = \mathbf{f} \cdot S_\mathbf{f} + \mathbf{d} \cdot S_\mathbf{d}. \tag{3}$$

We then compute μ as average of $\mathbf{m}$ and the RGB feature $\mathbf{f}$ for subsequent low-level structure detection.

$$\mu = \frac{\mathbf{f} + \mathbf{m}}{2}. \tag{4}$$

Besides the above way to aggregate the disparity estimation with image features, there are some other options. In this paper, we would investigate the different options and evaluate their performances. In Figure 2, we compare these options with the proposed disparity aggregation using diagrams. The first option is that we simply use it as multi-task without any aggregation, denoted as 'Multi-task' (see Figure 2a). The second option is to add the output of low-level structure detection into disparity, denoted as 'EA' (see Figure 2b). The third option is to have mutual aggregation between disparity and low-level structures, i.e., we add the output of the two branches into each other. In our implementation, we use the latest SA gate (Chen et al., 2020a) to achieve mutual aggregation, denoted as 'SA' (see Figure 2c). The forth option is to aggregate the disparity into low-level structure detection via simple concatenation, denoted as 'Concat' (see Figure 2d). Our experiments results in Section 4.5 show that the proposed disparity aggregation works better than other options.

3.3 LOW LEVEL STRUCTURES

One critical issue in the proposed framework is the ground truth used to train PRNet, i.e., low-level structure detection module. Using edges as additional supervision information is not a new concept. For example, Chen et al. (2016) have shown that incorporating edge detection via multi-task learning is helpful for segmentation task. However, in segmentation task where the segmentation mask is given, it is easy to compute the boundary of the mask as edge. In stereo matching, we do not have semantic labels and manual annotation of the edges is costly. In this paper, we explore a cheap way to obtain the edges. As shown in Figure 1, we compute edges using Canny operator(Canny, 1986). As the Canny edge is obtained in an unsupervised way, the PRNet can be considered as playing a role of self-supervised regularization. This is different from Chen et al. (2016) where the edge detection plays a role of filtering. Surprisingly, our results show that the Canny edge works similar to the ideal ground truth edge in improving the stereo matching in Scene Flow dataset. More details are given Section 4.5.

In this paper, we add the disparity output into the low-level structure detection module, as shown in Figure 2d. Noted that we do not add the low-level structure detection output into the disparity estimation for two reasons. First, in the current design, the PRNet can be discarded after the training. In another word, current method does not lead to any extra computation in the inference stage. Second, the color changes often appear in RGB images and they may not correspond to depth changes.

3.4 LOSS FUNCTION

As described in 3.3, we manage to obtain the supervisory edge information from the conventional Canny operator, trained in a self-supervised learning manner. To remedy the imbalanced samples in pseudo edge information, we adopt the focal loss to supervise the output of the PRNet, as shown in Equation (5):

$$F(p) = -(1-p)^{\gamma}\log(p), \tag{5}$$

where p defines the probability of the pixel being an edge pixel, and γ is empirically set as 2.

The overall loss is computed as the sum of the loss function L_o of the original stereo matching network and the new loss from PRNet, as shown in Equation (6):

$$L = L_o + \lambda \cdot F(p), \tag{6}$$

where λ controls the balance of the two items. In this paper, we use $\lambda = 0.01$ for ACVNet such that neither of the items would dominate the results.

4 EXPERIMENTAL RESULTS

4.1 DATASETS AND EVALUATION CRITERIA

In this paper, we use the following datasets to train and validate our algorithms, including Scene Flow (Mayer et al., 2016), KITTI 2012 (Geiger et al., 2012), and KITTI 2015 (Menze et al., 2015).

4.1.1 SCENE FLOW

Scene Flow (Mayer et al., 2016) is a large-scale synthetic dataset containing 35,454 training and 4,370 testing images with dimension 540×960. Following previous work, pixels with disparities larger than our limit $D = 192$ are excluded in the loss computation in our experiments. For evaluation in Scene Flow dataset, we use the commonly used end-point error (EPE) and percentage of disparity outliers $D1$ as evaluation criteria. The outliers are defined as the pixels whose disparity errors are greater than $\max(3px, 0.05d)$, where d denotes the ground-truth disparity.

4.1.2 KITTI 2012/2015

KITTI 2012 (Geiger et al., 2012) and KITTI 2015 (Menze et al., 2015) are real-world datasets with street views from driving cars. KITTI 2012 contains 194 training stereo image pairs with sparse ground truth disparities obtained using LiDAR and 195 testing image pairs. KITTI 2015 contains 200 training stereo image pairs and another 200 testing pairs. We further divided the whole training

	SceneFlow		KITTI 2012				KITTI 2015		
Method	EPE	D1	2-noc	2-all	3-noc	3-all	D1-bg	D1-fg	D1-all
RTNet	3.90	17.9	10.64	11.69	6.10	6.94	6.27	13.95	7.54
RTNet+PR	**3.27**	**15.7**	**5.92**	**6.77**	**3.72**	**4.34**	**4.41**	**10.98**	**5.50**
PSMNet	1.09	3.89	2.44	3.01	1.49	1.89	1.86	4.62	2.32
PSMNet+PR	**0.64**	**2.13**	**2.13**	**2.72**	**1.31**	**1.70**	**1.58**	**3.75**	**1.94**
GwcNet	0.76	2.71	2.16	2.71	**1.32**	**1.70**	1.74	3.93	2.11
GwcNet+PR	**0.65**	**2.23**	**2.12**	**2.69**	1.36	1.78	**1.58**	**3.69**	**1.93**
ACVNet	0.48	1.59	1.83	2.35	1.13	1.47	1.37	**3.07**	**1.65**
ACVNet+PR	**0.45**	**1.48**	**1.77**	**2.26**	**1.11**	**1.43**	**1.36**	3.09	**1.65**

Table 1: Effectiveness of physics regularization for RTNet, PSMNet, GwcNet and ACVNet in SceneFlow, KITTI2012 and KITTI2015.

data into a training set (80%) and a validation set (20%) to determine the hyper-parameters. For KITTI 2012, we report the percentage of pixels with errors larger than x disparities in both non-occluded (x-noc) and all regions (x-all). For KITTI 2015, we report $D1$ metric in background regions (bg), foreground areas (fg), and all.

4.2 Implementation Details

The proposed method was implemented using PyTorch. The training parameters are kept the same as the original network. We follow the setting in ACVNet (Xu et al., 2022). All models were trained with Adam (β_1 = 0.9, β_2 = 0.999). The maximum disparity was set to 192. The training images were randomly cropped to size 288×576. For the integration of PRNet with ACVNet, we follow the ACVNet and train the attention weights first with 64 epochs, followed by training of the entire model. For integration with other models, we trained the network end-to-end. The learning rate is set to 0.001 and decayed by half after epoch 20, 32, 40, 48 and 56. For Scene Flow, the trained model was directly used for testing. For validation in KITTI 2012 and KITTI 2015, we finetune the pretrained models for 500 epochs. The learning rate of the fine-tuning is set at 0.001 for the first 300 epochs and 0.0005 for the rest of 200 epochs. For KITTI submission, we fine-tuned the pre-trained model on the combination of KITTI 2012/2015 for 500 epochs. The learning rate of this fine-tuning began at 0.0005 for the first 300 epochs and 0.0001 for the remaining 200 epochs. The batch size was set to 8 for the training with two RTX3090 GPU cards. All the codes developed in this paper would be released upon the publication of the work.

4.3 Effectiveness of PRNet

To demonstrate the effectiveness and generality of PRNet, we integrate it with four different stereo matching networks, i.e. RTNet(Chang et al., 2020), PSMNet(Chang & Chen, 2018), GwcNet(Guo et al., 2019) and ACVNet(Xu et al., 2022), which are denoted as RTNet+PR, PSMNet+PR, GwcNet+PR, ACVNet+PR respectively. We first train the models in Scene Flow data and then finetune the models using KITTI 2012/2015 datasets. Table 1 shows all the results. As we can see from the table, the EPE in Scene Flow dataset is reduced by 16.2%, 41.3%, 14.5% and 6.3% for RTNet, PSMNet, GwcNet and ACVNet baselines respectively. In KITTI 2012 and KITTI 2015, PRNet also improves the performance in most scenarios. It is also noted that the improvement is relatively larger for approach with lightweight backbone such as RTNet. This is important as the some trade-off in performance is inevitable for practical deployment which requires high speed. The proposed method has good potential for such algorithms for real-world use. Figure 3 also show several examples for visual comparison. As we can see, the regularization improves the results with more smooth depths.

4.4 Comparison with other methods

In order to evaluate how the regularized model compares with other methods, we use the latest ACVNet (Xu et al., 2022) as the baseline and integrate it with the PRNet as a new proposed stereo matching method. Six different methods are compared, including GANet(Zhang et al., 2019a),

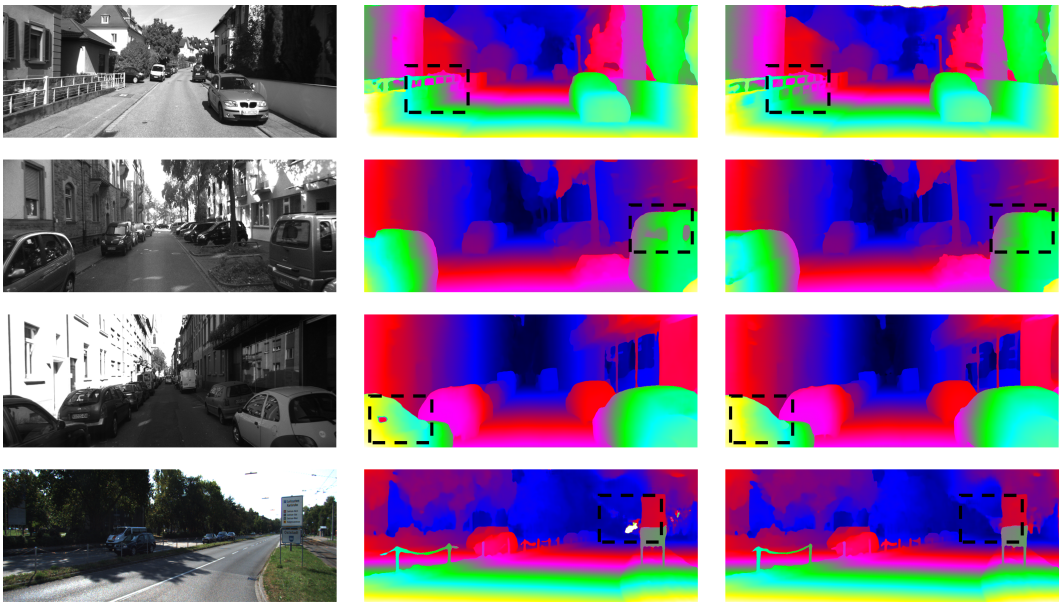

Figure 3: Sample Results. From left to right are (a) RGB image (b) disparity estimated by original network (c) disparity estimated by physics regularized network. From to bottom, the first two rows are based on GwcNet architecture, and the last two rows are based on ACVNet architecture.

	KITTI 2012				KITTI 2015			
Method	2-noc	2-all	3-noc	3-all	D1-bg	D1-fg	D1-all	run-time
GANet (2019)	1.89	2.50	1.19	1.60	1.48	3.46	1.81	1.8
AcfNet (2020)	1.83	2.35	1.17	1.54	1.51	3.80	1.89	0.48
HITNet (2021)	2.00	2.65	1.41	1.89	1.74	3.20	1.98	0.02
CFNet (2021)	1.90	2.43	1.23	1.58	1.54	3.56	1.88	0.18
LEAStereo (2020)	1.90	2.39	1.13	1.45	1.40	**2.91**	**1.65**	0.3
ACVNet(2022)	1.83	2.35	1.13	1.47	1.37	3.07	**1.65**	0.2
Proposed	**1.77**	**2.26**	**1.11**	**1.43**	**1.36**	3.09	**1.65**	0.2

Table 2: Comparison of proposed method (ACVNet+PR) vs. other methods in KITTI 2012 and KITTI 2015.

AcfNet(Zhang et al., 2019d), HITNet(Tankovich et al., 2021), CFNet(Shen et al., 2021), and LEAStereo(Cheng et al., 2020) and ACVNet(Xu et al., 2022). We focus on the real-world datasets and compare these methods in KITTI 2012 and KITTI 2015. Table 2 summarizes the results. As we can see, the proposed method outperforms these methods in most scenarios.

4.5 Ablation Studies

To validate the effectiveness of each component proposed PRNet, we conducted the following ablation studies on Scene Flow dataset. The baseline approach is the original stereo matching network, denoted as 'Baseline'. To validate the effectiveness of the proposed components, we validate them with four different stereo matching algorithms, including RTNet, PSMNet, GwcNet and ACVNet. We compare the baseline approach with 'Multi-task' to justify the benefit of low-level structure regularization. To justify the use of disparity aggregation, we compare the proposed disparity aggregation with 'EA', 'SA' and 'Concat'. The results from the ablation studies are shown in Table 3.

Method	RTNet	PSMNet	GwcNet	ACVNet
Baseline	3.90	1.09	0.76	0.48
Multi-task	3.38	0.70	0.67	0.46
Concat	3.32	0.68	0.66	0.46
EA	3.29	0.70	0.67	0.47
SA	3.35	0.75	0.70	0.50
Proposed	**3.20**	**0.65**	**0.64**	**0.45**

Table 3: Ablation study for PRNet in Scene Flow dataset

Method	RTNet	PSMNet	GwcNet	ACVNet
Canny Edge	3.265	0.644	0.650	0.449
Ground Truth Edge	3.200	0.649	0.641	0.448

Table 4: Comparison of regularization using ground truth edge vs. Canny edge as labels

As shown in Table 3, both the multi-task strategy and the disparity aggregation improve the performance of the stereo matching. Moreover, as we do not use the output of PRNet, the regularization can be discarded in the inference stage. Therefore, the proposed PRNet does not change the actual network architecture for disparity estimation. Therefore, it does not lead to extra computational cost in inference. 'EA' does not perform better than the proposed disparity aggregation, this is inline with our intuition that edge of disparity map is often the edge of RBG images but not the vice versa. Although 'SA' has shown to be promising for RGBD data, our experimental results show that it is not optimal for stereo matching. Its performance is slightly lower than that of simple multi-task without aggregation. This might be because SA is originally proposed for multi-modal aggregation where the RGB and the depth are obtained independently. However, our depth is estimated from RGB images.

We also investigate how the Canny operator affects the results. We compare the performances when we use the Canny edge as ground truth versus the case when the ideal ground truth is available. From the Scene Flow dataset, the object segmentation is available. Although the object segmentation labels are randomly assigned without consistency and cannot be used as semantic labels, they can be used to compute edges. We use these edges as the ideal ground truth edges. Table 4 compares the results. The performances of disparity estimation by the two different labels as ground truth are comparable. Although the idea to use Canny edge to replace ideal ground truth edge as training label is simple, the results are significant as Canny edge can be obtained efficiently without manual annotations.

5 Conclusions

The training of stereo matching network has been a challenging issue. In this paper, we propose a novel physics regularization framework, which can be applied on existing stereo matching algorithms to improve their performances. By combining the physics regularization with RTNet, PSMNet, GwcNet and ACVNet, we improve the performance of original stereo matching network consistently in Scene Flow, KITTI 2012 and KITTI 2015 datasets. Moreover, our experiments also show that the use of disparity estimation is better than other options of aggregation. We also find that by using pseudo ground truth edges extracted by Canny operator, we achieve comparable performances in stereo matching compared with those using ideal ground truth. This is significant as the Canny edge can be obtained automatically without manual annotation. A limitation is that this is only validated in the Scene Flow dataset which is a synthetic dataset instead of a real-world dataset. A further study on real-world data sets would be helpful. This could be investigated by improving the edge detection algorithms.

REFERENCES

Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars M. Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision*, 126:961–972, 2018.

John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. doi: 10.1109/TPAMI.1986.4767851.

Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418, 2018. doi: 10.1109/CVPR.2018.00567.

Jia-Ren Chang, Pei-Chun Chang, and Yong-Sheng Chen. Attention-aware feature aggregation for real-time stereo matching on edge devices. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

Liang-Chieh Chen, Jonathan T. Barron, George Papandreou, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4545–4554, 2016. doi: 10.1109/CVPR.2016.492.

Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *ECCV*, 2020a.

Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2020b.

Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Tom Drummond, Hongdong Li, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Pier Luigi Dovesi, Matteo Poggi, Lorenzo Andraghetti, Miquel Martí, Hedvig Kjellström, Alessandro Pieropan, and Stefano Mattoccia. Real-time semantic stereo matching. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10780–10787, 2020. doi: 10.1109/ICRA40945.2020.9196784.

Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4383–4392, 2019. doi: 10.1109/ICCV.2019.00448.

Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3141–3149, 2019. doi: 10.1109/CVPR.2019.00326.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3268–3277, 2019. doi: 10.1109/CVPR.2019.00339.

Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pp. 807–814 vol. 2, 2005. doi: 10.1109/CVPR.2005.56.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 66–75, 2017. doi: 10.1109/ICCV.2017.17.

Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 510–519, 2019. doi: 10.1109/CVPR.2019.00060.

Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers, 2021.

Michele Mancini, Gabriele Costante, Paolo Valigi, and Thomas A. Ciarfuglia. Fast robust monocular depth estimation for obstacle detection with fully convolutional networks. *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4296–4303, 2016.

Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015.

Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann, and Feng Liu. *Depth Conflict Reduction for Stereo VR Video Interfaces*, pp. 1–9. Association for Computing Machinery, New York, NY, USA, 2018. ISBN 9781450356206.

Ee Ping Ong, Jun Cheng, Damon W.K. Wong, Elton L. T. Tay, Hwei Yee Teo, Rosalyn Grace Loo, and Leonard W.L. Yip. Automatic glaucoma detection from stereo fundus images. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pp. 1540–1543, 2020. doi: 10.1109/EMBC44109.2020.9175923.

Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, S. Mattoccia, and Luigi di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. In *ACCV*, 2018.

Ashutosh Saxena, Jamie Schulte, and A. Ng. Depth estimation using monocular and stereo cues. In *IJCAI*, 2007.

Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47:7–42, 2002.

Akihito Seki and Marc Pollefeys. Sgm-nets: Semi-global matching with neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6640–6649, 2017.

Amit Shaked and Lior Wolf. Improved stereo matching with constant highway networks and reflective confidence learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6901–6910, 2017.

Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13906–13915, June 2021.

Irwin Sobel and Gary Feldman. A 3x3 isotropic gradient operator for image processing. *Pattern Classification and Scene Analysis*, pp. 271–272, 1973.

Xiao Song, Xu Zhao, Liangji Fang, Hanwen Hu, and Yizhou Yu. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. *International Journal of Computer Vision*, 128:910–930, 2020.

Vladimir Tankovich, Christian Häne, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14357–14367, 2021. doi: 10.1109/CVPR46437.2021.01413.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018. doi: 10.1109/CVPR.2018.00813.

Yan Wang, Zihang Lai, Gao Huang, Brian H. Wang, Laurens van der Maaten, Mark E. Campbell, and Kilian Q. Weinberger. Anytime stereo image depth estimation on mobile devices. *2019 International Conference on Robotics and Automation (ICRA)*, pp. 5893–5900, 2019.

Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Song Wang, and Lili Ju. Semantic stereo matching with pyramid cost volumes. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7483–7492, 2019. doi: 10.1109/ICCV.2019.00758.

Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12981–12990, 2022.

Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1956–1965, 2020. doi: 10.1109/CVPR42600.2020.00203.

Xiaowei Yang, Zhiguo Feng, Yong Zhao, Guiying Zhang, and Lin He. Edge supervision and multi-scale cost volume for stereo matching. *Image and Vision Computing*, 117:104336, 2022. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2021.104336. URL `https://www.sciencedirect.com/science/article/pii/S0262885621002419`.

Menglong Ye, Edward Johns, Ankur Handa, Lin Zhang, Philip Pratt, and Guang-Zhong Yang. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. *ArXiv*, abs/1705.08260, 2017.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1857–1866, 2018. doi: 10.1109/CVPR.2018.00199.

Chi Zhang, Zhiwei Li, Yanhua Cheng, Rui Cai, Hongyang Chao, and Yong Rui. Meshstereo: A global stereo model with mesh alignment regularization for view interpolation. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2057–2065, 2015. doi: 10.1109/ICCV.2015.238.

Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 185–194, 2019a.

Junming Zhang, Katherine A. Skinner, Ram Vasudevan, and Matthew Johnson-Roberson. Dispsegnet: Leveraging semantics for end-to-end learning of disparity estimation from stereo imagery. *IEEE Robotics and Automation Letters*, 4:1162–1169, 2019b.

Tianyang Zhang, Huazhu Fu, Yitian Zhao, Jun Cheng, Mengjie Guo, Zaiwang Gu, Bing Yang, Yuting Xiao, Shenghua Gao, and Jiang Liu. Skrgan: Sketching-rendering unconditional generative adversarial networks for medical image synthesis. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 777–785, Cham, 2019c. Springer International Publishing.

Youmin Zhang, Yimin Chen, Xiao Bai, Jun Zhou, Kun Yu, Zhiwei Li, and Kuiyuan Yang. Adaptive unimodal cost volume filtering for deep stereo matching. *arXiv preprint arXiv:1909.03751*, 2019d.

Kang Zhou, Jing Li, Yuting Xiao, Jianlong Yang, Jun Cheng, Wen Liu, Weixin Luo, Jiang Liu, and Shenghua Gao. Memorizing structure-texture correspondence for image anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021. doi: 10.1109/TNNLS.2021.3101403.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6619, 2017. doi: 10.1109/CVPR.2017.700.