Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/67/20240653

SGANnet: Super-resolution guided asymmetric stereo matching network

Jingyao Bao[1,3], Hongfei Yu[2,4,*], Yongjia Zou[3,5], Yang Cao[6]

[1]School of Artificial Intelligence and Software, Liaoning Shihua University, Fushun, China
[2]School of Information and Control Engineering, Liaoning Shihua University, Fushun, China

[3]779822989@qq.com
[4]yuhfln@163.com
[5]yzou406@gmail.com
[6]caoyang821313@163.com
*corresponding author

Abstract. With asymmetric resolution stereo images as input, existing stereo matching algorithms significantly decline in prediction performance. To address this, we introduce SGANet (Super-resolution Guided Asymmetric Stereo Matching Network), a model that employs unsupervised training methods to overcome the difficulty of acquiring ground truth disparity. For the lower resolution side, this paper designs a stereo guided super-resolution module (SGSR), where the network generates a super-resolved image enriched with details guided by the higher resolution side. Additionally, for this module, we propose a feature consistency loss based on the image's feature space to measure the similarity between the real and super-resolved images. Experimental results on the autonomous driving dataset KITTI demonstrate the effectiveness of the SGSR module and the feature consistency loss in improving the disparity prediction performance of asymmetric resolution stereo images.

Keywords: Stereo matching, Self-supervised, Asymmetric, Super-resolution, Autonomous driving.

1. Introduction

Binocular stereo vision has emerged as a hot research topic within the field of computer vision and has been widely applied in numerous areas such as autonomous driving and intelligent robotics. Utilizing visual information to estimate depth for road scene reconstruction presents a promising future alternative to expensive radar equipment. However, the currently mature supervised methods still rely on radar to obtain ground truth labels, and the performance of unsupervised binocular disparity estimation models lags significantly behind supervised models, leaving substantial room for development.

In various applications such as robotics, smartphones, and autonomous vehicles, multi-camera systems are a common configuration. These systems typically consist of camera modules with differing focal lengths and imaging qualities. Most stereo matching research focuses on predicting disparity from two consistent views. When input images are asymmetric, even if the same target object is imaged in

DOI: 10.54254/2755-2721/67/20240653

stereo views, this asymmetry makes it difficult for models to correctly match features, leading to significant declines in stereo matching model performance. When the inconsistency in stereo images reaches a higher degree, the models may even fail completely. Our experiments have confirmed this phenomenon and analyzed the impact of different magnifications of monocular blurring on the performance of stereo matching models. Long-short focal length camera modules are a common configuration in multi-camera systems, hence we assume different focal lengths for stereo cameras, meaning the same scene would be captured at different resolutions. In experiments, we downsampled the original images to simulate low-resolution images taken with short focal lenses, and used bilinear interpolation to upsample them to the same size. These asymmetric stereo images were then input into a stereo matching model, as shown in Figure 1, with the left side being the high-resolution left view and the right side being an 8x blurred right view. We tested with the advanced disparity prediction network CREStereo [1], and the results are shown in Figure 2, with (a) showing the symmetric original resolution disparity map, (b) the disparity results with a 2x blurred right view, (c) with a 4x blurred right view, and (d) with an 8x blurred right view. The visualized disparity results show that when monocular blurring is minor, it has a small impact on the performance of the stereo matching model. As the image quality gap between stereo views increases, the accuracy of the model significantly decreases and may even fail to predict.

Figure 1. Asymmetric resolution stereo images

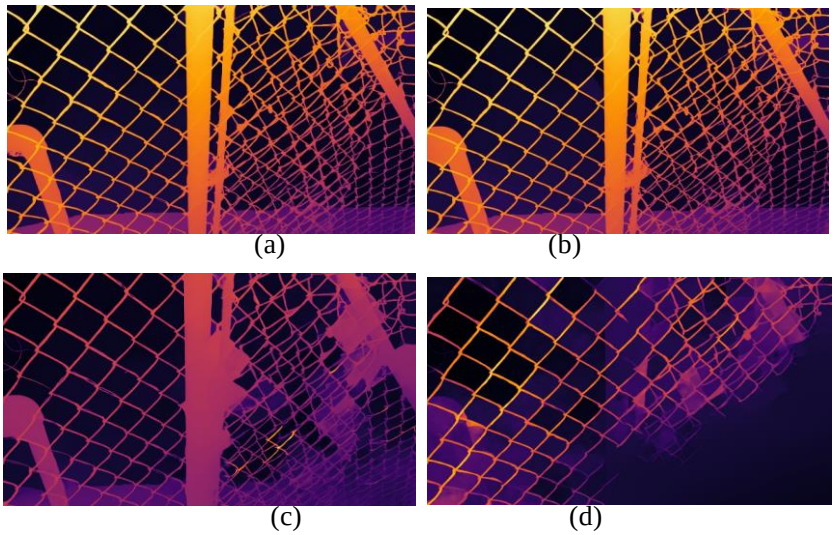

Figure 2. Visualization results of asymmetric resolution stereo matching

To address this challenge, we consider improving the binocular disparity estimation algorithms under the condition of using asymmetric resolution stereo images as input, to make them more broadly applicable in real-world scenarios. Furthermore, considering the difficulty and high cost of acquiring depth truth required by supervised models, we propose an asymmetric resolution self-supervised stereo matching framework to implement label-free training of the network. Specifically, for the lower resolution side, we employ advanced image super-resolution technology to obtain high-quality images.

DOI: 10.54254/2755-2721/67/20240653

To incorporate the rich detail information from the high-resolution side, we designed a stereo-guided super-resolution module (SGSR) to generate higher quality super-resolved images. Additionally, for this module, we did not use the common super-resolution image similarity loss, but rather proposed a feature consistency loss based on the feature space, to measure the similarity between the features of real and super-resolved images. Our method is capable of extracting rich detail information from low-resolution images and achieving higher consistency with features from the other view, thereby laying a solid foundation for subsequent prediction processes and solving the problem of disparity prediction failure in asymmetric stereo images.

2. Related Work

Current end-to-end disparity regression networks have achieved leading performance, with all steps of the traditional stereo matching algorithm process capable of being jointly trained within the same network. DispnetC [2] was the first end-to-end stereo matching network, capable of directly predicting disparity maps after inputting stereo images. GC-net [3] innovatively introduced the concept of a cost volume in the computation of matching costs, constructing a 3D cost volume by calculating the correlation between two pixel vectors. However, using the inner product to calculate correlation results in the loss of substantial spatial information, whereas PSMNet [4] does not compute but directly concatenates the left and right feature maps to build a 4D cost volume.

Supervised approaches have achieved significant research outcomes, yet the difficulty of obtaining ground truth labels has limited the development of depth estimation techniques. In recent years, unsupervised methods based on spatial transformations have made significant advancements. PASMnet [5] proposed a general disparity attention mechanism to learn the stereo matching relationships of image pairs with large disparity variations in an unsupervised manner. Zhou et al. [6] used the consistency constraint of left-right images to iteratively update network parameters. Godard et al. [7] introduced a new loss function that warps the predicted left and right disparity maps, enforcing consistency between the disparity maps.

Regarding the disparity estimation of asymmetric stereo images, Liu et al. [8] first considered unbalanced stereo matching and, to address the problem of stereo collapse caused by excessive differences, designed a guided view synthesis framework to restore corrupted views. Chen et al. [9] proposed feature metric consistency to address the situations where asymmetric self-supervised stereo matching would violate the photometric consistency assumption. Building on this, Song et al. [10] introduced spatially adaptive self-similarity, further enhancing feature consistency in loss computation.

3. Our Method

This paper focuses on stereo matching research for multi-camera systems commonly comprised of long-short focal length combinations, aiming to address the decline in prediction performance when asymmetric resolution stereo images are input. In this section, we detail our proposed Super-resolution Guided Asymmetric Stereo Matching Network (SGANet), starting with an overview of the model's structure, followed by descriptions of the stereo-guided super-resolution module and the super-resolution feature consistency loss function.

3.1. Model Overview

For disparity estimation of asymmetric resolution stereo images, we designed a self-supervised stereo matching network, SGANet, the overall structure of which is shown in Figure 3. Broadly, the model comprises two parts: an image quality restoration phase handled by the stereo-guided super-resolution module (SGSR) and a self-supervised stereo matching phase. Our approach achieves end-to-end super-resolution and disparity prediction; it inputs a pair of asymmetric resolution stereo images and outputs a predicted disparity map.

In the super-resolution component, which includes a stereo-guided super-resolution module (SGSR), this combines the detail information from the high-resolution side with the global structure information

DOI: 10.54254/2755-2721/67/20240653

from the low-resolution side to produce high-quality images, thus narrowing the feature space gap between the stereo images, detailed further in section 3.2.

In the disparity estimation component, the high-resolution left view I_L^H and the super–resolved right view I_R^{SR} are input. A shared-weight feature extractor Φ_F extracts features F_L and F_R^{SR} from the left and right views, respectively, and then constructs a 4D cost space(H×W×C×D). After passing through the cost aggregation network Φ_M, a predicted disparity map d is produced. This paper implements self-supervised training based on disparity reconstruction, using the disparity values to reproject each pixel in the super-resolved right image I_R^{SR} to obtain a reconstructed left image $I_{R\to L}^{SR}$ as shown in Equation (1). In the ideal scenario (zero error in predicted disparity), the reconstructed image should be consistent with the actual left image, thus minimizing their difference can constrain the network to generate more precise disparities.

$$I_{R\to L}^{SR} = Warp\left(d, I_R^{SR}\right) \tag{1}$$

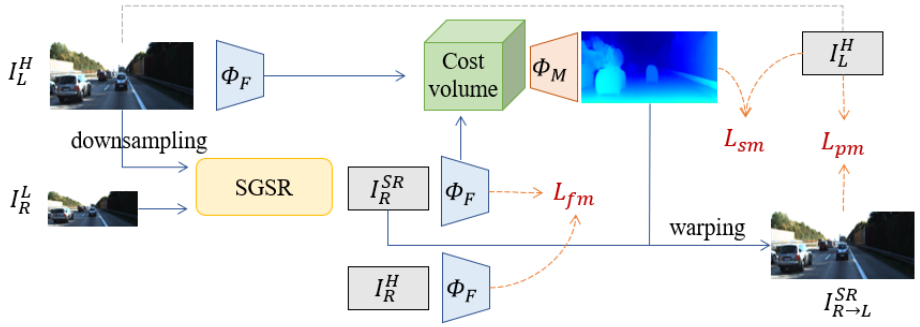

Figure 3. Overall structure of the Super-resolution Guided Asymmetric Stereo Matching Network (SGANet)

3.2. Stereo-Guided Super-Resolution

As mentioned earlier, asymmetric inputs can lead to significant degradation or even failure of disparity prediction models, even if the same imaging area is involved. Discrepancies in features extracted from images of inconsistent quality, as well as a lack of detail in low-resolution images, are among the factors that contribute to blurred similarity matching during the cost aggregation phase. The most direct method to ensure the imaging quality consistency of both views is to enhance the image quality of the lower resolution side. We consider employing advanced image super-resolution technology, inputting a low-resolution image and outputting a high-resolution image.

While single-image super-resolution technology continues to evolve and achieve good research results, it cannot utilize the complementary information from different views in stereo images, which greatly limits the performance of super-resolution models. Therefore, for tasks predicting disparity from asymmetric resolution stereo images, it is more ideal to perform super-resolution based on the global structural information of the low-resolution image while incorporating rich detail information from the high-resolution image. In our research, referencing the work of [11], we designed a stereo-guided super-resolution (SGSR) module to restore the quality and detail information of the low-resolution image.

As shown in Figure 4, this module consists of feature extraction, cross-view block (CVB), and spatial perception module (SPM). Inputting the downsampled high-resolution left image $I_L^{\downarrow}$ and low-resolution right image I_R^L, after feature extraction through a 3×3 convolution and two residual blocks, the cross-view block effectively learns global and local feature correlations, enhancing the representation of similar patterns in the image. Subsequently, the integrated features F_{CVB} are further extracted by a cascading spatial perception module SPM to provide hierarchical feature representations F_{SPM}, and the cascaded features are finally upsampled using PixelShuffle [12] to the same size as the high-resolution side, resulting in the super-resolved right image I_R^{SR}.

DOI: 10.54254/2755-2721/67/20240653

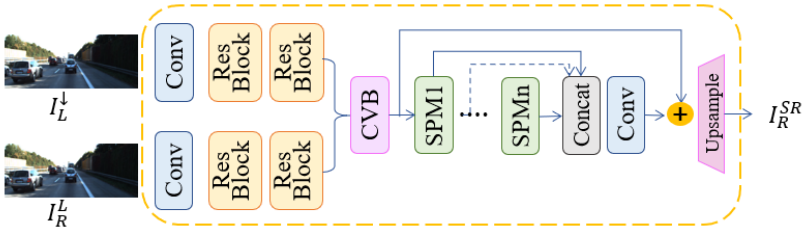

Figure 4. Stereo Guided Super-resolution Module (SGSR)

3.3. Super-resolution Feature Consistency Loss

To overcome the challenges and high costs associated with obtaining ground truth via LiDAR, we have constructed an unlabeled self-supervised stereo matching network, SGSRnet. Specifically, this model is based on the assumption of photometric consistency, where the reconstructed image and the real image should be consistent, typically formed by a weighted combination of L1 and SSIM distances to constitute the photometric loss function:

$$L_{pm} = (1-\alpha)\left\|I_L^H - I_{R\to L}^{SR}\right\|_1 + \alpha\left(1 - SSIM(I_L^H, I_{R\to L}^{SR})\right) \tag{2}$$

For super-resolution neural networks, the similarity between the super-resolved image and the actual high-resolution image is usually used as the loss function during training. In the asymmetric resolution disparity estimation studied in this paper, we utilize an SGSR module to restore the quality of low-resolution images. However, even the most advanced methods cannot completely accurately restore to high-resolution, and some inconsistency in the feature expression space of stereo images will still exist. In this paper, our ultimate goal is to predict accurate disparities rather than image super-resolution, making the consistency of extracted image features more important than the similarity of the images themselves. To enable the SGSR module to generate feature-consistent super-resolved images, we do not use image similarity as the loss; instead, we propose a super-resolution feature loss, calculating the L2 loss between the features of the super-resolved image and the real high-resolution image features:

$$L_{fm} = \left\|F_R^H - F_R^{SR}\right\|_2 \tag{3}$$

Additionally, to achieve a smoother disparity, we also incorporate an edge-aware smoothness loss [6], which encourages local smoothness in disparity by applying an L1 penalty to the gradient of disparity:

$$L_{sm} = |\partial_x^2 D_L^*|\mathrm{e}^{-\gamma|\partial_x I_L|} + |\partial_x^2 D_L^*|\mathrm{e}^{-\gamma|\partial_x I_L|} \tag{4}$$

Finally, our method's total loss is:

$$L_{final} = \lambda_{pm}L_{pm} + \lambda_{fm}L_{fm} + \lambda_{sm}L_{sm} \tag{5}$$

4. Experiments and Analysis

This section conducts experiments on the proposed method and provides comparative and analytical results. It details the datasets used, the methods for dataset generation, experimental setups, model parameter configurations, and includes a performance comparison and ablation study of the network.

4.1. Experimental Setup

Dataset: In this study, we used publicly available autonomous driving stereo image datasets, KITTI 2012 [13] and KITTI 2015 [14]. Each contains 198 and 200 pairs of training images with sparse disparity ground truth labels obtained from LiDAR, as well as 198 and 200 pairs of unlabeled test images, respectively. We follow the scheme from [9] and use the test dataset for network training and the train dataset for model performance evaluation. To simulate asymmetric resolution stereo image pairs, we

DOI: 10.54254/2755-2721/67/20240653

replaced the right images in the original stereo pairs with images downsampled by a factor of four, forming our training and test sets.

Evaluation Metrics: To quantitatively measure stereo matching performance, we referred to work in the field of asymmetric stereo matching [10], using End-point Error (EPE) and Three-pixel Error (3PE) as model evaluation metrics, as shown in Equations (6) and (7).

Given the total number of pixels N in the test set, the estimated disparity $\hat{d}_i$ and the ground truth disparity d_i for pixel i. EPE calculates the average Euclidean distance between the predicted disparity values and the true disparity values for all pixels, with smaller errors indicating higher matching accuracy. 3PE refers to the proportion of pixels whose absolute difference between the predicted disparity and the true disparity exceeds three pixels, with a higher proportion indicating more mismatched points and lower matching accuracy.

$$EPE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left(\hat{d}_i - d_i\right)^2} \tag{6}$$

$$3PE = \frac{1}{N}\sum_{i=1}^{N}\Phi(\left|\hat{d}_i - d_i\right|, 3), \quad \Phi(p,q) = \begin{cases}1, p > q\\ 0, p \le q\end{cases} \tag{7}$$

Implementation Details: The research was conducted using the deep learning framework Pytorch on a Windows 11 operating system. The CPU used was a 12th Gen Intel(R) Core(TM) i9-12900K 3.20 GHz, with 32GB RAM, and a GPU model GeForce RTX 3090 with 24G of VRAM. During training, we used the Adam gradient optimization algorithm with $\beta 1$ = 0.9 and $\beta 2$ = 0.999, and an initial learning rate of 0.0001. To accelerate model convergence, we adopted a phased training strategy; the initial SGSR module was pre-trained using an image L1 loss, with a batch size of 16 for 20 epochs, reducing the learning rate to 0.5 of its original at the 8th, 12th, 15th, and 17th epochs. For the disparity prediction part, we used bilinear interpolation to upsample the low-resolution side to the same dimensions as the network input, with a batch size of 8 for 20 epochs, reducing the learning rate at the same epochs as above. Finally, we refined the overall model training with a batch size of 4 for 20 epochs, again reducing the learning rate at the 8th, 12th, 15th, and 17th epochs. The parameter values in the model were: $\alpha = 0.85$, $\gamma = 10$, $\lambda_{pm} = 1.0$, $\lambda_{fm} = 0.5$, $\lambda_{sm} = 0.5$.

4.2. Model Analysis

To validate the effectiveness of the proposed SGSR module in restoring image quality, we conducted tests on the KITTI 2015 dataset. Figure 5 presents two sets of visual comparison results: the first row shows the low-resolution RGB images, the second row displays the results of image quality restoration using the SGSR module, the third row presents the results of directly inputting the first row images into our self-supervised stereo matching network, and the fourth row shows the disparity maps predicted by the super-resolution guided asymmetric stereo matching network. By comparing the disparity maps in rows three and four, it is evident that our stereo-guided super-resolution module significantly improves the performance of the stereo matching model for asymmetric resolution stereo images. Comparing the RGB images in rows one and two demonstrates that our method effectively enhances image quality. Since we did not perform image supervision on the output of the SGSR module, some noise pixels appear in low-texture, high-exposure areas of the image. However, the primary objective of this paper is not to obtain precise super-resolved images but to focus on improving disparity prediction accuracy. The experimental results fully demonstrate that the method proposed in this paper effectively resolves the performance degradation caused by asymmetric inputs.

DOI: 10.54254/2755-2721/67/20240653

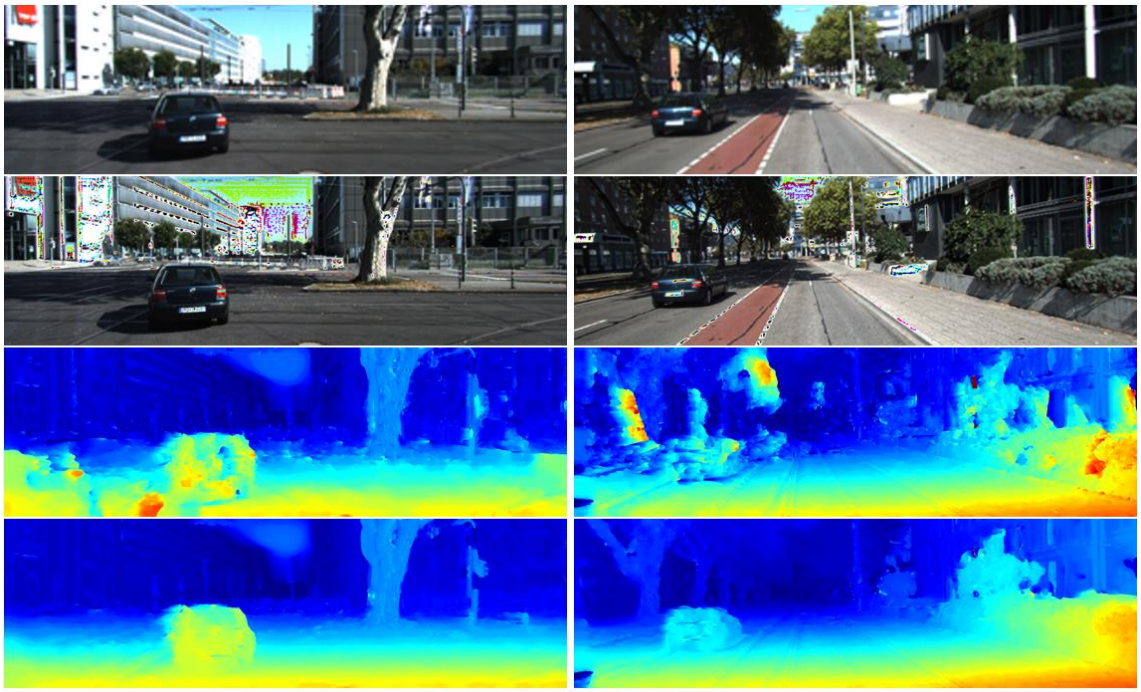

Figure 5. Visualization of Image Quality Restoration and Predicted Disparity Results

Regarding the selection of a super-resolution feature consistency loss function, we experimented with different computational methods. To analyze the impact of different loss functions on the performance of the stereo matching model, we selected several classic loss functions, namely L1, SSIM [42], and L2, and used them as the super-resolution feature consistency loss function to retrain our model. Additionally, we provided the results of models trained using image photometric loss (*PM_img*) as a control. As shown in Table 1, compared to *PM_img*, all three feature loss models yielded superior results, demonstrating that for stereo matching models using asymmetric images as inputs, super-resolution feature consistency constraints are more conducive to predicting accurate disparities. Among the three different feature loss functions, L2 demonstrated better prediction accuracy. To ensure optimal model performance, we selected L2 as the super-resolution feature consistency loss function for this paper.

Table 1. Impact of Different Loss Functions on Model Performance

	KITTI15		KITTI12	
Loss	**EPE**	**3PE**	**EPE**	**3PE**
PM_img	2.613	16.391	3.100	16.772
L1_fea	2.397	14.372	2.916	16.290
SSIM_fea	2.413	14.139	2.911	15.897
L2_fea	**2.387**	**12.324**	**2.697**	**11.927**

4.3. Ablation Study

To verify the effectiveness of our method for disparity prediction tasks on asymmetric resolution stereo images, we conducted ablation studies on the proposed SGSR module and the super-resolution feature consistency loss. (1) Baseline: Replacing the SGSR module with bilinear interpolation upsampling. (2) +SGSR: Adding the SGSR and using super-resolution image photometric loss. (3) +L_fm: Replacing the super-resolution image photometric loss with the super-resolution feature consistency loss proposed in this paper. The results of the ablation experiments are shown in Table 2. By comparing results (1) and

DOI: 10.54254/2755-2721/67/20240653

(2), the +SGSR model performed better on both datasets than the Baseline, thus demonstrating that the stereo-guided super-resolution module proposed in this paper can effectively restore image quality and significantly enhance stereo matching performance. Comparing (2) and (3), replacing the SGSR's super-resolution image photometric loss with super-resolution feature consistency loss allows the SGSR module to better learn how to restore images to achieve feature consistency, which is beneficial for cost aggregation in asymmetric stereo images, thus better predicting the disparity of each pixel in the image.

Table 2. Ablation Experiment Results Comparison

	KITTI15		KITTI12	
Method	**EPE**	**3PE**	**EPE**	**3PE**
Baseline	2.919	20.889	3.426	20.631
+SGSR	2.613	16.391	3.100	16.772
$+L_{fm}$	**2.387**	**12.324**	**2.697**	**11.927**

4.4. Comparison with Other Methods

Due to the scarcity of related research and the lack of open-source projects, we conducted the following comparative experiments: (1) *CVCnet+Baseline_sym*: Feeding CVCnet super-resolution results into a symmetric dataset trained version of the Baseline network; (2) Directly upsampling low-resolution images and using a self-supervised photometric loss trained stereo matching model; (3) *DAUS*: According to the original paper, retraining the model by replacing the photometric loss in the *Baseline* with feature metric consistency loss; (3) *Ours*: Our method. As shown in Table 3, our method achieved the best quantitative metrics for asymmetric resolution stereo disparity estimation.

Table 3. Comparison with Other Methods

	KITTI15		KITTI12	
Method	**EPE**	**3PE**	**EPE**	**3PE**
CVCnet[11]+ *Baseline_sym*	3.090	19.048	3.519	18.738
Baseline	2.919	20.889	3.426	20.631
DAUS [9]	2.702	19.973	2.904	18.671
Ours	**2.387**	**12.324**	**2.697**	**11.927**

Additionally, we also present a visual comparison of disparity prediction results with other methods. As shown in Figure 6, two sets of results are provided, from top to bottom representing: RGB images, and disparity maps predicted using *Ours, DAUS* [9], and *CVCnet* [11] +*Baseline_sym* methods, respectively. Through the comparison of the visualization results, it can be more intuitively shown that for asymmetric resolution stereo matching, our proposed method provides significant improvements in model prediction performance.

DOI: 10.54254/2755-2721/67/20240653

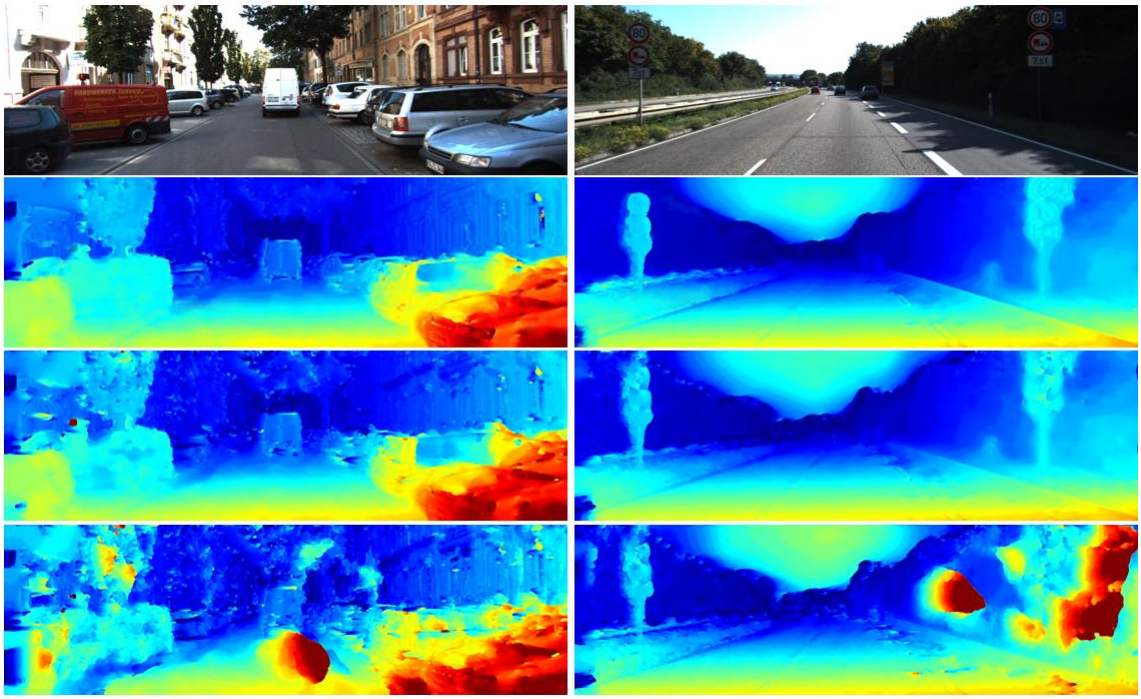

Figure 6. Comparison of Disparity Maps Predicted by Different Methods

5. Conclusion

This paper investigates stereo matching for asymmetric resolution stereo images. We proposed a stereo-guided super-resolution (SGSR) module and a super-resolution feature consistency loss to narrow the quality disparity between stereo image pairs and have empirically proven the effectiveness of both the module and the loss function. However, there are still shortcomings in this work. This paper only focuses on inconsistent image resolutions, while in reality, stereo image pairs may also exhibit inconsistencies in noise, color, and other aspects. In the future, we will consider a more diverse range of asymmetric conditions to make the model applicable to a broader array of scenarios. Moreover, although our model can achieve self-supervised training during the stereo matching phase, the SGSR module still requires supervision from original high-resolution images. Image degradation due to inconsistencies can adversely affect model accuracy, which is detrimental for practical applications. We will further improve the model structure and supervision methods to enhance its generalizability.

Funding Information

1. National Natural Science Foundation of China, Grant/Award Numbers: 61702247
2. Department of Education of Liaoning Province, Grant/Award Numbers: LJKMZ20220723, LJKMZ20220754

References

[1] Li, J., Wang, P., Xiong, P., Cai, T., et al.: Practical stereo matching via cascaded recurrent network with adaptive Correlation. IEEE Conference on Computer Vision and Pattern Recognition, 16263-16272 (2022)

[2] Mayer, N., Ilg, E., Hausser, P., et al.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. IEEE Conference on Computer Vision and Pattern Recognition, 4040-4048 (2016)

[3] Kendall, A., Martirosyan, H., Dasgupta, S., et al.: End-to-end learning of geometry and context for deep stereo regression. IEEE International Conference on Computer Vision, 66-75 (2017)

DOI: 10.54254/2755-2721/67/20240653

[4] Chang, J. R., Chen, Y. S.: Pyramid stereo matching network. IEEE Conference on Computer Vision and Pattern Recognition, 5410- 5418 (2018)
[5] Wang, L.G., Guo Y.L., et al.: Parallax attention for unsupervised stereo correspondence learning. IEEE Trans. Pattern Anal. Mach. Intell., early access, Sep. 25, (2020)
[6] Zhou, T., Brown, M., Snavely, N., et al.: Unsupervised learning of depth and ego-motion from video. IEEE Conference on Computer Vision and Pattern Recognition, 1851-1858 (2017)
[7] Godard, C., O. Aodha M, Brostow, G. J.: Unsupervised monocular depth estimation with left-right consistency. IEEE Conference on Computer Vision and Pattern Recognition, 270-279 (2017)
[8] Liu, Y. C., Ren, J., Zhang, J. W., et al.: Visually imbalanced stereo matching. IEEE Conference on Computer Vision and Pattern Recognition, 2029–2038 (2020)
[9] Chen, X. H., Xiong, Z. W., Cheng, Z. , et al.: Degradation-agnostic correspondence from resolution-asymmetric stereo. IEEE Conference on Computer Vision and Pattern Recognition, 12962–12971 (2022)
[10] Song, T., Kim, S., and Sohn, K.: Unsupervised Deep Asymmetric Stereo Matching with Spatially-Adaptive Self-Similarity. IEEE Conference on Computer Vision and Pattern Recognition, 13672-13680 (2023)
[11] Zhu, X. Y., Guo, K. H., Fang, H. , et al.: Cross view capture for stereo image super- resolution. IEEE Transactions on Multimedia, (2021)
[12] Wang, X., Girshick, R., Gupta, A., and He, K. : Non-local neural net-works. IEEE Conference on Computer Vision and Pattern Recognition, 7794–7803 (2018)
[13] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. IEEE Conference on Computer Vision and Pattern Recognition, 3354–3361 (2012)
[14] Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. IEEE Conference on Computer Vision and Pattern Recognition, 3061–3070 (2015)