FLICKR1024: A DATASET FOR STEREO IMAGE SUPER-RESOLUTION

Yingqian Wang¹, Longguang Wang¹, Jungang Yang¹, Wei An¹, Yulan Guo¹,²

¹. College of Electronic Science and Technology, National University of Defense Technology, Changsha, China
². School of Electronics and Communication Engineering, Sun-Yat-sen University, Guangzhou, China

ABSTRACT

With the popularity of dual cameras in recently released smartphones, a growing number of super-resolution (SR) methods have been proposed to enhance the resolution of stereo image pairs. However, the lack of high-quality stereo datasets has limited the research in this area. To facilitate the training and evaluation of novel stereo SR algorithms, in this paper, we propose a large-scale stereo dataset named Flickr1024. Compared to the existing stereo datasets, the proposed dataset contains much more high-quality images and covers diverse scenarios. We train two state-of-the-art stereo SR methods (i.e., StereoSR and PASSRnet) on the KITTI2015, Middlebury, and Flickr1024 datasets. Experimental results demonstrate that our dataset can improve the performance of stereo SR algorithms. The Flickr1024 dataset is available online at: https://yingqianwang.github.io/Flickr1024.

Index Terms — Dataset, image pairs, stereo vision, super-resolution (SR), diverse scenarios

1. INTRODUCTION

With recent advances in camera miniaturization, dual cameras are commonly adopted in commercial mobile phones. Using the complementary information provided by binocular systems, the resolution of image pairs can be enhanced by stereo super-resolution (SR) methods [1–3]. Nowadays, many top-performing SR methods [2–6] are built upon deep neural networks, and these data-driven SR methods can be enormously benefited from large-scale high-quality datasets such as DIV2K [7] and Vimeo-90K [8].

In the area of stereo vision, several datasets are currently available [9]. The KITTI stereo datasets [10,11] are mainly developed for autonomous driving. All images in the KITTI2012 [10] and KITTI2015 [11] datasets are captured by two video cameras mounted on the top of a car. The scenes in the KITTI datasets only include roads or highways from driving perspectives. Groundtruth disparity is provided for the training of stereo matching and visual odometry. The Middlebury stereo dataset consists of a series of sub-datasets, which are proposed in 2003 [12], 2005 [13], 2006 [14], and 2014 [15], respectively. The Middlebury dataset is recorded in the laboratory, and its scenes only cover close-shots of different objects. Note that, 55 of the total 65 image pairs are attached with groundtruth disparity for stereo matching. The ETH3D stereo dataset is a part of the ETH3D benchmark [16]. Groundtruth depth is provided for visual odometry and 3D reconstruction. Note that, images on the ETH3D dataset are of gray scale, of low resolution, and with limited scenarios.

Since the task of stereo vision can vary significantly, existing stereo datasets are unsuitable for stereo SR due to the insufficient number of images and limited types of scenarios. To design, train, and evaluate novel stereo SR methods, a large-scale and high-quality stereo dataset with diverse scenarios is highly needed.

In this paper, we propose a novel Flickr1024 dataset (see Fig. 1) for stereo SR. The Flickr1024 dataset consists of 1024 high-quality image pairs and covers diverse scenarios. Moreover, we train two state-of-the-art learning-based stereo SR methods (i.e., StereoSR [2] and PASSRnet [3]) on the proposed dataset and two existing stereo datasets (i.e., KITTI2015 [11] and Middlebury [12,15]). Experimental results demonstrate that algorithms trained on our dataset achieve better performance than those trained on the two existing datasets.
The contributions of this paper can be summarized as:

- We release the largest stereo dataset for stereo SR. This dataset contains 1024 high-quality images and covers various scenarios.

- The scenarios covered by the Flickr1024 dataset are highly consistent with real cases in daily photography (see Fig. 2). That is, algorithms developed on the Flickr1024 dataset can easily be adopted in real-world applications such as mobile phones.

- Experimental results show that our dataset can help to improve the performance of stereo SR methods, which benefits both research and industrial communities.

2. DATA ACQUISITION AND PROCESSING

To generate the Flickr1024 dataset, we manually collected 1024 RGB stereo photographs from albums on Flickr\(^1\) with the permissions of photograph owners. Since all images collected from Flickr are in cross-eye pattern for 3D visualization, their optical axes should be corrected to be parallel. As shown in Fig. 3, the processing pipeline can be summarized as follows:

1. We cut each cross-eye photograph into a stereo image pair. Note that, to transform a cross-eye photograph into an image pair with parallel optical axis, the left and right images in the stereo image pair need to be exchanged.

2. We check each pair of stereo images to ensure that they are vertically rectified (i.e., image pairs has horizontal disparities only). In practice, most image pairs have already been calibrated in vertical direction by the photo owners to achieve 3D visual effect. For these images without vertical calibration, we simply discard them from our dataset.

3. We crop the left and right images to remove black (or white) margins and to make zero disparity corresponding to infinite depth. Note that, regions with infinite depth are unavailable for close-shot images. We therefore, crop these image pairs to ensure that the minimum disparity is larger than a certain value (set to 40 pixels in our dataset).

Finally, we randomly split our dataset to generate 800 training image pairs, 112 validation image pairs, and 112 test image pairs.

\(^1\)https://www.flickr.com/
Table 1: Main characteristics of several stereo datasets. Both average value and standard deviation are reported. Among all the compared datasets, the Flickr1024 dataset achieves promising scores in image pairs, resolution, and perceptual image quality.

| Datasets          | Image Pairs | Resolution (†) | Entropy (†) | BRISQE (†) | SR-metric (†) | ENIQA (†) |
|-------------------|-------------|----------------|-------------|------------|---------------|-----------|
| KITTI2012 [10]    | 389         | 0.46 (±0.00) Mpx | 7.12 (±0.30) | 17.49 (±6.56) | 7.15 (±0.63) | 0.097 (±0.028) |
| KITTI2015 [11]    | 400         | 0.47 (±0.00) Mpx | 7.06 (±0.00) | 23.79 (±8.51) | 7.06 (±0.51) | 0.169 (±0.030) |
| Middlebury [12]   | 65          | 3.59 (±2.06) Mpx | 7.55 (±0.20) | 26.85 (±13.30) | 6.01 (±1.08) | 0.270 (±0.120) |
| ETH3D [15]        | 47          | 0.38 (±0.08) Mpx | 7.24 (±0.43) | 27.95 (±12.06) | 5.99 (±1.52) | 0.195 (±0.073) |
| Flickr1024        | 1024        | 0.73 (±0.33) Mpx | 7.23 (±0.64) | 19.40 (±13.77) | 7.12 (±0.67) | 0.065 (±0.073) |
| Flickr1024 (Train)| 800         | 0.74 (±0.34) Mpx | 7.23 (±0.65) | 19.10 (±13.69) | 7.12 (±0.66) | 0.063 (±0.074) |
| Flickr1024 (Validation) | 112 | 0.72 (±0.23) Mpx | 7.26 (±0.54) | 20.03 (±12.54) | 7.13 (±0.70) | 0.074 (±0.084) |
| Flickr1024 (Test) | 112         | 0.72 (±0.32) Mpx | 7.22 (±0.60) | 20.97 (±15.40) | 7.12 (±0.67) | 0.076 (±0.087) |

Note: Mpx denotes megapixels per image. The best scores are in bold and the second best scores are underlined.

3. COMPARISONS TO OTHER DATASETS

In this section, statistical comparisons are performed to demonstrate the superiority of the Flickr1024 dataset. The main characteristics of the Flickr1024 dataset and four existing stereo datasets are listed in Table 1. Following [7], we use entropy to indicate the amount of information included in each dataset, and use three no-reference image quality assessment (NRIQA) metrics (i.e., blind/referenceless image spatial quality evaluator (BRISQE) [17], SR-metric [18], and entropy-based image quality assessment (ENIQA) [19]) to assess the perceptual image quality. It is demonstrated in [18] that in the area of image quality assessment, these NRIQA metrics are proved superior to many fully-referenced measures (e.g., PSNR, RMSE, and SSIM), and highly correlated to human perception. For all of the NRIQA metrics presented in this paper, we run the codes provided by their authors under their original models and default settings. For BRISQE [17] and ENIQA [19], a small score indicates a high image quality. For SR-metric [18], a large score indicates a high image quality.

As listed in Table 1, the Flickr1024 dataset is larger than other datasets by at least 2.5 times. Besides, the image resolution of the Flickr1024 dataset also outperforms that of the KITTI2012, KITTI2015, and ETH3D datasets. Although the Middlebury dataset has the highest image resolution, the number of image pairs in this dataset is limited. The entropy values of all datasets are comparable, while the entropy of the KITTI datasets is relatively low. That is, the diversity of images in the KITTI datasets is smaller than that of other datasets. For perceptual image quality assessment, both the Flickr1024 and the KITTI2012 datasets achieve promising scores. Specifically, the Flickr1024 dataset has the best score in ENIQA, and has the second best scores in both BRISQE and SR-metric. Since these metrics are influenced by the brightness and textures of tested images, the Flickr1024 dataset has higher standard deviations than existing datasets due to its diverse scenarios. These assessments indicate that images in Flickr1024 are of relatively high perceptual quality and suitable for stereo SR.

It is also notable that, comparable scores of these metrics can be achieved on the subsets (i.e., training set, validation set, and test set) of the Flickr1024 dataset, as shown in Table 1. That means, a good balance is achieved with random partition, and the bias between the training and the test process is relatively small.

4. CROSS-DATASET EVALUATION

To investigate the potential benefits of a large-scale dataset to the performance improvement of learning-based stereo SR methods, experimental results are provided in this section. Besides, a cross-dataset evaluation is performed to further demonstrate the superiority of the Flickr1024 dataset.

4.1. Implementation Details

We use two state-of-the-art stereo SR methods (i.e., StereoSR [2] and PASSRnet [3]) in this experiment. These two methods are first trained on the KITTI2015, Middlebury, and Flickr1024 datasets, and then tested on the above three datasets and the ETH3D dataset. For simplification, only 4× SR models are investigated. That is, the stereo image pairs are first down-sampled by a factor of 4, and then super-resolved to their respective original resolutions. We compare the reconstructed image with the original image, and use PSNR and SSIM for performance evaluation.

We used the codes of StereoSR [2] and PASSRnet [3] released by their authors. Since the StereoSR model trained on the Middlebury dataset is available, we directly use this model in our experiment. For the other 5 unavailable models, we retrain the two SR methods following the instructions in their papers.

4.2. Results

Tables 2 and 3 present the results of StereoSR and PASSRnet trained with fixed training epochs. We can observe that both algorithms trained on the Flickr1024 dataset achieve the highest PSNR and SSIM values on all of the test sets as
compared to those trained on the KITTI2015 and Middlebury datasets. These results indicate that the Flickr1024 dataset can help to improve the performance of stereo SR algorithms.

Moreover, we train PASSRnet [3] with different training epochs, and further investigate the variation of PSNR and SSIM. The results are shown in Fig. 4, where each sub-figure illustrates the performance tested on a specific dataset. We can observe that the algorithm trained on the Flickr1024 dataset achieves the highest PSNR and SSIM values with arbitrary settings of training epochs. Compared to the models trained on the KITTI2015 dataset whose PSNR and SSIM curves suffer a downward trend, the models trained on the Flickr1024 dataset can achieve a gradually improved performance with increasing training epochs. That is, by using our dataset, a reasonable convergence can be steadily achieved, and the over-fitting issue can be well addressed.

### 5. CONCLUSION

In this paper, we introduce Flickr1024, a large-scale dataset for stereo SR. The Flickr1024 dataset consists 1024 high-quality images and covers diverse scenarios. Both statistical comparisons and experimental results demonstrate the superiority of our dataset. That is, the Flickr1024 dataset can be used to improve the performance of existing learning-based stereo SR methods. This dataset can also help to boost the research in stereo super-resolution.

### 6. ACKNOWLEDGMENT

The authors would like to thank Sascha Becher and Tom Bentz for the approval of using their cross-eye stereo photographs on Flickr. This work was partially supported in part by the National Natural Science Foundation of China (Nos. 61602499 and 61401474), the Hunan Provincial National Science Foundation (No. 2016JJ3025), and the Fundamental Research Funds for the Central Universities (No. 18lgzd06).
7. REFERENCES

[1] A. V. Bhavsar and A. Rajagopalan, “Resolution enhancement in multi-image stereo,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1721–1728, 2010.

[2] D. S. Jeon, S.-H. Baek, I. Choi, and M. H. Kim, “Enhancing the spatial resolution of stereo images using a parallax prior,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1721–1730.

[3] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo, “Learning parallax attention for stereo image super-resolution,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:1903.05784, 2019.

[4] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 2018, pp. 8–14.

[5] L. Wang, Y. Guo, Z. Lin, X. Deng, and W. An, “Learning for video super-resolution through HR optical flow estimation,” in *Proceedings of the Asian Conference on Computer Vision*, Perth, Australia, 2018.

[6] H. W. F. Yeung, J. Hou, X. Chen, J. Chen, Z. Chen, and Y. Y. Chung, “Light field spatial super-resolution using deep efficient spatial-angular separable convolution,” *IEEE Transactions on Image Processing*, 2018.

[7] E. Agustsson and R. Timofte, “Ntire 2017 challenge on single image super-resolution: Dataset and study,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, vol. 3, 2017, p. 2.

[8] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, “Video enhancement with task-oriented flow,” *arXiv preprint arXiv:1711.09078*, 2017.

[9] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang, “Learning for disparity estimation through feature constancy,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2811–2820.

[10] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[11] M. Menze and A. Geiger, “Object scene flow for autonomous vehicles,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3061–3070.

[12] D. Scharstein and R. Szeliski, “High-accuracy stereo depth maps using structured light,” in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1. IEEE, 2003, pp. I–I.

[13] D. Scharstein and C. Pal, “Learning conditional random fields for stereo,” in *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[14] H. Hirschmuller and D. Scharstein, “Evaluation of cost functions for stereo matching,” in *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[15] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, “High-resolution stereo datasets with subpixel-accurate ground truth,” in *German Conference on Pattern Recognition*. Springer, 2014, pp. 31–42.

[16] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, “A multi-view stereo benchmark with high-resolution images and multi-camera videos,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2017, 2017.

[17] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[18] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, “Learning a no-reference quality metric for single-image super-resolution,” *Computer Vision and Image Understanding*, vol. 158, pp. 1–16, 2017.

[19] X. Chen, Q. Zhang, M. Lin, G. Yang, and C. He, “No-reference color image quality assessment: From entropy to perceptual quality,” *arXiv preprint arXiv:1812.10695*, 2018.