This ICCV Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Flickr1024: A Large-Scale Dataset for Stereo Image Super-Resolution

Yingqian Wang[1], Longguang Wang[1], Jungang Yang[1*], Wei An[1], and Yulan Guo[1]
[1]College of Electronic Science and Technology, National University of Defense Technology, China
{wangyingqian16, yangjungang}@nudt.edu.cn

Abstract

With the popularity of dual cameras in recently released smart phones, 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 present a large-scale stereo dataset named Flickr1024, which contains 1024 pairs of high-quality images and covers diverse scenarios. We first introduce the data acquisition and processing pipeline, and then compare several popular stereo datasets. Finally, we conduct cross-dataset experiments to investigate the potential benefits introduced by our dataset. Experimental results show that, as compared to the KITTI and Middlebury datasets, our Flickr1024 dataset can help to handle the over-fitting problem and significantly improves the performance of stereo SR methods. The Flickr1024 dataset is available online at: https://yingqianwang.github.io/Flickr1024.

Figure 1: The *Flickr1024* dataset.

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 [2, 6, 17]. Nowadays, many top-performing SR methods [6, 16, 17, 19] are built upon deep neural networks. These data-driven SR methods can be enormously benefited from large-scale high-quality datasets such as *DIV2K* [1] and *Vimeo-90K* [18].

In the area of stereo vision, several datasets are currently available. The *KITTI* stereo datasets [4, 10] are mainly developed for autonomous driving. All images in the *KITTI2012* [4] and *KITTI2015* [10] 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 introduced in 2003 [14], 2005 [13], 2006 [5], and 2014 [12]. The *Middlebury* dataset is acquired in the laboratory, its scenes only cover close-shots of different objects. Note that, 55 of its 65 image pairs are provided with groundtruth disparity for stereo matching. The *ETH3D* stereo dataset is a part of the *ETH3D* benchmark [15]. Groundtruth depth is provided for visual odometry and 3D reconstruction. Note that, images in the *ETH3D* dataset are of gray scale, of low resolution, and with limited scenarios.

Since the task of stereo vision can vary significantly (e.g., stereo matching [8], stereo segmentation [7]), 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 present a new *Flickr1024* dataset (see Fig. 1) for stereo SR. In summary, the contributions of our dataset can be listed as follows:

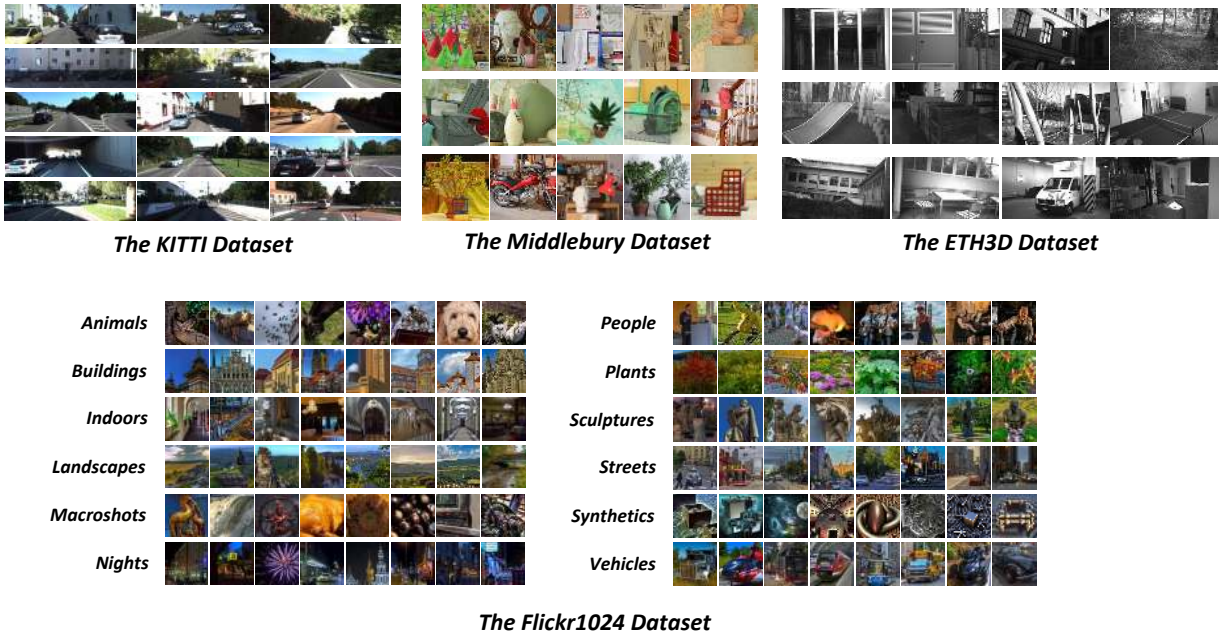


Figure 2: Example images sampled from several popular stereo datasets: *KITTI* [4, 10], *Middlebury* [5, 12–14], *ETH3D* [15], and *Flickr1024*. It can be observed that our *Flickr1024* dataset covers significantly more diverse scenarios as compared to the existing stereo datasets.

- It is the largest dataset for stereo SR to date, which contains 1024 high-quality image pairs and covers diverse scenarios.
- The scenarios covered by our dataset are highly consistent with real cases in daily photography (see Fig. 2). Consequently, algorithms developed on the *Flickr1024* dataset can be easily adopted in real-world applications such as mobile phones.
- Experimental results demonstrate that our dataset can help to address the over-fitting problem and significantly improve the performance of different stereo SR methods. That is, our dataset can benefit both industrial and research communities in stereo SR.

The *Flickr1024* dataset was initially introduced in our previous conference paper [17] and used as the augmented training data for our *PASSRnet* algorithm. However, in our preliminary work, both details and effectiveness of this dataset have not been investigated. In this paper, we deeply investigate the *Flickr1024* dataset and make several additional contributions, which can be summarized as follows:

- We provide more details in data acquisition and processing. The pipelines and processing methods used in this paper can help the community to develop new datasets for their own research.
- We comprehensively compare the *Flickr1024* dataset to several existing stereo datasets [4, 5, 10, 12–15] using different objective metrics. Evaluation results have demonstrated the superiority of our dataset.
- We conduct cross-dataset experiment to study the performance gains introduced by the *Flickr1024* dataset. Experimental results have demonstrated the effectiveness of our dataset in performance promotion and over-fitting elimination.

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.

[1] https://www.flickr.com/

Table 1: Main characteristics of several popular 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 (↓) [11]	SR-metric (↑) [9]	ENIQA (↓) [3]
KITTI2012 [4]	389	0.46 (±0.00) Mpx	7.12 (±0.30)	**17.49** (±6.56)	**7.15** (±0.63)	<u>0.097</u> (±0.028)
KITTI2015 [10]	<u>400</u>	0.47 (±0.00) Mpx	7.06 (±0.00)	23.79 (±5.81)	7.06 (±0.51)	0.169 (±0.030)
Middlebury [5, 12–14]	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	<u>7.24</u> (±0.43)	27.95 (±12.06)	5.99 (±1.52)	0.195 (±0.073)
Flickr1024	**1024**	<u>0.73</u> (±0.33) Mpx	7.23 (±0.64)	<u>19.40</u> (±13.77)	<u>7.12</u> (±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 <u>underlined</u>.

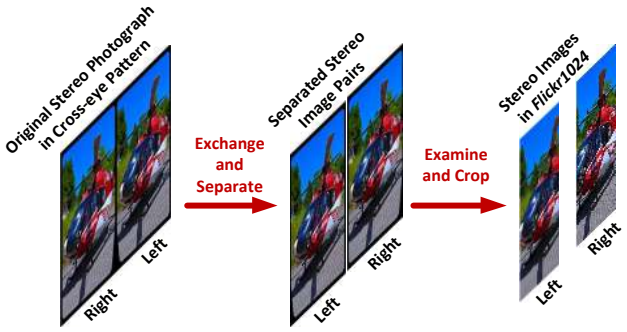


Figure 3: The processing pipeline to generate the *Flickr1024* dataset.

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. 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.

3. Comparisons to Existing 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 [4, 5, 10, 12–15] are listed in Table 1. Following [1], we use *entropy* to measure the amount of information included in each dataset, and use three no-reference image quality assessment (NRIQA) metrics to assess the perceptual image quality, including blind/referenceless image spatial quality evaluator (BRISQE) [11], SR-metric [9], and entropy-based image quality assessment (ENIQA) [3]. For image quality assessment, these NRIQA metrics are superior to many full-referenced measures (e.g., PSNR, RMSE, and SSIM), and highly correlated to human perception [9]. For all the NRIQA metrics presented in this paper, we run the codes provided by their authors under their original models and default settings. Note that, small values of BRISQE [11] and ENIQA [3], and large values of SR-metric [9] represent 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 highest ENIQA score, and the second highest BRISQE and SRmetric scores. 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 relatively high in perceptual quality and suitable for stereo SR.

Table 2: PSNR and SSIM values achieved by *StereoSR* [6] for 4× SR with 60 training epochs.

Dataset	*KITTI2015* (Test)	*Middlebury* (Test)	*Flickr1024* (Test)	*ETH3D* (Test)
KITTI2015 (Train)	24.28 / 0.741	26.27 / 0.749	21.77 / 0.617	29.63 / 0.831
Middlebury (Train)	23.64 / 0.743	26.62 / 0.773	21.64 / 0.646	28.66 / 0.843
Flickr1024 (Train)	**25.08 / 0.779**	**27.85 / 0.807**	**22.64 / 0.692**	**30.55 / 0.860**

Table 3: PSNR and SSIM values achieved by *PASSRnet* [17] for 4× SR with 80 training epochs.

Dataset	*KITTI2015* (Test)	*Middlebury* (Test)	*Flickr1024* (Test)	*ETH3D* (Test)
KITTI2015 (Train)	23.13 / 0.703	25.42 / 0.762	21.31 / 0.600	26.95 / 0.789
Middlebury (Train)	25.18 / 0.774	28.08 / 0.853	22.54 / 0.676	31.39 / 0.864
Flickr1024 (Train)	**25.62 / 0.791**	**28.69 / 0.873**	**23.25 / 0.718**	**31.94 / 0.877**

It is also notable that, comparable scores of these metrics can be achieved on three different 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 fully demonstrate the superiority of the *Flickr1024* dataset.

4.1. Implementation Details

We use two state-of-the-art stereo SR methods (i.e., *StereoSR* [6] and *PASSRnet* [17]) 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, stereo image pairs are first down-sampled by a factor of 4, and then super-resolved to their 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* [6] and *PASSRnet* [17] 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 these two SR methods.

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. Specifically, the *Flickr1024* dataset outperforms the second best datasets (*KITTI2015* in Table 2 and *Middlebury* in Table 3) by 1.04 and 0.58 in average PSNR values, respectively. These results demonstrate that the *Flickr1024* dataset can help to significantly improve the performance of stereo SR algorithms.

Moreover, we train *PASSRnet* [17] with different training epochs, and further investigate the variation of PSNR and SSIM values. 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 any number of training epochs. For the models trained on the *KITTI2015* dataset, their PSNR and SSIM curves suffer a downward trend. In contrast, the models trained on the *Flickr1024* dataset can achieve a gradually improved performance with an increasing number of training epochs. These results demonstrate that, by using our dataset, a reasonable convergence can be steadily achieved and the over-fitting issue can be well addressed.

5. Conclusion

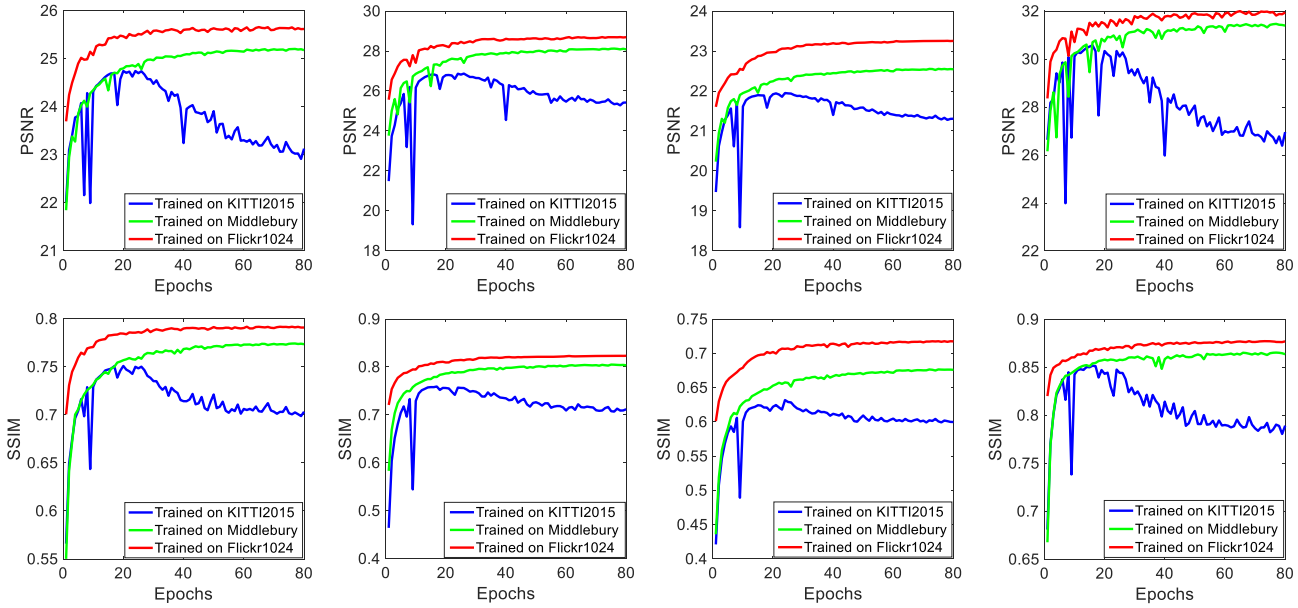
In this paper, we introduce *Flickr1024*, a large-scale dataset for stereo SR. The *Flickr1024* dataset consists 1024 high-quality images and covers diverse scenarios. Both statistical comparisons and cross-dataset experiments demonstrate the effectiveness of our dataset. That is, the *Flickr1024* dataset can be used to improve the performance of learning-based stereo SR methods. The *Flickr1024* dataset can also help to boost the reseach 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*.

References

[1] E. Agustsson and R. Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The*



(a) *KITTI2015* (b) *Middlebury* (c) *Flickr1024* (d) *ETH3D*

Figure 4: PSNR and SSIM values achieved by *PASSRnet* [17] with different numbers of training epochs for 4× SR. Note that, the performance is evaluated on the test sets of (a) *KITTI2015*, (b) *Middlebury*, (c) *Flickr1024*, and (d) *ETH3D*, respectively.

IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, volume 3, page 2, 2017. 1, 3

[2] A. V. Bhavsar and A. Rajagopalan. Resolution enhancement in multi-image stereo. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 32(9):1721–1728, 2010. 1

[3] 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. 3

[4] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2, 3

[5] H. Hirschmuller and D. Scharstein. Evaluation of cost functions for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 1, 2, 3

[6] 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 (CVPR)*, pages 1721–1730, 2018. 1, 4

[7] X. Li, H. Huang, H. Zhao, Y. Wang, and M. Hu. Learning a convolutional neural network for propagation-based stereo image segmentation. *The Visual Computer*, pages 1–14, 2018. 1

[8] 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 (CVPR)*, pages 2811–2820, 2018. 1

[9] 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*, 158:1–16, 2017. 3

[10] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 1, 2, 3

[11] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 3

[12] 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*, pages 31–42. Springer, 2014. 1, 2, 3

[13] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 1, 2, 3

[14] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 1, 2, 3

[15] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2017, 2017. 1, 2, 3

[16] 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 (ACCV), Perth, Australia*, 2018. 1

[17] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 4, 5

[18] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman. Video enhancement with task-oriented flow. *arXiv preprint arXiv:1711.09078*, 2017. 1

[19] S. Zhang, Y. Lin, and H. Sheng. Residual networks for light field image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11046–11055, 2019. 1