Holopix50k: A Large-Scale In-the-wild Stereo Image Dataset

Yiwen Hua[0000−0003−2334−7157], Puneet Kohli[0000−0002−4817−8436], Pritish Uplavikar*[0000−0001−7187−8652], Anand Ravi*[0000−0002−3656−601X], Saravana Gunaseelan[0000−0002−3315−7464], Jason Orozco[0000−0002−0638−3407], and Edward Li[0000−0003−2651−1100]

Leia Inc.

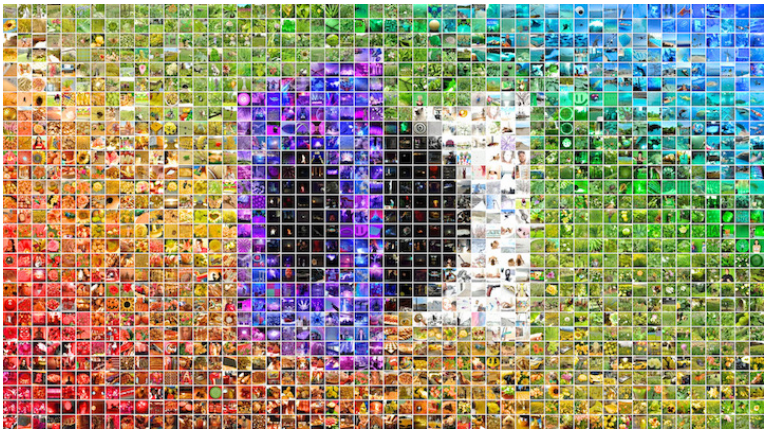


Fig. 1. Samples from the Holopix50k stereo image dataset

Abstract. With the mass-market adoption of dual-camera mobile phones, leveraging stereo information in computer vision has become increasingly important. Current state-of-the-art methods utilize learning-based algorithms, where the amount and quality of training samples heavily influence results. Existing stereo image datasets are limited either in size or subject variety. Hence, algorithms trained on such datasets do not generalize well to scenarios encountered in mobile photography. We present Holopix50k, a novel in-the-wild stereo image dataset, comprising 49,368 image pairs contributed by users of the Holopix™ mobile social platform. In this work, we describe our data collection process and statistically compare our dataset to other popular stereo datasets. We experimentally show that using our dataset significantly improves results for tasks such as stereo super-resolution and self-supervised monocular depth estimation. Finally, we showcase practical applications of our dataset to motivate novel works and use cases. The dataset is available at `http://github.com/leiainc/holopix50k`.

现存数据集少，多样性差

49368张图像对

可以大幅改善立体超像素和自监督单目深度估计结果

* indicates equal contribution

Keywords: Stereo vision · stereo image dataset · 3D photography · stereo super-resolution · disparity estimation

1 Introduction

Most mobile phones now come with two or more cameras, enabling consumer applications such as artificial depth of field (portrait mode), AI-based photo optimization, and 3D photography [29]. Through dual-cameras, mobile phones can perceive depth and lighting of a scene using complementary information from the stereo image pairs. Various applications utilize stereo imagery such as Holopix[1] to produce immersive 3D Photography experiences. More recently, larger platforms such as Facebook and Snap Inc. have also demonstrated unique and exciting user experiences enabled via stereo imagery.

Advancements in self-driving cars and robotics technologies with multi-camera configurations have accelerated research and development of computer vision algorithms [19][56], many of which utilize additional stereo information [40][17][30]. The mass-market adoption of dual-cameras in mobile phones has led to an increased need for these algorithms to work in casual in-the-wild scenarios encountered during mobile photography. A majority of state-of-the-art methods in this domain are based on deep learning methods, where accuracy and quality scale with the amount of data available for training. Current datasets either cover a limited subset of real-world scenarios [16][8] or are taken in a laboratory setting [48][9]. Further, there is an increased need for a large-scale stereo dataset representative of the diversity of real-world scenarios to enable generalization. Though recent works have been able to collect stereo pairs [55][57] from various internet platforms, they may not be diverse or extensive enough to represent in-the-wild scenarios captured from mobile devices.

In this work, we present a new large-scale stereo image dataset that is completely user-generated and captured mostly from mobile phone cameras in-the-wild. In our novel *Holopix50k* dataset, we present 49,368 (roughly 50k) rectified stereo image pairs collected from the mobile social media platform Holopix. This is larger than comparable stereo image datasets by an order of magnitude. To our knowledge, this is the first large-scale stereo image dataset collected from a mobile-based social platform where users curate and upload stereo images. Figure 1 shows samples from the dataset demonstrating the diversity of scenes present, arranged to depict the Holopix platform's logo.

In the next section, we explore existing stereo datasets and their applications to stereo vision tasks. In the following sections, we explore our data collection and filtering process. Further, we compare our dataset with existing stereo datasets on various metrics. Next, we demonstrate how our dataset helps improve existing state-of-the-art methods for stereo super-resolution and self-supervised monocular depth estimation tasks. Finally, we showcase the qualitative results of various disparity models we have trained using our dataset to motivate further

[1] Holopix™. The leading mobile social network for sharing lightfield imagery (`https://www.holopix.com`).

novel works and use cases.

Summarizing our key contributions below,

1. Our new dataset, Holopix50k, is the largest in-the-wild stereo image dataset to date with 49,368 high-quality rectified stereo image pairs covering a wide variety of realistic scenarios found in mobile photography.
2. Our experimental results demonstrate that our dataset improves a variety of stereo vision tasks in both quantitative and qualitative performance.

2 Related Work

2.1 Stereo Datasets

Real-world Datasets. There are a variety of real-world stereo datasets currently available. Some of the more popular ones, such as KITTI [16][38], Middlebury [48], and the NYU Indoor [41] datasets, all provide ground truth depth or disparities for the stereo scenes. However, these datasets either have a limited number of images, or the scenes are for a specific domain. For example, the KITTI datasets were mainly developed for self-driving vehicle use-cases, while the few scenes in Middlebury are all captured in a laboratory setting. Other real-world stereo datasets include Make3D [47], ETH3D [49], CMLA [9], and Cityscapes [8] — each focused on a specific domain. More recent datasets such as Flickr1024 [57] and WSVD [55] provide more diverse scenes, however, Flickr1024 is relatively small compared to our dataset and WSVD images score significantly lower on quality metrics, as shown in Table 1. While our dataset also comprises real-world scenarios, we demonstrate the superiority of our dataset in terms of the number and diversity of images present, as well as various metric scores.

Synthetic Datasets. The challenge with real-world stereo datasets is collecting ground truth information at-scale that covers a large variety of complex scene types, lighting effects, and motion scenarios. It is also difficult to collect data with different camera baselines and depth ranges. Recent works introduce synthetic datasets that take advantage of the rise of efficient, high-quality rendering techniques. Examples include MPI Sintel [4], SceneFlow [35], UnrealStereo [61], and the 3D Ken Burns [42] datasets. The challenge with such datasets is that techniques based on these datasets still suffer from the domain shift problem [46] when adapting to real-world scenarios. Furthermore, significant effort may be required to cover the diversity of real-world scenarios using synthetic techniques. We envision our dataset improving the robustness and generalization capabilities of learning-based models trained on these synthetic datasets by providing a large distribution of real-world scenarios.

虚拟数据存在域变换的问题

2.2 Stereo Image Tasks

Stereo Disparity Estimation. Traditional stereo disparity estimation algorithms rely on block-matching methods [22]. More recently, there has been a

传统算法BM

significant interest in learning-based techniques due to advances in neural networks [60][59][36][27]. Most learning-based techniques require ground-truth disparity (or depth) for supervision. Since most datasets with ground-truth disparity are limited to specific domains, recent works use self-adaption approaches to estimate disparity [52][53][43]. The diversity of our dataset can aid in the generalization of these unsupervised techniques. 自适应法估计视差

Monocular Depth Estimation. Monocular depth estimation aims to estimate the depth of a scene from a single image [18][12]. This is an ill-posed problem as image textures do not directly correspond to depth [2]. Monocular techniques exist using self-supervision with a stereo input, and some are entirely unsupervised [18][37]. Similar to stereo disparity tasks, the datasets used are limited in either domain or size. Hence, these techniques fail to generalize to tasks found in-the-wild. Our dataset has a large variety of diverse scenarios at a scale no other dataset currently provides, which significantly improves the generalizability of models trained for these tasks.

Stereo Super-Resolution. Super-resolution is a well-known task in computer vision that aims to construct a high-resolution image from their low-resolution versions [7][11][13]. Stereo super-resolution techniques aim to incorporate a second low-resolution image to increase the quality of the super-resolved high-resolution image. Recently, there has been an increased interest in learning-based stereo super-resolution methods [56][25][31] due to the advancement in stereo photography on mobile devices. The sheer size of our dataset is well-suited to improve the results of stereo super-resolution.

There are a variety of other stereo vision tasks such as stereoscopic style transfer [5][20][28], flow estimation [50][58], and stereo segmentation [44], which could benefit from a large-scale dataset such as ours. We do not demonstrate the improvements that our dataset brings to such tasks in this work.

3 The Holopix50k Dataset

Holopix50k is the largest in-the-wild stereo image dataset crowd-sourced from a social media platform to date, containing 49,368 stereo pairs (98,736 images in total). In this section, we discuss our data collection method and also shed light on the diversity of content present in our dataset. 揭示

3.1 Data Acquisition and Processing

The Holopix platform (visualized in Figure 2) is the only major social media platform built specifically for sharing 3D Photography where multiple viewpoints of a scene are captured and viewed with a parallax effect [29]. Combined with Lightfield displays, such as those made by Leia Inc. [14], there is an added 获得多视角场景图像并以视差效果观看，结合光场显示

depth [23] and multi-view parallax effect [10] that enhances the experience. Due to Holopix's unique position as the first Lightfield-enabled social media application, it has attracted a significant user-base who regularly upload photos to the platform. The platform takes in two or more input views and renders a multi-view image on Lightfield devices, or a motion-based animation on regular devices.

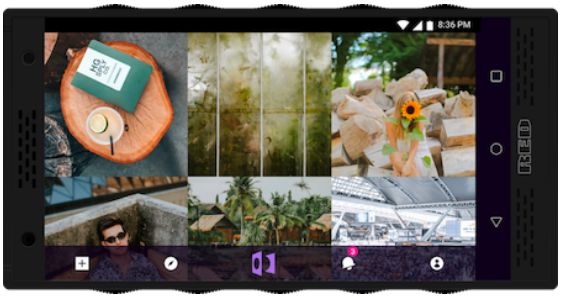


Fig. 2. The Holopix Social Platform

Statistically, the vast majority of images on Holopix are captured on the RED Hydrogen One[1] mobile phone, which is one of the first consumer-grade Lightfield devices, offering a 4-view display. The stereo camera application on this device rectifies images and converges on the mean disparity of the scene. Note that in works such as Flickr1024 [57], the stereo pairs are cropped to re-converge the stereo pairs at infinity, as a pre-processing step. To maintain the initially captured pair as-is, we do not re-converge the images via cropping.

Flickr将图像对剪切使得立体图像收敛在无穷远处，次数保留原始数据

Data Collection. We initially collect about 70k stereo image pairs from the Holopix platform that were suitable for use in the dataset. In the case of mismatching resolution in the left and right stereo images, we downscale the larger of the two images to match the resolution of the smaller image for consistency. The downscaling is done by first applying Gaussian smoothing to the image, followed by bicubic interpolation, as Holopix also used this method to downscale the images.

为防止左右图误匹配，缩小图像进行匹配，先进性高斯平滑，再做双三次插值

Removing Vertical Disparity. As our dataset originates from cameras that are horizontally aligned, the stereo pairs are expected only to have horizontal disparity. As such, stereo pairs with even slight vertical disparity can cause severe failures in stereo algorithms such as stereo block-matching. [15][51]. After filtering out stereo pairs with vertical disparity, we are left with roughly 60k images in our dataset.

移除垂直视差剩下6万张

[1] RED Hydrogen One. https://en.wikipedia.org/wiki/Red_Hydrogen_One

Disparity-based Filtering. Given that most images from our dataset are captured on the Hydrogen One, which has a small stereo baseline of 12mm for the rear camera and 5mm for the front camera, we expect certain scene conditions to produce limited stereo information. We filter out these images by analyzing the disparity maps produced by our stereo disparity network (see Section 5.2) and removing the images that produce extreme disparity failures. By doing this, we bias our dataset towards scenes which have useful disparity information. We provide more details in the supplemental material.

后置基线12mm，前置5mm，希望特定场景产生有限的立体信息

分析视差，移除失败例子（有超过0.95的匹配失败），这样保证数据视差信息都是有用的

Our filtering process removes up to 95% of image pairs with poor stereo characteristics, after which, our final dataset includes 49,368 stereo pairs.

3.2 Dataset Exploration

As an exploration of our dataset, we run the Mask-RCNN [21] object detector on the left image of each stereo pair. The detector we use is pre-trained on the MS COCO dataset [33], containing 91 object categories for common objects. We retain only the bounding boxes with ≥ 0.7 confidence. From the word-cloud representation in Figure 3, we can see that there is a good mixture of common objects including people, animal species, plants, vehicles, furniture, electronics, and food types. As Holopix is a social media platform, it is expected that a majority of the images would contain humans, which indeed dominates our dataset with approximately 21k humans detected.

保留0.7以上的bbox，发现场景中有人、动物、植物、交通工具、家具、电器和食物

21k有人的图片

Fig. 3. Word-cloud depicting common objects found in the Holopix50k dataset using the Mask-RCNN [21] object detector trained on the COCO dataset [33]

Our dataset also contains both landscape as well as portrait orientation photography. Although a small portion of images are in portrait orientation (roughly 4%), this adds additional diversity to the dataset in terms of unique content such as stereoscopic selfies. Figure 4 shows hand-picked samples from our dataset that demonstrate even more diversity in terms of scene categories.

4%是肖像图片

4 Comparison with Existing Datasets

In this section, we show the results of statistical comparison of the Holopix50k dataset to other popular stereo datasets. Table 1 shows the results of comparison against KITTI2012 [16], KITTI2015 [38], Middlebury [48], ETH3D [49], Flickr1024 [57], and Web Stereo Video (WSVD) [55] datasets. Note that WSVD

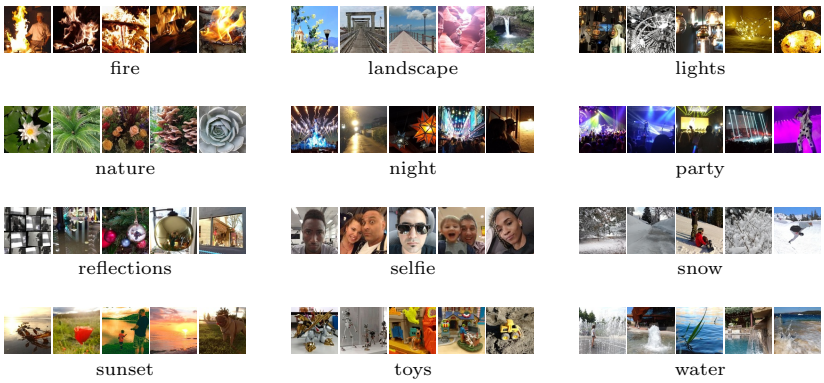


Fig. 4. Diversity of content found in the Holopix50k dataset

contains stereo videos collected from YouTube and Vimeo, filtered out into unique clips. Out of these, Flickr1024 and WSVD are ones that we consider to be in-the-wild datasets similar to ours, whereas the others are captured under specific environmental settings. We compare the amount of information present in the datasets using the entropy measure [54], and use BRISQUE [39], SR-metric [34], and ENIQA [6], in order to assess the no-reference perceptual quality of the datasets.

Table 1. Mean and standard deviation of main characteristics of popular stereo datasets. Holopix50k is by-far the largest stereo dataset with top 3 scores in resolution, entropy and SR-metric

Dataset	Image Pairs	Resolution (↑)	Entropy (↑)	BRISQUE (↓)	SR-metric (↑)	ENIQA (↓)
KITTI 2012 [16]	389	0.46 (±0.00) Mpx	7.12 (±0.30)	**17.49** (±6.56)	7.15 (±0.63)	0.097 (±0.028)
KITTI 2015 [38]	400	0.47 (±0.00) Mpx	7.06 (±0.00)	23.79 (±5.81)	7.06 (±0.51)	0.169 (±0.030)
Middlebury [48]	65	**3.59** (±2.06) Mpx	**7.55** (±0.20)	26.85 (±13.30)	6.01 (±1.08)	0.270 (±0.120)
ETH3D [49]	47	0.38 (±0.08) Mpx	7.24 (±0.43)	27.95 (±12.06)	5.99 (±1.52)	0.195 (±0.073)
Flickr1024 [57]	1024	0.73 (±0.33) Mpx	7.23 (±0.64)	19.40 (±13.77)	7.12 (±0.67)	**0.065** (±0.073)
WSVD [55][2]	10250	0.77 (±0.29) Mpx	7.13 (±0.59)	44.93 (±12.66)	4.98 (±1.54)	0.293 (±0.112)
Holopix50k SD	13068	0.23 (±0.00) Mpx	7.42 (±0.40)	19.34 (±11.94)	**8.19** (±0.95)	0.069 (±0.079)
Holopix50k HD	36300	0.92 (±0.00) Mpx	7.36 (±0.50)	19.54 (±11.77)	6.62[1](±0.98)	0.212 (±0.116)
Holopix50k	**49368**	0.74 (±0.30) Mpx	7.38 (±0.48)	19.49 (±11.81)	7.28[1](±1.24)	0.174 (±0.125)

Table adapted from [57]. **Bold** values indicate best scores, underlined values indicate second-best scores. For all metrics, higher values are better, except BRISQUE and ENIQA.
[1] Only *left* images from Holopix HD were used to compute SR-metric due to computation constraints. We assume trivial perceptual differences between *left* and *right* images.
[2] Only the middle frame per clip is considered for our analysis, similar to the original paper [55].

We divide our dataset into two subsets for analysis, *HD* and *SD*, containing the 720p and 360p images, respectively. From Table 1, we observe that the lower

高清720P，低清360P，低清有较高的SR度量和第二好的BRISQUE、ENIQA

resolution *SD* subset performs well in the perceptual metrics, with the best score in SR-metric and second-best in BRISQUE, ENIQA, and entropy. Although the higher resolution *HD* subset does not fare as well in these metrics, it has the second-highest resolution of all stereo datasets available, after Middlebury. Both of our subsets individually contain a larger number of images than any other dataset.

When we analyze the dataset as a whole, Holopix50k is the largest stereo image dataset, beating the second-largest dataset (WSVD) by roughly five times, and the third-largest (Flickr1024) by almost fifty times. Compared to other datasets, our complete dataset has the highest SR-metric, which indicates that our dataset contains higher quality images with respect to human visual perception [34]. Our dataset also scores second-highest in entropy, which shows that, on average, the amount of information present in images of our dataset is high.

说明质量较高，图像信息较高

These results showcase the superiority of our dataset on various fronts. First, our dataset has the most extensive selection of *HD* stereo image pairs. Second, it has the most extensive selection of *SD* stereo image pairs having top-2 scores across all metrics. Finally, as a whole, our dataset is orders of magnitudes larger than the rest, yet it performs relatively well in most perceptual metrics compared to the smaller datasets.

We additionally randomly split the dataset into train, test, and validation subsets based on an 85:5:10 ratio and for each subset to observe near-identical metric scores to the entire dataset, showing negligible selection bias (see supplemental material for details).

5 Experiments with Holopix50k

5.1 Stereo Super-Resolution

In this section, we demonstrate how the Holopix50k dataset helps boost the performance on the current state-of-the-art method, namely PASSRNet [56], for stereo super-resolution. The PASSRNet model is originally trained on the Flickr1024 [57] training dataset. We re-train it for the 4x super-resolution task in a variety of experiments using the training code provided by the authors. For these experiments, we use bicubic downsampling [3] to create the low-resolution inputs needed to maintain parity with the Flickr1024 training. We compare the 4x super-resolved images to their original counterparts to calculate the PSNR and SSIM full-reference metrics for evaluating the performance of the models.

将Flickr作为baseline模型

Figure 5 shows the qualitative results of the 4x super-resolution (SR) task on test images from various datasets. On visual inspection, it is evident that our Holopix50k model can resolve fine details and textures much better than both the Flickr1024 model and standard bicubic upsampling [26].

比Flickr和双三次上采样好

From Table 2, we observe that using only 1k images from our dataset does not outperform the Flickr1024 model. However, when we start increasing the number of training samples, we note that our models outperform the Flickr1024 model on all test sets under consideration, except Flickr1024's own test set. These results

我们模型在所有测试集上都比Flickr好，除了Flickr自己的测试集，这说明泛化性不够，防止过拟合倒是有

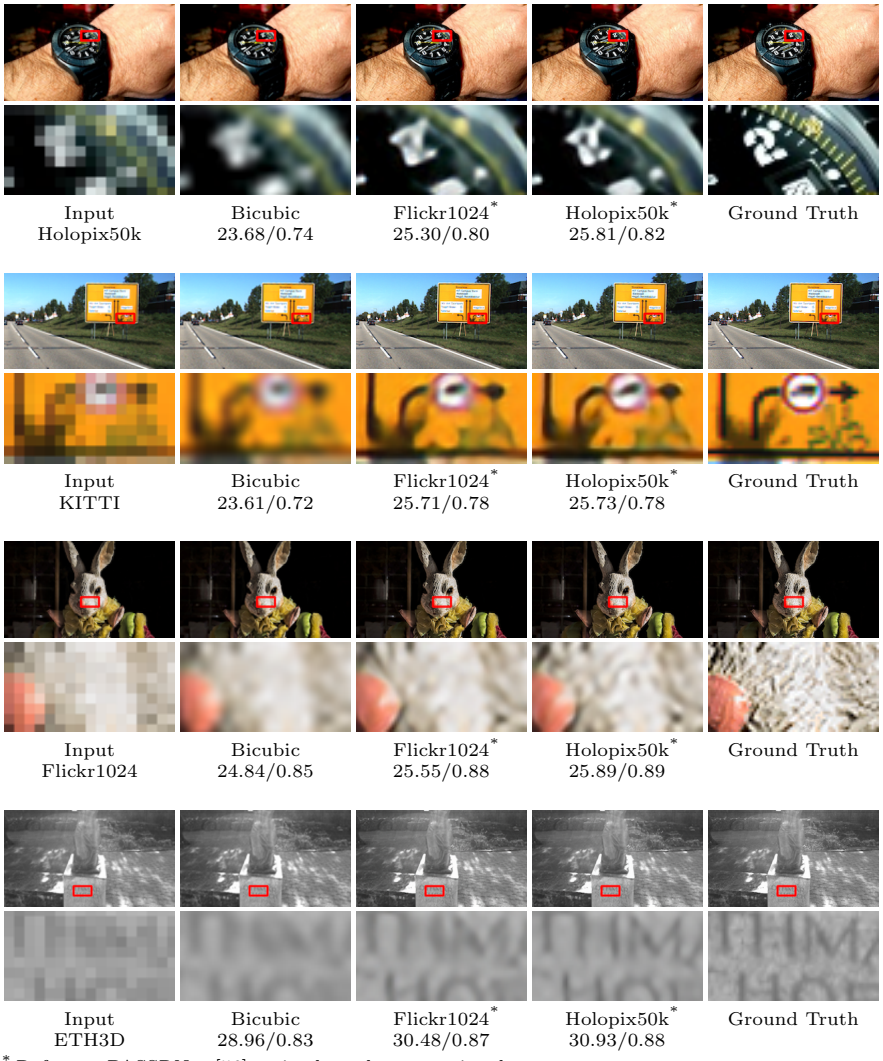


Fig. 5. Qualitative results of different super-resolution methods for 4x SR on images from a variety of datasets. Note that results produced by PASSRNet [56] trained on Holopix50k have the sharpest results and highest PSNR / SSIM scores

show that our data improves generalization and prevents over-fitting. The model trained on the full Holopix50k training set outperforms all other models, further validating the need for such a large-scale stereo image dataset.

Table 2. PSNR (dB) and SSIM values from training PASSRNet [56] on various datasets for 4x SR with 80 training epochs

Dataset	KITTI2015 (Test)	Middlebury2014 (Test)	Flickr1024 (Test)	ETH3D (Test)	Holopix50k (Test)
Flickr1024 (Train)	25.62 / 0.79	36.54 / 0.94	**23.25 / 0.72**	31.94 / 0.88	26.96 / 0.79
Holopix50k (1K)	24.97 / 0.76	34.93 / 0.92	22.29 / 0.66	31.13 / 0.86	26.15 / 0.77
Holopix50k (5K)	25.58 / 0.79	36.01 / 0.93	22.83 / 0.70	32.06 / 0.88	27.10 / 0.80
Holopix50k (10K)	25.71 / 0.79	36.55 / 0.94	22.98 / 0.71	32.29 / 0.88	27.34 / 0.81
Holopix50k (Train)	**25.78 / 0.79**	**36.92 / 0.94**	23.13 / 0.71	**32.64 / 0.89**	**27.56 / 0.81**

Bold values indicate best scores, underlined values indicate second-best scores. For all metrics, higher values are better.

5.2 Self-supervised Monocular Depth Estimation

We use the Holopix50k dataset to fine-tune a self-supervised monocular depth estimation model, namely Monodepth2 [19]. The model is originally trained on the KITTI dataset [16] and predicts a dense depth map from a single image. We use the training code provided by the authors with minor modifications for compatibility with our dataset. To compute the image reconstruction loss for training, we warp images using disparity-based backward mapping [1] as we do not have camera parameters or pose information.

Fig. 6. Disparity maps produced by Monodepth2 [19] models on samples from the Middlebury [48] (Left) and MPI Sintel [4] (Right) datasets, respectively. (a) Trained on KITTI, (b) Trained on Holopix50k

Figure 6 shows qualitative results on images from the Middlebury and MPI Sintel datasets. The vanilla model produces blurry depth maps with only a few of the fine details from the original images intact. Results from fine-tuning on our dataset are sharper and maintain edges from the original images more consistently on both datasets.

原始模型预测模糊，该模型结果锐利

Table 3. Disparity comparison of running Monodepth2 [19] models on the Middlebury visual benchmark and MPI Sintel test tests, respectively

Dataset	Middlebury Test				MPI Sintel Test			
	Abs Rel	Sq Rel	RMSE	RMSE log	Abs Rel	Sq Rel	RMSE	RMSE log
KITTI	0.590	15.968	20.891	0.308	0.337	31.726	41.597	0.173
Holopix50k (1K) FT	0.421	7.528	14.224	0.203	0.239	17.062	25.882	0.114
Holopix50k (5K) FT	0.420	7.387	14.136	0.202	**0.235**	**16.474**	**25.300**	0.114
Holopix50k (10K) FT	**0.416**	**7.343**	**14.100**	**0.201**	0.237	17.265	25.668	0.114
Holopix50k (Train) FT	0.417	7.373	14.112	0.201	0.235	16.810	25.302	**0.113**

Bold values indicate best scores, underlined values indicate second-best scores. For all metrics, lower values are better. *FT* refers to fine-tuned model originally trained on the KITTI dataset.

Table 3 shows the quantitative results of running the original model and multiple fine-tuned models on the Middlebury and MPI Sintel test sets. We compare the disparity value by applying a per-image normalization and median scaling [62]. On the Middlebury dataset, our models perform better in all four error metrics. Specifically, the relative squared error is less than half of the original model, even with using 1k images from our dataset. On the MPI Sintel test set, the metrics are distributed between our models, with all the errors being significantly lower than the original KITTI model. From these results, we can conclude that our dataset helps a model trained on road scenes (from KITTI) generalize to entirely new scenarios that are also not present in our dataset.

5.3 Disparity Models Trained using Holopix50k

In this section, we briefly describe models we have trained using the Holopix50k dataset for different disparity estimation tasks and share qualitative results. Note that we do not go into the implementation details of any of the networks or their training procedures. This section aims to motivate a variety of practical use cases and applications.

Stereo Disparity Estimation. We use the Holopix50k dataset to train a stereo disparity estimation network for mobile inference. We jointly train a forward warping network as a part of our training pipeline. We use the disparity maps generated via semi-global block-matching [22] with additional filtering as supervised input during training. Our network follows a U-Net like architecture [45] and has a computation footprint of ~340k parameters (~1.5 GFLOPS). When using our Holopix50k dataset for this task, we observe that results have sharp edge detail (See Figure 7) and maintain stereo consistency. Our network also generalizes very well to a variety of real-world as well as synthetic scenarios. Images uploaded to Holopix are passed through this network to generate disparity maps. The final multi-view imagery users see on Holopix is synthesized using a combination of the original images and the generated disparity.

用SGBM+滤波作为监督

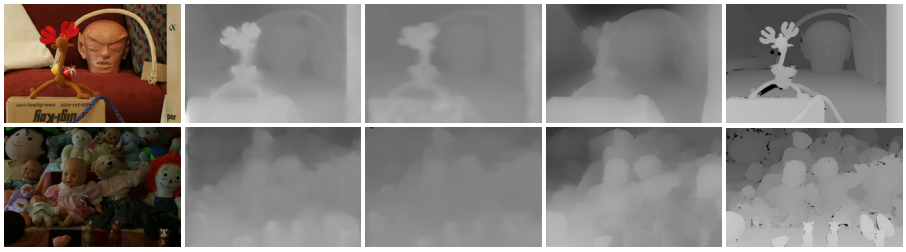


Fig. 7. Left image disparity maps produced by our models on samples from the Middlebury dataset [48]. (a) shows left disparity from our stereo disparity model, (b) shows result of our real-time disparity estimation, and (c) shows relative depth from our monocular depth estimation model a立体视差模型，b实时视差估计，c单目估计相对深度

Real-time Disparity Estimation. We follow the same training pipeline and methodology as our stereo disparity estimation network to train a real-time disparity estimation network. The real-time network generates a single disparity map corresponding to both the left and right images. This network is optimized to have a computation footprint of ∼15k parameters (∼0.15 GFLOPS). Though results are significantly smoother and have fewer edge details as seen in Figure 7, user testing demonstrates its feasibility for real-time mobile use-cases such as Lightfield camera previews, video calling, and video playback.

Monocular Depth Estimation. We also train a monocular depth estimation network that is semi-supervised using a combination of the Holopix50k and Megadepth datasets [32]. The dense depth maps required in training are pseudo-labeled using our stereo neural network and the Megadepth model, respectively. We model this as a generative image translation task conditioned by stereo depth information using an architecture similar to Pix2pix [24] with the PatchGAN discriminator [24]. When using only the Megadepth dataset, we observe the model does not perform well on close-up to mid-range scenes, and scenes containing human-like subjects. By using our sufficiently large and diverse data set, our model learns to generate relative depths quite well from far away to close-up scenes, which has proven to be useful in a variety of practical scenarios. From Figure 7, we can see that the model correctly identifies the various depth layers present in the scene without any stereo supervision.

半监督训练单目深度估计模型，用GAN，只用MegaDepth对近到中距离场景和人像场景不好，该模型对近距离很好

6 Future Work

The Holopix50k dataset is our first release of a large-scale dataset scraped from the Holopix platform. As the platform continues to grow and increase its user-base, we plan to continue supporting the dataset with multiple future iterations. We also plan to share more details of the models trained using the Holopix50k

dataset that we briefly cover in Section 5.2. Finally, since we have used the output disparity maps from our stereo disparity network as a reference for filtering the Holopix50k dataset, releasing these disparity maps as a pseudo-labeled ground truth dense disparity is another direction we can consider, based on the interest from the research community.

7 Conclusion

In this work, we present *Holopix50k*, a large-scale in-the-wild stereo image dataset that contains a variety of diverse scene types and semantic image content. To the best of our knowledge, this is the largest publicly released stereo image dataset collected from a mobile social media application. We showcase the superiority of our dataset, as a whole as well as its subsets, compared to other stereo datasets. We also demonstrate how using our dataset improves the performance of stereo image tasks, such as stereo super-resolution and self-supervised monocular depth estimation. We also briefly describe the models we trained using this dataset and the tasks they perform. We hope that this dataset encourages more work in the field of in-the-wild stereo imagery, both in areas we have discussed, as well as new research areas that are yet to be explored.

References

1. Beier, T., Neely, S.: Feature-based image metamorphosis. In: Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques. p. 3542. SIGGRAPH 92, Association for Computing Machinery, New York, NY, USA (1992). https://doi.org/10.1145/133994.134003, https://doi.org/10.1145/133994.134003
2. Bhoi, A.: Monocular depth estimation: A survey. CoRR **abs/1901.09402** (2019), http://arxiv.org/abs/1901.09402
3. de Boor, C.: Bicubic spline interpolation. Journal of Mathematics and Physics **41**(1-4), 212–218 (1962). https://doi.org/10.1002/sapm1962411212, https://onlinelibrary.wiley.com/doi/abs/10.1002/sapm1962411212
4. Butler, D., Wulff, J., Stanley, G., Black, M.: Mpi-sintel optical flow benchmark: Supplemental material. In: MPI-IS-TR-006, MPI for Intelligent Systems (2012. Citeseer (2012)
5. Chen, D., Yuan, L., Liao, J., Yu, N., Hua, G.: Stereoscopic neural style transfer. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6654–6663 (June 2018). https://doi.org/10.1109/CVPR.2018.00696
6. Chen, X., Zhang, Q., Lin, M., Yang, G., He, C.: No-reference color image quality assessment: from entropy to perceptual quality. EURASIP Journal on Image and Video Processing **2019** (12 2019). https://doi.org/10.1186/s13640-019-0479-7
7. Chen, Y., Wang, J., Chen, X., Zhu, M., Yang, K., Wang, Z., Xia, R.: Single-image super-resolution algorithm based on structural self-similarity and deformation block features. IEEE Access **7**, 58791–58801 (2019). https://doi.org/10.1109/ACCESS.2019.2911892
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
9. Dagobert, T.: The production of ground truths for evaluating highly accurate stereovision algorithms. Image Processing On Line **8**, 1–23 (2018)
10. Dodgson, N.A., Moore, J., Lang, S.: Multi-view autostereoscopic 3d display. In: International Broadcasting Convention. vol. 2, pp. 497–502. Citeseer (1999)
11. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2), 295–307 (Feb 2016)
12. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. pp. 2366–2374 (2014)
13. Esmaeilzehi, A., Ahmad, M.O., Swamy, M.N.S.: Srsubbandnet: A new deep learning scheme for single image super resolution based on subband reconstruction. In: 2019 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1–5 (May 2019). https://doi.org/10.1109/ISCAS.2019.8702351
14. Fattal, D., Peng, Z., Tran, T., Vo, S., Fiorentino, M., Brug, J., Beausoleil, R.G.: A multi-directional backlight for a wide-angle, glasses-free three-dimensional display. Nature **495**(7441), 348–351 (2013). https://doi.org/10.1038/nature11972, https://doi.org/10.1038/nature11972
15. Foroosh, H., Zerubia, J.B., Berthod, M.: Extension of phase correlation to subpixel registration. IEEE Transactions on Image Processing **11**(3), 188–200 (March 2002). https://doi.org/10.1109/83.988953

16. Geiger, A.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 33543361. CVPR 12, IEEE Computer Society, USA (2012)
17. Gennery, D.B.: A stereo vision system for an autonomous vehicle. In: Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 2. p. 576582. IJCAI77, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1977)
18. Godard, C., Aodha, O., Brostow, G.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (07 2017). https://doi.org/10.1109/CVPR.2017.699
19. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth prediction. The International Conference on Computer Vision (ICCV) (October 2019)
20. Gong, X., Huang, H., Ma, L., Shen, F., Liu, W., Zhang, T.: Neural stereoscopic image style transfer. CoRR **abs/1802.09985** (2018), `http://arxiv.org/abs/1802.09985`
21. He, K., Gkioxari, G., Dollr, P., Girshick, R.: Mask r-cnn. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (Oct 2017). https://doi.org/10.1109/ICCV.2017.322
22. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(2), 328–341 (Feb 2008). https://doi.org/10.1109/TPAMI.2007.1166
23. Howard, I.P., Rogers, B.J., et al.: Binocular vision and stereopsis. Oxford University Press, USA (1995)
24. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CoRR **abs/1611.07004** (2016), `http://arxiv.org/abs/1611.07004`
25. Jeon, D.S., Baek, S.H., Choi, I., Kim, M.H.: Enhancing the spatial resolution of stereo images using a parallax prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1721–1730 (2018)
26. Keys, R.: Cubic convolution interpolation for digital image processing. IEEE Transactions on Acoustics, Speech, and Signal Processing **29**(6), 1153–1160 (December 1981). https://doi.org/10.1109/TASSP.1981.1163711
27. Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J., Izadi, S.: Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 573–590 (2018)
28. Kohli, P., Gunaseelan, S., Orozco, J., Hua, Y., Li, E., Dahlquist, N.: Gpu-accelerated mobile multi-view style transfer (2020)
29. Kopf, J., Alsisan, S., Ge, F., Chong, Y., Matzen, K., Quigley, O., Patterson, J., Tirado, J., Wu, S., Cohen, M.F.: Practical 3d photography (2019)
30. Kriegman, D.J., Triendl, E., Binford, T.O.: Stereo vision and navigation in buildings for mobile robots. IEEE Transactions on Robotics and Automation **5**(6), 792–803 (Dec 1989). https://doi.org/10.1109/70.88100
31. Li, J., You, S., Robles-Kelly, A.: Stereo super-resolution via a deep convolutional network. In: 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1–7 (Nov 2017). https://doi.org/10.1109/DICTA.2017.8227492
32. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. CoRR **abs/1804.00607** (2018), `http://arxiv.org/abs/1804.00607`

33. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)
34. Ma, C., Yang, C.Y., Yang, X., Yang, M.H.: Learning a no-reference quality metric for single-image super-resolution. Computer Vision and Image Understanding (12 2016). https://doi.org/10.1016/j.cviu.2016.12.009
35. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2016), `http://lmb.informatik.uni-freiburg.de/Publications/2016/MIFDB16`, arXiv:1512.02134
36. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4040–4048 (2016)
37. Mehta, I., Sakurikar, P., Narayanan, P.J.: Structured adversarial training for unsupervised monocular depth estimation. In: 2018 International Conference on 3D Vision (3DV). pp. 314–323 (Sep 2018). https://doi.org/10.1109/3DV.2018.00044
38. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
39. Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR). pp. 723–727 (Nov 2011). https://doi.org/10.1109/ACSSC.2011.6190099
40. Murray, D., Little, J.: Using real-time stereo vision for mobile robot navigation. Auton. Robots **8**, 161–171 (04 2000). https://doi.org/10.1023/A:1008987612352
41. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
42. Niklaus, S., Mai, L., Yang, J., Liu, F.: 3d ken burns effect from a single image. ACM Transactions on Graphics (TOG) **38**(6), 1–15 (2019)
43. Pang, J., Sun, W., Yang, C., Ren, J., Xiao, R., Zeng, J., Lin, L.: Zoom and learn: Generalizing deep stereo matching to novel domains. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
44. Ping An, Chaohui Lu, Zhaoyang Zhang: Object segmentation using stereo images. In: 2004 International Conference on Communications, Circuits and Systems (IEEE Cat. No.04EX914). vol. 1, pp. 534–538 Vol.1 (June 2004). https://doi.org/10.1109/ICCCAS.2004.1346183
45. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
46. Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S., Chellappa, R.: Unsupervised domain adaptation for semantic segmentation with gans. CoRR **abs/1711.06969** (2017), `http://arxiv.org/abs/1711.06969`
47. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. IEEE transactions on pattern analysis and machine intelligence **31**(5), 824–840 (2008)
48. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: Jiang, X., Hornegger, J., Koch, R. (eds.) Pattern Recognition. pp. 31–42. Springer International Publishing, Cham (2014)

49. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
50. Schuster, R., Bailer, C., Wasenmüller, O., Stricker, D.: Combining stereo disparity and optical flow for basic scene flow. CoRR **abs/1801.04720** (2018), `http://arxiv.org/abs/1801.04720`
51. Stone, H.S., Orchard, M.T., Ee-Chien Chang, Martucci, S.A.: A fast direct fourier-based algorithm for subpixel registration of images. IEEE Transactions on Geoscience and Remote Sensing **39**(10), 2235–2243 (Oct 2001). https://doi.org/10.1109/36.957286
52. Tonioni, A., Poggi, M., Mattoccia, S., Di Stefano, L.: Unsupervised adaptation for deep stereo. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
53. Tonioni, A., Tosi, F., Poggi, M., Mattoccia, S., Di Stefano, L.: Real-time self-adaptive deep stereo. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
54. Tsai, D.Y., Lee, Y., Matsuyama, E.: Information entropy measure for evaluation of image quality. Journal of Digital Imaging **21**, 338–347 (2007)
55. Wang, C., Lucey, S., Perazzi, F., Wang, O.: Web stereo video supervision for depth prediction from dynamic scenes. In: 2019 International Conference on 3D Vision (3DV). pp. 348–357. IEEE (2019)
56. Wang, L., Wang, Y., Liang, Z., Lin, Z., Yang, J., An, W., Guo, Y.: Learning parallax attention for stereo image super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
57. Wang, Y., Wang, L., Yang, J., An, W., Guo, Y.: Flickr1024: A large-scale dataset for stereo image super-resolution. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
58. Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U., Cremers, D.: Stereoscopic scene flow computation for 3d motion understanding. International Journal of Computer Vision **95**, 29–51 (10 2011). https://doi.org/10.1007/s11263-010-0404-0
59. Yin, Z., Darrell, T., Yu, F.: Hierarchical discrete distribution decomposition for match density estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
60. Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: 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 (2019)
61. Zhang, Y., Qiu, W., Chen, Q., Hu, X., Yuille, A.: Unrealstereo: Controlling hazardous factors to analyze stereo vision. In: 2018 International Conference on 3D Vision (3DV). pp. 228–237. IEEE (2018)
62. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1851–1858 (2017)

Supplementary Material

Holopix50k: A Large-Scale In-the-wild Stereo Image Dataset

Yiwen Hua, Puneet Kohli, Pritish Uplavikar*, Anand Ravi*, Saravana Gunaseelan, Jason Orozco, and Edward Li

Leia Inc.

Abstract. This supplementary material adds discussions and results that were not included in the main paper. We additionally show results from experiments that were omitted from the main paper to ensure brevity.

1 About Holopix

Holopix was created in 2018 by Leia Inc. as a Lightfield image-sharing social media platform. It was launched along with the Hydrogen One Lightfield mobile device, created in partnership with the industrial camera manufacturer RED and pre-loaded on these devices. The user experience of this app is similar to other mobile photo-sharing platforms, such as Instagram, where users upload photos to the platform for others to view and interact within feeds. The app curates the top images to showcase to other users on the platform. The key difference compared to traditional image-sharing platforms is that on Holopix, the images uploaded are multi-view images or are converted to multi-view images by the Holopix platform. The multi-view images allow for both stereoscopic depth and a parallax effect depending on the user's device and viewing method. The conversion to multi-view is usually done by disparity estimation, followed by novel view-synthesis, and finally, image in-painting.

2 Data Filtering

2.1 Vertical Disparity

As our dataset is collected with cameras that are horizontally aligned, the stereo pairs are expected only to have horizontal disparity. Hence images with even slight vertical disparity can cause severe failures in matching algorithms such as stereo block-matching. Figure 1 shows the results of applying a horizontal stereo block-matching algorithm to an image pair with and without vertical disparity. As it is visible, introducing vertical disparity causes complete failure of the algorithm.

* indicates equal contribution

The reason for having vertical disparity could range from a variety of factors such as physically misaligned cameras, calibration errors, or synchronization issues. We filter out stereo pairs with any vertical disparity by calculating the vertical phase correlation between the left and right image pair.

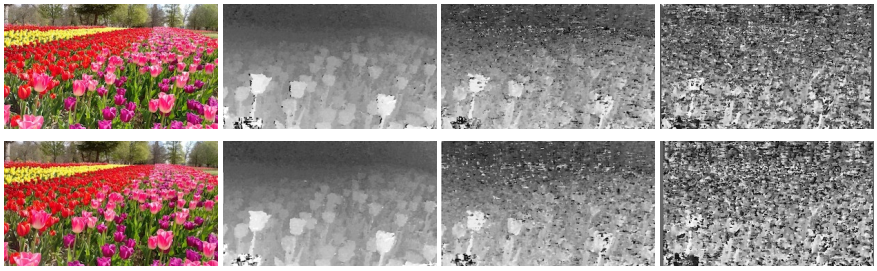


Fig. 1. (a) Input stereo pair (b) Disparity maps calculated using stereo block-matching (c) Disparity maps calculated with 2 pixels of vertical offset (d) Disparity maps calculated with 5 pixels of vertical offset. Top row corresponds to the left image and bottom row corresponds to the right image.

2.2 Disparity-based Filtering

We filtered the dataset based on the disparity maps generated from our stereo neural network. In cases such as extremely far-away scenes, where there is no stereo difference between left and right images, we would see apparent artifacts in the generated disparity map. There are other cases where an object is extremely close to the stereo cameras, causing significant visual differences in the stereo pair. A typical example of this is a photographer's finger partially covering the cameras. Additionally, as users on Holopix are free to upload stereo pairs from any source, there are also a variety of computer-generated pairs that do not exhibit stereo characteristics, or completely invalid stereo pairs where the left and right image do not match. These pairs are also flagged in our disparity-based filtering process and subsequently removed from the dataset. Note that the process involves human inspection of the stereo pairs and the disparity images to determine what should be filtered out. We also applied a low-standard deviation filter to filter out flat scenes that are generally texture-less or monochromatic. Figure 2 shows some examples of filtered stereo pairs and the calculated disparity map that was used to determine that these should be filtered out.

2.3 Dataset Indexes

We release 49,368 image pairs, which are indexed from posts on Holopix. If a user decides to delete their post from Holopix, the previously indexed image will

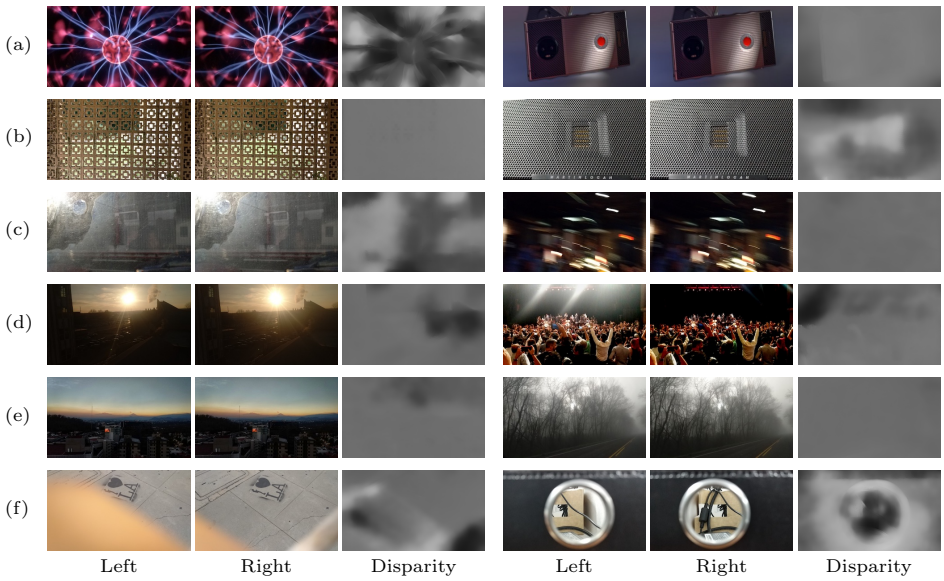


Fig. 2. Examples of images filtered out from the Holopix50k dataset. The original stereo pair and the left disparity image generated by our stereo disparity model is shown. (a) Computer-generated images, (b) Images that are flat or have repeating textures, (c) Images with strong lens flare or motion blur, (d) images with significant lighting differences, (e) images of far-away scenes, (f) images of extremely close objects

no longer be available as a part of our dataset. As a result, the available portion of the dataset may shrink over time. This is similar to other datasets that are indexed from user-generated sources such as Flickr or Youtube.

3 Dataset Comparison

In addition to the image metric comparisons in the main paper, we provide additional dataset property comparisons in Table 1. Apart from being the largest stereo image dataset, we see that our dataset contains both indoor and outdoor in-the-wild scenes similar to Flickr1024 and WSVD. In contrast, the KITTI dataset contains only outdoor street scenes, and the Middlebury dataset contains indoor scenes in a laboratory setting. ETH3D contains varied scenes that consist of a limited number of indoor and outdoor imagery of buildings, parking lots, and some natural scenes. Our dataset also contains a mixture of both portrait and landscape orientation images unlike other datasets, except Flickr1024. Similar to the other in-the-wild stereo datasets, we do not provide ground truth disparity.

Table 2 shows the metric scores for the randomly sampled train, test, and validation subsets of our datasets. It can be observed that all the subsets show sim-

Table 1. Comparison of properties of popular stereo image datasets

Dataset	Image Pairs	Resolution	Ground Truth	Orientation	Setting
KITTI 2012	389	0.46 Mpx	✓	Landscape	Street scenes
KITTI 2015	400	0.47 Mpx	✓	Landscape	Street scenes
Middlebury	65	3.59 Mpx	✓	Landscape	Laboratory
ETH 3D	47	0.38 Mpx	✓	Landscape	Varied
Flickr1024	1024	0.73 Mpx		Landscape + Portrait	In-the-wild
WSVD	10250	0.77 Mpx		Landscape	In-the-wild
Holopix50k	49368	0.74 Mpx		Landscape + Portrait	In-the-wild

Table 2. Mean and standard deviation of main characteristics of Holopix50k train, test and validation subsets

Dataset	Image Pairs	Resolution (↑)	Entropy (↑)	BRISQUE (↓)	SR-metric (↑)	ENIQA (↓)
Holopix50k	49368	0.74 (±0.30) Mpx	7.38 (±0.48)	19.49 (±11.81)	7.28[1](±1.24)	0.174 (±0.125)
Holopix50k (Train)	41954	0.74 (±0.30) Mpx	7.38 (±0.48)	19.49 (±11.83)	7.27[1](±1.24)	0.175 (±0.125)
Holopix50k (Test)	2471	0.73 (±0.31) Mpx	7.38 (±0.50)	19.52 (±11.94)	7.30[1](±1.25)	0.170 (±0.123)
Holopix50k (Val)	4943	0.74 (±0.31) Mpx	7.37 (±0.48)	19.46 (±11.64)	7.29[1](±1.23)	0.175 (±0.124)

For all metrics, higher values are better, except BRISQUE and ENIQA.
[1] Only *left* images from Holopix HD were used to compute SR-metric due to computation constraints. We assume trivial perceptual differences between *left* and *right* images.

ilar performance across all the metrics. Additionally, these scores are all nearly identical to the dataset as a whole, indicating a negligible selection bias.

4 Stereo Super-Resolution

In addition to training PASSRNet models from scratch with different amounts of samples from our Holopix50k dataset, we also fine-tuned the model pre-trained on the Flickr1024 dataset. The PSNR and SSIM values of these experiments are reported in Table 3. In contrast to the from-scratch results reported in the main paper, we see that with the pre-trained weights, even a small sample (1k) of our dataset boosts performance on a variety of test sets. When using the entire dataset, the performance of the fine-tuned and the from-scratch models are quite similar.

Figure 3 graphically shows the change in PSNR and SSIM scores with respect to the number of training images, across our different PASSRNet models, as well as the original model trained on Flickr1024. We note that fine-tuning generally

Table 3. PSNR (dB) and SSIM values from training PASSRNet on various datasets for 4x SR with 80 training epochs

Dataset	KITTI2015 (Test)	Middlebury (Test)	Flickr1024 (Test)	ETH3D (Test)	Holopix50k (Test)
Flickr1024 (Train)	25.62 / 0.79	36.55 / 0.94	**23.25 / 0.72**	31.94 / 0.88	26.96 / 0.79
Holopix50k (1K) FT	25.46 / 0.78	36.25 / 0.94	22.93 / 0.70	32.10 / 0.88	27.14 / 0.80
Holopix50k (5K) FT	25.66 / 0.79	36.58 / 0.94	23.05 / 0.71	32.29 / 0.89	27.38 / 0.81
Holopix50k (10K) FT	25.68 / 0.79	36.65 / 0.94	23.09 / 0.71	32.37 / 0.89	27.42 / 0.81
Holopix50k (Train) FT	25.60 / 0.79	**36.86 / 0.94**	23.15 / 0.71	32.45 / 0.89	**27.58 / 0.81**
Holopix50k (Train)	**25.80 / 0.80**	**36.86 / 0.94**	23.10 / 0.71	**32.56 / 0.89**	27.51 / 0.81

Note that *FT* above stands for Fine-tuned. **Bold** values indicate best scores, underlined values indicate second best scores.

produces better results for a smaller number of samples. As we increase the training data, the fine-tune and from-scratch models converge. As expected, using more training data improves results across the board.

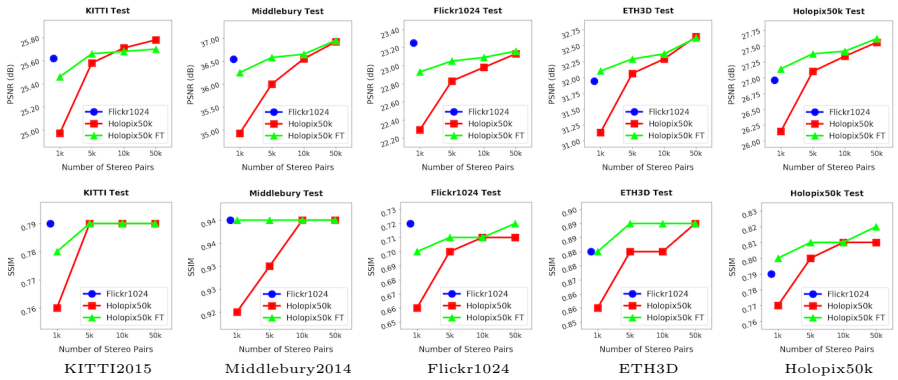
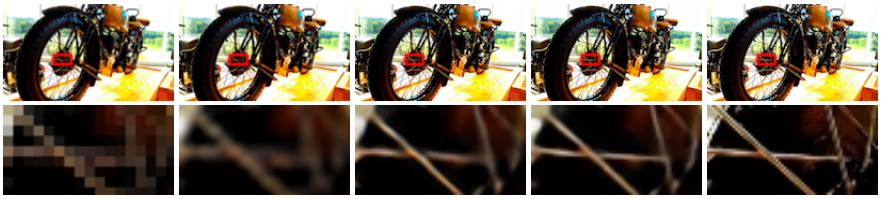


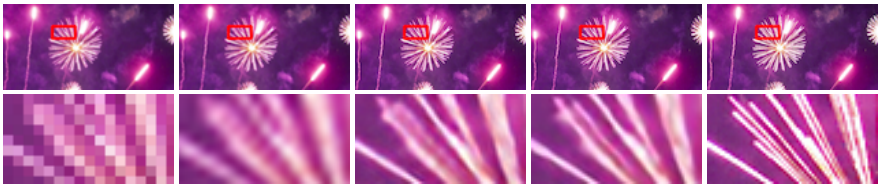
Fig. 3. PSNR and SSIM values for 4x SR tasks obtained by PASSRNet models trained on different number of stereo pairs

In Figure 4, we show additional qualitative results of the PASSRNet model trained on Holopix50k, compared to other super-resolution methods. In-line with the results we show in the main paper, we observe sharper details and higher quality textures in samples super-resolved with our model.

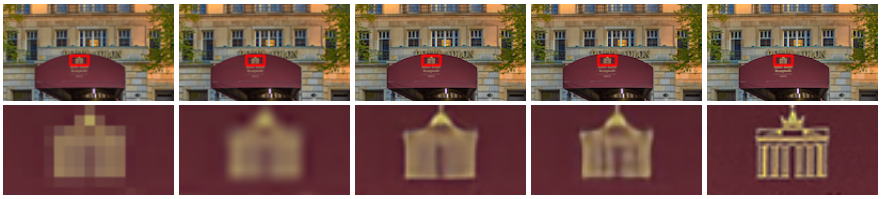


Input Holopix50K	Bicubic 19.59/0.69	Flickr1024* 21.74/0.79	Holopix50k* 22.32/0.82	Ground Truth

Input Holopix50K	Bicubic 24.92/0.78	Flickr1024* 27.45/0.85	Holopix50k* 28.44/0.87	Ground Truth



Input Flickr1024	Bicubic 26.55/0.86	Flickr1024* 27.47/0.87	Holopix50k* 27.79/0.88	Ground Truth



Input Flickr1024	Bicubic 21.35/0.58	Flickr1024* 23.06/0.70	Holopix50k* 22.80/0.70	Ground Truth

Input KITTI2015	Bicubic 25.72/0.82	Flickr1024* 27.30/0.82	Holopix50k* 27.38/0.82	Ground Truth

* Refers to PASSRNet trained on the respective datasets

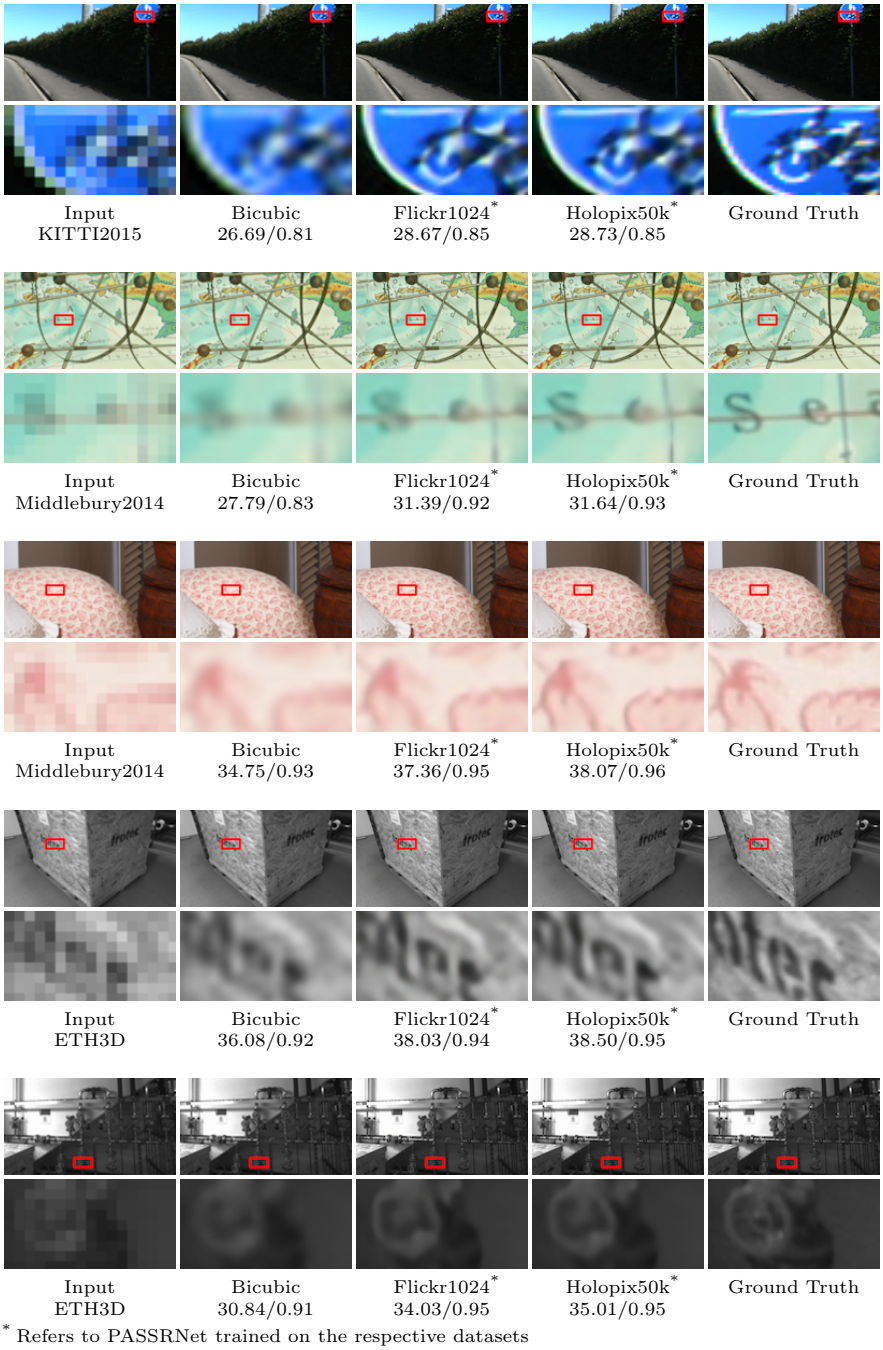


Fig. 4. Additional results of SR methods for 4x SR on samples from a variety of datasets.

5 Self-Supervised Depth Estimation

5.1 Implementation Details

In the original implementation of Monodepth2, the authors use a reprojection-based image warping technique, which requires camera pose information. As the stereo pairs in our dataset are already rectified, we modify the image warping to use a disparity-based backward warping technique instead, which does not require pose information. We additionally skip the disparity to depth conversion during training as we do not have ground truth depth to bound our depth scale. Finally, we change the smoothness term λ to 1.0 as we empirically found that it generates the best results for our experiments.

5.2 Additional Results

We report the fine-tune results of training the Monodepth2 model in our main paper, as we focus on the generalization our dataset empowers. Specifically, in the self-supervised case, the generalization of a model trained on street scenes (KITTI) to laboratory (Middlebury), and synthetic (MPI Sintel) scenes.

Table 4. Disparity comparison of running Monodepth2 models on the Middlebury visual benchmark and MPI Sintel test tests, respectively

Dataset	Middlebury Test				MPI Sintel Test			
	Abs Rel	Sq Rel	RMSE	RMSE log	Abs Rel	Sq Rel	RMSE	RMSE log
KITTI	0.590	15.968	20.891	0.308	0.337	31.726	41.597	0.173
Holopix50k (1K)	0.420	7.502	14.268	0.23	0.236	17.040	**25.279**	0.113
Holopix50k (5K)	0.422	7.585	14.302	0.204	0.240	18.075	25.747	0.114
Holopix50k (10K)	0.420	7.501	14.224	0.202	0.236	<u>17.018</u>	25.335	0.113
Holopix50k (Train)	**0.414**	**7.269**	**14.032**	**0.200**	<u>0.235</u>	17.155	25.357	**0.113**
Holopix50k (Train) FT	<u>0.417</u>	<u>7.373</u>	<u>14.112</u>	<u>0.201</u>	**0.235**	**16.810**	<u>25.302</u>	<u>0.113</u>

Bold values indicate best scores, <u>underlined</u> values indicate second-best scores. For all metrics, lower values are better. *FT* refers to fine-tuned model originally trained on the KITTI dataset.

In addition to fine-tuning the Monodepth2 model that was originally trained on the KITTI dataset, we also train the model from scratch with different amounts of samples from our dataset. The results are reported in Table 4. Interestingly, we observe that the model does not always perform better with the increase in the amount of training data. For example, the 1k model has better results than the 5k model. We believe this is due to the challenging images our dataset contains for the monocular depth estimation task, and more data is required for the model to learn the distribution of our data. We also observe that

the from-scratch models out-perform the fine-tune models. We theorize this is because of the difference in distribution between KITTI and Holopix50k datasets, and perhaps the pre-trained model serves as a less-than-ideal initialization for the Holopix50k data.

Fig. 5. Disparity maps produced by Monodepth2 models on samples from the Holopix50k test set. (a) Input image, (b) Trained on KITTI, (c) Trained on Holopix50k

Figure 5 shows additional qualitative results of the Monodepth2 models on samples from the Holopix50k test set. Similar to the results from Middlebury and MPI Sintel test sets, the results are sharper and respect edges better when trained on Holopix50k. We do not show ground truth disparities as we do not have the ground truth disparities for our dataset.

6 More Disparity Results

In Figure 6 we show additional results from running our various disparity estimation models. Note that strong light reflections are modeled as Lambertian effects. Hence disparity maps should display far back (dark) values for light reflections.

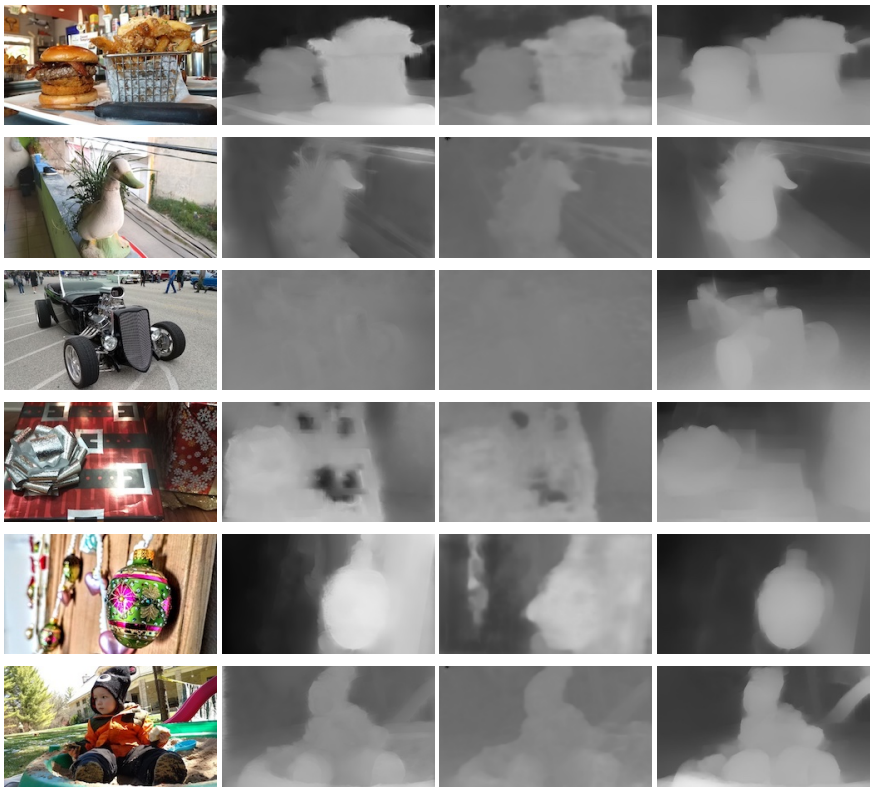


Fig. 6. Left image disparity maps produced by our models on samples from Holopix50k. (a) shows left disparity from our stereo disparity model, (b) shows result of our real-time disparity estimation, and (c) shows relative depth from our monocular depth estimation model

7 Dataset Examples

In Figure 7 we show sample stereo pairs present in the Holopix50k dataset.

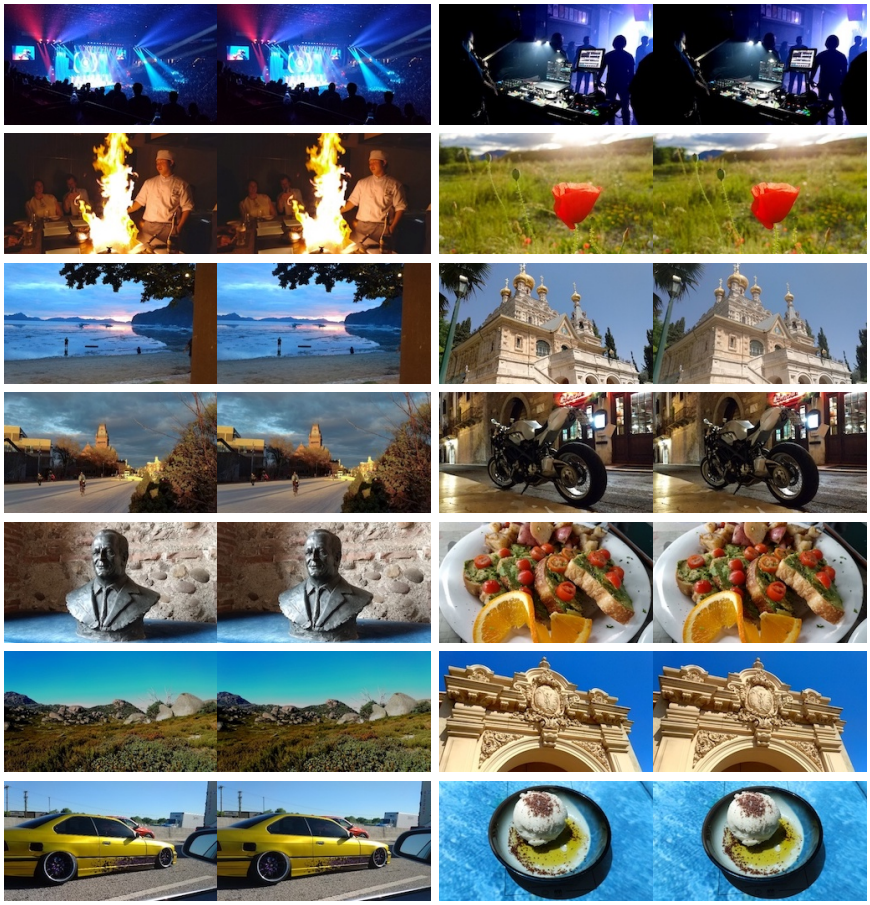


Fig. 7. Stereo images present in the Holopix50k dataset