MIPI 2023 Challenge on RGB+ToF Depth Completion: Methods and Results

Qingpeng Zhu1 Wenxiu Sun1 Yuekun Dai2 Chongyi Li2 Shangchen Zhou2 Ruicheng Feng2 Qianhui Sun3 Chen Change Loy2 Jinwei Gu1 Yi Yu1 Yangke Huang Kang Zhang Meiya Chen1 Yu Wang1 Yongchao Li1 Hao Jiang1 Amrit Kumar Muduli1 Vikash Kumar2 Kunal Swami1 Pankaj Kumar Bajpai1 Yunchao Ma1 Jiajun Xiao1 Zhi Ling1

Abstract

Depth completion from RGB images and sparse Time-of-Flight (ToF) measurements is an important problem in computer vision and robotics. While traditional methods for depth completion have relied on stereo vision or structured light techniques, recent advances in deep learning have enabled more accurate and efficient completion of depth maps from RGB images and sparse ToF measurements. To evaluate the performance of different depth completion methods, we organized an RGB+ sparse ToF depth completion competition. The competition aimed to encourage research in this area by providing a standardized dataset and evaluation metrics to compare the accuracy of different approaches. In this report, we present the results of the competition and analyze the strengths and weaknesses of the top-performing methods. We also discuss the implications of our findings for future research in RGB+sparse ToF depth completion. We hope that this competition and report will help to advance the state-of-the-art in this important area of research. More details of this challenge and the link to the dataset can be found at https://mipi-challenge.org/MIPI2023/.

1. Introduction

RGB+ sparse ToF depth completion is a novel approach to depth estimation in computer vision that combines RGB images with sparse depth measurements obtained from time-of-flight (ToF) sensors. Depth estimation is a fundamental problem in computer vision with numerous applications, such as robotics, autonomous driving, and augmented reality. However, estimating depth from a single modality can be challenging due to noise, occlusions, and other factors. RGB+sparse ToF depth completion has the potential to improve the accuracy and robustness of depth estimation in various real-world scenarios.

The task involves predicting a dense depth map from a single RGB image and a sparse set of depth measurements. By combining RGB images and sparse depth measurements, RGB+ sparse ToF depth completion aims to leverage the complementary information provided by both modalities to produce accurate and detailed depth maps. RGB images capture color information, while the sparse depth measurements provide direct distance information about the scene. This allows for more robust and accurate depth estimation, particularly in challenging scenarios such as low-light conditions, texture-less regions, and scenes with reflective surfaces.

This challenge is a part of the Mobile Intelligent Photography and Imaging (MIPI) 2023 workshop and challenges that emphasize the integration of novel image sensors and imaging algorithms, which is held in conjunction with CVPR 2023. It consists of four competition tracks:

- **Nighttime Flare Removal** is to improve nighttime image quality by removing lens flare effects.
- **RGB+ToF Depth Completion** uses sparse, noisy ToF depth measurements with RGB images to obtain a complete depth map.
- **RGBW Sensor Re-mosaic** converts RGBW RAW data into Bayer format so that it can be processed with standard ISPs.
- **RGBW Sensor Fusion** fuses Bayer data and

---

1 Qingpeng Zhu1 (zh qingpeng@tetras.ai), Wenxiu Sun1 (sunwx@tetras.ai), Yuekun Dai2, Chongyi Li2, Shangchen Zhou2, Ruicheng Feng2, Qianhui Sun3, Chen Change Loy2, Jinwei Gu1, Yi Yu1, Yangke Huang Kang Zhang Meiya Chen1 Yu Wang1 Yongchao Li1 Hao Jiang1 Amrit Kumar Muduli1 Vikash Kumar2 Kunal Swami1 Pankaj Kumar Bajpai1 Yunchao Ma1 Jiajun Xiao1 Zhi Ling1

2863
monochrome channel data into Bayer format to increase SNR and spatial resolution.

2. Challenge

2.1. Problem Definition

Depth completion [15, 9, 11, 13, 5, 14, 18, 3, 19, 20, 6, 12, 4, 2] aims to recover dense depth from sparse depth measurements. Earlier methods concentrate on retrieving dense depth maps only from the sparse ones. However, these approaches are limited and not able to recover depth details and semantic information without the availability of multi-modal data. In this challenge, we focus on the RGB+ToF sensor fusion, where a pre-aligned RGB image is also available as guidance.

2.2. Dataset: TetrasRGBD [22]

The training data contains 7 image sequences of aligned RGB and ground-truth dense depth from 7 indoor scenes (20,000 pairs of RGB and depth in total). For each scene, the RGB and the ground-truth depth are rendered along a smooth trajectory in our created 3D virtual environment. RGB and dense depth images in the training set have a resolution of 640×480 pixels. We also provide a function to simulate the sparse depth maps that are close to the real sensor measurements. A visualization of an example frame of RGB, ground-truth depth, and simulated sparse depth is shown in Fig. 1.

The testing data contains, a) Synthetic: a synthetic image sequence (500 pairs of RGB and depth in total) rendered from an indoor virtual environment that differs from the training data; b) iPhone dynamic: 24 image sequences of dynamic scenes collected from an iPhone 12Pro (600 pairs of RGB and depth in total); c) iPhone static: 24 image sequences of static scenes collected from an iPhone 12Pro (600 pairs of RGB and depth in total); d) Modified phone static: 24 image sequences of static scenes (600 pairs of RGB and depth in total) collected from a modified phone. Please note that depth noises, missing depth values in low reflectance regions, and mismatch of the field of views between RGB and ToF cameras could be observed from this real data. RGB and dense depth images in the entire testing set have a resolution of 256×192 pixels. RGB and spot depth data from the testing set are provided and the GT depth is not available to participants. The depth data in both training and testing sets are in meters.

2.3. Challenge Phases

The challenge consisted of the following phases:

1. Development: The registered participants get access to the data and baseline code, and are able to train the models and evaluate their running time locally.

2. Validation: The participants can upload their models to the remote server to check the fidelity scores on the validation dataset, and to compare their results on the validation leaderboard.

3. Testing: The participants submit their final results, code, models, and factsheets.

2.4. Performance Evaluation

2.4.1 Objective Evaluation

We define the following metrics to evaluate the performance of depth completion algorithms.

- Relative Mean Absolute Error (RMAE), which measures the relative depth error between the completed depth and the ground truth, i.e.

\[
RMAE = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{|D(m, n) - \hat{D}(m, n)|}{D(m, n)},
\]

where \( M \) and \( N \) denote the height and width of depth, respectively. \( D \) and \( \hat{D} \) represent the ground-truth depth and the predicted depth, respectively.

- Edge Weighted Mean Absolute Error (EWMAE), which is a weighted average of absolute error. Regions with larger depth discontinuity are assigned higher weights. Similar to the idea of Gradient Conduction Mean Square Error (GCMSE) [17], EWMAE applies a weighting coefficient \( G(m, n) \) to the absolute error between pixel \((m, n)\) in ground-truth depth \( D \) and predicted depth \( \hat{D} \), i.e.

\[
EWMAE = \frac{1}{M \cdot N} \left[ \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} G(m, n) \cdot |D(m, n) - \hat{D}(m, n)|}{\sum_{m=1}^{M} \sum_{n=1}^{N} G(m, n)} \right],
\]

where the weight coefficient \( G \) is computed in the same way as in [17].
Figure 1. Visualization of an example training data. Left, middle, and right images correspond to RGB, ground-truth depth, and simulated spot depth data, respectively.

RMAE and EWMAE will be measured on the testing data with GT depth. We will rank the proposed algorithms according to the score calculated by the following formula, where the coefficients are designed to balance the values of different metrics,

\[
\text{Objective score} = 1 - 1.8 \times \text{RMAE} - 0.6 \times \text{EWMAE}.
\]  
(3)

For each dataset, we report the average results over all the processed images belonging to it.

2.4.2 Subjective Evaluation

For subjective evaluation, we adapt the commonly used Mean Opinion Score (MOS) with blind evaluation. The score is on a scale of 1 (bad) to 5 (excellent). We invited 13 expert observers to watch videos and give their subjective scores independently. The scores of all subjects are averaged as the final MOS.

2.5. Running Time Evaluation

The proposed algorithms are required to be able to process the RGB and sparse depth sequence in real time. Participants are required to include the average run time of one pair of RGB and depth data using their algorithms and the information in the device in the submitted readme file. Due to the difference of devices for evaluation, we set different requirements of running time for different types of devices according to the AI benchmark data from the website.

3. Challenge Results

Among 77 registered participants, 4 teams successfully submitted their results, code, and factsheets in the final test phase. Table 1 summarizes the final test results and rankings of the teams. Team MGTV shows the best overall performance, followed by Team MiMicAlgo and Team DIntel. The proposed solutions are described in Section 4 and the team members and affiliations are listed in Appendix A.

4. Challenge Methods

4.1. MGTV

Model. This team has proposed a model architecture, as illustrated in Figure 2, that incorporates ResNet34 [7] for the Encoder module, while the Decoder module is composed of a sequence of 5 upsample blocks, each of which integrates a convolutional layer, a pixel shuffle layer, and a channel attention branch [10]. The Local Network is a concatenation of three consecutive convolutional layers, while the Refined Network capitalizes on the FCSPNet [8] for optimal performance.

Loss. This team employed three loss functions jointly for training, namely L1 loss, RMAE loss, and Gradient loss proposed by [8]. In the early stage, in order to ensure the stability of model training, the L1 losses were calculated for global depth, local depth, and refined depth simultaneously. After the model training became stable, only the L1 loss for refined depth was calculated to obtain optimal results.

Data augmentation. This team observed that the sparse depth inputs in the testing data were not all sampled on a regular grid. During training, this team randomly applied local and circular removal operations to the sparse depth inputs with a certain probability, to simulate the input distribution of sparse depth data in the testing phase. This approach not only enhanced the generalization capability of the model but also improved its performance on such testing data.

Outlier handling. This team observed that the sparse depth inputs in the testing data have some out-

https://ai-benchmark.com/ranking/deeplearning_detailed.html
Table 1. MIPI 2023 RGB+ToF Depth Completion challenge results and final rankings.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team name</th>
<th>RMAE</th>
<th>EWMAE</th>
<th>Objective score</th>
<th>Subjective score</th>
<th>Final score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MGTV</td>
<td>0.02164</td>
<td>0.13257</td>
<td>0.88151</td>
<td>3.60577</td>
<td>0.80133</td>
</tr>
<tr>
<td>2</td>
<td>MiMcAlgo</td>
<td>0.01921</td>
<td>0.13462</td>
<td>0.88465</td>
<td>3.38846</td>
<td>0.78117</td>
</tr>
<tr>
<td>3</td>
<td>Dintel</td>
<td>0.02702</td>
<td>0.13569</td>
<td>0.86995</td>
<td>3.06731</td>
<td>0.74171</td>
</tr>
<tr>
<td>4</td>
<td>Chameleon</td>
<td>0.02232</td>
<td>0.13305</td>
<td>0.87999</td>
<td>2.90385</td>
<td>0.73038</td>
</tr>
</tbody>
</table>

liers which will introduce bias in the predictions around them. So this team removed the outliers beyond 3 standard deviations from the mean during the inference phase.

4.2. MiMcAlgo

Figure 3. Overview of EMNLP-TN.

Method. The team has developed an effective and efficient method for depth completion called Enhancing Multi-Scale Non-Local Propagation with Transformer Network (EMNLP-TN), illustrated in Figure 3.

Coarse-to-fine prediction: The proposed method, EMNLP-TN, employs a two-stage approach for depth completion, which shares similarities with the method outlined in penet [11]. However, EMNLP-TN utilizes a coarse-to-fine framework that addresses the complexity of depth completion in two sub-problems: structure prediction and detail refinement. In both stages, the same encoder and decoder network are utilized. Specifically, the second stage takes the coarse depth prediction from the first stage as input and applies several non-local spatial propagation networks [19] to enhance the prediction.

Artifacts-free up-sample: From the perspective of super-resolution, completion can also be seen as the process of interpolating missing depth information using RGB image guidance. This makes depth completion an equivalent process to depth super-resolution. To achieve this, the team used the pixel-shuffle[1] operation from super-resolution to bridge the gap between downscaled and full-sized depth data.

Self-attention: Depth completion is distinct from stereo depth because it does not rely on depth cues from the input. Instead, it leverages sparse depth information to

Figure 4. Bad point result.

4.3. DIntel

Figure 5. Overview of DIntel architecture

Model. This team uses a cascaded residual refinement approach for the depth completion task. In this

Training loss. Each branch is trained with

\[ \text{Loss} = \alpha \times \text{RMAE} + \beta \times \text{EWMAE}. \]  (4)

Specifically, the loss shares the same formulation as RMAE and EWMAE metrics.

The total loss of EMNLP-TN can be described as follow

\[ \text{Loss}_{\text{total}} = \text{Loss}_{\text{corase}} + \text{Loss}_{\text{fine}}. \]  (5)
cascaded style network architecture, the authors have used two network modules: Base Network and Residual Refinement Network (RRN), as shown in Figure 5. Inspired from [8], the base network’s architecture gives local and global depth predictions, which are then fused and passed through a few more layers to get the affinity matrix for funnel convolutional spatial propagation network (FCSPN) module, which in turn gives the initial refined depth map prediction. The output of the base network, along with RGB image and spot depth, is then fed as input to the RRN (a ResNet-34 [7] based architecture), whose objective is to rectify the incorrect depth predictions from the base network.

Training. The base network was trained similarly to [8]. After training the base network, the weights of the base network were frozen, and only the RRN was trained. RRN was trained for 42 epochs with a batch size of 16. The initial learning rate was 0.001 with a decay factor of {1.0, 0.2, 0.04} at epochs {10, 15, 20}. To tackle the missing sparse depth challenge in the test dataset, the authors trained the model by removing the spot depth from random areas of the image.

Loss. In the base network and RRN, the authors have used a combination of $L_1$, $L_2$ and Corrected Gradient Loss (CGDL) [8] with adaptive weights of 1, 1 and 0.7, respectively. While $L_1$ and $L_2$ losses helped get accurate depths, the main objective of CGDL loss was to keep the depth prediction output smooth and edges sharp.

4.4. Chameleon

Model. This team proposes an effective and efficient depth completion method for depth completion called Bilateraly-Aware Depth Completion via Cascaded Dynamic Graph Convolutional Network (CDGNN-BA), illustrated in Figure 6. The proposed method, CDGNN-BA, outputs dense depth and follows the Convolutional Spatial Propagation Network, originally introduced in [4]. In particular, this team starts out with Dynamic Graph Convolutional Network design as in GraphCSPN [16] and propose modifications for further improvement as below:

Bilateraly-Aware: GraphCSPN utilizes an encoder-decoder to jointly learn the initial depth map and affinity matrix, which is sampled and reshaped into sequences of patches and concatenated with 3D position embeddings. Then the model estimates the neighbors of different patches on the basis of geometric constraints, and performs spatial propagation leveraging dynamic graph convolution networks with self-attention mechanism. After the graph is generated, the Top k patches are judged, and the weighted sum is obtained to predict the final depth. However, the way of calculating the similarity only considers the physical distance between patches while the heterochromia from various objects is ignored. In our method, the RGB information is considered based on the fact that patches from the same object have a higher correlation; Additionally, the naive 3D reconstruction requires camera internal parameters which are hard to obtain in several scenarios. Instead of using these parameters, the authors utilized the distances between pixels and depth distance from the depth map. In particular, the above distance weights are calculated using bilateral weight:

$$d = c \cdot \frac{\left( \frac{(d_{rgb} - n_{rgb})^2}{2\sigma_{rgb}} \right) - \left( \frac{(d_{pixel} - n_{pixel})^2}{2\sigma_{pixel}} \right) - \left( \frac{(d_{depth} - n_{depth})^2}{2\sigma_{depth}} \right)}{}.$$  

Cascaded refinement: GraphCSPN re-samples the affinity matrix by 3-time losses the information of original features. For compensating this deficiency, the cascade structure is proposed. First, 4-time down-sampling is performed. After graph propagation, coarse depth is obtained, and then 2-time down-sampling is applied to obtain refined depth.

Loss. This team uses the same loss function as Equation 4.

Training Setup. For training, this team randomly cropped and resized 192 × 256 patches from the training images as inputs. The mini-batch size is set to 4 and the whole network is trained for 200 epochs. The learning rate is initialized as $1 \times 10^{-3}$, decayed with a LinearLR schedule.

5. Conclusions

In this report, we review and summarize the methods and results of MIPI 2023 challenge on RGB+ToF Depth Completion. The participants were provided with a high-quality training/testing dataset, TetrasRGBD, which is now available for researchers to download for future research. We are excited to see the new progress contributed by the submitted solutions in such a short time, which are all described in this paper. The challenge results are reported and analyzed. De-
tetailed descriptions of the submitted solutions are also provided in this report. For future works, there is still plenty of room for improvements including dealing with depth outliers/noises, precise depth boundaries, high depth resolution, and high temporal stability, etc.

6. Acknowledgements

We thank Shanghai Artificial Intelligence Laboratory, Sony, Nanyang Technological University and The Chinese University of Hong Kong to sponsor this MIPI 2023 challenge. We thank all the organizers for their contributions to this workshop and all the participants for their great work.

References


A. Teams and Affiliations

**MGTV team**

**Title:** MangoTV  
**Members:**  
YiYu¹ (yuyi@mgtv.com)  
YangkeHuang¹(yangke5@mgtv.com)  
KangZhang¹(zhangkang@mgtv.com)  

**Affiliations:**  
¹ MGTV (https://w.mgtv.com/)

**MiMcAlgo team**

**Title:** Enhancing Multi-Scale Non-Local Propagation with Transformer Network.  
**Members:**  
Meiya Chen (chenmeiya@xiaomi.com)  
Yu Wang (wangyu50@xiaomi.com)  
Yongchao Li (liyongchao1@xiaomi.com)  
Hao Jiang (jianghao11@xiaomi.com)  

**Affiliations:**  
Xiaomi Inc.

**DIntel team**

**Title:** Efficient and Refined Multi-modal Depth Completion  
**Members:**  
Amrit Kumar Muduli (amrit.muduli@samsung.com)  
Vikash Kumar (vikash.k7@samsung.com)  
Kunal Swami (kunal.swami@samsung.com)  
Pankaj Kumar Bajpai (pankaj.b@samsung.com)  

**Affiliation:**  
Samsung R & D Institute India - Bangalore

**Chameleon team**

**Title:** Bilateraly-Aware Depth Completion via Cascaded Dynamic Graph Convolutional Network  

**Members:**  
Yunchao Ma (mayunchao@megvii.com)  
Jiaojun Xiao (xiaojiajun@megvii.com)  
Zhi Ling (lingzhi@megvii.com)