Improving Real-Time Aerial 3D Reconstruction: Towards Fusion of a Hand-Crafted SfM Algorithm with a Data-Driven Deep Neural Network

Jack Akers, Andrew Buck,
Derek Anderson, James Keller
Dept. of Electrical Engineering and Computer Science
University of Missouri
Columbia, MO

Raub Camaioni, Matthew Deardorff,
Robert Luke III
U.S. Army DEVCOM C5ISR Center
Fort Belvoir, VA

***Abstract*—Depth estimation from imagery is a crucial component of numerous real-time systems, such as unmanned aerial vehicles. Traditional approaches, like structure from motion (SfM), rely on stereo cameras or sequential images from a single moving camera to infer depth. However, these methods often yield sparse or incomplete depth maps. Moreover, their accuracy is contingent on factors like baseline and platform motion. Recent advancements leveraging deep neural networks have exhibited remarkable outcomes in predicting dense depth from a single image. Nonetheless, the reliability of these black box models is a concern, and the depth approximations they provide often require refinement. In this study, we delve into the merits and drawbacks of each approach, exploring the potential synergies that could arise from their combination. The initial approach, serving as our baseline, relies solely on sparse SfM depth estimates. In contrast, the second method employs spatial interpolation to fill in missing SfM depths. These methodologies are then contrasted with the performance of two deep nets—one trained on simulated data and the other on real data. The validation of all four approaches is conducted against a gold standard generated using the Unreal Engine.**

***Keywords*—** structure from motion, monocular depth estimation, single image depth estimation, fusion

I. Introduction

In this research endeavor, our overarching objective is to achieve real-time 3D reconstruction of an environment with minimal errors and maximal completeness. We operate under the assumption of having access to a live feed of color imagery and reliable extrinsic information. However, a critical challenge lies in performing accurate depth estimation using only these inputs. Existing literature, coupled with insights from our prior work, presents a plethora of solutions to this predicament. Yet, none of these individual solutions has proven sufficient for our specific application. Typically, these methods fall into one of two categories: Structure from Motion (SfM) or Single-Image Depth Estimation (SIDE). In this context, we propose and investigate the integration of post-processing and fusion techniques to leverage the strengths inherent in both SfM and SIDE approaches.

II. Related Work

There exists a massive body of related work in the area of three dimensional reconstruction from imagery, with many overlapping fuzzy boundaries. Readers can find a comprehensive recent overview of Structure from Motion (SfM) and Multi-View Stereo (MVS) in [1], delve into deep learning for SfM [2], or explore Simultaneous Localization and Mapping (SLAM) in [3]. Most relevant to this article, early investigations into Single Image Depth Estimation (SIDE) [4] focused on tasks like semantic segmentation and they distinguished between indoor and outdoor environments. In SIDE's evolution, its scope was expanded to include elements like log space, spatial coordinates, global context, considerations of relative vs absolute depth, and more. The latest and most impressive generation includes deep neural networks, which includes but is not limited to unsupervised depth and ego-motion learning from video [5], Monodepth [6], Monodepth2 [7], DPT [8, 9], and unsupervised monocular depth learning from unknown cameras [10]. Notably, these algorithms operate without access to ground truth depth, because of real data, and as a result they rely on self-supervised learning via video and motion exploitation. Most of these networks adopt some form of a pose and depth architecture, using self-supervised learning from video data to predict the pose of consecutive images. This pose estimation is then employed to train a network, mapping a single input image to depth. During evaluation, the pose

net is typically bypassed, and images are directly mapped to depth. Arguably, these contemporary network-based estimators may not strictly adhere to the traditional definition of SIDE. Last, in a departure from the popular self-supervised paradigm, our recent work, coined simulation driven passive ranging (SimPR) [11], leveraged simulation (SIM) to enable accurate evaluation and SIM has the advantage that it obviates the need for self-supervised learning. As a result, if SIM is a close enough photorealistic match, then SIDE models can be bootstrapped or directly learned. The next two subsections detail the traditional and SIDE approaches explored herein.

A. Hand Crafted SfM via EpiDepth

Numerous small Unmanned Aerial Vehicles (UAVs) come equipped with a monocular image sensor and GPS/IMU/-magnetometer, providing color imagery and camera extrinsics. This data can be leveraged by an SfM algorithm to estimate depth and ultimately three dimensional scenes. Our EpiDepth algorithm, Camaioni et al., [12] takes a continuous stream of images as input and it generates dense, per-pixel depth estimates. EpiDepth is unique in the respect that it is designed for real-time operation on embedded hardware, offering adjustable parameters for scale and quality optimization. For the sake of this article, let I_t be an image at time step t. Epidepth is a function, $E(I_t, I_{t+k})$, which produces a depth image D_t, such that $d_{i,j} \in \Re^+$ is the depth associated with pixel (i, j) in a real-world system (e.g., lat/lon coordinates). The only other information required for this article is EpiDepth's ability to estimate an uncertainty value per-pixel. In [13], we showed how to calculate such a "confidence" value, $U_{i,j} \in [0, 1]$; where 0 means do not trust pixel (i, j) and 1 is fully trust. Figure 1 is an example of EpiDepth, which results in an unoccupied, free, and occupied three dimensional map (UFOMap) when multiple depth images are fused [14], or a fuzzy voxel space if $U_{i,j}$ is utilized [13].

B. Deep Neural Network-Based SIDE

In our previous work [11], we (Buck et al.) introduced SIM passive ranging (SimPR) for two main purposes. Firstly, SimPR was designed to generate precise ground truth for depth assessment. In [15], we (Akers et al.) explored various voxel space metrics to understand, evaluate, and compare different 3D estimation algorithms. Additionally, in [11], we utilized SimPR for in-depth performance analysis, allowing filtering based on user-defined factors like range intervals, per-pixel object labels, image features (e.g., mixed pixel edges), and camera specifics like field of view.

The second advantage of employing SimPR is to acquire photorealistic imagery with associated metadata and truth across diverse environments and contexts for training a SIDE algorithm. This is challenging, if not impossible, to achieve densely and at scale in the real world. However, our later work [16] revealed that while SIM truth extraction surpasses real-world capabilities, it is not flawless. Specifically, in the realm of computer vision, defining a pixel becomes a nuanced challenge. For a detailed exploration of trust extraction and AI bias mitigation, readers can refer to [16]. In this context, we refer to SIM as a "gold-standard" versus absolute truth

Figure 3 summarizes the use of SimPR in the current article. Specifically, we crafted a SIM environment where an agent, represented by a low-altitude drone, navigates randomly in a scene. At each moment, an image, its ground truth, and metadata is extracted. Subsequently, SimPR computes a depth estimate, which is compared to the extracted golden standard. Our assessment involves per-pixel error, which is used to iteratively update our network weights.

In this initial investigation, SimPR ran continuously for multiple days with the primary aim of constructing an exceptionally accurate model in a single map. Our deliberate strategy involved overtraining our SIDE network, pushing it to its upper performance limit. We chose this approach strategically, especially considering our subsequent comparisons with SimPR findings versus DPT. DPT, a model trained on millions of real images with questionably acquired supervised depth, is anticipated to perform suboptimally compared to our overtrained SimPR. This expectation stems from DPT's limited exposure to aerial contexts and its training data, which was predominantly at relatively close range compared to our longer-range SIM environment. For the sake of completeness, we also included an undertrained variant of SimPR that had only been exposed to a building-free variant of the sim environment under different weather and flight conditions. We employed Unreal Engine 4.27, used Microsoft's AirSim plugin [17], and used the Mountain Grassland Environment [18], which is available on the UE Marketplace (see Figure 2).

III. Methodologies

The subsequent section delineates our initial foray into various approaches for integrating a SIDE network with a meticulously crafted SfM algorithm. In Method 1 (M1), detailed in subsection III-A, we contemplate the concept of discovering a linear affine transformation to register the output of a SIDE network based on a select set of dependable SfM anchor points. Conversely, in M2, covered in subsection III-B, we entertain the opposing notion of not fusing but rather filling in the gaps within our SfM algorithm.

A. Method 1: Linear Fit

Our initial attempt to integrate a deep neural net depth map with a SfM estimate involves assuming a linear relationship

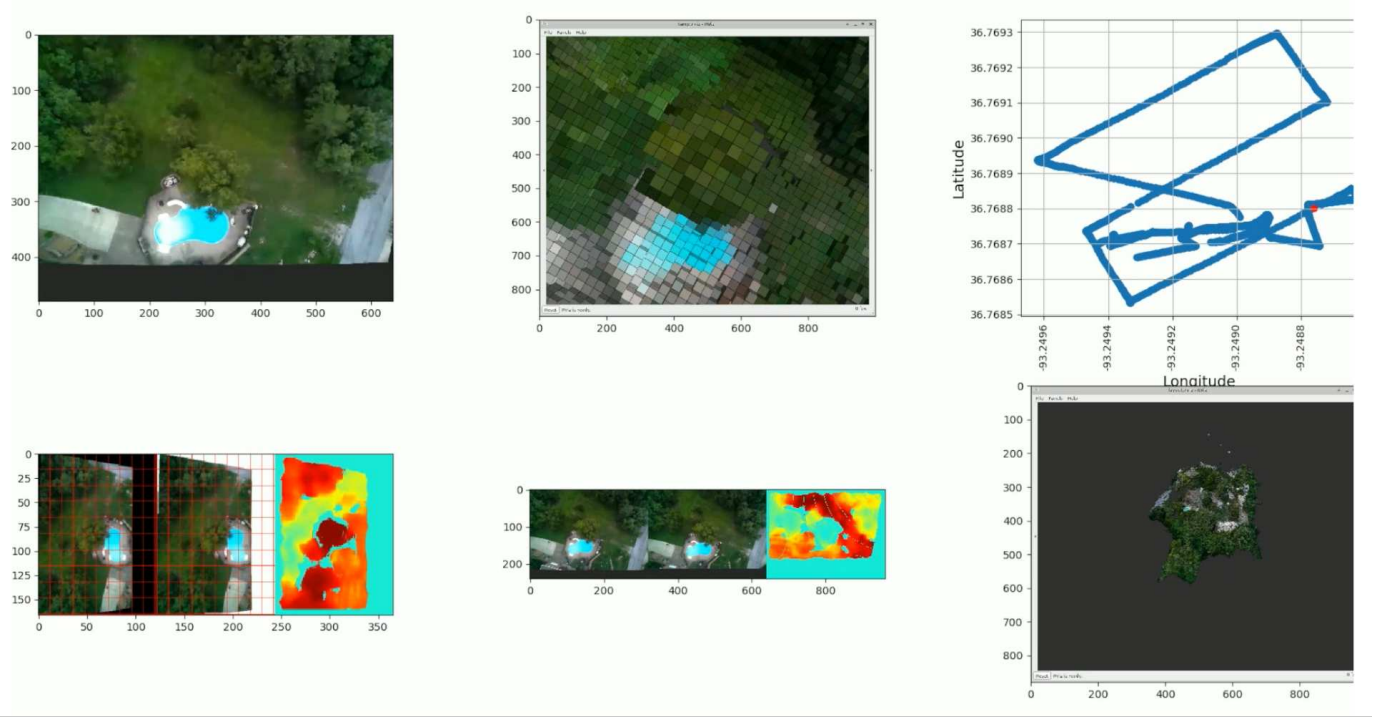

Fig. 1. EpiDepth algorithm applied to real world micro drone data: (Upper left) current input, (upper center) synthetically rendered drone position in voxel space (via UFOMap), (upper right) drone's current position in flight path, (lower left) epipolar space, (lower center) unwarped image pairs, and (lower right) complete 3D voxel space (via UFOMap).

between the output of EpiDepth and the DNN depth net. Notably, EpiDepth benefits from already existing in an absolute coordinate space, while our depth net offers advantages in terms of completeness and sharper local structure. Operating under the assumption of a linear relationship, we propose scaling the relative output of the depth net to align with the absolute output of EpiDepth. However, it's essential to acknowledge that EpiDepth is susceptible to errant outliers, especially with sub-optimal frame selection [19]. To address this, we leverage methods for computing per-pixel confidence within EpiDepth [13], allowing us to filter EpiDepth results based on this confidence before selecting minimum/maximum points for scaling.

The plots in the bottom-left corner of Figure 5 illustrate this process, with green points representing all pixels with both EpiDepth and DNN depth values. The blue line denotes the chosen linear fit. It's noteworthy that the two plots correspond to different DNNs (DPT and SimPR), and the linear transformation identified for DPT exhibits a negative slope, while the one found for SimPR is positive.

B. Method 2: SfM Interpolation

Our second approach involves applying a spatial interpolation operation to the output of EpiDepth. Initially, we utilized the widely known iterative closest point (ICP) algorithm for this purpose. However, we encountered significant drawbacks, such as its sluggish performance even with moderately sized images and the undesirable tendency to extrapolate and fill the entire image rather than focusing on interpolation within the voids present in EpiDepth's output.

Consequently, we opted for the Natural Neighbor Interpolation algorithm. This method constructs a triangulated mesh with all known points as vertices and subsequently computes a weighted average for each intermediate pixel based on its proximity to the vertices forming the triangle containing it. Given the scarcity and inconsistency of academic literature discussing this algorithm, we have provided the implementation details in Figure 4.

It's important to note that this approach does not qualify as a fusion technique since it doesn't depend on the output of the depth net in any way. Nevertheless, it represents an intriguing and practical method that significantly aligns with the motivation behind this investigation. Its principal purpose lies in enhancing the completeness of EpiDepth's results without incurring a substantial increase in error, a demonstration of which will be presented in the subsequent results section of this paper. While we employ this method in isolation within this context, its utility suggests that it may prove valuable, or perhaps even indispensable, as a precursor

Fig. 2. The (left) Mountain Grassland Environment used and (right) extracted SIM gold-standard ("truth") in UFOMap at 1m resolution.

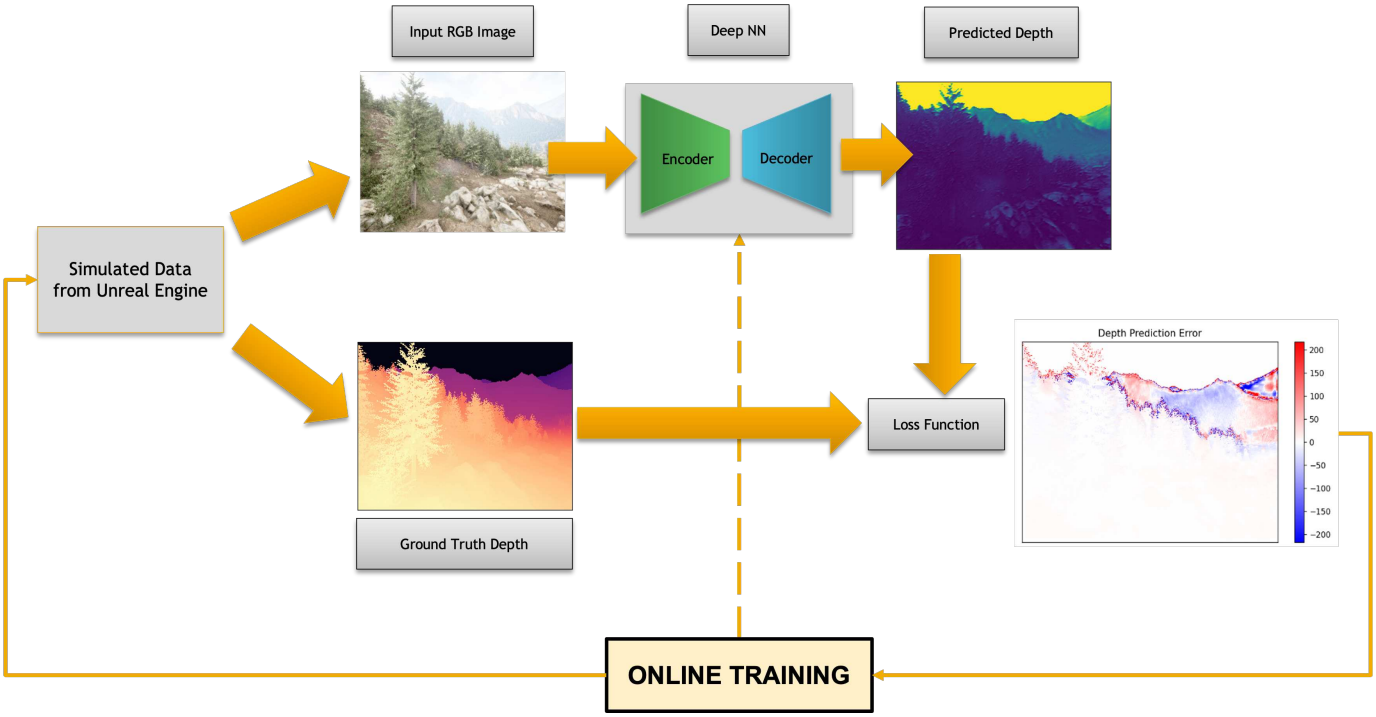

Fig. 3. Visualizing our SimPR framework: the simulator, leveraging Unreal Engine, generates data, metadata, and ground truth fed into our SIDE algorithm. Following prediction, the error between our estimate and the ground truth is computed. SimPR operates in a continuous online learning mode, gathering data for ongoing training of our depth prediction network.

for more sophisticated fusion algorithms that necessitate high completeness in their inputs.

IV. Examples

During our testing phase, we conducted a simulated flight within the Mountain Village Environment [18], featuring a 45-degree slant look angle and a moderate altitude suitable for micro UAV operations. This configuration was chosen to establish favorable operating conditions for both EpiDepth, which performs optimally with nadir (or 0-degree) flights, and off-the-shelf deep nets like DPT. Given the self-driving-focused training data of DPT, we anticipated superior performance with a 90-degree look angle, making the 45-degree slant a reasonable compromise.

During the simulated flight, we extracted a video stream containing both color imagery and ground-truth depth. The color imagery underwent processing through the EpiDepth

```python
import cv2
import numpy as np
import scipy
def interpolate_natural_neighbor(image):
    x, y = np.where(image > 0)
    values = image[image > 0]

    triangulation = scipy.spatial.Delaunay(np.column_stack((x, y)))
    interpolator = scipy.interpolate.CloughTocher2DInterpolator(
        triangulation, values
    )

    x_grid, y_grid = np.meshgrid(
        np.arange(image.shape[0]), np.arange(image.shape[1])
    )
    interpolated_image = interpolator(x_grid, y_grid)

    interpolated_image = interpolated_image.T
    interpolated_image[np.isnan(interpolated_image)] = -1.0
    return interpolated_image
```

Fig. 4. Our Python 3.8.10 implementation of the Natural Neighbor sparse interpolation algorithm using NumPy, SciPy, and (optionally) OpenCV. The input image is a 2D NumPy array, which is the representation of an image used by OpenCV. Note that OpenCV is not strictly necessary to load or display the image, as many other popular libraries (ex. Matplotlib) accept the same format.

pipeline, incorporating a frame selection algorithm that transformed the video stream into pseudo-stereo pairs. The aggregation of these pairs yielded an EpiDepth result comprising 3D points. This result was then projected back into the camera space of the first image to generate a 2D depth map. Concurrently, we applied the same first image to the SIDE algorithms (DPT and SimPR) to obtain alternative depth maps in the same camera space.

These "raw" results are then fed into the linear fit and spatial interpolation approaches discussed above to yield seven candidate results, not including the ground truth. All of these candidates are then evaluated using a suite of metrics. While we have used 3D metrics in our prior work [15], we chose to limit the scope of this investigation to 2D given the lack of accumulation over time and the native 2D output of SIDE. The metrics we chose are completeness (percentage of pixels for which an estimate was produced), root-mean-square error (RMSE) $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(gt_i - pred_i)^2}$, RMSE of logs $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(ln(gt_i) - ln(pred_i))^2}$, mean-relative-square error $\frac{1}{n}\sum_{i=1}^{n}\frac{(gt_i - pred_i)^2}{gt_i}$, and mean-relative-absolute error $\frac{1}{n}\sum_{i=1}^{n}\frac{abs(gt_i - pred_i)}{gt_i}$. These "relative" errors are intended to reduce the penalty for errors in distant scenery, as they are inherently more difficult to predict with the same precision as nearer objects. This process produces a verbose set of results for every frame pair in the dataset, which would be impossible to showcase here. As such, we instead present two prototypical examples: one where EpiDepth is performing well, and another where EpiDepth is performing poorly.

Our initial results, depicted in Figure 5 (refer to the caption for detailed information on each plot), showcase an optimal frame pair selection for EpiDepth. This selection yields over 80% completeness and competitive performance across all error metrics. The application of natural-neighbor interpolation to EpiDepth significantly enhances completeness, accompanied by a negligible increase in error. The resubstituted/overfit variant of SimPR demonstrates top-notch performance across all metrics. However, the more realistic depth nets—SimPR (non-resubstituted) and linearly-fit DPT—display unsatisfactory results, with RMSE scores ranging between 40 and 60 meters. While a linear fit is essential to enable the use of DPT (as evidenced by its negative slope and exaggerated scale), this adjustment exacerbates the performance of SimPR.

It's noteworthy that despite these objectively poor results for SIDE, the visual outcomes exhibit robust segmentation performance, presenting qualitatively strong results when lacking the context of ground-truth. The non-resubstituted SimPR's notable shortcomings are observed in areas with buildings, as

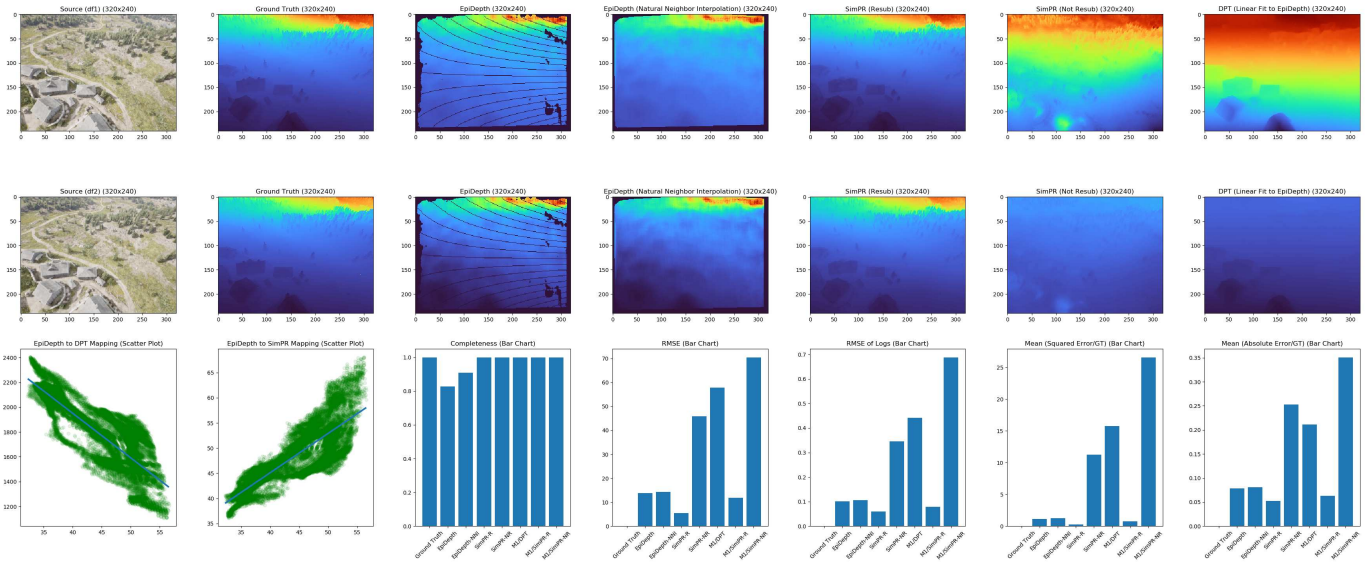

Fig. 5. An example of each method being applied to a well-selected image pair from the Mountain Village Environment [18] using UE4 + AirSim. The two frames are shown on the left (note that only the top image is used for the SIDE methods). The methods in the first and second row are identical, with the only distinction being that the first row uses per-image color scaling while the second row uses a common color scale across all images. From left-to-right: Ground-Truth/Gold-Standard Depth extracted from AirSim, EpiDepth, EpiDepth with Natural Neighbor Interpolation, SimPR (Resubstituted), SimPR (Not Resubstituted), and DPT (Linear Fit to EpiDepth). The third row shows the linear fits used for DPT and for SimPR (Resubstituted), then a suite of methods evaluating each method with respect to the ground truth. The metrics present each method in the same order as the images above, but with the addition of Linear Fits to EpiDepth for SimPR (Resubstituted) and SimPR (Not Resubstituted). These results were visually indistinguishable from the raw data (assuming a positive slope, the color scaling used for visualization is itself a linear fit), and thus their images were omitted for the sake of brevity.

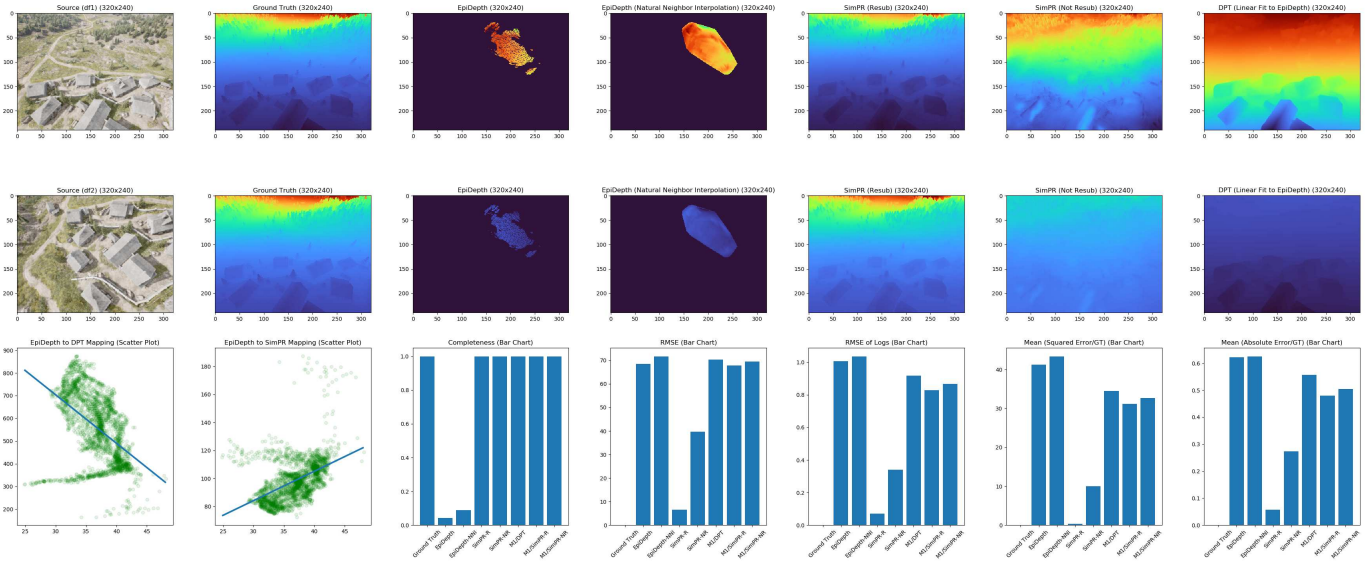

Fig. 6. Same methods and metrics as Figure 5, except for a poor frame pair selection causing an incomplete and inaccurate EpiDepth result.

it was not originally trained on such structures. Conversely, DPT excels at identifying buildings but struggles with the recognition of distant trees and topography, likely stemming from its limited training data in rural or aerial contexts. Consequently, a promising avenue for future research involves retraining these off-the-shelf networks with more diverse and contextually appropriate training data for this application.

In our subsequent example, depicted in Figure 6, we illustrate an alternative scenario where inadequate frame selection results in severely compromised EpiDepth performance. The EpiDepth result, quite fittingly, resembles a coffin and showcases dismal completeness, falling well below 10%. Although natural-neighbor interpolation offers a marginal improvement, the only effective solution in this case is the resubstituted/overfit variant of SimPR, with the non-resubstituted variant of SimPR trailing at a considerable distance. Similar to the previous example, SimPR excels at identifying trees, while DPT excels at recognizing buildings, but the reverse is not true. Despite EpiDepth's dismal performance in this instance, it exhibits a notable advantage over SIDE approaches: its failures are relatively straightforward to identify and explain, facilitating rapid diagnosis and improvement.

For instance, consider the second case illustrated in Figure 6, where there is a noticeable and abrupt change in camera pitch between the two frames. EpiDepth's frame selection pipeline should have recognized and rejected this pair as a match, a relatively straightforward filter to manually implement. In both Figure 5 and 6, peculiar "blotches" appear over areas with buildings, and the best assumption we can make is that this is a consequence of insufficient training data. While the EpiDepth failure could have been automatically identified in real-time due to its low completeness score (coupled with camera extrinsics [19] and other internal metrics within EpiDepth beyond the scope of this research), the failures in SimPR are undetectable without manual review or ground-truth. Consequently, even with the potential for enhanced Deep Neural Network (DNN) results in the future, we maintain that a fusion of these approaches will be essential to attain reliable outcomes in the field for an autonomous system.

V. Conclusions and Future Work

A. Conclusions

In summary, though SfM and deep neural network approaches show promise for real-time 3D reconstruction depth estimation, existing off-the-shelf deep neural networks fall short of the accuracy required for standalone use. While post-processing and fusion techniques present compelling and potentially necessary strategies for addressing these limitations, their practical implementation hinges on the improvement of inputs to achieve higher quality and quantifiable uncertainty.

In Method 1, we endeavored to execute an uncertainty-sensitive scaling to align the outputs of deep nets with the global space of EpiDepth. Although this approach succeeded in making certain deep nets, like DPT, applicable for 3D reconstruction, it fell short of producing results with the requisite precision for practical utilization.

Method 2, on the other hand, was a strong success, increasing completeness with a measurable but tolerable increase in error. However, while this method may be useful for future fusion techniques where a dense baseline is necessary, it currently only takes into account the results of EpiDepth. This is an important step towards our goal of dense, real-time, high-fidelity 3D reconstruction, but it falls well short of our goal of using fusion to mitigate the shortcomings of any individual approach, EpiDepth very much included.

B. Future Work

This work represents an early stage, laying the foundation for numerous potential avenues in future research. These possibilities include the exploration of more sophisticated fusion techniques, the development of data-driven conditional fusion and model switching strategies, the application of transfer learning specifically tailored for aerial domains, and the design of novel deep network architectures capable of more effectively harnessing the available information in this context (such as trusted extrinsics, a reasonably reliable prior estimate, and a stream of pseudo-stereo pairs in lieu of single images). In the pursuit of these approaches, we aim to expand our metrics suite to encompass a measure of resiliency against simulated data degradation, encompassing factors like imprecise/binned/noisy extrinsics, image compression, rolling shutter effects, and more.

References

[1] J. L. Schoenberger and M. Pollefeys, "Colmap - structure-from-motion and multi-view stereo," *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016, available online: https://demuc.de/colmap/ (accessed Dec. 20, 2022).

[2] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, "Sfm-net: Learning of structure and motion from video," *arXiv*, Apr 2017.

[3] J. A. Placed and et al., "A survey on active simultaneous localization and mapping: State of the art and new frontiers," *arXiv*, Jul 2022, accessed: Dec. 20, 2022. [Online]. Available: http://arxiv.org/abs/2207.00254

Corresponding author: Jack Akers, email:jdapm8@missouri.edu

[4] A. Mertan, D. J. Duff, and G. Unal, "Single image depth estimation: An overview," *Digital Signal Processing*, vol. 123, p. 103441, Apr 2022.

[5] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," *arXiv*, Jul 2017, accessed: Dec. 20, 2022. [Online]. Available: http://arxiv.org/abs/1704.07813

[6] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," *arXiv*, Apr 2017, accessed: Dec. 20, 2022. [Online]. Available: http://arxiv.org/abs/1609.03677

[7] C. Godard, O. Mac Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," *arXiv*, Aug 2019, available online: http://arxiv.org/abs/1806.01260.

[8] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," *ArXiv preprint*, 2021.

[9] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[10] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," *arXiv*, Apr 2019, accessed: Dec. 20, 2022. [Online]. Available: http://arxiv.org/abs/1904.04998

[11] A. Buck, J. Kerley, D. T. Anderson, and J. Keller, "Simulated data to train and evaluate deep learning-based passive monocular vision algorithms at medium to long ranges," in *MSS*, 2023.

[12] R. Camaioni, R. H. Luke, A. Buck, and D. T. Anderson, "EpiDepth: a real-time monocular dense-depth estimation pipeline using generic image rectification," in *SPIE*, 2022.

[13] A. Buck, D. T. Anderson, R. Camaioni, J. Akers, R. H. Luke, and J. Keller, "Capturing uncertainty in monocular depth estimation: Towards fuzzy voxel maps," in *FUZZ-IEEE*, 2023.

[14] D. Duberg and P. Jensfelt, "UFOMap: An Efficient Probabilistic 3D Mapping Framework That Embraces the Unknown," *arXiv:2003.04749 [cs]*, Mar. 2020.

[15] J. Akers, A. Buck, R. Camaioni, D. T. Anderson, R. H. Luke, J. M. Keller, M. Deardorff, and B. Alvey, "Simulated gold-standard for quantitative evaluation of monocular vision algorithms," in *SPIE Defense + Commercial Sensing*, 2023.

[16] A. R. Buck, D. T. Anderson, J. Fraser, J. Kerley, and K. Palaniappan, "Ignorance is bliss: Flawed assumptions in simulated ground truth," in *SPIE Defense + Commercial Sensing*, 2023.

[17] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "AirSim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*, 2017. [Online]. Available: https://arxiv.org/abs/1705.05065

[18] FreshCan, "Mountain village environment," Apr 2021. [Online]. Available: https://www.unrealengine.com/marketplace/en-US/product/mountain-village-environment

[19] A. R. Buck, J. D. Akers, D. T. Anderson, R. Camaioni, M. Deardorff, R. H. L. III, and J. M. Keller, "Frame selection strategies for real-time structure-from-motion from an aerial platform," in *2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2023.