Learning Event-based Height from Plane and Parallax

Kenneth Chaney Alex Zihao Zhu Kostas Daniilidis

Abstract

In this work, we propose a fast method to perform event-based structure estimation for vehicles traveling in a roughly 2D environment (e.g. in an environment with a ground plane). Our method transfers the method of plane and parallax to events, which, given the homography to a ground plane and the pose of the camera, generates a warping of the events which removes the optical flow for events on the ground plane, while inducing flow for events above the ground plane. We then estimate dense flow in this warped space using a self-supervised neural network, which provides the height of all points in the scene. We evaluate our method on the Multi Vehicle Stereo Event Camera dataset, and show its ability to rapidly estimate the scene structure both at high speeds and in low lighting conditions.

1. Introduction

In this paper, we propose a novel structure estimation method for event cameras that is suitable for on autonomous driving in real world scenarios. We utilize recent advances in self-supervised learning methods for event cameras to train a convolutional neural network to learn height and depth from a loss that utilizes Plane and Parallax (P+P) [6] principles, Wulff et al. [7] show that this method effectively reduces the complexity of computing optical flow on static scenes while simultaneously providing a direct metric representation of the magnitude of the computed flow. Our network takes as input raw events, and predicts the ratio between the height of each point above the ground plane, and the depth in the camera frame. We show that this ratio can be used to compute the optical flow between a pair of images warped using P+P, and apply a semi-supervised loss to minimize the photometric error between the images.

In order to accurately predict metric depth directly from a scene, a network must learn to make a large number of assumptions about objects in the scene such as cars, pedestrians, and buildings. As such, these networks have a hard time generalizing to other contexts. Predicting relative factors up to a scale that represent the structure of the scene, but need to be scaled or otherwise decoded, allows networks to generalize better. Our method leverages this by predicting the ratio of height and depth, alone this provides a relative measurement, but when coupled with a camera to ground calibration, it allows for the system to recapture the full metric information of the scene, in a similar manner to the monocular implementation that accompanies Geiger et al. [2]. Our method runs at 75Hz on a modern high grade GPU, and can estimate scene height and depth in low-light and high speed driving scenarios, making it suitable for night time autonomous driving. We evaluate our method on the Multi Vehicle Stereo Event Camera (MVSEC) dataset [10], and demonstrate our network's ability to accurately predict the heights and depths of objects in the scene. We further show an application of these predictions towards accurately segmenting free space on the ground plane. In all experiments, we demonstrate superiority over image input and depth prediction baselines.

The technical contributions of the paper are as follows:

- A novel loss that leverages P+P to isolate static scene components from the motion of the event camera.
- A novel pipeline that trains a neural network to learn the ratio between the height of a point from the ground plane and its depth in the camera frame, using a self-supervised loss, where camera pose and the ground normal is used at training time, but only the ground normal is needed at test time.
- Evaluation on challenging high-speed and low-light night time MVSEC dataset scenes.

2. Method

In this section, we will describe our proposed pipeline, which is summarized in Fig. 1.

2.1. Input Representation

Prior works summarize the event stream into an image [5, 9, 4, 8]. These representations inherently lose much of the high temporal resolution of the events by throwing away most of the timestamps. To resolve this issue, we adopt the 3D spatiotemporal volume used by Zhu et al. [11]. For a

The authors are with the University of Pennsylvania, Philadelphia, PA, USA. Emails are: {chaneyk, alexzhu, kostas}@seas.upenn.edu.

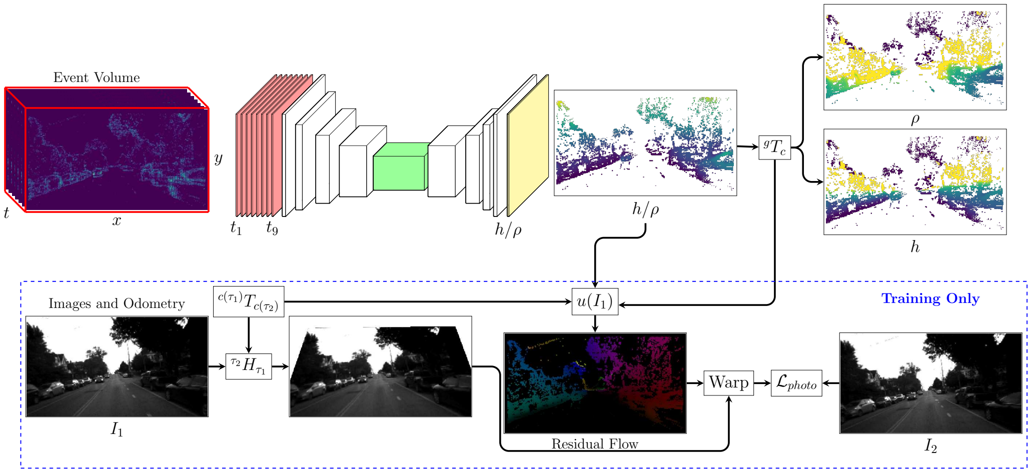

Figure 1: Our network inputs a raw event volume and predicts the scene structure, comprised of the ratio of height (h) over depth (ρ). Given the ground plane calibration, we can recover depth and height independently. During training (blue box), odometry and the predicted scene structure is used in a two stage warping process to warp I_1 into I_1^w. First, a homography, ${}^{c(\tau_2)}H_{c(\tau_1)}$, warps and aligns the ground plane between the images. Second, the scene structure is used to compute the residual flow to warp any portions of the scene not on the ground plane. I_1^w is compared to I_2 in $\mathcal{L}_{photo}$. At test time, only the calibration from the camera to the ground plane, ${}^gT_c$, is used to compute ρ and h.

set of N events, denoted as $\{x_i, y_i, t_i, p_i\}$, consisting of the x_i, y_i pixel position, timestamp t_i and polarity, p_i, we first discretize the time dimension into B bins. However, simply rounding the events and inserting them into the volume, V, would lose a significant amount of information, and so we instead insert events using trilinear interpolation. First, the range of the event timestamps is scaled to the number of bins: $t_i^* = (B-1)(t_i - t_1)/(t_N - t_1)$.

The volume is then generated as follows:

$$V(x, y, t) = \sum_i p_i k_b(x - x_i) k_b(y - y_i) k_b(t - t_i^*) \quad (1)$$

where: $k_b(a) = \max(0, 1 - |a|)$.

2.2. Self-Supervision from Plane and Parallax

Plane and parallax (P+P) methods warp images (or individual points) through a common reference plane to create parallax between images. The warping will exactly register points that lie on the plane, while points above or below will have some residual flow, which can be parameterized by a rigid structure parameter and camera motion. The P+P warping can be represented as the homography, ${}^{c(\tau_2)}H_{c(\tau_1)}$, which transforms the $c(\tau_1)$ frame to the common ground plane and finally to the $c(\tau_2)$ frame.

At training time, we apply a P+P warping on the image immediately before the event volume, I_{τ_1}, to the one immediately after, I_{τ_2}. To generate the homography, we assume that the fixed transformation between the camera and the ground frame, ${}^cT_g$, and the relative pose between the camera frames, ${}^{c(\tau_2)}T_{c(\tau_1)}$, are known. T is composed of the homogeneous form of the rotation, R, and translation, t: $T = [R, t; [0, 0, 0, 1]]$. The relative pose between camera frames can be decomposed as two transformations from the camera to the ground at the respective times:

$$\begin{aligned} {}^{c(\tau_2)}T_{c(\tau_1)} &= {}^cT_g {}^gT_c {}^{c(\tau_2)}T_{c(\tau_1)} && (2) \\ {}^{c(\tau_2)}T_g &= {}^cT_g {}^g && (3) \\ {}^{c(\tau_1)}T_g &= ({}^gT_c {}^{c(\tau_2)}T_{c(\tau_1)})^{-1} && (4) \end{aligned}$$

The homography, ${}^{c(\tau_2)}H_{c(\tau_1)}$, that passes through the ground plane, can then be generated as the composition of two homographies to the ground plane, ${}^{c(\tau_2)}H_{c(\tau_1)} = {}^{c(\tau_2)}H_g {}^{c(\tau_1)}H_g^{-1}$. Each homography from a camera plane to the ground plane, ${}^{c(\tau_i)}H_g$, is defined as: ${}^{c(\tau_i)}H_g = {}^{c(\tau_i)}R_g + \begin{bmatrix} 0 & 0 & {}^{c(\tau_i)}t_g \end{bmatrix}$. Given ${}^{c(\tau_2)}H_{c(\tau_1)}$, every pixel, p, in the previous image, I_1, can be warped according to the following equation to generate the warped image, I_1^w:

$$I_1^w\left({}^{c(\tau_2)}H_{c(\tau_1)} p\right) = I_1(p) \quad (5)$$

2.3. Residual Flow Loss

After P+P, the remaining differences between the images I_1^w and I_2 correspond to a residual flow induced by the height of the point off the ground plane. We train our network to learn a rigid structure parameter which can be used to recover the flow, which is used to further warp I_1^w. This residual flow, $u(\vec{x})$, can be written as:

$$\begin{aligned} u(\vec{x}) &= \frac{A(\vec{x})b}{A(\vec{x})b - 1}(e - \vec{x}) && (6) \\ A(\vec{x}) &= \frac{h(\vec{x})}{\rho(\vec{x})}, b = \frac{{}^{c_2}t_{c_1}(3)}{{}^ct_g(3)} && (7) \end{aligned}$$

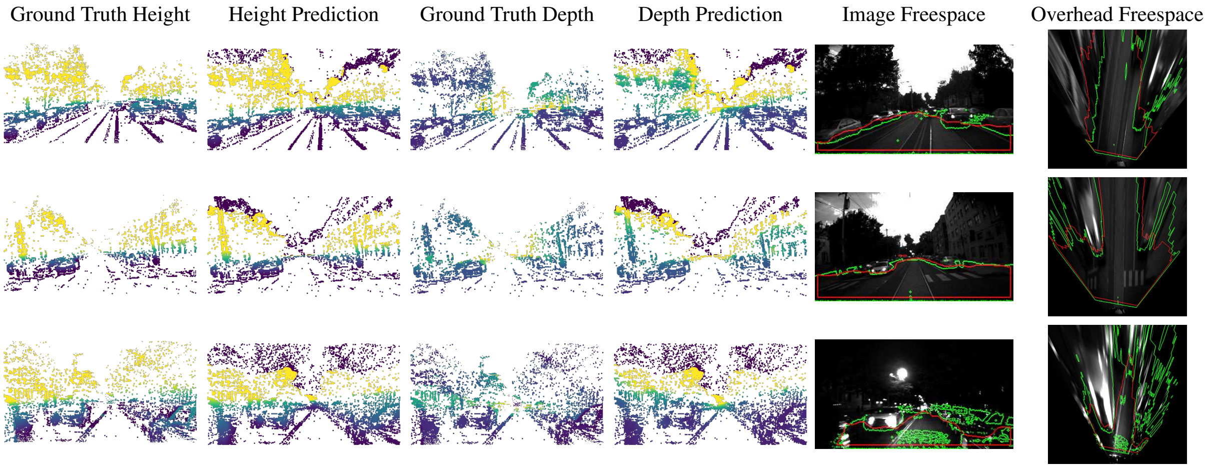

Figure 2: Left to right: (a-d) The ground truth and predictions of height and depth at pixels with events over the time window in which events were collected (e) The grayscale image overlayed with the ground truth (green) and network (red) free space regions in the camera frame (f) The freespace image projected into the ground frame.

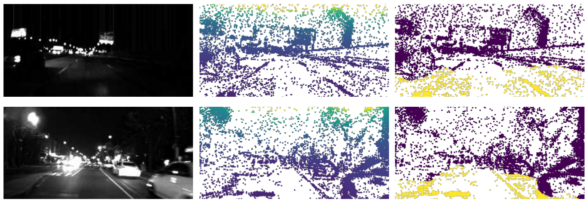

Figure 3: Qualitative results from the motorcycle sequence. Left to right: Grayscale image, predicted height over pixels with events, freespace mask (yellow is free).

where $h(\vec{x})$ and $\rho(\vec{x})$ are the height and depth of point $\vec{x}$, respectively, e is the epipole in the image, ${}^{c_2}t_{c_1}(3)$ is the camera translation along the Z axis and ${}^{c}t_{g}(3)$ is the height of the camera c above the ground plane. We refer to Wulff et al. [7] for the full derivation of this equation.

After the warping, the flow for each pixel can be fully parameterized by a single scalar, composed of a rigid structure component, $A(\vec{x})$ and a time varying component, b, which we assume is known. Our network learns the structure component, $A(\vec{x})$, from which we can exactly recover the height, $h(\vec{x})$, and depth, $\rho(\vec{x})$, of each point:

$$\rho(\vec{x}) = \frac{{}^{g}t_{c}(3)}{A(\vec{x}) - {}^{g}R_{c}(3,:)\vec{x}} \tag{8}$$

$$h(\vec{x}) = A(\vec{x})\rho(\vec{x}) \tag{9}$$

Using (6), we can estimate the optical flow at every pixel. This flow is used to further warp I_1^w towards I_2, generating $\hat{I}_2$ to remove the residual flow. We train our network using the photometric and smoothness losses employed by Zhu et al. [9], applied to I_2 and $\hat{I}_2$.

3. Results

3.1. Implementation Details

Our network was trained on the outdoor_day2 sequence from MVSEC, where we only use the left camera's events and images. The events and images are cropped to 176×336 pixels to remove the hood of the car in the images. The ground to camera extrinsic calibration was computed by applying a RANSAC plane fit to each ground truth depth image in outdoor_day2 and taking the median plane. This calibration was used for all other experiments.

3.2. Depth and Height Evaluation

We compare our model, which uses an event volume as input, against an implementation that uses images as input instead of events, which we label Image P+P. Additionally we compare against a weakly supervised network which directly predicts depth from events, which we label Event Depth. This model utilizes the odometry feed and predicted depths to warp between images for a photometric loss, through the standard motion field equations [3]. Qualitative results from these experiments can be found in Fig. 2, these include qualitative results on freespace detection as an application. We defined freespace as points > 0.1m from the ground plane. The architecture was kept constant between models except where noted.

Table 1 outlines the results from the given methods on the test sequences that we choose from MVSEC. Each network constructs depth and height maps which is compared to the provided ground truth. For the Event Depth network, heights can be estimated using the given depth map through the ground plane calibration, $X(\vec{x}) = {}^{g}R_{c(\tau_2)}\rho(\vec{x})K^{-1}\vec{x} + {}^{g}t_{c(\tau_2)}$, and the height component is then, $h(\vec{x}) = X(\vec{x})(3)$. Our method, Event P+P, provides results close to or bet-

		Average Depth Error (m)			Average Height Error (m)		
	Threshold	ρ <10m	ρ <20m	ρ <100m	-0.5m< h <5.0m	0.1m< h <5.0m	1.0m< h <5.0m
outdoor_day1	Event P+P	6.26	8.37	**10.77**	0.55	0.69	1.00
	Image P+P	**4.15**	**7.45**	12.06	1.09	1.18	1.44
	Event Depth	14.49	13.83	13.37	**0.48**	**0.59**	**0.81**
outdoor_night1	Event P+P	**2.93**	**4.30**	**6.36**	**0.41**	**0.42**	**0.50**
	Image P+P	4.57	7.33	10.61	1.00	1.15	1.44
	Event Depth	15.62	15.88	15.27	0.48	0.53	0.62
outdoor_night2	Event P+P	**2.89**	**5.07**	**6.94**	**0.37**	**0.40**	**0.55**
	Image P+P	4.41	7.75	10.58	1.27	1.58	2.03
	Event Depth	10.8	11.09	10.82	0.40	0.47	0.60
outdoor_night3	Event P+P	**3.24**	**5.61**	**7.64**	**0.41**	**0.47**	0.70
	Image P+P	4.45	8.01	10.97	1.41	1.84	2.35
	Event Depth	13.17	12.39	11.50	0.42	0.51	**0.66**

Table 1: Results of the baseline networks against our network on all testing scenes. For all evaluation, only pixels with events during the relevant time window are evaluated. The thresholds for depth and height are applied to the ground truth depth and height images to create a additional mask to evaluate within.

ter than the baselines. In addition, the network is able to generalize to the night time sequences where there are significantly more noisy events, as well as to a change in environment in outdoor_day1, which is inside an office park as opposed to the suburban roads in the training set. The depth network performs significantly worse than the P+P methods across the depth metrics. One possible explanation is that the P+P loss could be more robust to error in the ground truth provided by lidar odometry.

3.3. High Speed Tests

In order to demonstrate the ability of our proposed pipeline on different vehicles and fast motions, we tested our network on the motorcycle sequence from MVSEC. This sequence contains a motorcycle driving at night on surface streets and highway, with speeds up to 140km/hr. Ground truth is not available for this sequence, but we provide qualitative results in Fig. 3. Due to the lack of ground truth, the camera to ground calibration was only roughly tuned manually. In particular, these results show that it is possible to accurately segment other vehicles from free space, by thresholding points with height <0.1m as freespace.

Acknowledgements: We gratefully appreciate support through the following grants: NSF-IIP-1439681 (I/U-CRC), NSF-IIS-1703319, NSF MRI 1626008, ARL RCTA W911NF-10-2-0016, ONR N00014-17-1-2093, ARL DCIST CRA W911NF-17-2-0181, the DARPA-SRC C-BRIC, and by Honda Research Institute.

References

[1] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014.

[2] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE Intelligent Vehicles Symposium*, Baden-Baden, Germany, June 2011. 1

[3] B. Horn, B. Klaus, and P. Horn. *Robot vision*. MIT press, 1986. 3

[4] A. I. Maqueda, A. Loquercio, G. Gallego, N. Garcıa, and D. Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5419–5427, 2018. 1

[5] D. P. Moeys, F. Corradi, E. Kerr, P. Vance, G. Das, D. Neil, D. Kerr, and T. Delbrück. Steering a predator robot using a mixed frame/event-driven convolutional neural network. In *Event-based Control, Communication, and Signal Processing (EBCCSP), 2016 Second International Conference on*, pages 1–8. IEEE, 2016. 1

[6] H. S. Sawhney. 3d geometry from planar parallax. In *CVPR*, volume 94, pages 929–934, 1994. 1

[7] J. Wulff, L. Sevilla-Lara, and M. J. Black. Optical flow in mostly rigid scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 7. IEEE, 2017. 1, 3

[8] C. Ye, A. Mitrokhin, C. Parameshwara, C. Fermüller, J. A. Yorke, and Y. Aloimonos. Unsupervised learning of dense optical flow and depth from sparse event data. *arXiv preprint arXiv:1809.08625*, 2018. 1

[9] A. Zhu, L. Yuan, K. Chaney, and K. Daniilidis. EV-Flownet: Self-supervised optical flow estimation for event-based cameras. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018. 1, 3

[10] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. 1

[11] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. *arXiv preprint arXiv:1812.08156*, 2018. 1