Learning Monocular Dense Depth from Events

PR-SMARCLE

김찬영

SMARCLE

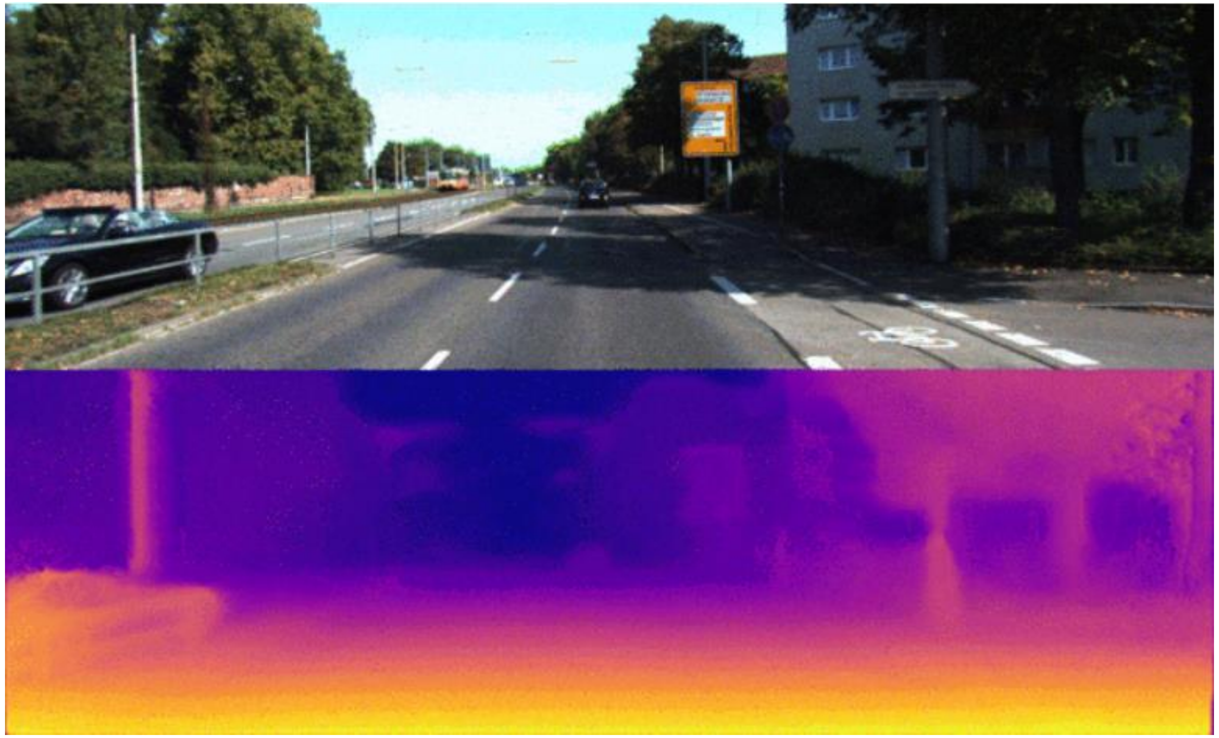
Content

- Introduction to Depth Estimation
- Introduction to Event Data
- Introduction to E2Depth
- Input Data
- E2Depth Model
- Training
- Experiments

SMARCLE

Introduction to Depth Estimation

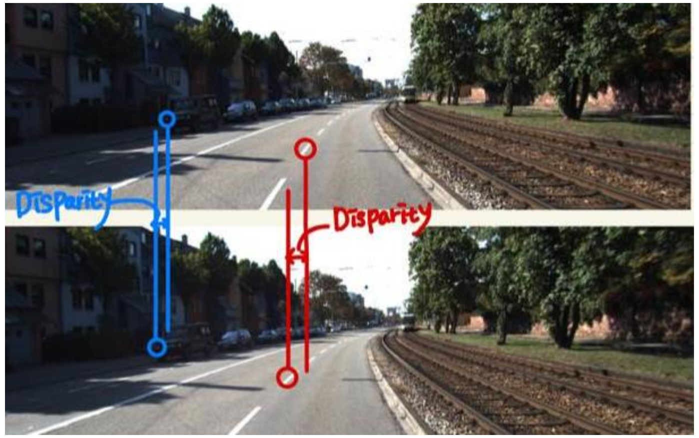
- Depth Estimation : 깊이 추정



- Monocular VS Stereo

- Monocular : 학습을 통한 depth estimation (지도 학습)

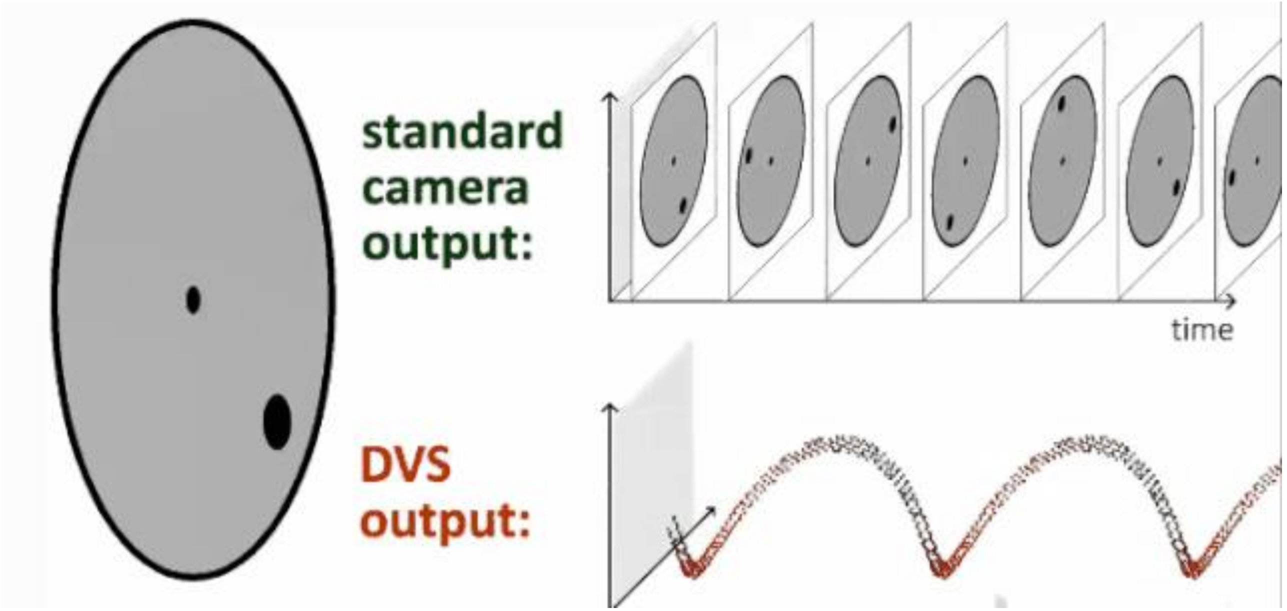
- Stereo : 두 이미지간의 시차를 통한 depth estimation
  - 시차가 작으면 멀리있는 물체
  - 시차가 크면 가까이 있는 물체



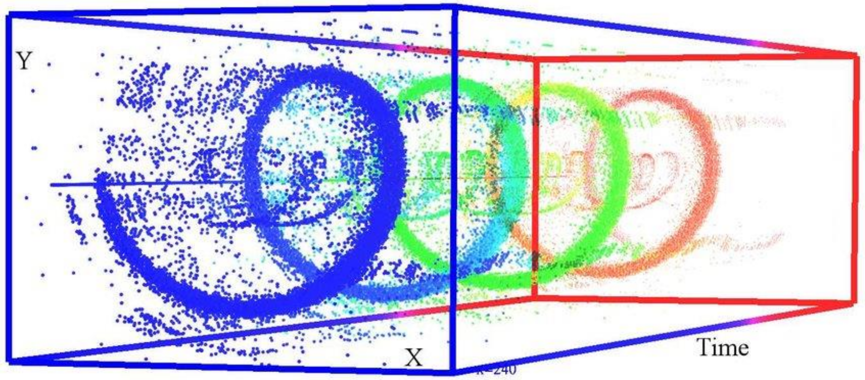
SMARCLE

Introduction to Event Camera

- 각각의 독립적인 pixel의 밝기 변화를 감지하는 센서
- scene에서 motion이 있을 경우에만 검출됨



- (x, y, t, p) 로 구성된 데이터

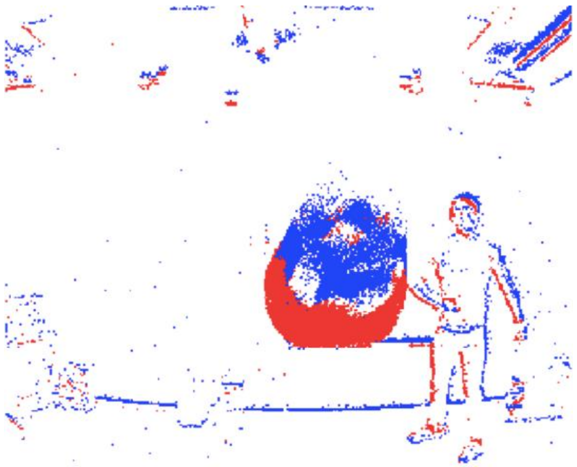


SMARCLE

Gallego, Guillermo, et al. "Event-based vision: A survey." *arXiv preprint arXiv:1904.08405* (2019).

Introduction to Event Camera

- Event Camera의 장점
  - High Temporal Resolution
    - Can capture very fast motions without suffering from motion blur
  - Low Latency
    - Each pixel works independently without waiting for a global exposure : change applies in 10µs
  - Low Power : power is only used to progress changing pixels
  - High Dynamic Range
    - HDR over 120dB. Works well in very dark or bright condition



SMARCLE

Gallego, Guillermo, et al. "Event-based vision: A survey." *arXiv preprint arXiv:1904.08405* (2019).

Introduction to E2Depth

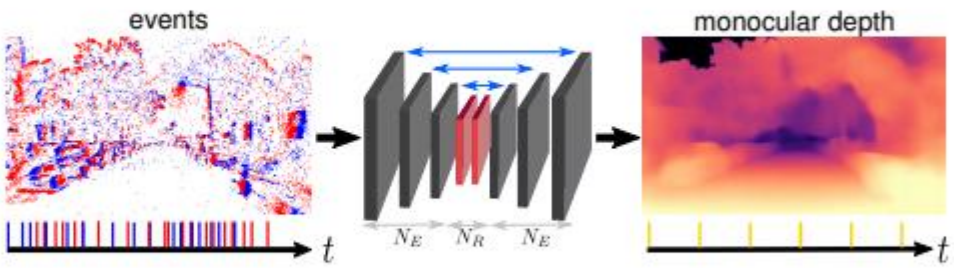


Figure 1: Method overview, the network receives asynchronous events inputs and predicts normalized log depth $\hat{\mathcal{D}}_k$. Our method uses N_R recurrent blocks to leverage the temporal consistency in the events input.

- event data로부터 dense depth map을 예측하는 모델
- input data : event voxel grid
- output : dense depth map

SMARCLE

Hidalgo-Carrió, Javier, Daniel Gehrig, and Davide Scaramuzza. "Learning monocular dense depth from events." *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020.

Introduction to E2Depth

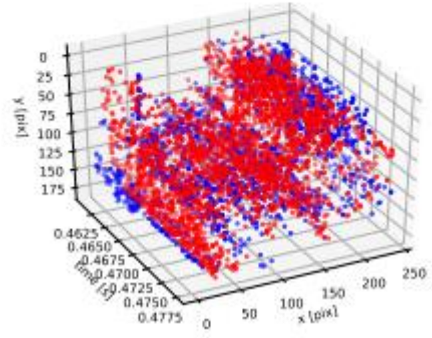
Contribution

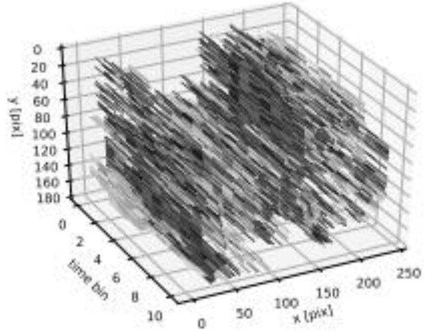
- 단안 이벤트 카메라를 이용하여 픽셀 단위의 세밀한 Depth를 추정하는 Recurrent Network 제시
- CARLA simulator의 event camera plugin 구현 (ESIM)
- DENSE(Depth Esitimation oN Synthetic Evnets) 데이터셋 제시 → Synthetic events와 GT 제공
- 우리의 방법을 MVSEC 데이터셋에 적용시키고 그 결과를 보임으로써 SOTA임을 증명

SMARCLE

Hidalgo-Carrió, Javier, Daniel Gehrig, and Davide Scaramuzza. "Learning monocular dense depth from events." *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020.

Input Data





- Raw Event Data
- Event Voxel-Grid

$$\mathbf{E}_k(\mathbf{u}_k, t_n) = \sum_{e_i} p_i \delta(\mathbf{u}_i - \mathbf{u}_k) \max(0, 1 - |t_n - t_i^*|) \quad (1)$$

$$t_i^* = \frac{B-1}{\Delta T}(t_i - t_0)$$

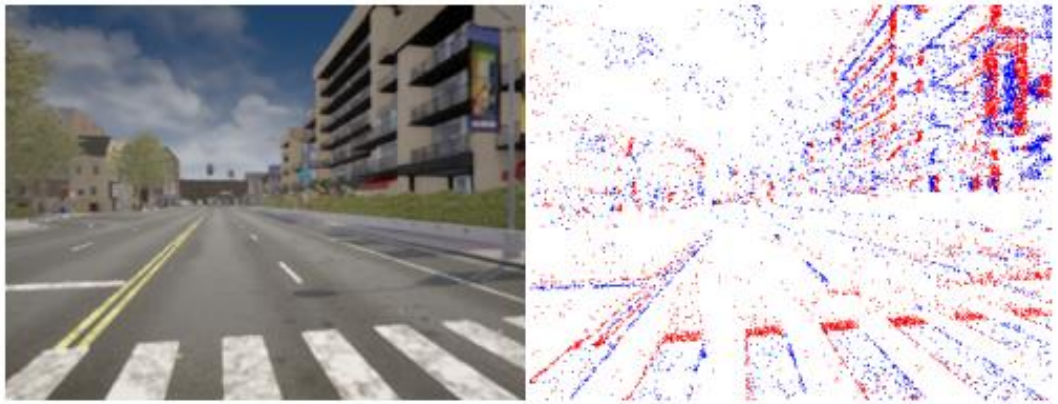
ΔT = 50ms
B = 5

- sparse한 event data의 특성상 spatio-temporal voxel-grid로 변환해 input으로 가져감
- 2D grid에서 이벤트들의 충돌을 방지하면서도 시간적 정보를 가져감

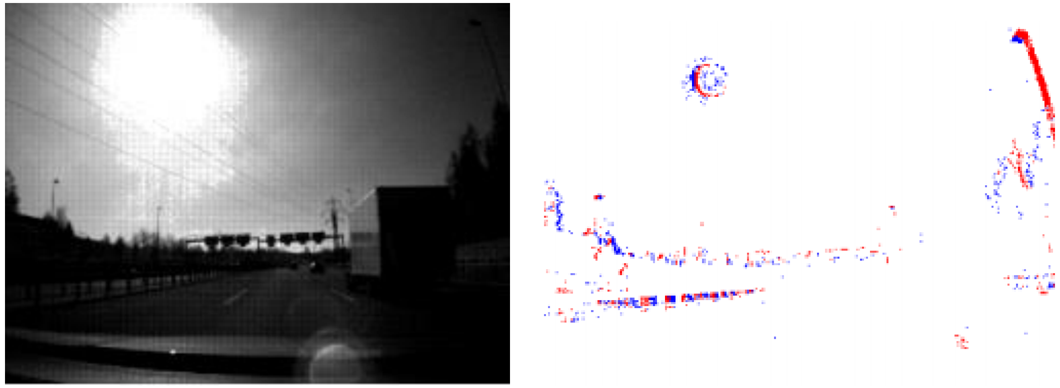
SMARCLE

Hidalgo-Carrió, Javier, Daniel Gehrig, and Davide Scaramuzza. "Learning monocular dense depth from events." *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020.

Input Data



- CARLA Synthetic Data



- MVSEC Real Data

- Train on CARLA -> finetuned on MVSEC

SMARCLE

Hidalgo-Carrió, Javier, Daniel Gehrig, and Davide Scaramuzza. "Learning monocular dense depth from events." *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020.

E2Depth Model

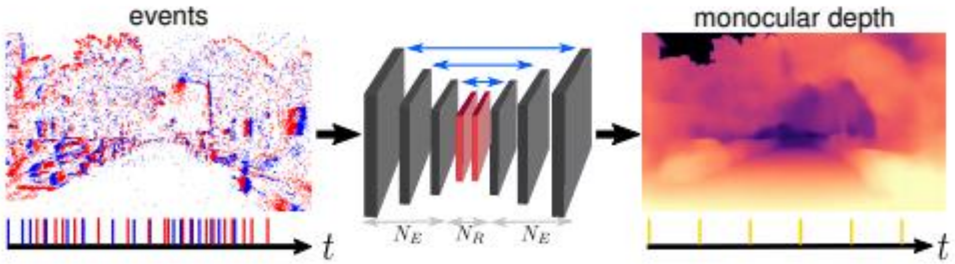


Figure 1: Method overview, the network receives asynchronous events inputs and predicts normalized log depth $\hat{\mathcal{D}}_k$. Our method uses N_R recurrent blocks to leverage the temporal consistency in the events input.

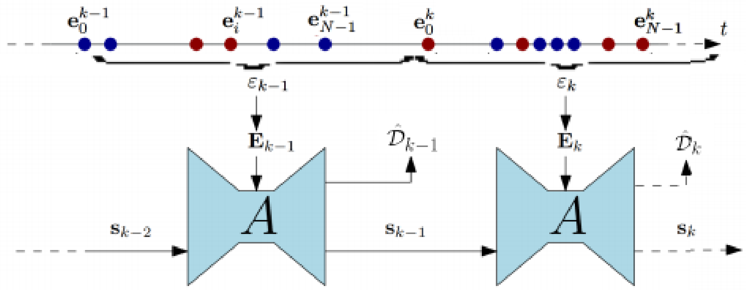


Figure 2: Our network architecture, image adapted from [27]. The event stream is grouped into non-overlapping windows of events and converted to tensor-like voxel grids [40]. These voxel grids are passed to our recurrent fully convolutional neural network to produce normalized log depth predictions.

LSTM recurrent fully convolutional neural network based on UNet

- 3 Encoder layers(kernel 5) followed by ConvLSTM
- 2 Recurrent block
- ReLU activation function except for the prediction layer(sigmoid)

Hidalgo-Carrió, Javier, Daniel Gehrig, and Davide Scaramuzza. "Learning monocular dense depth from events." *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020.

Training

$$\mathcal{R}_k = \hat{\mathcal{D}}_k - \mathcal{D}_k$$ ground truth depth maps $\{\mathcal{D}_k\}$

$$\mathcal{L}_{k,\mathrm{si}} = \frac{1}{n}\sum_{\mathbf{u}}(\mathcal{R}_k(\mathbf{u}))^2 - \frac{1}{n^2}\left(\sum_{\mathbf{u}}\mathcal{R}_k(\mathbf{u})\right)^2,$$

- Scale-invariant loss

$$\mathcal{L}_{k,\mathrm{grad}} = \frac{1}{n}\sum_{s}\sum_{\mathbf{u}}|\nabla_x\mathcal{R}_k^s(\mathbf{u})| + |\nabla_y\mathcal{R}_k^s(\mathbf{u})|.$$

- multi-scale scale-invariant loss

$$\mathcal{L}_{\mathrm{tot}} = \sum_{k=0}^{L-1}\mathcal{L}_{k,\mathrm{si}} + \lambda\mathcal{L}_{k,\mathrm{grad}}.$$

- Resulting loss, $\lambda = 0.5$

- encourages smooth depth changes and enforces sharp depth discontinuities in the depth map prediction

- batch size = 20
- learning rate = 0.0001
- optimizer = Adam

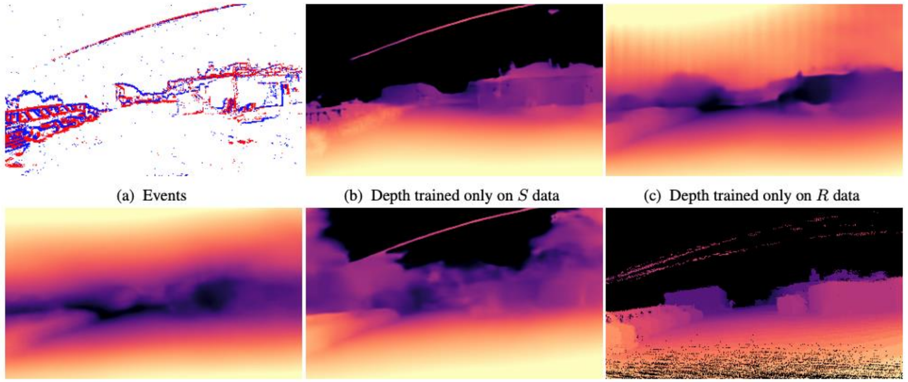
SMARCLE

Hidalgo-Carrió, Javier, Daniel Gehrig, and Davide Scaramuzza. "Learning monocular dense depth from events." *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020.

Experiments

Training set	Dataset	Abs Rel↓	Sq Rel ↓	RMSE↓	RMSE log↓	SI log↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
S	outdoor day1	0.698	3.602	12.677	0.568	0.277	0.493	0.708	0.808
R		0.450	0.627	9.321	0.514	0.251	0.472	0.711	0.823
S* → R		0.381	**0.464**	9.621	0.473	0.190	0.392	0.719	0.844
S* → (S+R)		**0.346**	0.516	**8.564**	**0.421**	**0.172**	**0.567**	**0.772**	**0.876**
S	outdoor night1	1.933	24.64	19.93	0.912	0.429	0.293	0.472	0.600
R		0.770	3.133	**10.548**	0.638	0.346	0.327	0.582	0.732
S* → R		**0.554**	**1.798**	10.738	**0.622**	**0.343**	0.390	0.598	0.737
S* → (S+R)		0.591	2.121	11.210	0.646	0.374	**0.408**	**0.615**	**0.754**
S	outdoor night2	0.739	3.190	13.361	0.630	0.301	0.361	0.587	0.737
R		0.400	0.554	**8.106**	**0.448**	**0.176**	0.411	0.720	0.866
S* → R		0.367	**0.369**	9.870	0.621	0.279	0.422	0.627	0.745
S* → (S+R)		**0.325**	0.452	9.155	0.515	0.240	**0.510**	**0.723**	**0.840**
S	outdoor night3	0.683	1.956	13.536	0.623	0.299	0.381	0.593	0.736
R		0.343	0.291	**7.668**	**0.410**	**0.157**	0.451	0.753	**0.890**
S* → R		0.339	0.230	9.537	0.606	0.258	0.429	0.644	0.760
S* → (S+R)		**0.277**	**0.226**	8.056	0.424	0.162	**0.541**	**0.761**	**0.890**

Table 2: Ablation study and evaluation of MVSEC. All rows are the same network with the change in the training set. The Training set is denoted with S (synthetic data from the DENSE training split), R (real data from the training split in *outdoor day2* sequence), S^* (first 1000 samples of the DENSE training split), $S^* \rightarrow R$ (pretrained on S^* and retrained on R), $S^* \rightarrow (S+R)$ (pretrained on S^* and retrained on both datasets). ↓ indicates lower is better and ↑ higher is better. The results are the driving sequences of MVSEC (except for *outdoor day2*). Best values are shown in bold.



(a) Events (b) Depth trained only on S data (c) Depth trained only on R data

(d) Depth trained on $S^* \rightarrow R$ (e) Depth trained on $S^* \rightarrow (S+R)$ (f) Depth ground truth

Figure 3: Ablation study of our method trained with different training sets (see Table 2). Fig. 3a shows the events, from Fig. 3b to Fig. 3e the predicted dense monocular depth using different training sets. Fig. 3f depicts the corresponding ground truth. The depth maps are shown in logarithmic scale and correspond to sample 3562 in the *outdoor day1* sequence of MVSEC.

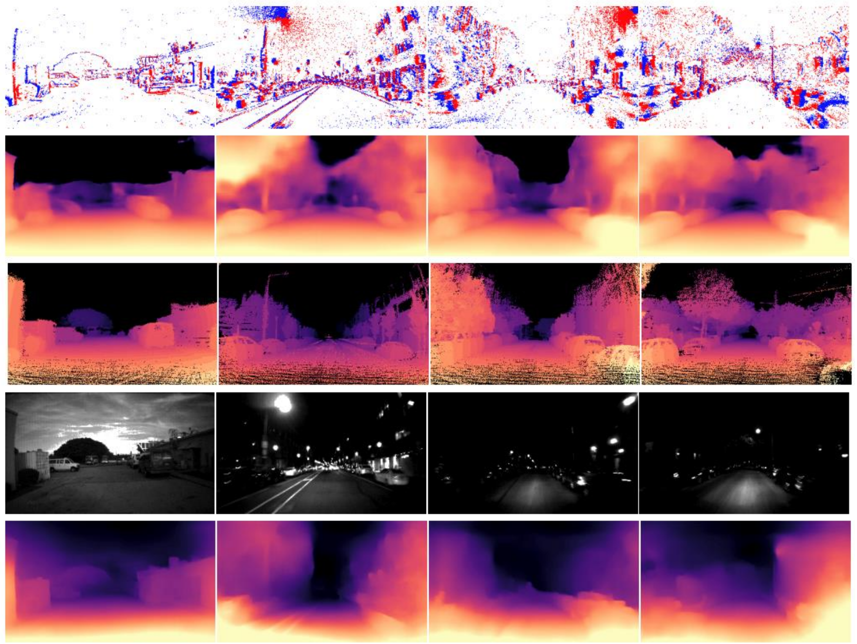
SMARCLE

Hidalgo-Carrió, Javier, Daniel Gehrig, and Davide Scaramuzza. "Learning monocular dense depth from events." *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020.

Experiments

Dataset	Distance	Frame based			Event based				
		MonoDepth [10]	MegaDepth [18]	MegaDepth^{+} [18]	Zhu et al. [40]	Ours^{S}	Ours^{R}	Ours$^{S^{*}\rightarrow R}$	Ours$^{\#}$
outdoor day1	10m	3.44	2.37	3.37	2.72	4.60	2.70	2.13	**1.85**
	20m	7.02	4.06	5.65	3.84	5.66	3.46	2.68	**2.64**
	30m	10.03	5.38	7.29	4.40	6.10	3.84	3.22	**3.13**
outdoor night1	10m	3.49	2.54	**2.40**	3.13	10.36	5.36	3.31	3.38
	20m	6.33	4.15	4.20	4.02	12.97	5.32	**3.73**	3.82
	30m	9.31	5.60	5.80	4.89	13.64	5.40	**4.32**	4.46
outdoor night2	10m	5.15	3.92	3.39	2.19	6.14	2.80	1.99	**1.67**
	20m	7.80	5.78	4.99	3.15	8.64	3.28	3.14	**2.63**
	30m	10.03	7.05	6.22	3.92	9.57	3.74	4.14	**3.58**
outdoor night3	10m	4.67	4.15	4.56	2.86	5.72	2.39	1.76	**1.42**
	20m	8.96	6.00	5.63	4.46	8.29	2.88	2.98	**2.33**
	30m	13.36	7.24	6.51	5.05	9.27	3.39	3.98	**3.18**

Table 3: Average absolute depth errors (in meters) at different cut-off depth distances (lower is better). MegaDepth^{+} refers to MegaDepth [18] using E2VID [27] reconstructed frames and Ours$^{\#}$ refers to our method trained using $S^{*} \rightarrow (S + R)$. Our results outperform state of the art image-based monocular depth prediction methods [10, 18] while outperforming state of the art event-based methods [40].



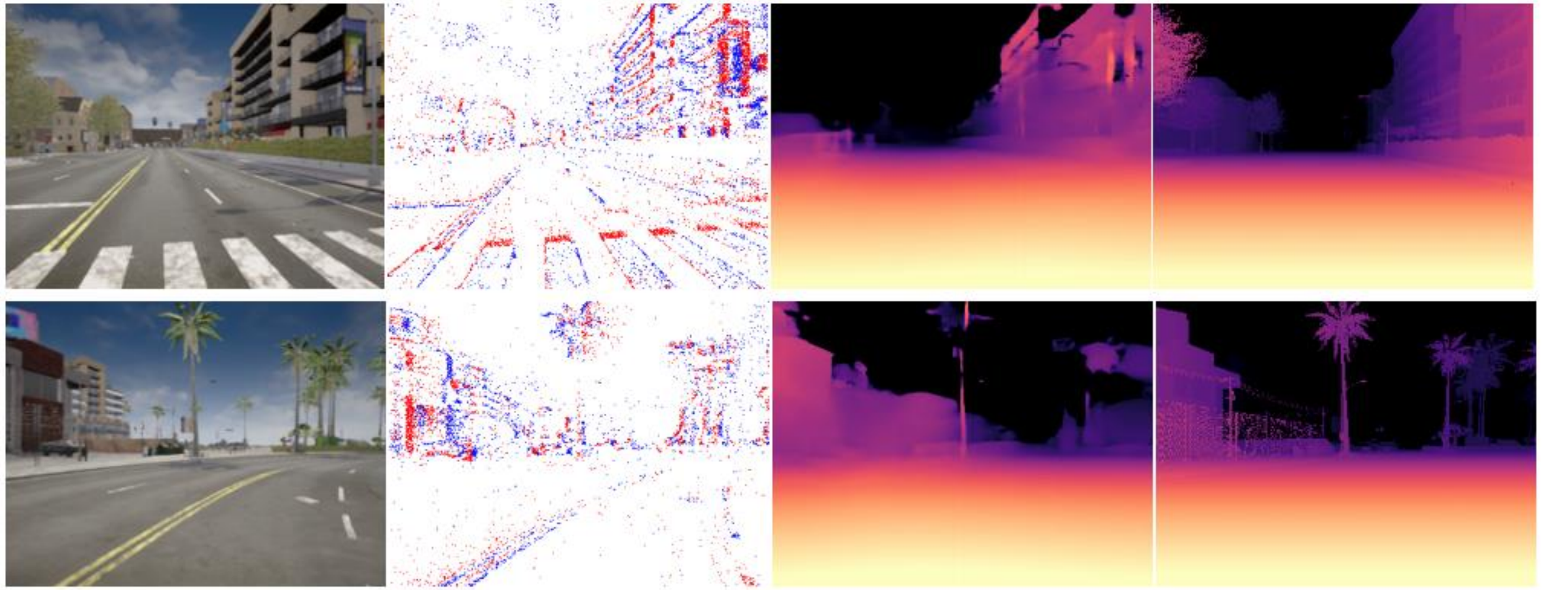
(a) outdoor day1 (b) outdoor night1 (c) outdoor night2 (d) outdoor night3

Hidalgo-Carrió, Javier, Daniel Gehrig, and Davide Scaramuzza. "Learning monocular dense depth from events." *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020.

Results

Dataset	Abs Rel↓	Sq Rel ↓	RMSE↓	RMSE log↓	SI log↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	Avg. error 10m↓	Avg. error 20m↓	Avg. error 30m↓
Town06	0.120	0.083	6.640	0.188	0.035	0.855	0.956	0.987	0.31	0.74	1.32
Town07	0.267	0.535	10.182	0.328	0.098	0.774	0.878	0.927	1.03	2.35	3.06
Town10	0.220	0.279	11.812	0.323	0.093	0.724	0.865	0.932	0.61	1.45	2.42

Table 4: Quantitative results on the DENSE dataset. We train the network only on synthetic events from the training split S. The first two sequences are used for validation and the Town10 sequence for testing.



(a) Frame (b) Events (c) Ours on events (d) Ground truth

Figure 5: Qualitative results on DENSE for the Town10 sequence. The first row corresponds to sample 143 and the second row to sample 547 in the sequence.

SMARCLE

Hidalgo-Carrió, Javier, Daniel Gehrig, and Davide Scaramuzza. "Learning monocular dense depth from events." *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020.

감사합니다

PR-SMARCLE

김찬영