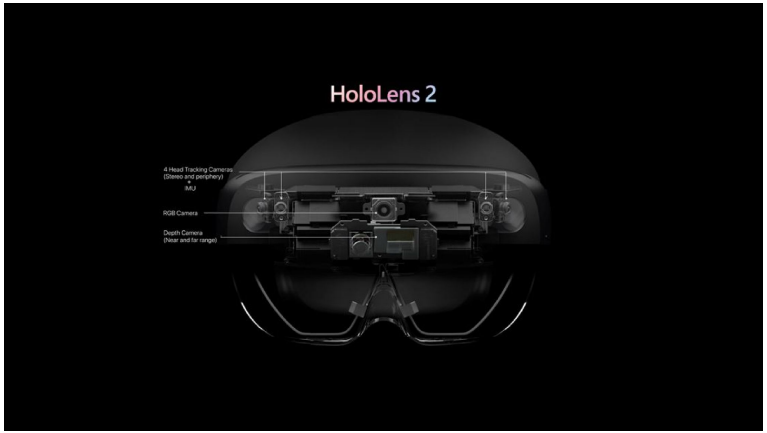
Self-Supervised Event-Intensity Stereo Matching

EI 2023, San Francisco, United States

Jinjin Gu, Jinan Zhou, Ringo Sai wo Chu, Yan Chen, Jiawei Zhang,
Xuanye Cheng, Song Zhang, Jimmy S. Ren

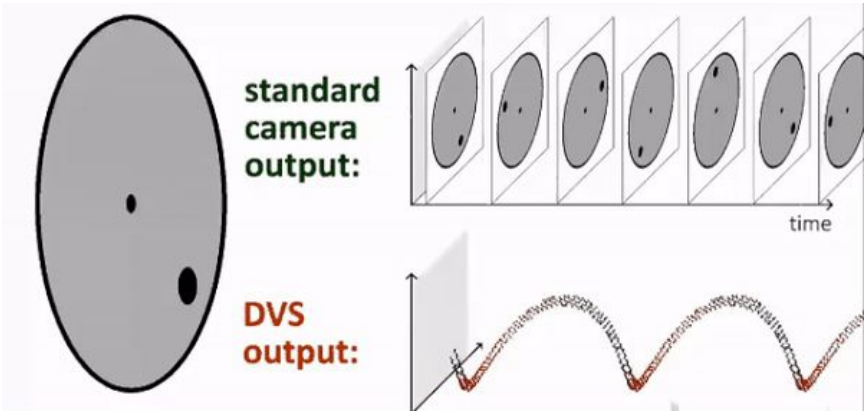
What is Stereo Matching?

- A process to **infer depth** from two or more cameras.
- Require rectifying stereo images, compute depth from matched pixels
- Application including AR/VR, Self-Driving Cars



Event Camera

- Dubbed '**Silicon Retina**', as Event Cameras mimic the human visual system
- Bio-inspired vision sensors that output pixel-level brightness changes instead of standard intensity frames



Why Event Camera?

🙂 High Dynamic Range
🙂 No Motion Blurring
🙂 Low Latency
🙂 High Temporal Resolution

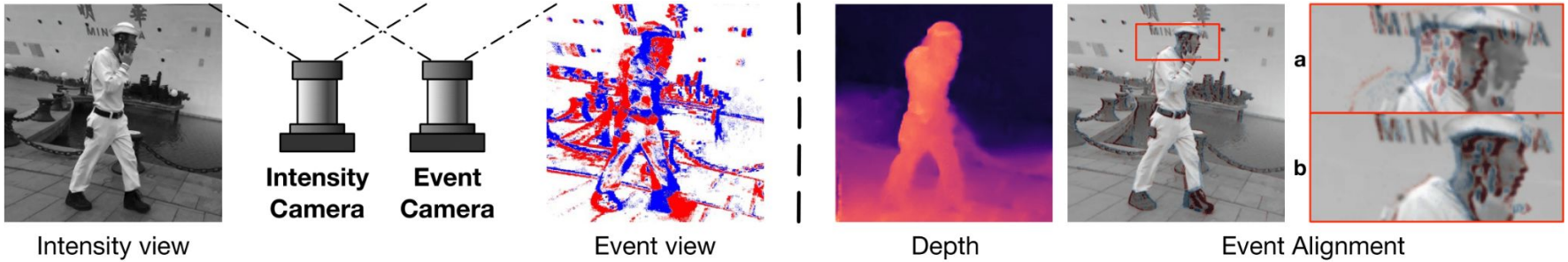
But...

😟 Traditional frame-based algorithm does not apply, due to asynchronous pixels and no intensity information

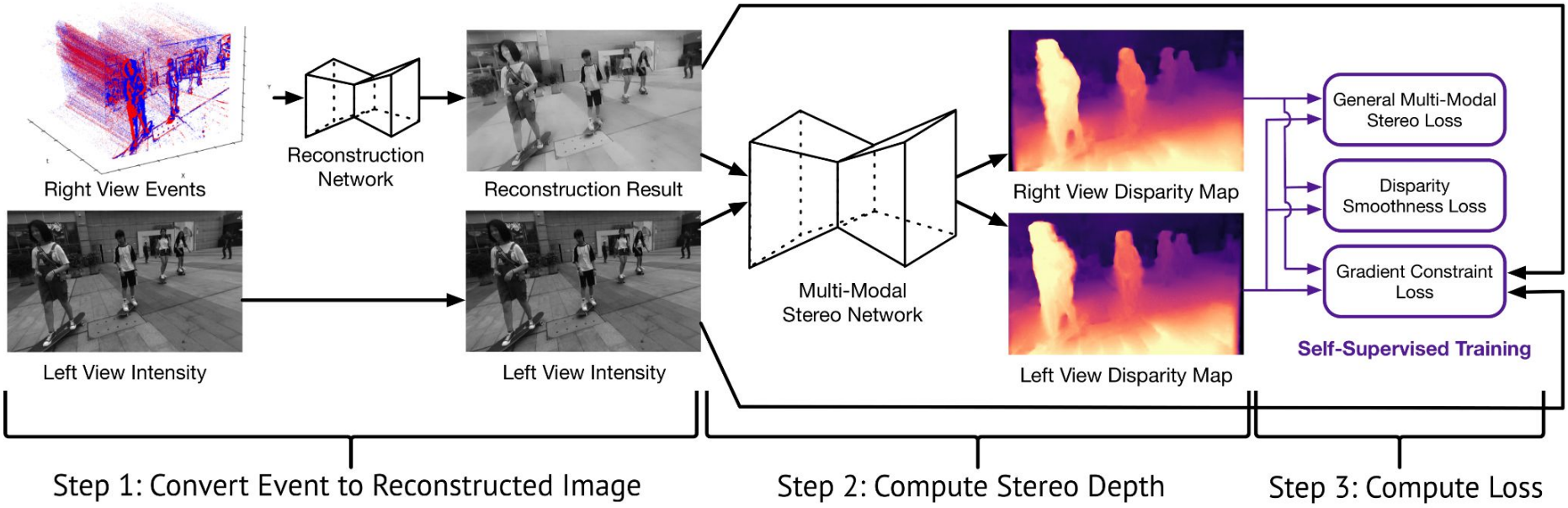
Motivation

- **Event-Intensity Stereo**: Combining a Frame camera and an Event Camera for Stereo
  - 👉 Absorb the advantages from both modality.

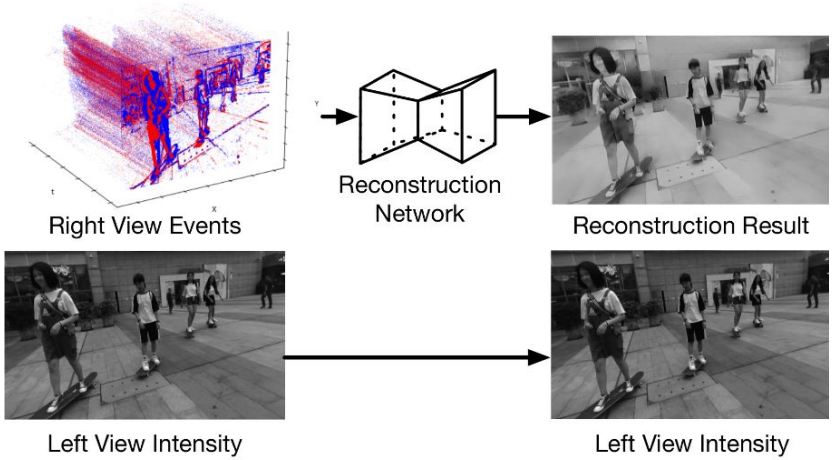
- A Self-Supervised Learning Paradigm
  - 👉 Mitigate multi-modal data collection and processing.



Our Method Overview

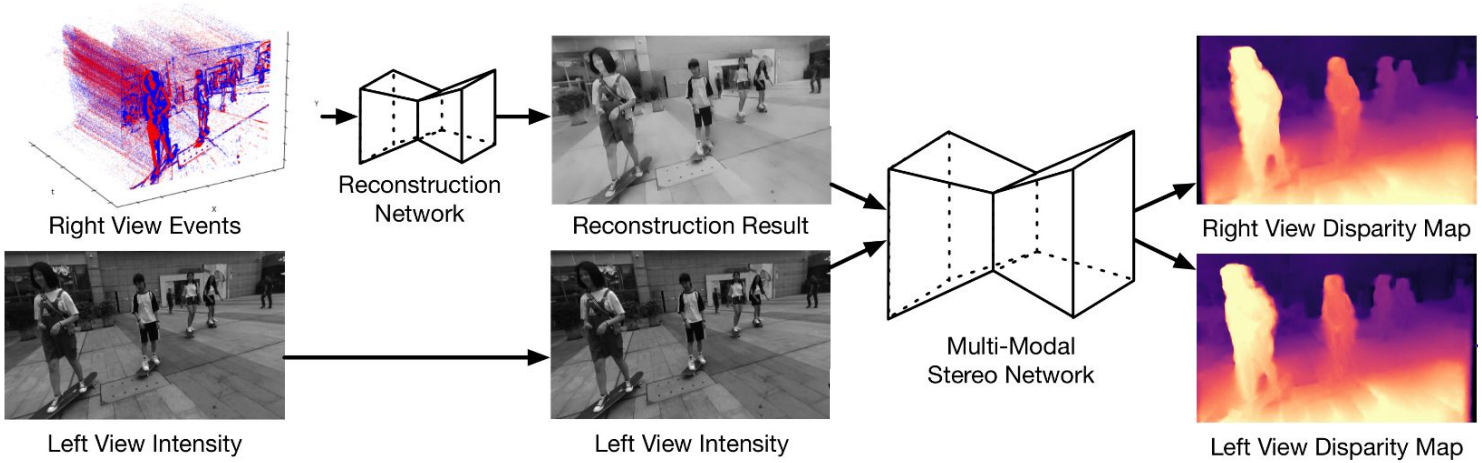


Our Method Overview - Event Reconstruction



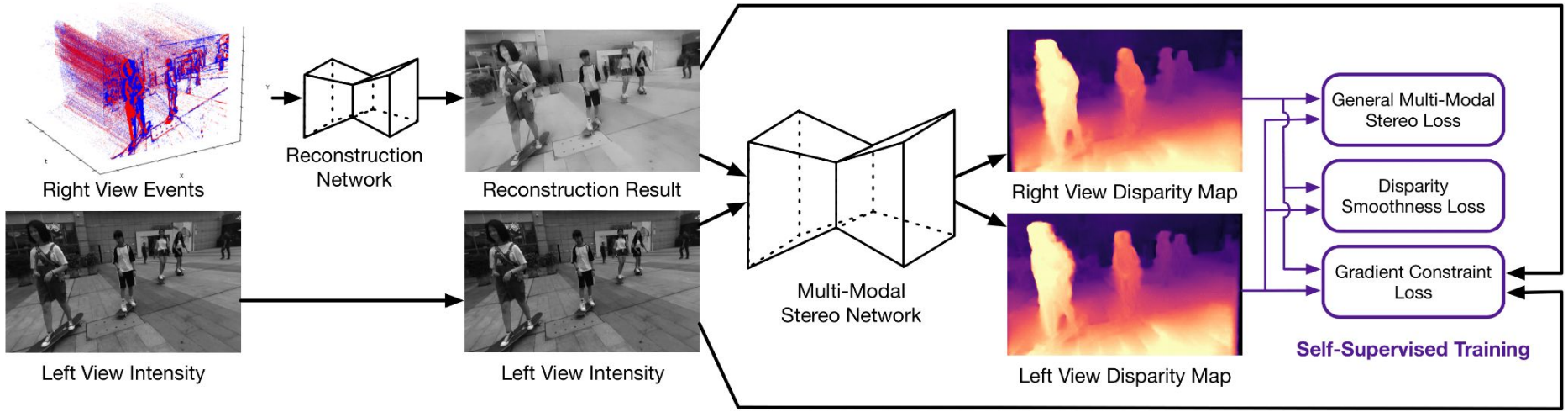
- Obtain a coarse image reconstruction from events.
- Off-the-shelf reconstruction model. Such as FireNet or E2VID

Our Method Overview - Stereo Matching



- Off-the-Shelf Stereo matching models, with minor changes
- Instead of weight sharing backbone, we use different backbones for modalities

Our Method Overview - Self-Supervised Training



- **Self-Supervised Training Loss:** Gradient Structure Loss + Disparity Smoothness Loss + General-Modal Stereo Loss

Self-Supervised Loss Function - Part 1

- **Image Structure Loss** - Use image gradient for structure information

$$\mathcal{L}_{gd} = 1 - \frac{2\mu_{G^l}\mu_{G^r} + c_1}{\mu_{G^l}^2 + \mu_{G^r}^2 + c_1} \times \frac{2\sigma_{G^l G^r} + c_2}{\sigma_{G^l}^2 + \sigma_{G^r}^2 + c_2}$$

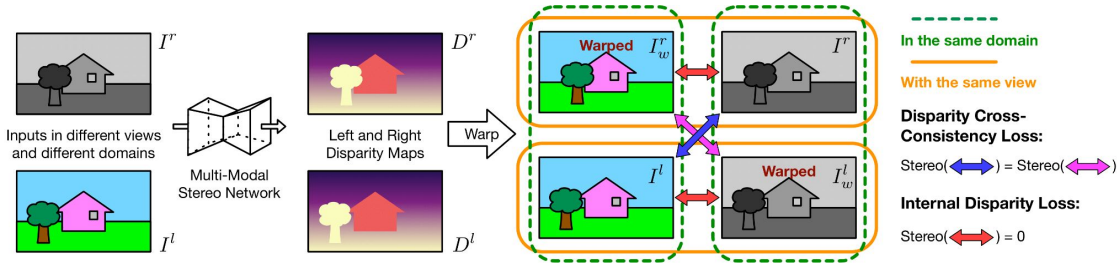
Self-Supervised Loss Function - Part 2

- **Disparity Smoothness Loss**

$$\mathcal{L}_{\text{sm}} = \frac{1}{N} \sum_{i,j} |\nabla_x D_{ij}| e^{-|\nabla_x I_{ij}|} + |\nabla_y D_{ij}| e^{-|\nabla_y I_{ij}|}$$

Self-Supervised Loss Function - Part 3

- **General Multi-Modal Stereo Losses**



With the above information, we propose cross-consistency loss,

$$\mathcal{L}_{cc} = \frac{1}{N}\sum_{i,j}\Big|\,|D^l| - |D^r_{\mathrm{w}}|\,\Big| + \Big|\,|D^r| - |D^l_{\mathrm{w}}|\,\Big|$$

And internal disparity Loss, D^r_{itn} from I^r, I^r_{wrap} and D^l_{itn} from I^l I^l_{wrap}

$$\mathcal{L}_{\mathrm{itn}} = \frac{1}{N}\sum_{i,j}|D^r_{\mathrm{itn}}| + |D^l_{\mathrm{itn}}|$$

Experiment Setup

- **Synthetic Dataset** -

  Taken from Stereo Blur Dataset - Argument with frame interpolation from 60FPS to 2400FPS, followed by V2E event simulator for event generation.
- **Real Dataset** -

  MVSEC - contains the stereo intensity images and events captured by DAVIS 240C

Results

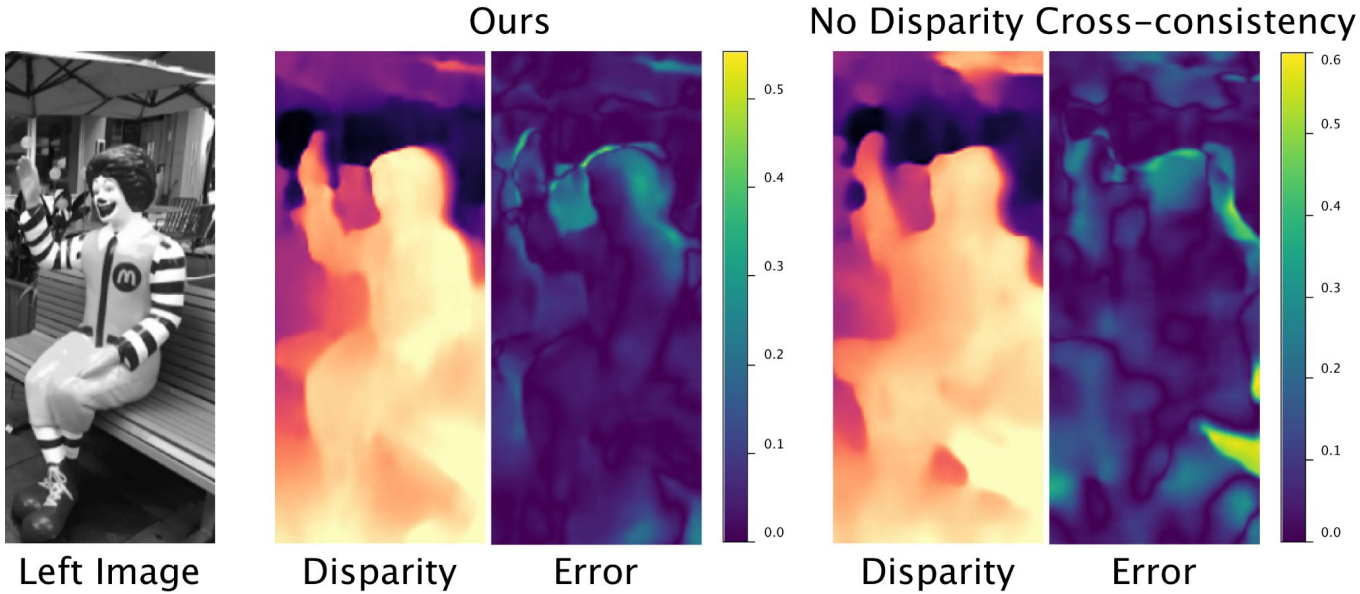
Model	EPE ↓	Bad Pixels ↓		
		$\delta > 1$	$\delta > 3$	$\delta > 5$
Monodepth2	8.849	0.953	0.781	0.648
DeepPruner (upper bound)	0.712	0.123	0.027	0.015
FireNet+AANet (baseline)	4.811	0.649	0.419	0.336
E2VID+AANet (baseline)	5.154	0.673	0.440	0.379
FireNet+DeepPruner (baseline)	10.29	0.417	0.226	0.181
E2VID+DeepPruner (baseline)	6.386	0.381	0.184	0.140
FireNet+AANet* ($\mathcal{L}_{gd}$ and $\mathcal{L}_{sm}$)	1.591	0.366	0.139	0.088
E2VID+AANet* ($\mathcal{L}_{gd}$ and $\mathcal{L}_{sm}$)	1.496	0.351	0.123	0.075
FireNet+DeepPruner* ($\mathcal{L}_{gd}$ and $\mathcal{L}_{sm}$)	1.336	0.355	0.123	0.068
E2VID+DeepPruner* ($\mathcal{L}_{gd}$ and $\mathcal{L}_{sm}$)	1.321	0.355	0.116	0.068
FireNet+AANet (all losses)	1.988	0.409	0.189	0.134
E2VID+AANet (all losses)	1.775	0.378	0.166	0.117
FireNet+DeepPruner (all losses)	1.626	0.377	0.147	0.097
E2VID+DeepPruner (all losses)	1.57	0.368	0.143	0.094
FireNet+AANet* (all losses)	1.201	0.306	0.110	0.065
E2VID+AANet* (all losses)	1.101	0.287	0.094	0.057
FireNet+DeepPruner* (all losses)	0.971	0.317	0.087	0.049
E2VID+DeepPruner* (all losses)	0.913	0.289	0.074	0.042

On synthetic dataset 👍

Model	EPE ↓	Bad Pixels ↓		
		$\delta > 1$	$\delta > 3$	$\delta > 5$
Monodepth2	10.235	0.914	0.844	0.768
E2VID+AANet (baseline)	11.332	0.954	0.864	0.776
E2VID+AANet (all losses)	5.830	0.736	0.660	0.434
E2VID+DeepPruner (all losses)	4.979	0.673	0.581	0.384
E2VID+AANet* (all losses)	2.734	0.653	0.330	0.197
E2VID+DeepPruner* (all losses)	2.397	0.601	0.268	0.164

On real dataset 👍

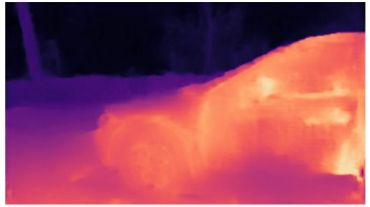
Visualisation

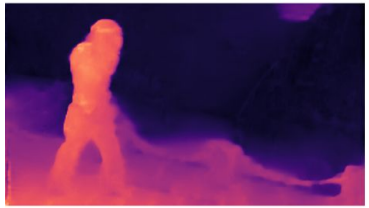


Visualisation

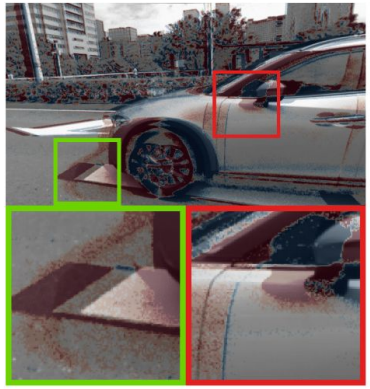
Visualisation

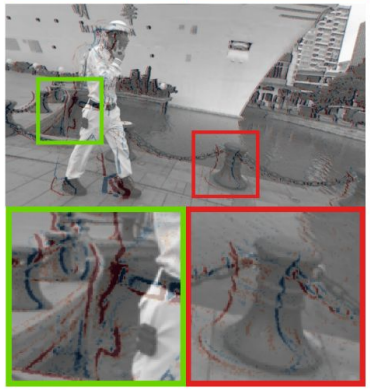
Disparity Map



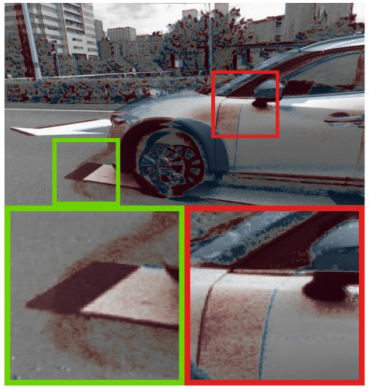


Before Warping





Warping Event to Intensity



Frame Interpolation using Events

Limitation

- Performance degradation when Event Intensity Reconstruction produces low-quality results.

Conclusion

- We propose Event-Intensity Stereo- A novel multi-modal stereo setup with standalone event and frame camera.

- We propose a self-supervised loss formulated from image gradient structure loss, disparity smoothness loss, cross-consistency and internal consistency

- Our method is robust to synthetic and real dataset.

Thank You!