Zero-Shot
Event-Intensity Asymmetric Stereo
via Visual Prompting from Image Domain

Hanyue Lou[#1] Jinxiu Liang[#1] Minggui Teng[1] Bin Fan[1] Yong Xu[2] Boxin Shi[1]
[1] Peking University [2] South China University of Technology
{hylz, cssherryliang, minggui_teng, binfan, shiboxin}@pku.edu.cn yxu@scut.edu.cn

北京大学 PEKING UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

華南理工大學 South China University of Technology

Stereo vision

- Stereo vision estimates depth by mimicking human binocular vision.
- It computes the disparity between images captured by each camera to estimate the distance of objects.
- Applications: 3D reconstruction, robotics, autonomous driving……

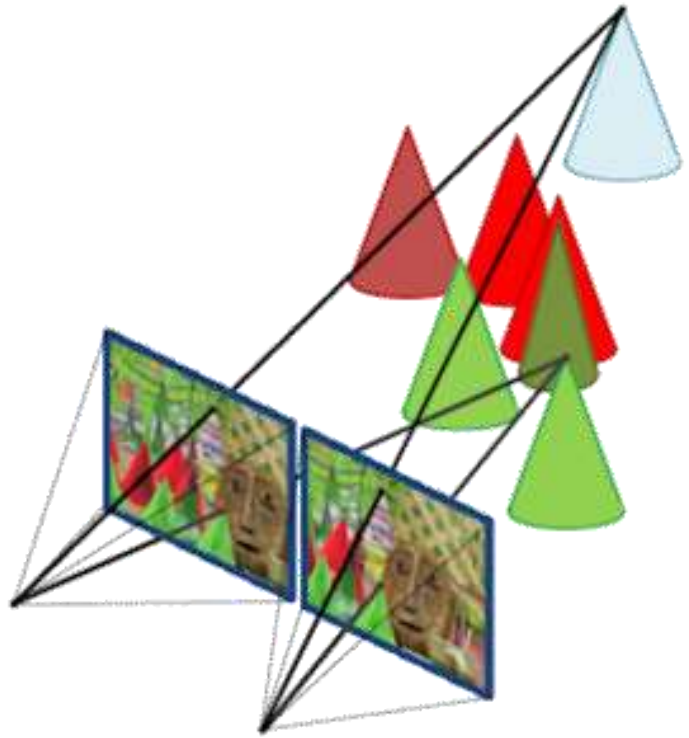


Image source: en.ids-imaging.com

Event-based vision

- Event cameras are novel visual sensors that report intensity changes, providing high dynamic range and high temporal resolution.
- However, no event signals are triggered when the scene is static or lacks texture.

Intensity images & Event signals

*Video courtesy of Elias Mueggler

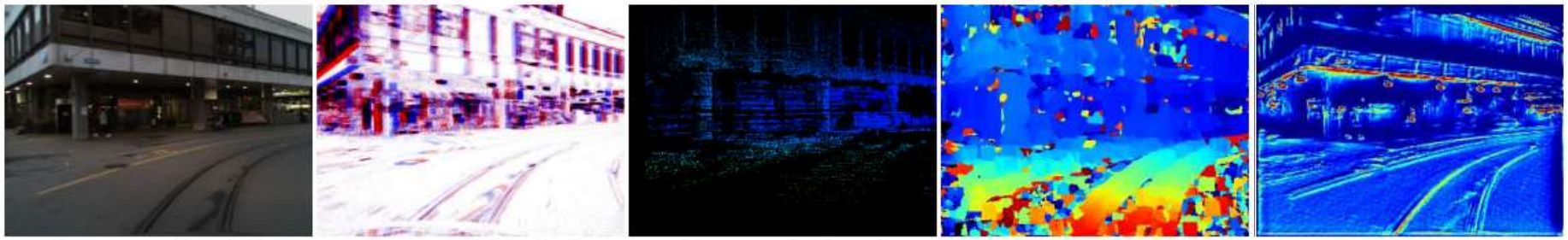
Event-Image Stereo

- Image-based cameras suffer from low dynamic range and low temporal resolution. However, they always provide spatially dense information.
- Events and images provide complementary information, making them a good combination for a stereo system.

(Left view)  +  (Right view)

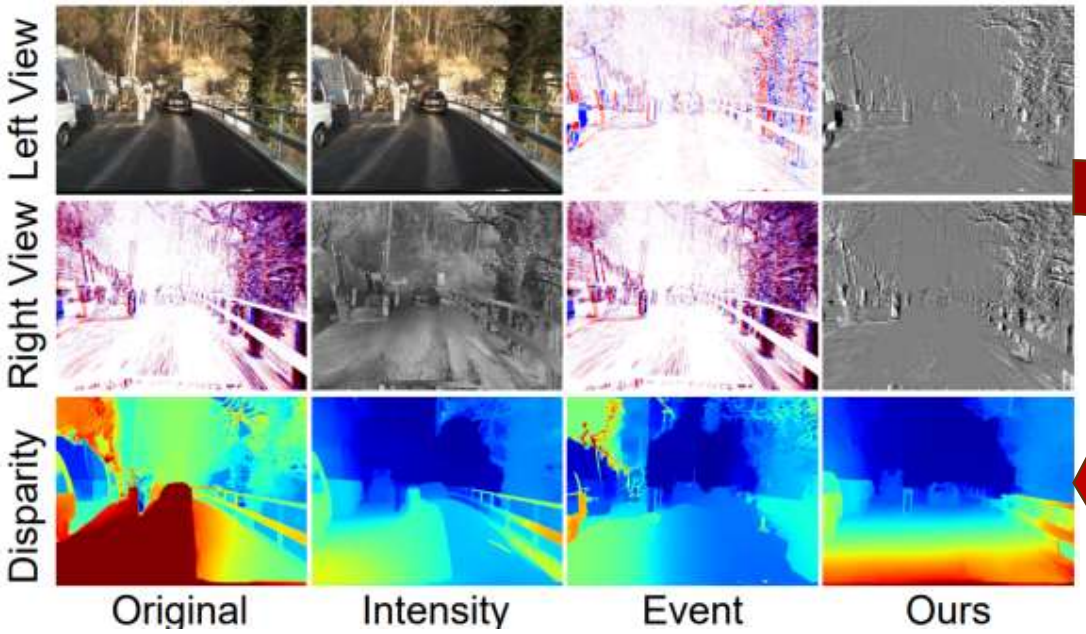
Challenge

- Problem: Existing E.-I. A. S. datasets are not sufficient.
- Traditional methods → Limited performance
- Data-driven methods → Overfit on DSEC / MVSEC



Motivation

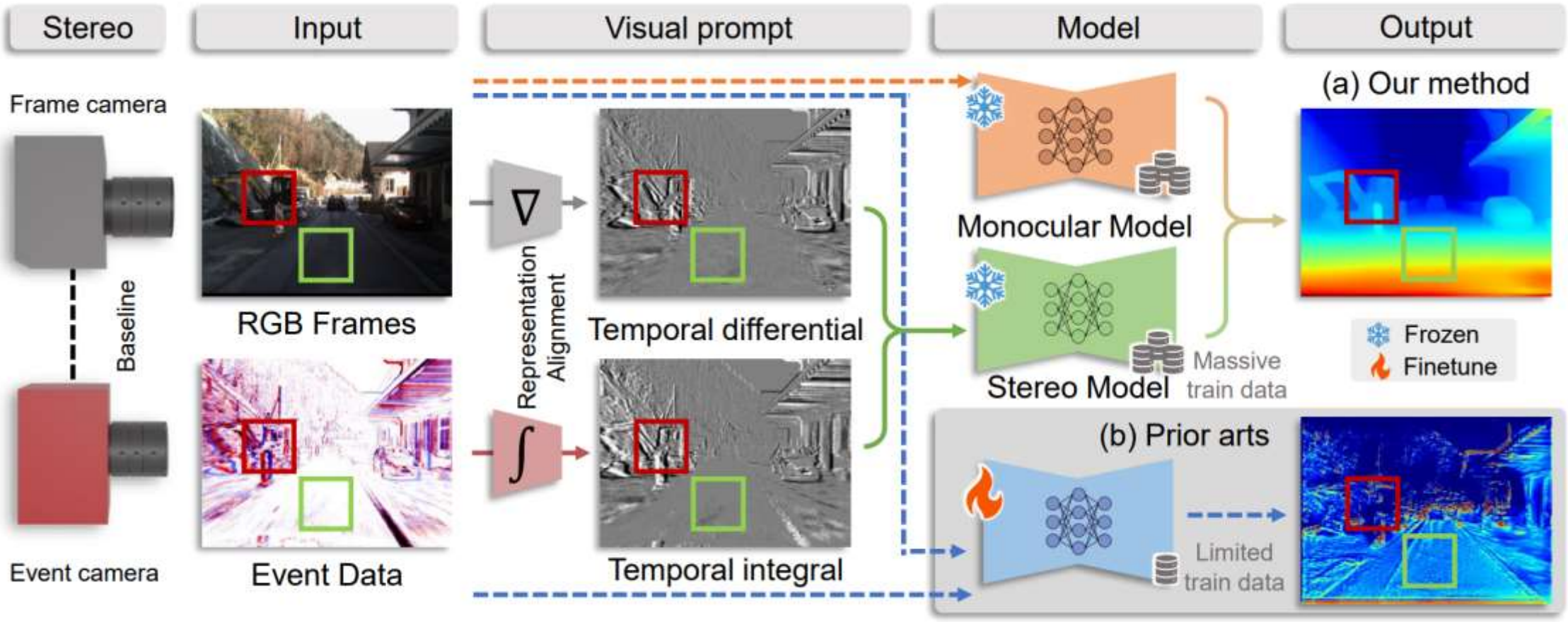
- Surprisingly: Existing frame-based methods work **very** well!
- Frame-based SOTA methods can generalize to non-natural images and even overcome large differences in appearance.



Frame-based stereo methods

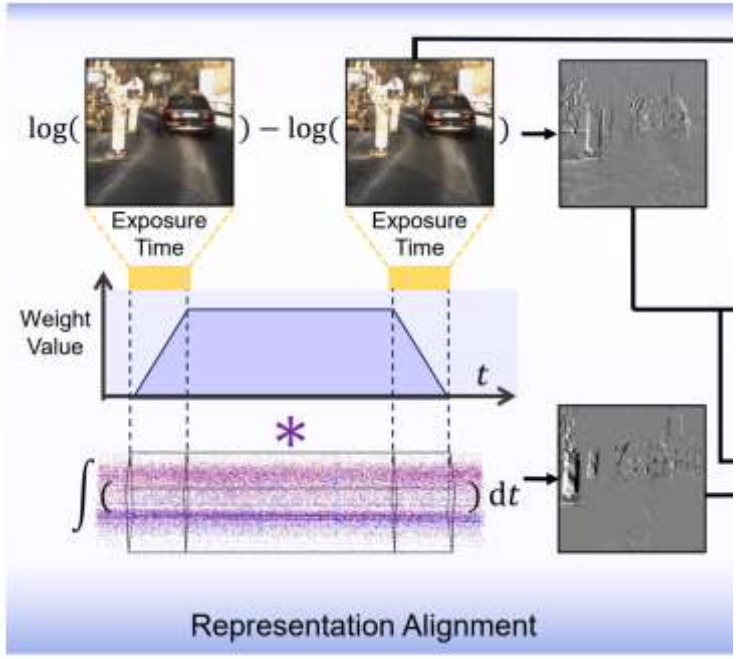
E-I asymmetric stereo methods

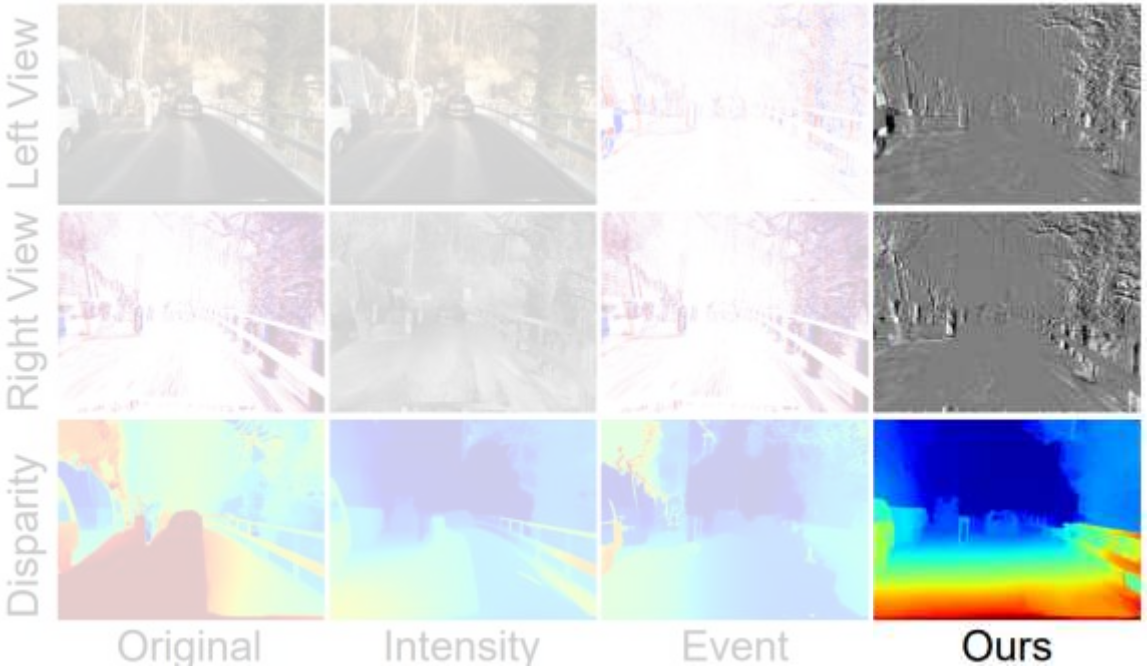
Method


Stereo
Input
Visual prompt
Model
Output
Frame camera
Baseline
Event camera
RGB Frames
Event Data
Representation Alignment
Temporal differential
Temporal integral
Monocular Model
Stereo Model
Massive train data
(a) Our method
Frozen
Finetune
(b) Prior arts
Limited train data

Method

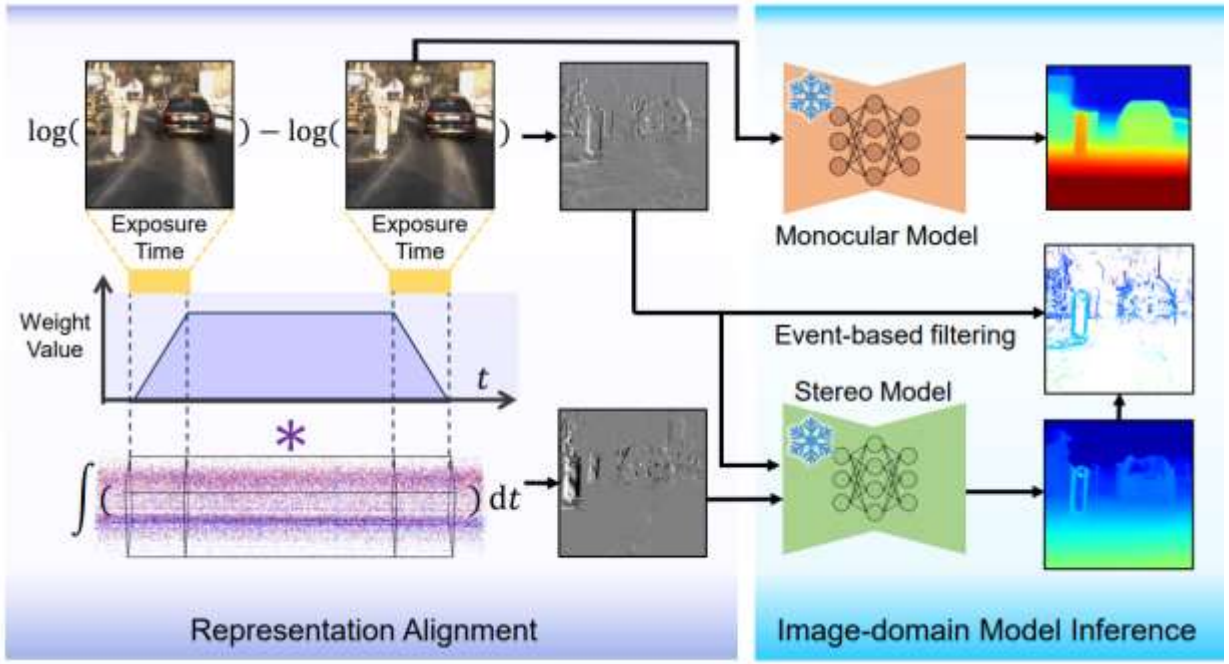
- We design a “visual prompt”: an intermediate representation that minimizes the appearance gap between images and event streams.

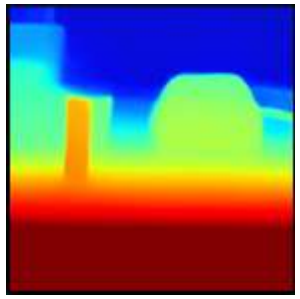




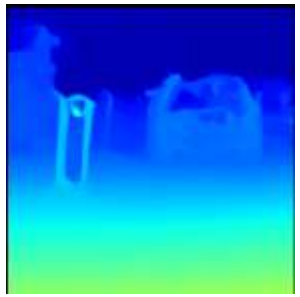
Method

- However, the stereo method is erroneous where events are sparse. A frame-based monocular depth estimation model can provide complementary information, but its output is relative.





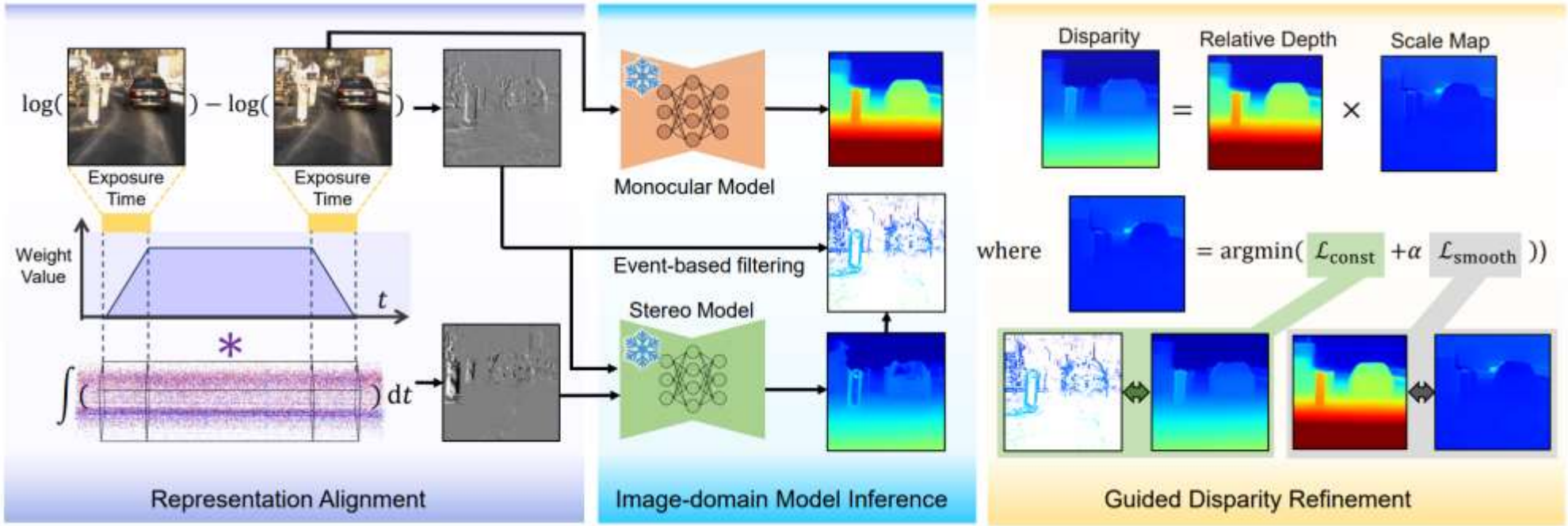
Monocular results
😊 Dense
😕 Relative depth



Monocular results
😊 Absolute depth
😕 Sparse

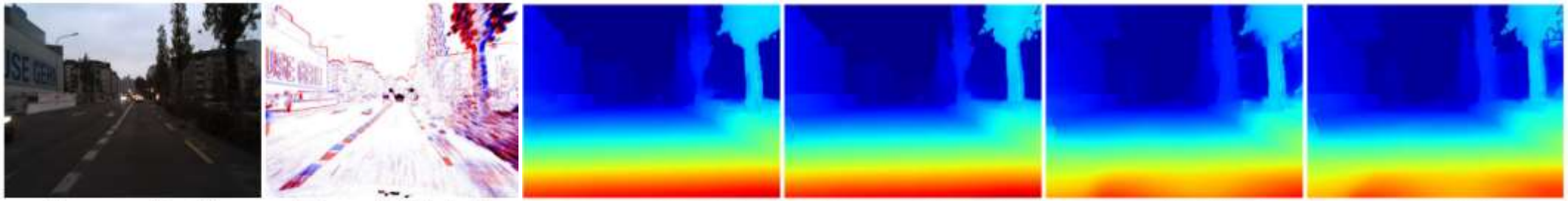
Method

- We solve an optimization problem to fuse the relative monocular results and the absolute stereo results. The refined disparity only follows the stereo where events are dense.



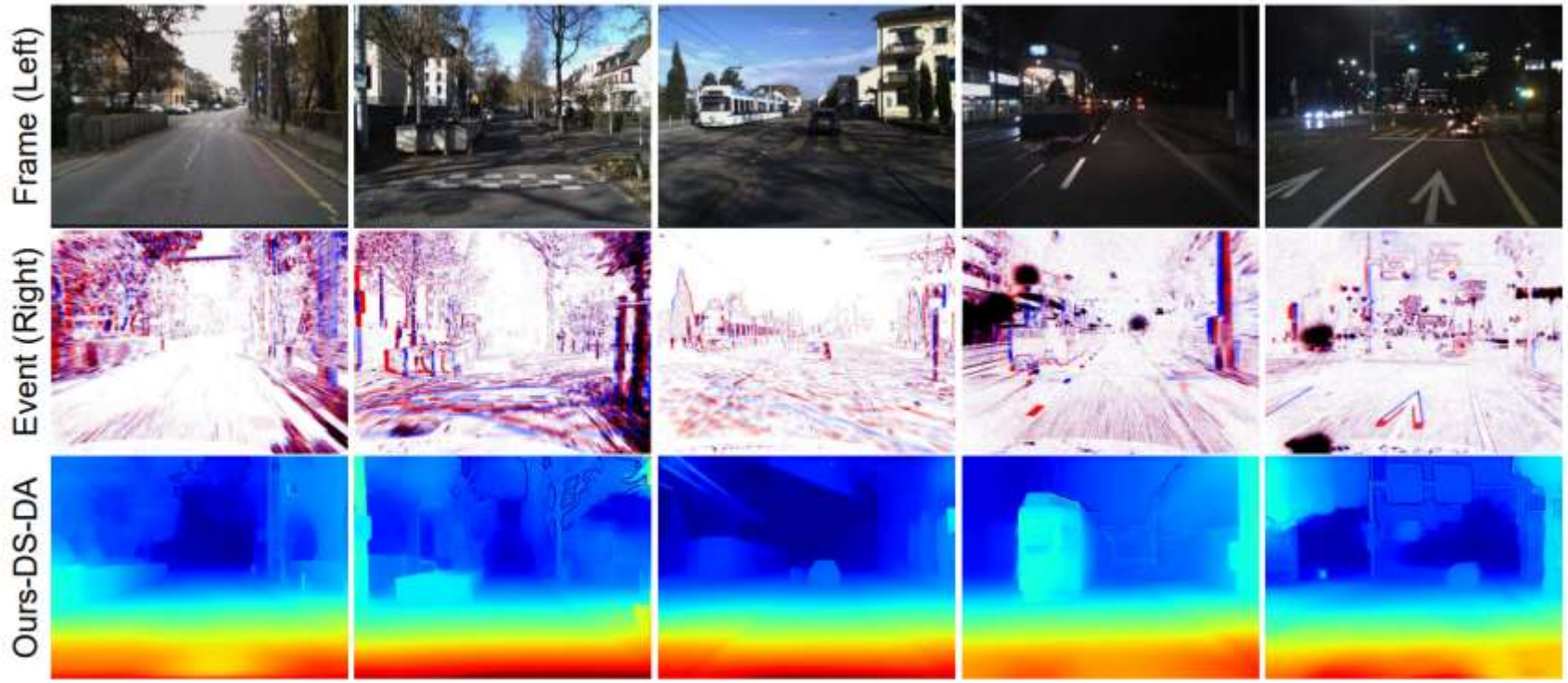
Results

- Our framework achieves state-of-the-art performance among all zero-shot solutions.
- The frame-based stereo and monocular models used in the framework can be seamlessly changed without any finetuning, allowing for flexible upgrades as related fields advance.



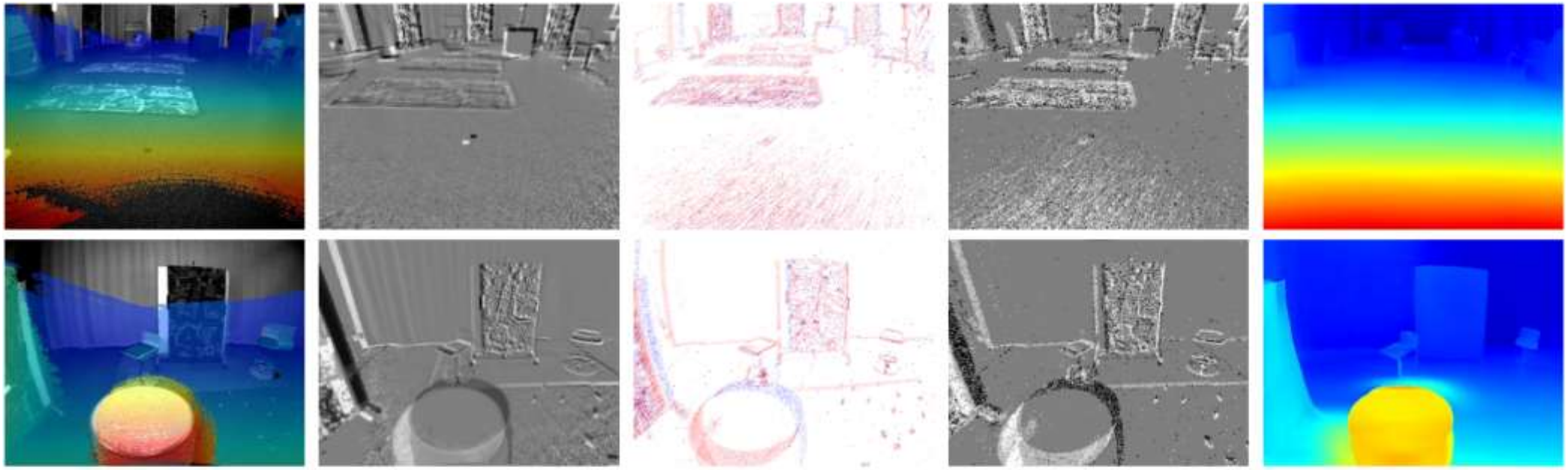
Results

- Our method performs robustly in diverse scenarios and datasets.



Results

- Our method performs robustly in diverse scenarios and datasets.



Camera Intelligence
A Computational Photography Lab @ PKU

北京大學
PEKING UNIVERSITY

華南理工大學
South China University of Technology

NEURAL INFORMATION PROCESSING SYSTEMS

Thank You!

Lab page

https://camera.pku.edu.cn

Code available

https://github.com/HYLZ-2019/ZEST