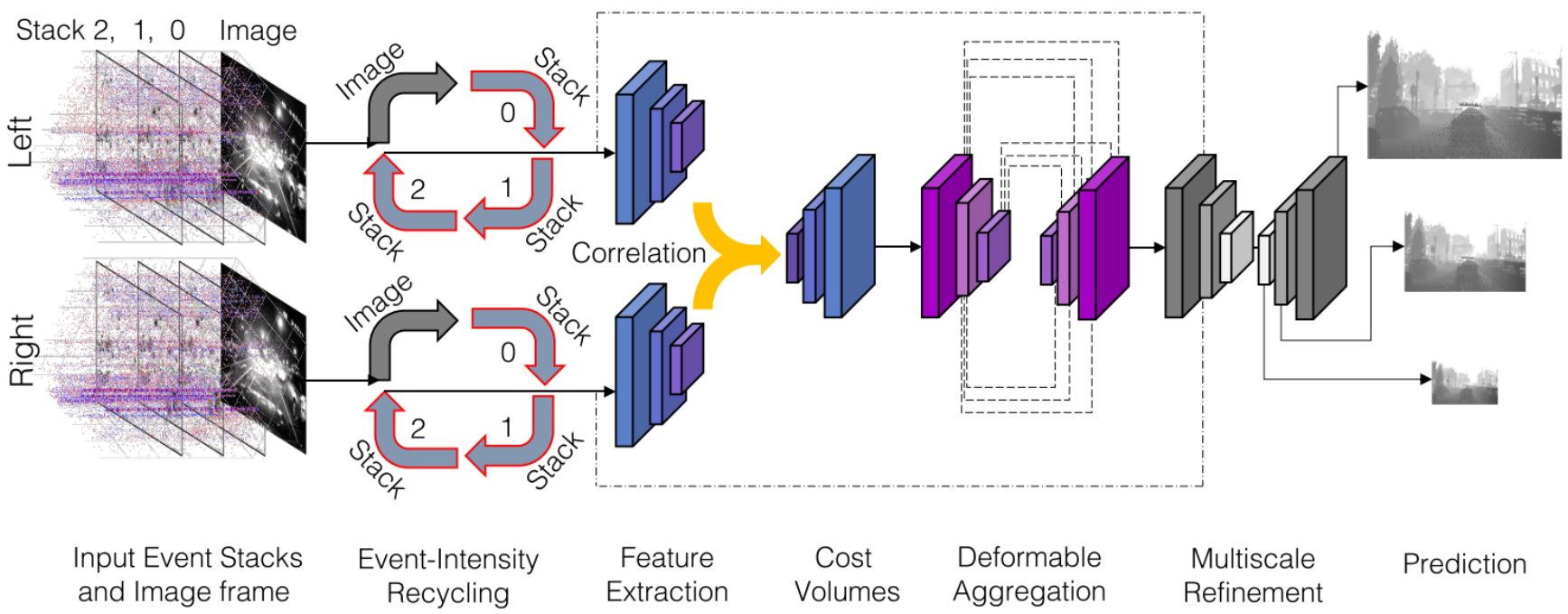
10.31-11.06 (1)

Event-Intensity Stereo: Estimating Depth by the Best of Both Worlds(ICCV2021)

（双目，events and intensity images）

S. Mohammad Mostafavi I. GIST, Kuk-Jin Yoon KAIST South Korea,Jonghyun Choi† GIST, South Korea （与Stereo Depth from Events Cameras: Concentrate and Focus on the Future（CVPR 2022）同一作者，笔记中标记为A文）



Approach: Event-Intensity Stereo

1. Event Representation

采用的**SBN**（A文中提到的SBN有很多的缺点，并进行了改进）

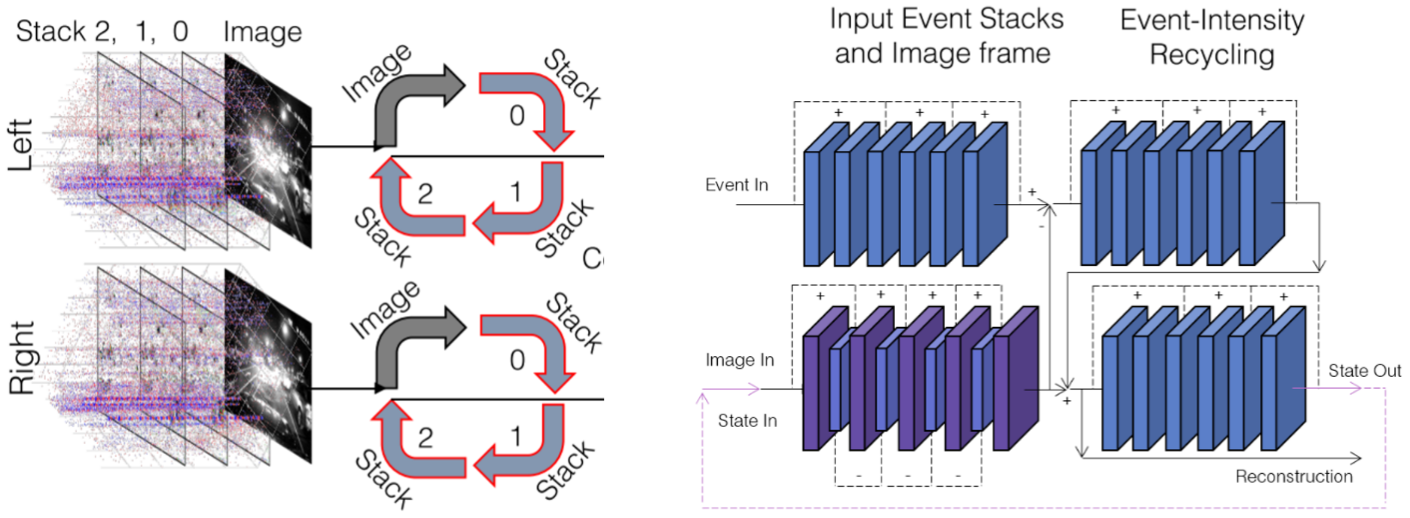
事件相机的异步事件流转化为一种更适合深度估计等任务的事件堆栈。

事件堆栈的大小为W×H×C的张量，其中W表示宽度，H表示高度，C表示通道数。将初始张量值设置为 128。对于每个传入事件，将其位置更新为 0（负事件）或 256（正事件）C取3，每一个stack包含3000个事件。

事件数据的堆叠过程以生成事件张量。堆叠强度帧之前的事件，使用SBN（基于数量的堆叠）的方法。事件流中的正事件（红色）和负事件（蓝色）可视化显示，强度信息来自APS（活动像素传感器）位于事件流的末尾，按照一系列的顺序创建，并与APS帧的时间戳对齐。

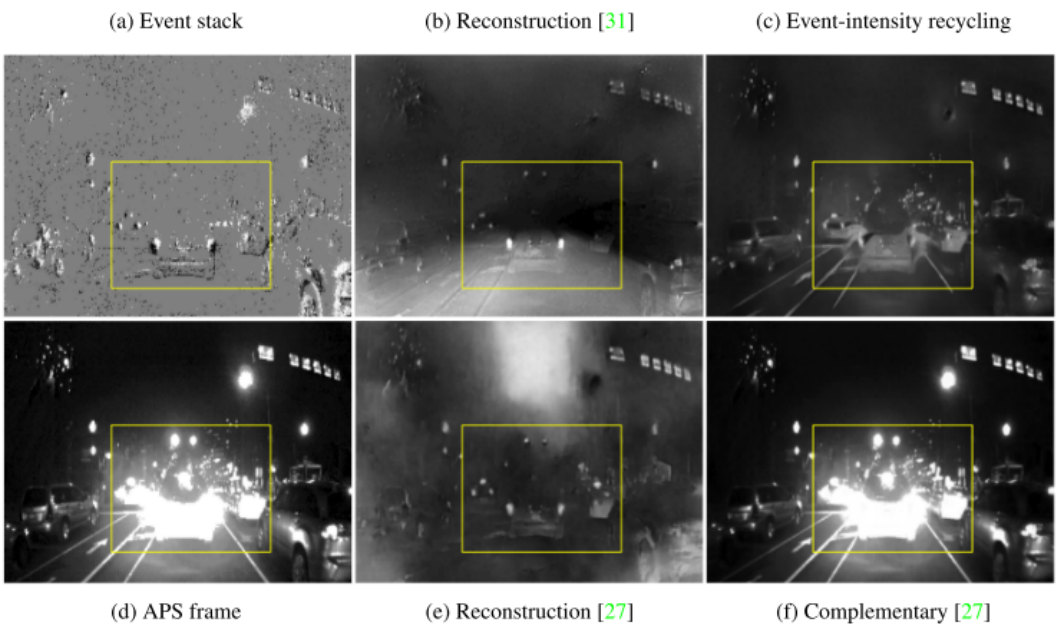
其中，事件数据分为两种类型：正事件和负事件。根据事件的类型，被映射到一个张量中，不同的事件类型（正事件和负事件）在张量中有不同的表示。

2. Event-Intensity Recycling



这个过程是一个连续的循环，其中事件堆栈和状态之间交互，将事件数据和图像数据结合起来，以生成事件强度输出。每个事件堆栈依次用于事件输入，同时前一个状态被回收并用于状态输入，创建新的隐藏状态。

事件强度重建结果可以包含有关事件的各种信息，如事件的位置、强度、频率等。**作者提到目标不是重建强度图像，而是将事件数据和APS图像进行有效整合，以便进行立体匹配，以获得更多的深度信息。**在训练中直接利用清晰的强度帧，采用了多种强度图像，包括模糊的强度图像和动态范围发生变化的图像，以使神经网络能够在不同条件下适应事件数据和强度图像之间的差异，让网络学会从事件和图像中重建结构细节。



Event-Intensity Recycling能够更好地保留场景中的结构信息。

3.Deformable Aggregation

1. 特征提取和代价体积：采用**Res-Net**架构进行特征提取，使用特征金字塔和特征相关性创建代价体积。
2. 聚合模块：使用内尺度和跨尺度聚合模块，对三个金字塔级别的代价体积进行聚合。
3. 可变形卷积：引入可变形卷积，以自适应地聚合代价体积，包括本地和全局聚合。这有助于处理事件相机感知的边缘和强度相机感知的多区域信息。

> 事件相机主要感知边缘，而强度相机感知多个区域，因此使用可变形卷积来适应性地聚合这些信息，以提高深度估计的准确性。

4. 采用金字塔结构：通过金字塔结构的代价体积聚合，可以更好地处理对象边界和细微结构，以提高深度估计的质量。
5. 跨尺度聚合：在深度估计中，跨尺度聚合有助于在粗略尺度上搜索对应关系，特别是在低纹理或无纹理区域更具有辨别性。

4. Disparity Estimation

通过计算视差来估计深度。

采用**soft argmin**方法，该方法用于从cost volume中找出每个像素的最佳视差值。

> **Soft argmin**：一种数学计算方法，用于估计某些数值的最小值或最大值的位置。在深度估计的背景下，soft argmin 用于找到代价体积（cost volume）中每个像素的最佳视差值，以便确定场景中不同物体或表面的深度信息。与传统argmin不同，soft argmin 使用数学函数（通常是 softmax 函数）来平滑化计算，从而减少了估计中的硬性边界。这使得结果更具连续性，有助于更好地捕捉深度变化的细节和连续性。在计算机视觉和深度估计领域，soft argmin 经常用于提高深度估计的质量和准确性

> **Cost volume（代价体积）：**计算机视觉和深度估计领域。**输出一个三维数据结构，**通常表示为 D(x, y, d)，其中 x 和 y 是图像中的像素坐标，d 是深度或视差的值（d包含了几个不同值）。**其中包含了在不同深度或视差值下每个像素位置的匹配代价。这些匹配代价表示了图像中的像素与另一视图中不同深度或视差值下的像素之间的相似性或差异度**。代价体积用于存储不同深度或视差值之间像素之间的相似性或差异度，通常是通过一些匹配度量来计算的。这样的匹配度量可以基于像素的颜色、纹理、亮度或其他特征。**在深度估计的任务中，代价体积通常是通过将左视图和右视图的像素之间的相似性计算出来，以确定不同深度或视差值下的匹配程度**。通过计算代价体积，可以找到在每个像素位置处的最佳深度或视差值，从而实现深度估计。

5. Learning objectives

作者提到，由于最终的目的是估计深度，实际上不需要网络来重建图像，因此在中途停止使用图像重建损失，只使用端点误差（EPE）作为的主要损失函数。

End-point-error. L1 loss（EPE）

$$\mathcal{L}_{EPE}(d_v, \hat{d}_v) = \frac{1}{V}\sum_{v=0}^{V}|d_v - \hat{d}_v| \qquad (1)$$

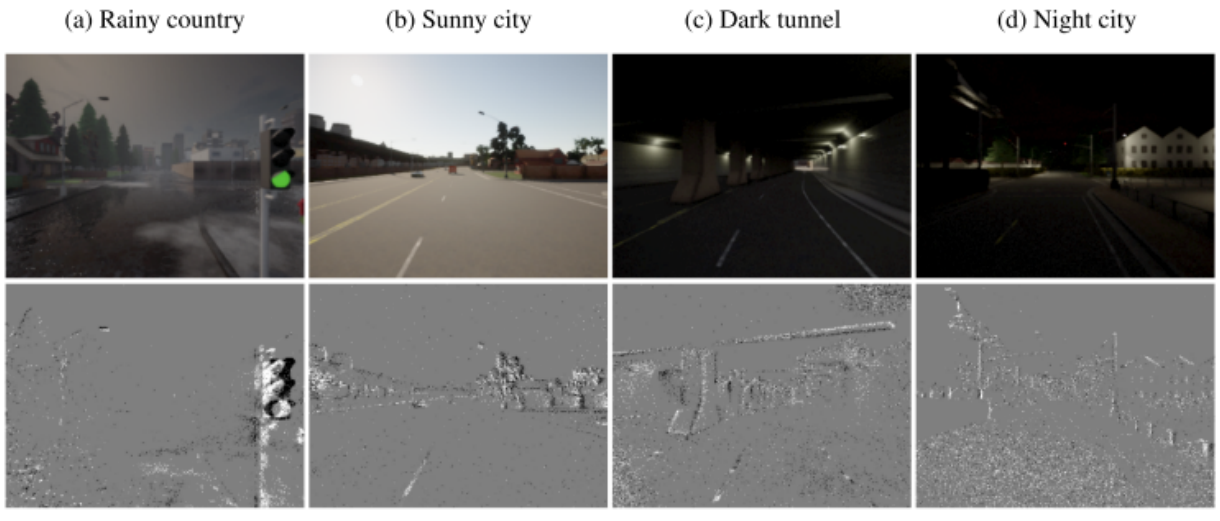
对于中间图像重建（前几个起始时期），利用 L1 损失和学习的感知相似性损失（**LPIPS**）

最终损失（L）

$$\mathcal{L} = \begin{cases} \mathcal{L}_{EPE} + \lambda_1 \mathcal{L}_{LPIPS} + \lambda_2 \mathcal{L}_{L_1}, & epoch < E \\ \mathcal{L}_{EPE}, & epoch \geq E \end{cases} \qquad (2)$$

Experiments and Analysis

1. Datasets



ECAR dataset.

使用两个主要数据集：MVSEC和DSEC代表真实世界事件数据，以及ECAR代表模拟数据。

- MVSEC包含多辆车辆上装有DAVIS相机的数据，覆盖了不同的白天和夜晚光照情况，但其地面真实深度信息不总是与图像或事件数据对齐。MVSEC 有两个 DAVIS摄像机，提供图像帧和事件流
- DSEC是一个大规模室外立体事件摄像机数据集，具有不同分辨率和基线的相机对，涵盖更多照明条件。事件和图像来自不同相机对。
- ECAR数据集是通过CARLA和ESIM模拟器生成的，覆盖了多种驾驶场景和真实世界相机可能的变化。通过模拟，生成了立体事件和图像以及GT depth。

2. Experimental Setup

设计了三种不同的深度估计方法：事件-强度立体法（EIS），事件立体法（ES）和强度立体法（IS）。其中ES使用额外的事件堆栈，而IS使用额外的强度图像。所有方法都利用过去和现在的数据，是因果关系的。

3. Quantitative and Qualitative Analysis

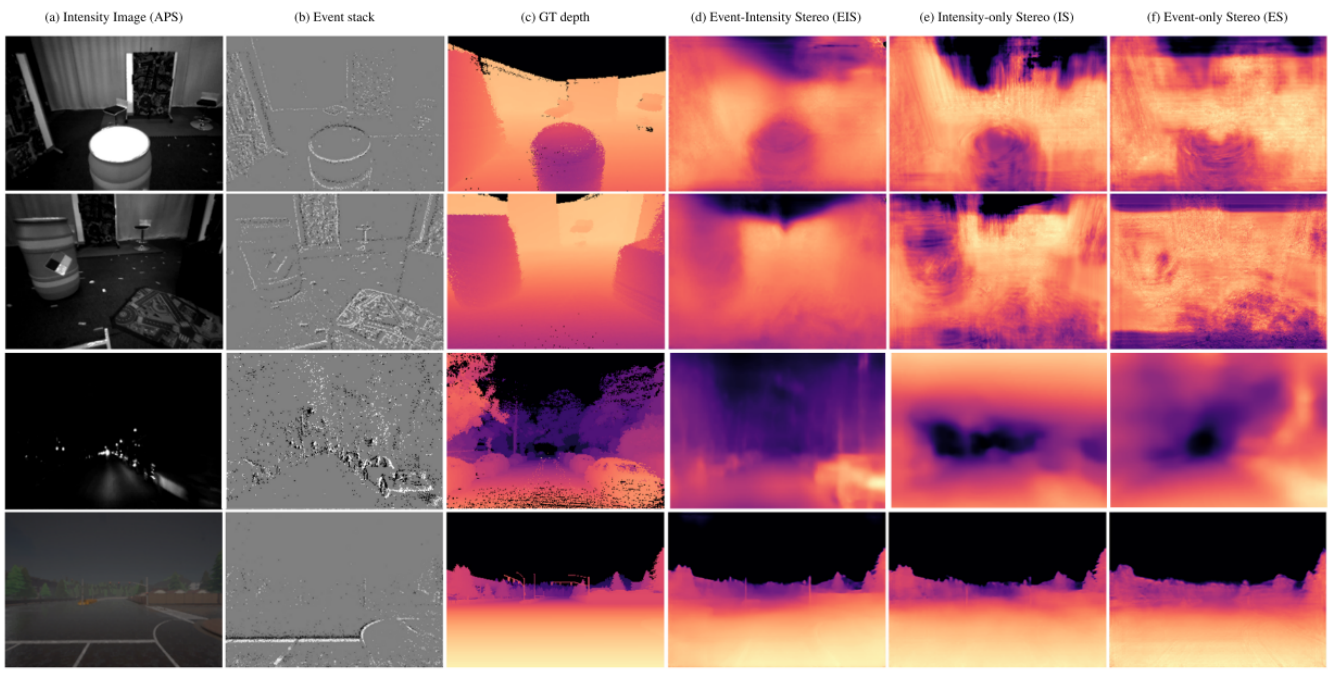
1. **数据集使用：**
   - 使用MVSEC，DSEC和ECAR数据集，进行定性和定量分析。
   - 采用MVSEC的训练和验证协议（MVSEC拆分1-3），执行不同组合的性能比较。
2. **MVSEC数据集结果分析：**
   - 使用平均深度误差和一个像素误差进行定性比较。以ground truth像素中视差误差小于一个像素的百分比进行定量比较。在验证集中找到最低端点误差的结果。
   - EIS方法在大多数数据拆分上在平均深度误差和1PE上优于ES和IS。
3. **ECAR数据集和MVSEC拆分分析：**
   - ECAR数据集涵盖了各种照明条件，MVSEC拆分1-3只涵盖室内飞行场景。
   - IS通常比ES具有较少的误差，但在MVSEC拆分1中，ES表现优于EIS。（由于训练数据中没有包含足够的高度变化的样本）
   - 拆分1中的训练序列与测试集不同，导致网络不能很好地泛化。
4. **DSEC数据集和结果：**
   - 报告了两个像素误差、均方根误差（RMSE）和平均绝对误差（MAE）。
   - EIS方法明显优于DSEC挑战基线。

Split	Mean depth error [cm]			One pixel error [%]		
	ES	IS	EIS	ES	IS	EIS
MVSEC Split 1	**13.27**	14.12	13.74	80.6	71.7	**89.0**
MVSEC Split 2	25.18	23.24	**18.43**	73.0	67.3	**85.2**
MVSEC Split 3	25.72	23.78	**22.36**	68.3	53.8	**88.1**
ECAR	22.3	18.7	**11.8**	67.7	79.5	**81.7**

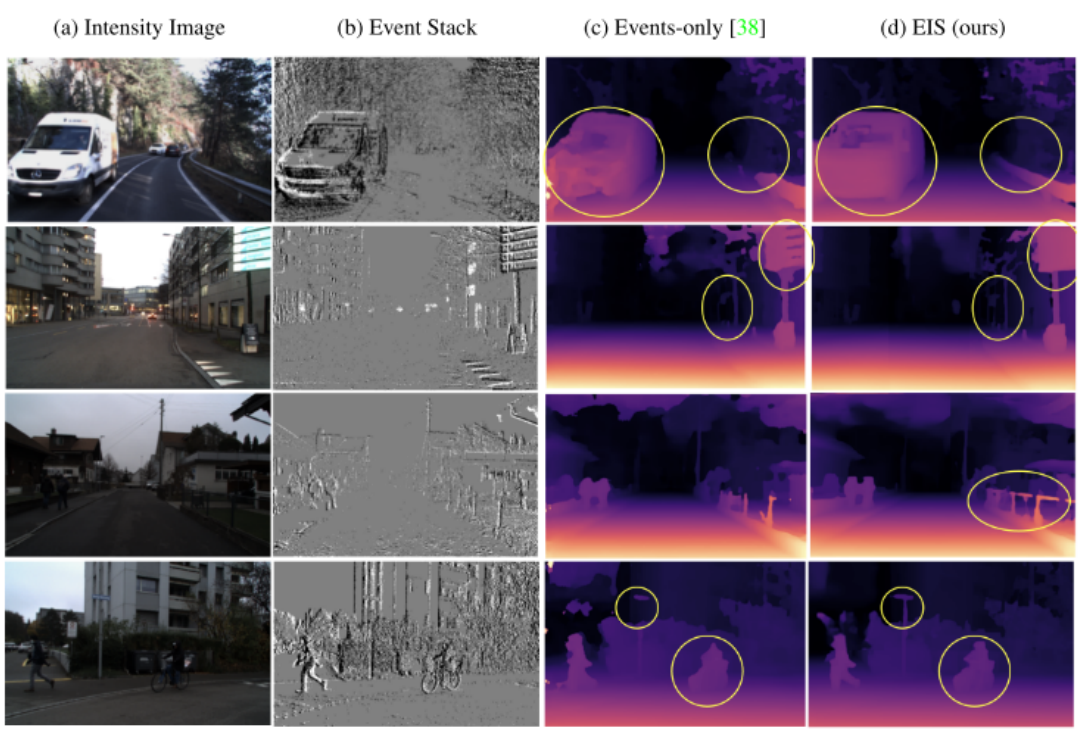
MVSEC和ECAR 数据集上使用密集的ground truth进行性能评估

	MAE	1PE	2PE	RMSE
Events-only Baseline [38]	0.576	10.915	2.905	1.386
Event Stereo ES	0.529	9.958	2.645	1.222
Event-Intensity Stereo EIS	**0.396**	**5.814**	**1.055**	**0.905**

使用**DSEC 数据集**将EIS 和 ES 方法与baseline（events-only中的先进方法）



定量比较EIS，IS, ES



与event-only baseline比较（GT未公布）

4.Ablation study

删除不同的网络组件，并使用 ECAR 的子集对其进行评估。

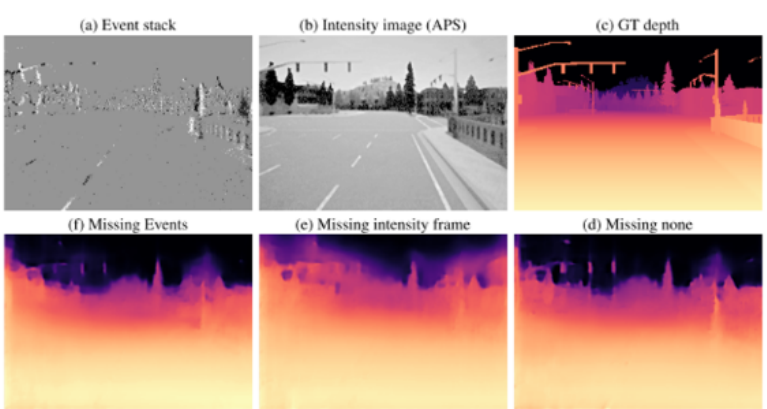
Table 4. Ablating the effect network components on the depth.

Network	MDE	1PE
Full network (FN)	8.3	78.6
FN - {Feature pyramid net.}	37.8	64.1
FN - {Deformable aggregation net.}	23.4	70.3
FN - {Multi-scale refinement}	13.8	67.2

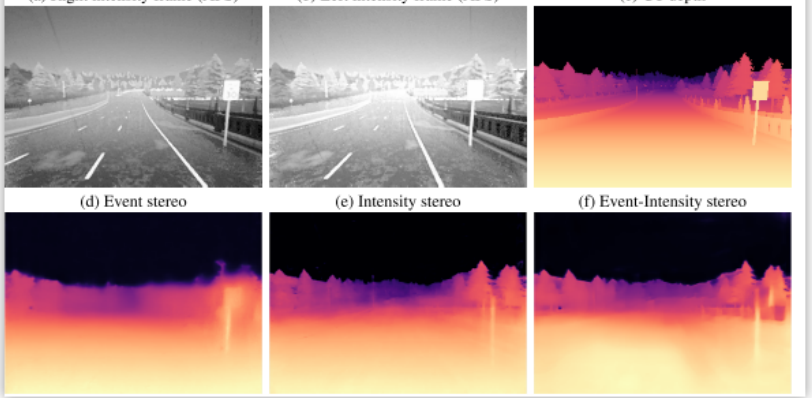
5. Extensions

文中还提到缺失事件或者强度图像以及左右光照不同（Inconsistent Left-Right pairs.）的情况下进行深度估计的比较

Train	Test	MDE	1PE
Normal	Full modality	10.6	85.5
Normal	No left event stack	87.5	8.1
Normal	No left APS frame	91.7	11.3
Missing	Full modality	16.8	84.9
Missing	No left event stack	17.9	82.2
Missing	No left APS frame	20.3	76.5



Train and Test	MDE	1PE	Train and Test	MDE	1PE
Consistent IS	17.3	79.5	Consistent EIS	10.6	85.5
Inconsistent IS	69.0	57.2	Inconsistent EIS	32.0	65.7



Conclusion

1. 提出了一种端到端的神经网络，将事件数据和图像用于立体匹配，并通过可变形聚合来充分利用事件强度立体框架的优势。
2. 扩展了方法的可靠性，使其能够进行缺失数据的立体深度估计以及左右摄像机对不一致的立体深度估计。
3. 提到了未来方向，包括使用脉冲神经网络来实现更快速和完全异步的设计。