10.24-10.30

Combining Events and Frames using Recurrent Asynchronous Multimodal Networks for Monocular Depth Prediction

本文的事件表征方法，深度表征方法以及训练方法同Learning Monocular Dense Depth from Events 10.17-10.23

Robotics and Perception Group

https://rpg.ifi.uzh.ch/

GitHub - uzh-rpg/rpg_ramnet: Code and datasets for the paper "Combining Events and Frames using Recurrent Asynchronous Multimodal Networks for M

Code and datasets for the paper "Combining Events and Frames using Recurrent Asynchronous Multimodal Networks for Monocular Depth Prediction" (

https://github.com/uzh-rpg/rpg_ramnet

Daniel Gehrig*, Michelle R¨uegg*, Mathias Gehrig, Javier Hidalgo-Carri´o, Davide Scaramuzza

简介

> 动态和主动像素视觉传感器（DAVIS）：将标准图像传感器与异步事件传感器组合在同一像素阵列上，有两种不同类型的传感器：一种用于捕捉传统的图像帧，另一种用于捕捉事件数据。

1. **传统方法 vs. 数据驱动方法**：

- 传统方法，如卡尔曼滤波器，通常需要精确的专家设计的模型，但可能不适用于不符合其基本假设的数据。
- 数据驱动方法，如LSTM，可以直接从数据中学习模型，适用于更复杂的问题，但要求输入数据以固定速率和与辅助输入同步呈现。

1. **事件和帧数据融合的两种方法：**

- 基于模型的方法使用生成模型来融合事件和帧数据，但由于事件生成模型中的非线性和不确定性，这些方法较脆弱，对超参数调整敏感，且在某些复杂像素级任务中表现不佳。
- 基于学习的方法使用大型数据集来生成更准确的预测，但目前的学习方法限制了事件和帧数据的同步堆叠，牺牲了事件的异步性和高时间分辨率，同时采用简单的前馈神经网络而不是循环神经网络（RNN）。
- 将帧和事件结合起来可以充分利用两者的优势。

> **重要性**：帧数据可以在几乎没有自我运动或很少触发事件的情况下提供重要的参考信息。这是因为事件数据通常只捕捉到亮度变化，而帧数据可以提供更多关于场景的详细信息，有助于更好地理解和处理那些相对静止的情境。

3. **引入Recurrent Asynchronous Multimodal（RAM）网络**：

- 传统的循环神经网络（RNN）不适合处理来自多个传感器的异步和不规则数据。
- RAM网络利用事件和帧数据的互补性，同时考虑事件和帧数据的异步性和不规则性，通过循环性集成多传感器的时间上下文。
- RAM网络受传统RNN启发，保持内部状态的异步更新，可以在任何时刻进行预测。
- 这一新型网络架构应用于基于事件和帧的单目深度预测任务，这在现代算法中用于障碍物避免、路径规划和3D地图构建等应用中非常重要，尤其是在高速和高动态范围环境下，事件相机的优势能够提高算法的鲁棒性。

用于学习深度的高质量**事件**和基于**帧**的数据集仍然很少，发布了 **EventScape数据集**，记录在 CARLA 模拟器中，包括不同汽车场景中的事件、**强度帧**、语义标签、深度图和车辆导航参数。

RELATED WORK

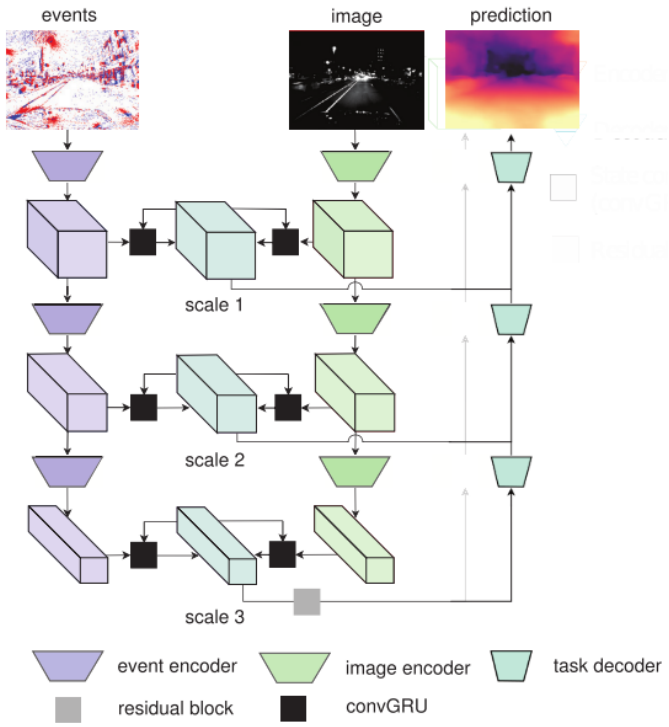
一、介绍事件和帧数据的互补性，以及多种算法通过融合它们来充分利用**两种数据模态**的优势的方式。

1. **事件和帧数据融合的趋势：** 由于事件和帧数据的互补性，各种算法试图通过融合它们来实现不同应用，尤其是自从DAVIS引入以来，趋势变得更加显著。
2. **算法应用领域：** 这些融合算法的应用领域包括同时定位与地图构建（SLAM）、特征跟踪、高动态范围强度重建和图像去模糊等。
3. **依赖模型的限制：** 许多融合算法依赖于事件和帧数据融合的底层模型，这些模型通常是理想的传感器模型，而在非理想条件下，如噪声和动态效应，这会导致性能下降。
4. **数据驱动方法：** 复杂任务如图像重建或单目深度估计通常使用数据驱动方法，其中许多采用循环架构，能够利用事件的长时间窗口来改善预测性能。
5. **事件和图像融合方法：** 尽管存在许多纯事件驱动的学习方法，但很少涉及图像和事件的融合。这些方法通过同步和连接两种输入，然后传递给标准前馈网络，但这种策略会丧失事件的异步性和高时间分辨率。
6. **传统方法和异步输入的挑战：** 传统方法如RNN在采样率高度可变和异步输入的情况下面临困难，解决方案通常涉及填充、复制或采样率转换，但可能导致频率降低或输入数据的时间错位。
7. **RAM网络的解决方案：** RAM网络克服了这些限制，通过（i）利用成熟的数据驱动模型如RNN来实现复杂任务的稳健预测，（ii）通过循环性最大化来自事件的时间上下文，（iii）引入适用于每种输入模态的个别和异步状态更新规则，保留事件的异步性和高时间分辨率。

二、介绍基于事件单目深度估计

少有使用**事件和帧数据融合的方法。**

METHOD



A. Recurrent Asynchronous Multimodal Networks

这部分公式繁多比较难看懂，后续慢慢补充

当处理来自多个传感器的数据以进行单目深度估计时，RAM Networks的工作原理：

1. **多传感器数据融合**：RAM Networks的任务是将来自多个传感器的数据合并。每个传感器在不同时间戳下提供测量数据。
2. **数据编码**：首先，通过传感器特定的可学习编码器对数据进行处理，将其转换为中间特征。这有助于更好地理解和处理不同传感器的数据。
3. **异步和可变数据速率**：生成的特征序列具有异步性，这意味着不同传感器的特征可以以不同的顺序出现在序列中。此外，数据速率是可变的，即特征之间的时间间隔会随时间变化。
4. **特征融合**：为了正确融合这些异步和可变速率的特征，RAM Networks使用传感器特定的状态组合运算符。这些运算符接受特征并更新潜在变量，以保留特征的有用信息。这确保了来自不同传感器的特征被适当地整合。
5. **循环更新**：特征的组合可以通过不同方式执行，例如求和、串联或使用卷积门控循环单元（convGRU）。这有助于使用编码器-解码器结构，包括跳连接，以处理特征的时序性。
6. **任务变量预测**：最终，RAM Networks使用潜在变量来预测与任务相关的变量，例如单目深度估计。这确保了RAM Networks是马尔可夫的，即它们的输出仅依赖于当前潜在变量，而不受传感器测量历史的影响。

通过这个过程，RAM Networks能够更好地融合来自不同传感器的异步和异速率的数据，以进行各种复杂任务的预测和估计。

B. Network Architecture for monocular depth estimation

组件/特点	描述
网络架构基于 U-Net	基本架构灵感来自 U-Net，以前已用于事件的单目深度估计。
Skip 连接	在每个尺度上，使用 skip 连接作为中间特征。

ConvGRU 层	中间特征与 Σj-1 通过 ConvGRU 层结合，生成 Σj。
低尺度的潜在变量	最低尺度的潜在变量直接输入到残差块，然后进入三个解码器级别。
Skip 连接中的解码器输入	在 skip 连接中，解码器输入通过求和与相应阶段的状态结合。
编码器	每个编码器包括降采样卷积，内核大小为 5，步幅为 2。
残差块	残差块由两个内核大小为 3 的卷积组成。
解码器	解码器使用双线性上采样，然后是内核大小为 5 的卷积。
激活函数	所有层均使用 ReLU 激活函数。

C. Event Generation Model

事件相机的工作原理和事件触发的条件：

- 事件相机包含独立的像素 u，对亮度信号的对数变化L(u,t)作出反应。
- 如果自上一个事件以来的对数亮度变化超过阈值 C，则在像素位置 u =(xk,yk)T处触发新事件 ek =(xk,yk,tk,pk)。
- 事件的极性pk可以取值 {-1,+1}，具体取决于亮度变化的方向。
- 对于理想传感器的生成事件模型，当满足以下条件时，带有极性pk 的事件会在像素 uk和时间戳tk处触发：

$$\Delta L(\mathbf{u}_k, t_k) = p_k \left(L(\mathbf{u}_k, t_k) - L(\mathbf{u}_k, t_k - \Delta t_k) \right) \geq C$$

其中Δtk是相同像素位置上的上一个事件发生之后到当前事件发生的时间间隔，即两次亮度变化事件之间的时间间隔。

D. Event Representation

事件表征方式完全同10.17-10.23

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

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

E. Depth Representation

首先，将度量深度转换为归一化的对数深度图（同10.17-10.23），有助于学习大范围深度变化。深度图是逐像素构建的。对数深度的计算公式如下：

$$\widehat{D}_k = \frac{1}{\alpha} \log \frac{\widehat{D}_{m,k}}{D_{\max}} + 1$$

提到，对于单目深度估计，只能按比例估计深度。一些方法通过使用来自立体设置、IMU 数据或相机姿势的附加信息来规避这个问题，但是，使用循环深度学习架构并手动设置数据集中观察到的最大深度可以获得令人满意的结果，而无需额外的数据。

F. Training

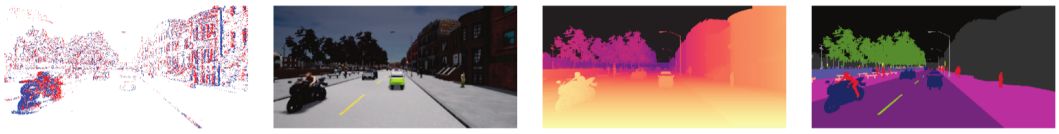
同10.17-10.23

EXPERIMENTAL SETUP

在合成的 EventScape 数据集上验证 RAM Net，并与仅事件、仅帧以及基于事件和帧的基线进行比较。

A. Datasets

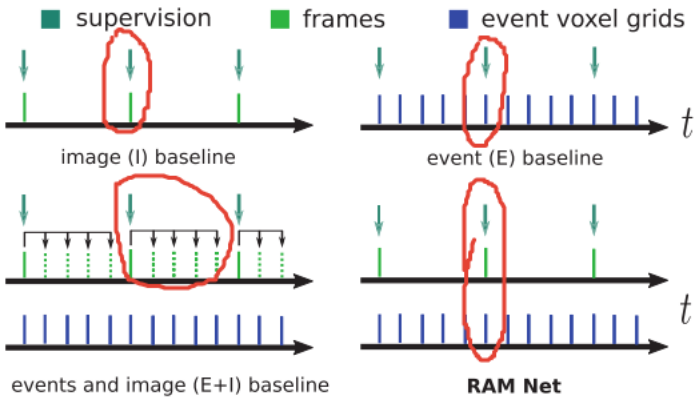
数据集特点	EventScape	MVSEC
数据来源	CARLA模拟器	真实世界
数据类型	合成数据	真实数据
数据种类	事件、图像、深度、语义分割、车辆控制	事件、图像、激光雷达深度
数据数量	743序列	多个日夜驾驶序列和室内序列
标签数量	171,000	深度地图、灰度图像
数据时长	约2小时	多个驾驶和室内序列
数据帧率	图像、深度、语义分割：25Hz，车辆控制：1000Hz	灰度图像：10Hz（夜晚）、45Hz（白天） 深度：20Hz
事件生成方式	使用CARLA模拟器插件和ESIM事件相机模拟器，渲染图像 500Hz 后转换为异步事件（添加refractory period 100 μs）	真实双DAVIS事件相机
数据同步和对齐性	图像、深度、语义分割、事件同步和对齐（事件和帧是同步和对齐的，类似于 DAVIS 传感器的输出）	-
训练数据集城镇	CARLA城镇1、2和3	-
验证和测试数据集	从CARLA城镇5选择，地理上分隔的区域	-



(a) events (b) images (c) depth maps (d) segmentation labels

EventScape的数据，图像、深度图和分割标签以 25Hz 提供，而车辆导航参数以 1000Hz 提供。事件是通过使用 CARLA中的事件相机插件将 500Hz 图像转换为事件来生成的，该插件基于事件相机模拟器。

B. Baselines



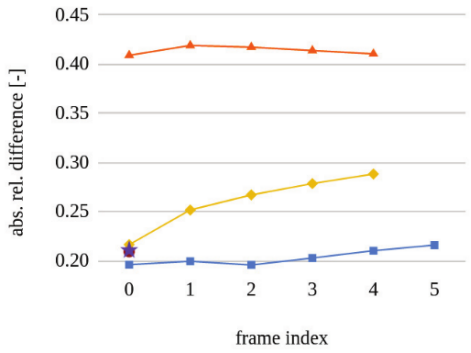
(a) baselines and our method

1. **E** 是一个基准模型，用于理解多模态性的影响。只接收体素网格作为输入，而不使用图像数据。
2. **I**也是一个基准模型，同样用于理解多模态性的影响。只接收灰度图像作为输入，而不使用事件数据。
3. **E+I**模型用于评估 RAM 网络的异步更新方案。它接收体素网格和图像的堆叠作为输入，取消了 RAM 网络中提出的异步更新方案以作为基准。结构不包括与 RAM 网络中的循环状态组合操作符（用于将不同传感器模式（事件和图像数据）的状态进行组合）相同的内容，而是使用了不同的递归 convLSTM 编码器。可以访问与文中方法相同数量的信息，但事件和图像数据之间存在时间不匹配。
4. **无循环性的事件与图像模型（E+I no recurrency）**：基于简单的前馈网络，将 convLSTM 块替换为简单的卷积层。该模型训练使用 5 Hz 的对应图像堆叠 200 毫秒事件数据。

基准模型的设计和使用旨在帮助研究人员理解 RAM 网络的性能、**多模态性的影响**以及 **RAM 网络中异步更新方案的作用**。

RESULTS

A. Effect of Input Modality and Architecture



(b) Images at 5Hz, events at 25Hz

图像以 5 Hz 的频率出现，深度标签和事件数据以 25Hz 的频率提供，训练网络

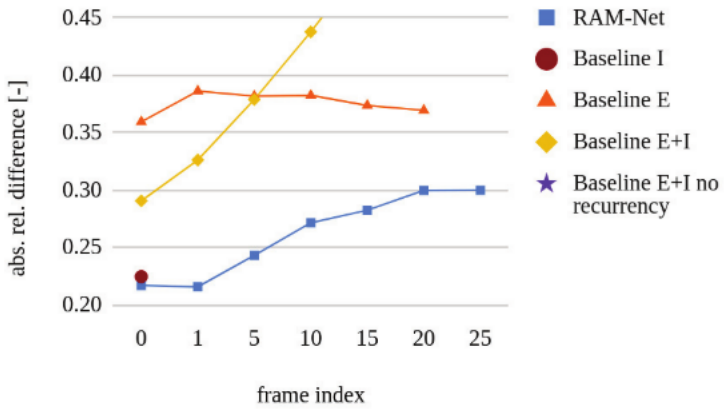
通过对不同模型和方法的比较来探索了RAM网络在事件和图像深度估计方面的性能。分别评估了各种基线模型，包括**仅事件、仅图像、事件和图像结合的多模态模型，以及不具备RAM网络的异步更新机制的模型。**

首先在EventScape数据集上进行了训练和测试。实验结果表明，在处理深度估计时，RAM网络的性能优于仅图像或仅事件的基线模型。此外，RAM网络还优于结合了事件和图像的多模态基线。实验结果呈现了RAM网络相对其他模型的明显性能优势。

实验证明了：(i) 结合图像和事件的好处，突出了它们的互补性，(ii) 对于循环方法的时间对齐效应(iii) 循环的重要性，它使得在图像帧之间实现高质量的深度预测成为可能。

B. Generalization to Different Data Rates

分析其泛化到新帧速率的能力。



(c) Images at 1Hz, events at 25Hz

在训练过程中，对 1Hz 的图像和 25Hz 的事件进行**泛化**研究。

1. I基线表现稍差，观察到的图像之间的时间间隔增加。这表明I不是非常依赖于循环结构。
2. RAM Net的预测质量在图像输入后也有所下降。和之前一样，RAM Net的预测准确性随着帧索引的增加而逐渐降低，最终在20索引处趋于稳定。然而，RAM Net的误差仍然相对较低。
3. E+I的误差在大时间间隔后急剧增长，甚至超过E。这是因为对于较大的时间间隔，事件和先前帧的副本之间存在显著的不对齐。这表明RAM Net在较大时间跨度之后仍然保留了来自图像的附加信息。
4. RAM Net能够在不同频率的数据输入下实现良好的泛化能力，这表明它对于不同类型的输入数据具有较强的适应性。

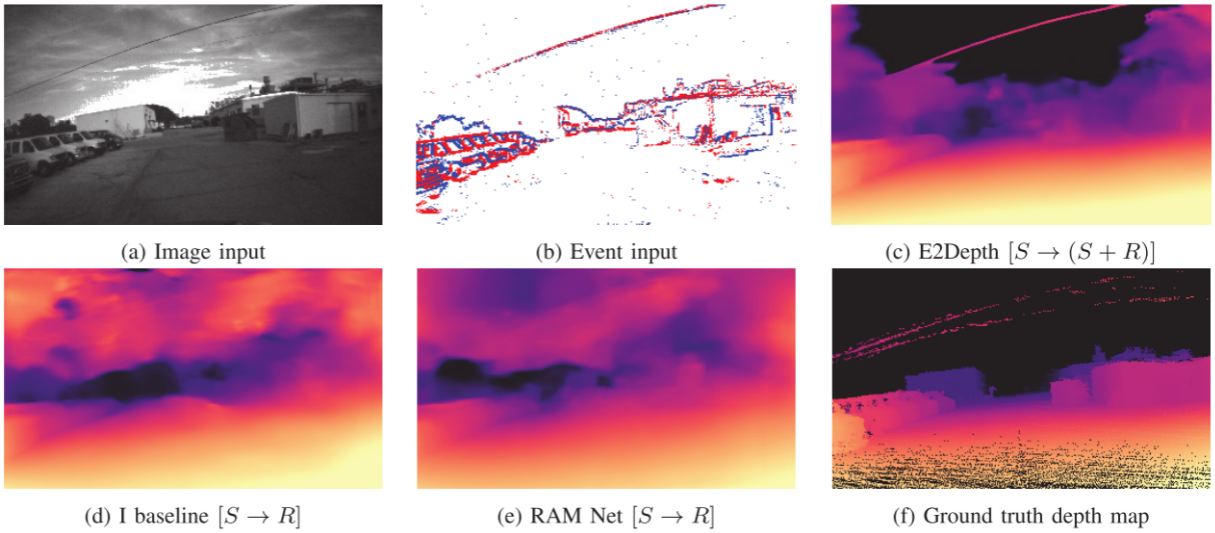
总的来说，RAM Net在不同输入模式和频率下都表现出较好的性能和泛化能力，尤其是在长时间跨度和异步输入的情况下。这强调了RAM Net的多模态性和高效性。

C. Real World Experiments

在**MVSEC**真实数据集上的实验过程：

1. **数据准备**：MVSEC数据集中的深度和图像数据不再是同步的，因此需要将标签和网络预测进行时间对齐。为了做到这一点，首先将事件数据分割成包含10,000个事件的等大小数据包，并在需要时均匀分配多余的事件。接下来，从这些数据包生成体素网格，时间戳对应于数据包中的最后一个事件。将这些不规则的体素网格序列与帧序列一起输入网络，根据它们的出现交替传入。
2. 评估：首先，在不重新训练的情况下，对MVSEC数据集中的四个夜间室外序列和一个白天室外序列进行测试，报告网络预测与ground truth之间的平均绝对深度误差([S])。同时，还与不同的基于事件的和基于帧的最先进的方法进行比较，包括一些在MVSEC数据集上微调的方法（[S→R]）。
3. 方法比较：两个基于事件的方法分别采用了前馈网络和循环架构。而基于帧的方法仅处理图像数据，其中一种方法使用前馈网络，另一种采用循环架构。此外，还有一个只基于图像的基线方法。
4. 微调：为了提高预测质量，对网络和基线方法进行了对MVSEC数据集中的白天室外序列的微调。

result：



(a) Image input (b) Event input (c) E2Depth $[S \rightarrow (S+R)]$

(d) I baseline $[S \rightarrow R]$ (e) RAM Net $[S \rightarrow R]$ (f) Ground truth depth map

MVSEC 数据集上的定性性能比较。 (a) 和 (b) 显示图像和事件输入。 (c) 同时对模拟和真实数据进行训练时E2Depth 的结果。这种训练配置可以让天空获得更好的质量结果。 (d) 和 (e) I 基线和 RAM Net 根据真实世界数据进行了重新训练。天空的预测有误，但是地面上的物体是正确的。

1. 对于仅在合成数据上训练的方法（[S]）：RAM Net的泛化性能优于I。但与E2Depth方法相比，模型在从合成到真实数据的迁移方面的性能明显较差。这表明纯粹基于合成数据的训练可能不足以实现出色的泛化性能。
2. 对于使用真实数据进行微调的方法（[S→R]）：I和RAM Net的性能明显提高。I基线的表现突出，强调了图像数据对深度估计的重要性，甚至超过了E2Depth方法。然而，值得注意的是，纯粹基于帧的方法无法在帧之间的盲区中进行深度预测。
3. RAM Net在几乎所有数据集上都优于其他方法。与E2Depth方法相比，RAM Net能够重建更多的细节，但在天空区域观察到一些伪影。文中提到这些伪影可能是由于训练和测试数据集中的图像差异导致的。I 基线也受到这些伪影的影响，纯粹基于事件的方法（E2Depth ）不依赖于图像，因此不会遇到这些问题。
4. **最后，研究人员提到了一些潜在的改进方法，包括使用语义分割来检测天空并将最大深度添加到这些像素的深度标签，以减轻伪影问题。还提到了多任务学习和深度估计与语义分割相结合的方法来在测试时遮蔽天空区域。**

CONCLUSION

解决融合异步和多模态传感器数据的学习问题

主要贡献	概述
网络架构 **RAM Net** 引入	用于处理异步和不规则数据的多传感器数据的新型网络架构。
应用于单目深度估计任务	RAM Net 用于单目深度估计任务，展示了其在图像帧之间进行高质量深度预测的能力。
在合成数据 EventScape 和真实数据 MVSEC 上的性能实验	RAM Net 在合成数据和真实数据集上的实验中优于传感器单一模态基线，充分利用了事件和图像之间的互补性。
对比基于传统 RNN 的基线和现有最新方法的性能	RAM Net 更鲁棒地适应输入频率变化，因此优于多个基线和基于传统 RNN 的最新方法。
合成数据集 EventScape	EventScape 是一个大规模合成数据集，包含来自 CARLA 模拟器的事件、图像、语义标签、深度地图和车辆导航参数，可用于进一步研究事件的多模态学习。