


10.17-10.23

Learning **Monocular** Dense Depth from Events(3DV)

GitHub - uzh-rpg/rpg_e2depth: Code for Learning Monocular Dense Depth from Events paper (3DV20)

Code for Learning Monocular Dense Depth from Events paper (3DV20) - GitHub - uzh-rpg/rpg_e2depth: Code for Learning Monocular Dense Depth from Events paper (3DV20)

 https://github.com/uzh-rpg/rpg_e2depth

uzh-rpg/
rpg_e2depth

Code for Learning Monocular Dense Depth from Events paper (3DV20)


 1 Contributor  2 Issues  95 Stars  28 Forks

Learning Monocular Dense Depth from Events

<https://rpg.ifi.uzh.ch/E2DEPTH.html>

Multi Vehicle Stereo Event Camera Dataset

A data set for Stereo Event Cameras.

 <https://daniilidis-group.github.io/mvsec/>



Javier Hidalgo-Carri o, Daniel Gehrig and Davide Scaramuzza ,Robotics and Perception Group, University of Zurich, Switzerland

Learning **Monocular** Dense Depth from Events(3DV)

[简介](#)

[相关工作](#)

[Depth Estimation Approach](#)

[1.Event Representation](#)

[2.Network Architecture](#)

[3.Depth Map Post-processing](#)

[4.Training Details](#)

[Experimentence](#)

[Conclusion](#)

简介

基于事件的深度估计是预测图像平面中每个像素的场景深度的任务，大多数基于事件的学习方法是使用①**标准前馈架构**来生成网络预测，这不利用事件流中存在的时间一致性。文中提出了一种循环架构来解决此任务，并显示出相对于标准前馈方法的显著改进。且文中讨论的**基于事件的单目深度估计**。

前馈神经网络（Feedforward Neural Network）是一种基本的神经网络结构，也被称为多层感知机（Multilayer Perceptron, MLP）。它是一种由多个神经元层组成的网络，信息在网络中从输入层向输出层单向传递，没有循环连接。这意味着数据流在网络中只能朝一个方向前进，从输入层到输出层，而没有任何环路或反馈

基于事件的深度估计的先前工作，②它们只能可靠地预测稀疏或半密集深度图或依赖stereo setup来生成密集深度预测。

三种map：

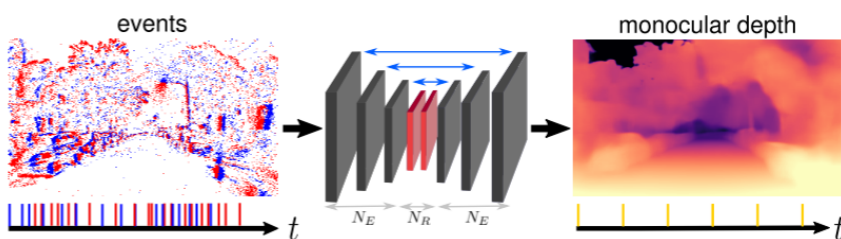
1. 稀疏深度图（Sparse Depth Map）：，稀疏深度图包含有限数量的深度值，通常只包括少数离散的像素位置的深度信息。这些深度值通常是通过深度传感器、立体视觉或其他深度估计方法获得的。**仅发生事件时像素处的深度**
2. 半密集深度图（Semi-Dense Depth Map）：半密集深度图在某些关键区域或特定像素周围提供深度值。**图像上重建边缘处的深度**
3. 密集深度图（Dense Depth Map）：密集深度图包括**图像中每个像素的深度值**，因此提供了完整的三维场景中物体距离的信息。（10.8也是Dense Depth）

stereo setup：即立体深度估计，双目方法，见10.8论文

本文讨论dense, monocular, and metric depth estimation 的事件相机深度估计

预测单个单目摄像头的密集度量深度。

metric depth estimation：度量深度估计，以度量单位（如米、厘米等）来表示物体在三维空间中的距离。而不是只提供深度的相对信息



网络接收异步事件输入并预测归一化对数深度 \hat{D}^k , 使用 N_R 循环块来利用事件输入中的时间一致性

contributions：

- A recurrent network that predicts dense per-pixel depth from a monocular event camera. (用循环网络做事件相机单目密集深度估计)
- The implementation of an event camera plugin in the CARLA simulator. (CARLA 是一个开源的自动驾驶汽车模拟器，为其开发了一个事件相机插件)
- **DENSE** - Depth Estimation on Synthetic Events: a new dataset with synthetic events and perfect ground truth. (开发了一个数据集)
- Evaluation of our method on the popular Multi-Vehicle Stereo Event-Camera (MVSEC) Dataset [39] where we show improved performance with respect to the state of the art (在多车辆立体事件相机数据集 (Multi-Vehicle Stereo Event-Camera, MVSEC) 上进行评估)

相关工作

1. 传统单目深度估计.....
2. 基于事件相机的深度估计，比较现有方法，**基于学习和基于模型的方法**：

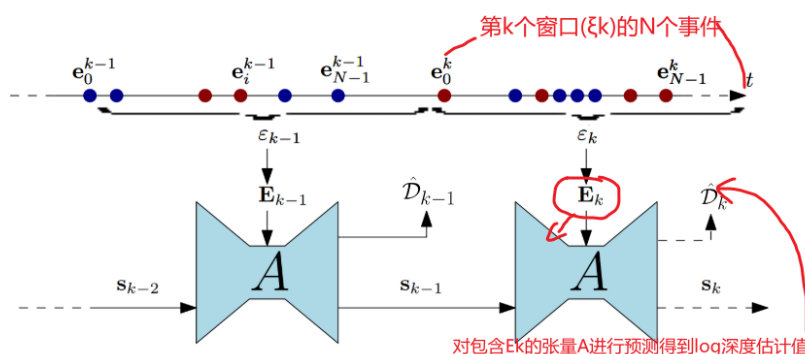
学习方法是指使用深度学习，从事件相机数据中学习深度估计模型，而基于模型的方法通常是依赖先验知识或特定模型来执行深度估计。（基于事件的深度估计的早期工作使用multi-view stereo、SLAM构建环境的表示，从而导出度量深度）
本文的方法属于学习方法，旨在预测单目摄像头的密集度量深度，通过利用事件流的时间一致性，采用递归卷积神经网络架构并使用合成和真实数据进行训练来实现这一目标。

Depth Estimation Approach

目标是从事件流中连续预测单目摄像头的密集深度，方法是以通过处理连续的**非重叠事件窗口**来实现深度估计。每个窗口包含一组事件，对于每个窗口，预测**对数深度图**，具体地，是以一个带有内部状态 s_k 的循环卷积神经网络来实现对数深度的预测。通过**监督学习**的方式进行网络训练，使用ground truth作为训练数据。在模拟环境中使用完美的真实数据和合成事件进行训练，然后在真实序列中进行微调。

内部状态 s_k ：在这个循环卷积神经网络中，有一个内部状态 s_k ，用来保存关于先前时间步的信息。这个内部状态有助于网络在处理事件窗口之间的数据时保持一些记忆或上下文信息。

1.Event Representation



事件相机捕获的一系列事件数据在时间和空间上是分散和不规则的，为了更好地处理这种**稀疏和异步**的事件数据，需要将其转换为一种更结构化的“张量”形式。文中提到的方式是“spatio-temporal voxel grid”，即 $B \times H \times W$ 的三维形式。其中 **B** 是时间维度。

时间窗口 ΔT 内的事件，根据以下公式收集到 **B** 个时间块中，总共有 B 个 k ：

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

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

结果表示了在指定时间块和时间戳内的事件数据的情况。这个元素本身是一个标量值，表示时间块内事件数量。

解释：

- $\delta(\mathbf{u}_i - \mathbf{u}_k)$ ：检查事件是否属于时间块 \mathbf{u}_k
- $\max(0, 1 - |t_n - t_i^*|)$ ：确定事件是否在时间块内。 t_i^* 表示事件的标准化时间戳

这里对比10.8论文的事件表征方式，是不是可以作一个结合

2. Network Architecture

网络结构：

- **网络架构**：网络采用了一个基于UNet架构的循环、全卷积神经网络。
- **网络组成**：网络由以下组件构成：
 - 输入数据首先经过头部层（H）进行处理。
 - 然后是NE个的重复编码器层（Ei）。
 - 每个编码器层之后跟随NR个残差块（Rj）。
 - 接着是NE个解码器层（DI）。
 - 最后的深度预测层（P）生成网络的输出，用于产生深度估计。
- **通道和特征图**：头部层生成具有Nb个通道的初始特征表示，而在每个编码器层中，通道数量会翻倍，从而产生了一个特征图，其输出通道的数量为Nb*2^NE。
- **深度预测**：深度预测层（P）执行深度卷积操作，生成一个具有一个输出通道的深度估计结果。
- **跳跃连接**：网络中使用了对称编码器和解码器层之间的跳跃连接，这有助于信息传递和特征共享。
- **激活函数**：网络中使用了ReLU激活函数（最后一层的深度预测使用了sigmoid激活函数）
- **编码器和解码器层**：编码器层采用降采样卷积操作，而解码器层采用双线性上采样操作，这些操作有助于提取和还原特征。
- **网络参数设置**：在这个工作中，设定了一些网络的超参数，包括编码器层数（NE = 3）、残差块数量（NR = 2）、初始通道数量（Nb = 32），以及网络在时间上展开的步数（L = 40）。

3. Depth Map Post-processing

训练网络预测归一化的log depth map（可以在紧凑的范围内表示大的深度变化，从而有利于学习）

恢复度量深度公式（Dk是模型预测深度；Dmax 是最大期望深度，表示场景中可能的最大深度； α ：当深度值为 0 时，对应于在观测数据中观测到的最小深度，文中选择 $\alpha = 3.7$ ，这意味着他们根据任务和观测数据的特性，使得深度值为 0 对应于实际观测到的最小深度为 2 米）：

$$\hat{D}_{m,k} = D_{\max} \exp(-\alpha(1 - \hat{D}_k))$$

4. Training Details

①损失函数

以一种监督学习的方式来训练神经网络。在训练过程中，通过**最小化尺度不变损失和多尺度尺度不变梯度匹配损失（Scale-Invariant Loss and Multi-Scale Scale-Invariant Gradient Matching Loss）**来优化网络的参数，这个训练流

程的目标是通过不同损失函数的组合来训练深度估计网络，以提高其在事件数据和ground truth之间的匹配，同时增加尖锐深度不连续性的感知。

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

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

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

总结：

概述	详细信息
训练方法	以监督学习方式训练深度估计网络，最小化每个时间步的尺度不变和多尺度尺度不变梯度匹配损失。
数据集	使用 合成数据(密集深度图像不易获得) 和来自MVSEC数据集的 真实事件数据 进行训练。
事件相机模拟器	基于ESIM，实现了一个事件相机传感器，在CARLA模拟器中生成事件数据。
事件数据模拟	事件相机传感器从渲染图像中计算每个像素的亮度变化，以模拟事件相机的操作。
数据集（DENSE）	划分为五个序列用于训练，两个序列用于验证，一个序列用于测试，总共八个序列。
样本	-一个RGB图像、两个连续图像之间的事件流数据、地面真实深度和分割标签的元组。
	-网络只使用事件数据和深度地图进行训练。RGB图像仅用于可视化目的，分割标签用于数据集信息。
优化器	Adam优化器
超参数	超参数 $\lambda=0.5$ 通过交叉验证选择。批量大小为20，学习率为 $10^{(-4)}$ 。
数据集场景划分	使用CARLA Towns 01到05进行训练，Town 06和07用于验证，测试序列由Town 10获取。
训练样本总数	5000个样本用于训练，2000个样本用于验证，1000个样本用于测试。
模型损失函数	尺度不变损失和多尺度尺度不变梯度匹配损失
多尺度尺度不变梯度匹配损失	鼓励深度变化的平滑和深度图预测中的尖锐深度不连续性。
残差定义	使用残差 $R_k = \hat{D}_k - D_k$ ，其中 \hat{D}_k 表示网络估计的深度， D_k 为地面真实深度。
损失计算	- 尺度不变损失：对每个像素 u 计算残差平方，并计算残差之和的平方。
	- 多尺度尺度不变梯度匹配损失：计算不同尺度下残差的L1范数。

三个损失函数:

1. 尺度不变损失 (Scale-Invariant Loss) :

这个损失函数用于测量深度预测的准确性。它基于ground truth与模型预测深度图之间的残差 $R_k(R_k = \hat{D}_k - D_k)$ ，即模型预测值减去真实值。损失函数包括两项：第一项是 R_k 中每个有效地面像素 u 的平方和的均值，用于衡量平均残差的大小；第二项是 R_k 中所有像素的平均值的平方，用于衡量残差的平均值。这些项的组合构成了尺度不变损失，其中 n 代表有效ground truth的数量。**对每个像素 u 计算残差平方，并计算残差之和的平方**

2. 多尺度尺度不变梯度匹配损失 (Multi-Scale Scale-Invariant Gradient Matching Loss) :

这个损失鼓励深度变化的平滑性并强制深度图预测中的尖锐深度不连续性。是在不同尺度 s 处的残差 $R_k s$ ，并使用 L1 范数来衡量这些残差的梯度。具体来说，它计算了 $R_k s$ 的 x 和 y 方向梯度的绝对值，然后对所有像素 u 和所有尺度 s 进行求和。有助于强化深度图中尖锐深度不连续性的信号。**计算不同尺度下残差的L1范数**

3. 总损失 (Total Loss) :

对于一系列 L 个depth map，整体损失由尺度不变损失和多尺度尺度不变梯度匹配损失组成。通过将尺度不变损失的总和与多尺度尺度不变梯度匹配损失的加权总和（加权参数为 λ ）相加而得出。这个总体损失函数用于优化深度预测模型，其中 λ 是一个通过交叉验证选择的超参数。

Experiment

1. 进行消融研究比较使用**合成训练数据**和**真实事件数据**的效果，进行定量和定性实验分析。

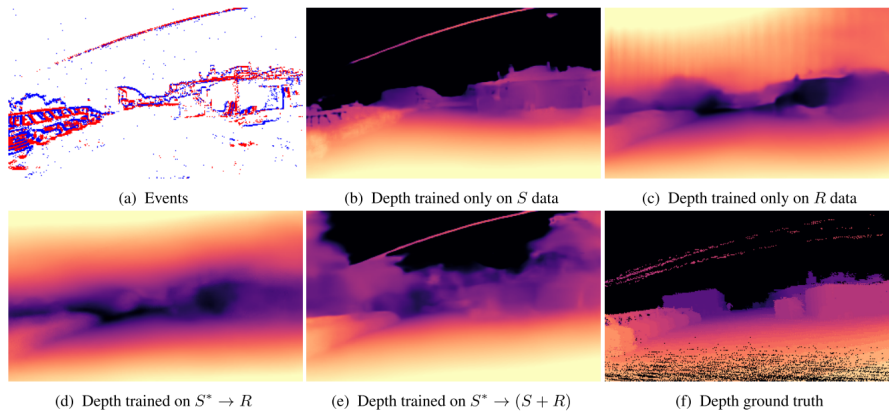
Training with **real** reduces the errors by predicting the correct metric while using **synthetic** data helps to estimate qualitatively better depth maps. (**S** \rightarrow (**S**+**R**))

合成数据：通过计算机模拟生成，生成的事件可以提供完美的depth map。上面提到的他们做的**DENSE数据集**（用 CARLA 记录的数据集，其中包括事件、强度帧、语义标签和深度图）。

真实数据：从**MVSEC数据集**中获取，代表实际场景中的事件数据。depth map与事件数据是同步获取的。通过其他技术来获取与事件数据同步的深度地图。

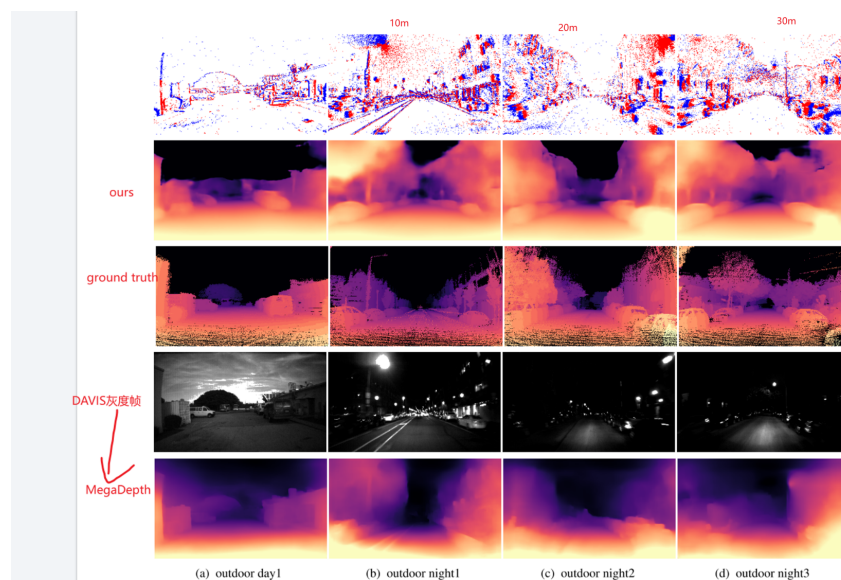
Training set	Dataset	Abs Rel \downarrow	Sq Rel \downarrow	RMSE \downarrow	RMSE log \downarrow	SI log \downarrow	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
S	outdoor day1	0.698	3.602	12.677	0.568	0.277	0.493	0.708	0.808
R		0.450	0.627	9.321	0.514	0.251	0.472	0.711	0.823
S* \rightarrow R		0.381	0.464	9.621	0.473	0.190	0.392	0.719	0.844
S* \rightarrow (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* \rightarrow R		0.554	1.798	10.738	0.622	0.343	0.390	0.598	0.737
S* \rightarrow (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* \rightarrow R		0.367	0.369	9.870	0.621	0.279	0.422	0.627	0.745
S* \rightarrow (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* \rightarrow R		0.339	0.230	9.537	0.606	0.258	0.429	0.644	0.760
S* \rightarrow (S+R)		0.277	0.226	8.056	0.424	0.162	0.541	0.761	0.890

MVSEC 的消融研究和评估



tip：两个数据集都模仿 DAVIS346B 传感器分辨率和焦距，因此可以在不改变损失函数的情况下组合合成训练数据和真实训练数据

2.与现有的单目深度估计方法进行了比较，包括两种基于图像的技术（MonoDepth 3和MegaDepth）以及基于事件的方法。特别是在暗光条件下体现了优越性（night2，3）



Conclusion

主要贡献	数据集	合成数据的优势	方法的性能
单目从事件数据中进行密集深度估计的首次尝试	MVSEC（多车辆立体事件摄像机数据集）	快速收敛、地面真实数据提高深度图质量、模拟多条件优势	相对于现有方法，生成更准确的密集深度图。

作者最后还给出一个观点：事件有足够的信息来估计密集的单目深度。

10.8的论文使用的双目方法，本文中多次强调这是第一次事件相机单目深度估计，但我还不太理解文中哪一部分内容是强调单目而非双目的解决方案。提到的数据表征的方式是不是针对单目的，与10.8（双目）的数据表征方式有何异同。