武汉大学学报(信息科学版)
Geomatics and Information Science of Wuhan University
ISSN 1671-8860,CN 42-1676/TN

《武汉大学学报(信息科学版)》网络首发论文

题目：基于图像和事件的无监督学习稠密连续光流估计
作者：胡建朗，郭迟，罗亚荣
DOI：10.13203/j.whugis20230390
收稿日期：2024-05-22
网络首发日期：2024-06-21
引用格式：胡建朗，郭迟，罗亚荣．基于图像和事件的无监督学习稠密连续光流估计[J/OL]．武汉大学学报(信息科学版). https://doi.org/10.13203/j.whugis20230390

网络首发时间：2024-06-21 14:46:27
网络首发地址：https://link.cnki.net/urlid/42.1676.tn.20240621.0927.007

DOI:10.13203/j.whugis20230390
引用格式：
胡建朗，郭迟，罗亚荣. 基于图像和事件的无监督学习稠密连续光流估计[J].武汉大学学报(信息科学版),2024,DOI: 10.13203/j.whugis20230390（HU Jianlang，GUO Chi，LUO Yarong. Unsupervised Dense and Continuous Optical Flow Estimation Based on Image and Event Data[J].Geomatics and Information Science of Wuhan University,2024,DOI: 10.13203/j.whugis20230390）

基于图像和事件的无监督学习稠密连续光流估计

胡建朗[1,2] 郭迟[1,2,3] 罗亚荣[1]
1 武汉大学卫星导航定位技术研究中心，湖北 武汉，430079
2 武汉大学人工智能研究院，湖北 武汉，430079
3 湖北珞珈实验室，湖北 武汉，430079

摘要： 为了获得稠密且连续的光流，并实现长时间间隔光流估计，本文提出了一种基于图像和事件的多模态多尺度递归光流估计网络，它能够融合从输入的单个图像和对应事件流中提取的多尺度特征，并以从粗糙到精细和迭代递归的方式对输出的光流进行优化。为了摆脱对光流标注数据的依赖，本文采用无监督学习的方式对网络进行训练，并设计了动态损失过滤机制，该机制能够自适应地过滤训练过程中的不可靠监督梯度信号，进而实现更加有效的网络训练。本文采用MVSEC数据集对本文提出的网络和策略进行综合对比分析，结果表明，本文方法具有更高的光流估计精度，尤其是在长时间间隔稠密光流估计方面，本文方法在三个室内序列上进行测试，并在平均端点误差和异常值百分比这两项指标上取得了最优结果，分别为1.43、1.87、1.68以及7.54%、14.36%、11.46%，均优于DCEIFlow方法，这说明本文方法不仅能够实现稠密且连续的光流估计，而且在长时间间隔光流估计方面也更加具有优势。
关键词： 光流估计；事件相机；多模态；无监督学习

Unsupervised Dense and Continuous Optical Flow Estimation Based on Image and Event Data

HU Jianlang[1,2] GUO Chi[1,2,3] LUO Yarong[1]
1. GNSS Research Center, Wuhan University, Wuhan 430079, China
2. Artificial Intelligence Institute, Wuhan University, Wuhan 430079, China
3. Hubei Luojia Laboratory, Wuhan 430079, China

Abstract: Objectives: Dense and continuous optical flow plays an important role in many applications, including robot navigation, autonomous driving, motion planning, visual odometry, etc. Current related works mainly utilize shutter cameras and event cameras to output optical flow. However, dense and continuous flow estimation is still a challenge due to the fixed frame rate of shutter camera and the sparse event data. In addition, existing related approaches focus on the way how to integrate images and event data, but neglect to deal with the long-time-interval optical flow estimation. **Methods**: To this end, we propose a multi-scale recurrent optical flow estimation framework fusing events and images. The network architecture contains three components: multi-scale feature extractor, image-event feature fusion module and flow recurrent updater. The multi-scale feature extractor is a CNN-based downsampler capable of mapping input image and event data into features at different scales. The image-event feature fusion module is applied to fuse features from two different modalities of data. The flow recurrent updater is a recurrent residual flow optimizer, incorporated the pyramid methods, estimating flow in coarse-to-fine way as well as performing flow feature refinement. Furthermore, to avoid expensive flow annotations and perform effective network training, we train network in the unsupervised way and design a novel training strategy, namely dynamic loss filtering mechanism, to filter out redundant and unreliable supervisory signals. **Results**: We conduct a series of experiments on the MVSEC dataset. The results show the proposed method performs well in both indoor and outdoor sequences. In particular, for long-time-interval dense optical flow estimation, the proposed method which tested on three indoor sequences achieves optimal performance in mean endpoint error and the percentage of anomalies, which are 1.43, 1.87, and 1.68, as well as 7.54, 14.36, and 11.46%, respectively. **Conclusions**: The proposed method not only can perform dense and continuous optical flow estimation, but also has a remarkable advantage on long-time-interval optical flow estimation.
Keywords: optical flow estimation, event camera, multimodal, unsupervised learning

收稿日期：2024-05-22
基金项目：湖北重大科技专项（2022AAA009）；中国博士后科学基金资助（2023TQ0248）；湖北珞珈实验室开放基金（230100007）。
第一作者：胡建朗，博士生，主要从事智能导航以及光流估计方面的研究。hujianlang123@whu.edu.cn
通讯作者：罗亚荣，博士后。yarongluo@whu.edu.cn

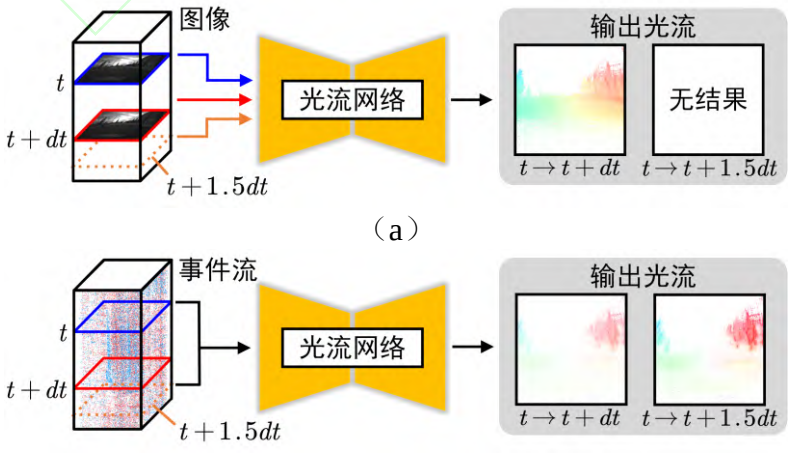
光流（Optical Flow, OF）反映了计算机视觉系统中图像像素的二维运动[1]。作为一项基本的视觉任务，光流估计技术在机器人导航[2]、自动驾驶[3]、障碍物检测[4]、运动规划[5]、视觉里程计[6]、行人运动估计[7]等领域发挥了非常重要的作用。光流估计任务中常用的视觉传感器包括快门相机（Shutter Camera）和事件相机（Event Camera）。受到快门相机的固定帧率限制，现有的基于图像的光流估计方法只能够估计整数图像帧间隔的光流。与此同时，事件相机输出的稀疏事件流增加了空间上下文信息的编码难度，使得基于事件的光流估计方法难以输出稠密的光流。也就是说，仅使用快门相机或者事件相机难以估计稠密且时间连续[8]的光流。

近年来有文献[9-11]指出，快门相机和事件相机在光流估计任务中是优势互补的。一方面，快门相机拍摄的图像具有丰富的场景纹理信息，可以补充事件数据编码缺失的空间上下文信息。另一方面，事件相机能够以高时间分辨率的形式异步地记录图像中每个像素的亮度变化[12]，进而增强模型输出光流的时间连续性。一般情况下，基于图像的模型以两个连续图像帧为输入，其输出的光流反映的是两个连续图像帧时间间隔的像素运动（dt表示间隔图像帧的数量，dt=1表示间隔图像帧数量为1）；而融合图像和事件的模型以图像和事件作为输入，能够通过调节输入的事件数量来估计任意时间间隔的光流[13]。在模型能够估计任意时间间隔光流这一前提下，如何准确地估计更长时间间隔的光流，同样是非常值得研究的问题。在本文中，本文将反映大于两个连续图像帧时间间隔的像素运动的光流称为长时间间隔光流。

由于传统的基于图像的光流估计方法在挑战性视觉条件、时间连续性等方面上难以取得令人满意的效果，加之大规模事件数据集的提出，目前有部分学者专注于研究基于学习的事件光流估计方法。这类方法一般使用卷积神经网络（Convolutional Neural Network, CNN）来处理数据，并使用真值数据直接对网络进行优化，比如E-RAFT[14]、STE-FlowNet[15]、TMA[16]等。囿于可供训练的标签数据不足等原因，一部分工作则专注于研究基于无监督学习的事件光流估计方法。EV-FlowNet[17]是第一个基于事件的端对端的无监督光流估计框架。在训练阶段，它基于亮度一致性假设[18]，通过测量由DAVIS相机[19]拍摄的原始图片和重构图片的差异来为网络训练提供监督信号。随后，文献[20]提出一种无监督光流网络训练方法，它只使用事件相机数据，通过最小化对比度最大化代理损失函数来最小化运动模糊，从而训练网络。除了使用CNN之外，由于脉冲神经网络（Spiking Neural Network, SNN）能够在专用的神经形态硬件上以节能的方式处理异步事件数据[21]，近来一部分工作专注于探索使用SNN来实现光流估计，比如Spike-FlowNet[22]、Fusion-FlowNet[11]、LIF-EV-FlowNet[23]。虽然这些只基于事件的方法能够估计更加连续的光流，但是这些方法难以在没有事件的区域捕获空间上下文，从而估计稠密光流。为了克服这一缺陷，学者们提出融合图像和事件的多模态光流估计方法。

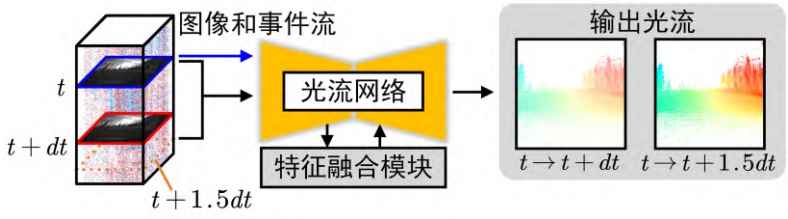
目前融合图像和事件的多模态光流估计方法主要分为基于优化（Optimization-based）和基于学习（Learning-based）两种方法。文献[9]和文献[10]是基于优化的方法，其中前者通过变分优化的方式，对未来的图像进行重建并估计相应的像素运动，后者则是构建了一个基于事件的亮度恒定性约束来描述光流和事件的关系，并通过变分优化的方式，联合实现光流估计和图像去模糊。然而这些方法在面对低纹理场景或者不稳定的噪声事件数据时的效果并不理想。受到基于学习的方法在各个计算机视觉任务中取得成功的影响，学者们也尝试使用神经网络来实现基于图像和事件的光流估计。Fusion-FlowNet[11]提出使用CNN和SNN分别对图像和事件进行特征编码，并在此基础上进行光流估计。DCEIFlow[13]利用事件构建伪图像特征，并与单个图像帧相结合，进而输出稠密且时间连续的光流。然而，这些光流估计方法都只关注如何融合图像数据和事件数据，而没有重视如何在图像和事件的基础上准确地估计长时间间隔光流这一问题。

为了估计出稠密且时间连续的光流，本文提出了一种基于图像和事件的多模态多尺度递归光流估计网络。该网络的架构包含了多尺度特征提取器和图像-事件特征融合模块，前者能够将输入的图像数据和事件数据映射为多尺度特征，后者能够以两种模态数据的特征融合的方式生成伪图像特征，进而为后续的稠密连续的光流估计提供线索。需要注意的是，本文的网络以单个图像帧和相对应的连续事件流为输入，网络输出的连续光流的开始时间与输入图像帧的时间戳相同。为了实现长时间间隔光流估计，网络中的多尺度光流迭代更新器在递归迭代精化（Recurrent Iterative Refinement）[14-15]的基础上引入了被证明有助于提升大位移估计（Large Displacement Estimation）[24]性能的多尺度光流回归方法[17,20]，这使得网络除了能够递归迭代地优化单一尺度的像素运动估计结果，还能够以从小尺寸到大尺寸、从粗糙到精细方式输出全尺寸的稠密且时间连续的光流。本文方法与基于图像以及基于事件的光流估计方法的对比如图 1所示。



（a）

（b）



（c）

图 1 本文方法与基于图像以及基于事件的光流估计方法的对比示意图，其中t表示图像帧的时间戳，dt表示两个连续图像帧的整数时间间隔，也就是 dt=1。（a）基于图像的光流估计方法以连续的图像帧作为输入，估计稠密但是时间离散的光流。（b）基于事件的光流估计方法以事件流作为输入，输出时间连续但是稀疏的光流。（c）本文提出的方法通过特征融合模块融合输入的单个图像帧和事件流数据，并估计稠密且时间连续的光流

Fig.1 Comparison of our optical flow estimation framework with the learning-based methods using only images or events. t refers to the timestamp of a frame, and dt denotes the integer time interval between two consecutive frames. (a) Conventional image-only methods estimate dense optical flow from consecutive images at integer discrete time intervals. (b) Event-only methods take advantage of event streams to produce sparse but continuous pixel motions from any start time to end time. (c) Our framework integrates a single image and event stream through the feature fusion module and generates dense and continuous optical flow

现阶段要准确地标注光流真值仍然较为困难，尤其是在非常复杂真实的自然环境中，因此具有准确的光流标注的事件数据集较少。为了摆脱对光流真值标注数据的依赖，本文以无监督学习的方式训练网络。然而受到稀疏的事件数据、事件噪声、现实环境复杂的光照变化等因素的影响，在无监督学习的过程中网络不一定能够得到有意义的监督梯度信号[25][26]。为了实现更加有效的网络训练，本文设计了动态损失过滤机制，该机制包含空间平滑过滤策略和自适应损失选择策略，前者能根据事件的触发位置以及对应的空间上下文来过滤不可靠的监督梯度信号，后者则根据损失函数数值自适应保留对网络训练更加有利的信号。此外，为了进一步提高网络的性能，本文还采用了特征相似性损失函数和双向光流训练策略。

本文在MVSEC数据集[27]上评估了所提出的方法。实验结果表明，本文的网络在不同时间间隔（dt=1和dt=4图像帧间隔）以及不同场景中具有更高的光流计算精度，并且能够输出更稠密且连续的光流。此外，本文通过消融研究对本文提出的方法进行深入分析，以证明它们是有效的。

本文的主要贡献如下：

1、 为了能够估计出更加稠密且连续的光流，本文提出了一个基于图像和事件的多模态多尺度递归光流估计网络，该网络能够在特征层面建立输入的图像数据和事件数据的联系，并以从粗糙到精细和递归迭代的方式输出光流。

2、 为了摆脱对光流真值标注数据的依赖，并实现更加有效的网络训练，本文设计了动态损失过滤机制，该机制能够自适应地过滤不可靠的监督信号。

3、 为了评估本文所提出的方法，本文在MVSEC数据集上进行了一系列实验，结果表明，本文提出的网络不仅能够在保证光流估计精度的同时输出更加稠密且连续的光流，并且在长时间间隔像素运动估计方面具有更加明显的优势。

1 事件表示

原始的事件数据包含事件触发时间戳、事件触发像素位置以及事件极性三项数据，其中事件极性记录的是像素亮度值的变化情况。若当前时刻的像素亮度值高于上一时刻的像素亮度值，且亮度变化量大于特定阈值，则表示为正极性，若当前时刻的像素亮度值低于上一时刻的像素亮度值，且亮度变化量大于特定阈值，则表示为负极性[12]。和文献[13]类似，为了从原始的事件数据中保留时空信息，并使事件数据能够被CNN高效处理，本文将原始事件数据编码为和输入图像具有相同空间尺寸的三通道事件体E（Event Volume），编码后的事件体的尺寸为$H\times W\times B$，其中H和W表示图像的高度和宽度，B表示事件数据按照时间戳顺序被划分的份数，每一份被称为一个时间格子（Temporal Bin）。根据文献[13][28]，为了保留亮度变化信息，本文对表示为正极性的事件和表示为负极性的事件进行分开编码，得到两个事件体$\tilde{E}$，其中各个极性的事件数据按照时间顺序划分为5份，即$B=5$。本文将编码后的两个极性的事件体沿着通道维度进行拼接，最后构成的事件体的通道数$B=10$。事件体E的编码方式如下式所示：

$$\tilde{E}(b,\ x_i,\ p_i)=\sum_{i=0}^{N}\max\left(0,\ 1-\left|b-\frac{t_i-t}{dt}(B-1)\right|\right) \tag{1}$$

$$E(x)=\left[\tilde{E}(x,\ p=1)\,\|\,\tilde{E}(x,\ p=-1)\right] \tag{2}$$

式中，$b\in[0,\ B)$表示时间格子的索引号，$\|$表示张量在通道维度上的拼接操作，x表示触发事件的位置，p表示事件的极性，其中$p=1$表示正极性，$p=-1$表示负极性。

2 网络架构

本文提出的网络框架主要包含三个部分：多尺度特征提取器、图像-事件特征融合模块以及多尺度光流递归更新器，如图 2所示。从t时刻到$t+dt$时刻，本文首先将原始的事件流编码为三通道事件体$E_{t\to t+dt}$，并将其连同t时刻的灰度图像I_t输入多尺度特征提取器中，从而生成多尺度的事件特征$F^{E}_{t\to t+dt}$和图像特征F^{I}_{t}。与RAFT[29]类似，图像也被用于提取上下文信息F^{con}_{t}。然后，图像-事件特征融合器会利用事件特征$F^{E}_{t\to t+dt}$和图像特征F^{I}_{t}来生成$t+dt$时刻的图像伪特征$\hat{F}^{I}_{t+dt}$，从而用

于后续的代价体构建。在光流回归阶段，多尺度光流递归更新器会以输入的多尺度图像特征、上下文信息、代价体等作为线索，对不同尺度的光流进行递归迭代更新，并最终以从粗糙到精细的方式输出全分辨率的光流 f 。

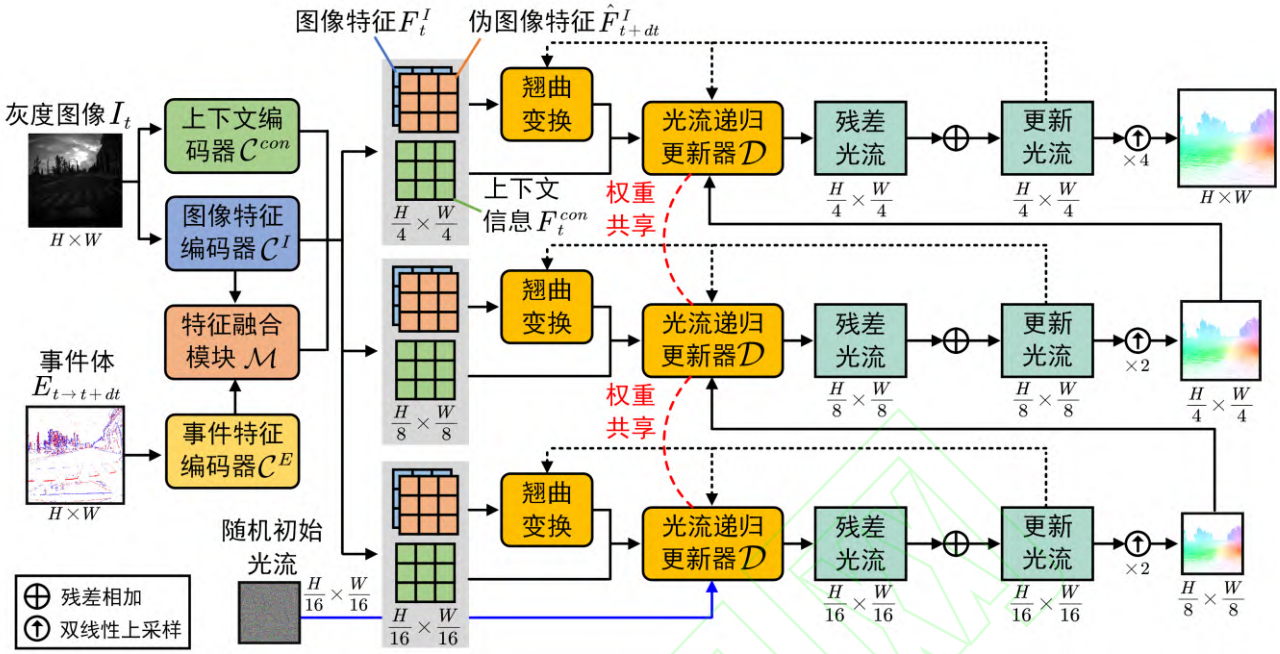


图 2 本文方法的架构示意图，其中主要包含三部分：1）多尺度特征提取器，它能够处理单个灰度图像和编码后的事件体，将其转换为多尺度特征和上下文；2）图像-事件特征融合模块，能够结合图像特征和事件特征来生成伪图像特征；3）光流递归更新器，能够利用输入的隐状态、前一次更新的光流（或者随机初始光流）、上下文信息、代价体以及残差光流，来迭代更新不同尺度的光流，并以从粗糙到精细的方式输出全分辨率的光流

Fig.2 The architecture of our method, which mainly contains three components: 1) The multi-scale feature extractor processes a single grayscale image and encoded event volumes into multi-scale features and context. 2) The image-event feature fusion module integrates image features and event features to generate pseudo image features. 3) The flow recurrent updater thoroughly exploits hidden states, former flow (or randomly initial flow), context, cost volumes and residual flow to refine flow iteratively at different scales, producing full-resolution optical flow in coarse-to-fine way

2.1 多尺度特征提取

多尺度特征提取器是一个基于CNN的特征编码器，分为图像特征编码器 $\mathcal{C}^I$ 和事件特征编码器 $\mathcal{C}^E$ ，能够分别对输入的图像和事件体进行编码，并映射为多尺度特征，在本文中若不进行额外说明，默认为三种尺度（1/4、1/8和1/16尺度）。此外本文还引入与多尺度特征编码器相同结构的上下文编码器 $\mathcal{C}^{con}$ 来从图像中提取上下文信息。被输入到特征提取器的数据会通过卷积层和残差层被分别下采样为1/4、1/8和1/16尺度的特征，并进一步地通过一个用于通道对齐的卷积核尺寸为1×1的独立卷积层。多尺度特征提取器输出的多尺度特征尺寸分别是 $H/4\times W/4\times C$ 、 $H/8\times W/8\times C$ 以及 $H/16\times W/16\times C$ ，其中 $C=256$ 。特征编码过程如下式所示：

$$F_t^I=\mathcal{C}^I(\theta^I,\ I_t) \tag{3}$$

$$F_{t\to t+dt}^E=\mathcal{C}^E(\theta^E,\ E_{t\to t+dt}^{weighted}) \tag{4}$$

$$F_t^{con}=\mathcal{C}^{con}(\theta^{con},\ I_t) \tag{5}$$

式中， θ^I 、 θ^E 和 θ^{con} 分别表示图像特征编码器 $\mathcal{C}^I$ 、事件特征编码器 $\mathcal{C}^E$ 和上下文编码器 $\mathcal{C}^{con}$ 的网络参数，且这些参数并不共享，这些多尺度特征将会被用于后续多尺度的图像-事件特征融合以及光流递归迭代更新。

2.2 图像-事件数据的特征融合

在基于图像的光流估计方法中，一般认为特征相关性非常重要[29]。受到该观点的启发，本文设计了图像-事件特征融合模块。该模块在网络训练和推理过程中，能够根据输入的 t 时刻的图像特征 F_t^I 和事件特征 $F_{t\to t+dt}^E$ ，以特征融合的方式构造 $t+dt$ 时刻的伪图像特征 $\hat{F}_{t+dt}^I$ ，以便于在光流回归阶段构造代价体。伪图像特征 $\hat{F}_{t+dt}^I$ 的构造如下式所示：

$$\hat{F}_{t+dt}^I=\mathcal{M}(\theta^{fuse},\ F_t^I,\ F_{t\to t+dt}^E) \tag{6}$$

式中， $\mathcal{M}$ 是图像-事件特征融合模块， θ^{fuse} 是对应的网络参数。需要注意的是，本文采用了双向光流训练技术[30]来对网络进行训练，因此在训练过程中，本文还需要将 $t+dt$ 时刻的图像特征 F_{t+dt}^I 和后向事件特征 $F_{t+dt\to t}^E$ 输入到模块中来生成 t 时刻的伪图像特征 $\hat{F}_t^I$ 。

图像-事件特征融合模块的结构如图 3所示，包含4个卷积层，其中前2个卷积层用于分别对图像特征和事件特征进行编码，后2个卷积层则用于融合两种模态数据的特征。此外，本文发现当图像特征编码分支和事件特征编码分支都是可微的时候，模型的性能会下降。本文推测是基于稀疏的事件数据编码得到的事件特征对伪图像特征构建具有不利的影响，因此本文

在图像特征编码分支处应用了梯度停止策略，使梯度传递到事件特征编码分支，后续本文通过实验验证了这一做法的有效性。

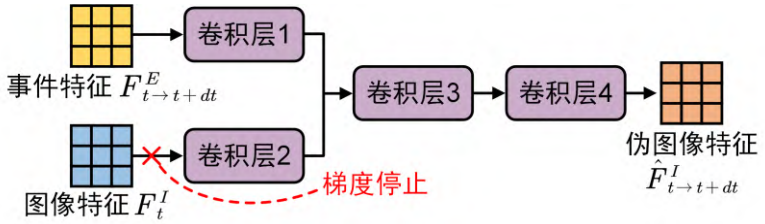


图 3 图像-事件特征融合模块的细节，其中主要包含了 4 个卷积层。需要注意的是，在图像特征分支使用了梯度停止策略

Fig.3 The details of the image-event feature fusion module, which mainly contains four convolution layers. Note that gradient stopping is applied on the image feature branch

2.3 多尺度光流递归更新

为了解决长时间间隔光流估计的问题，本文将传统的被证明能够提升大位移估计[24]性能的基于金字塔的光流估计模式[17,20]引入到递归迭代精化模式[14-15]中，设计了多尺度光流递归更新器。它基于ConvGRU模块[29]，能够在单一尺度上生成残差流，从而对当前预测的光流进行更新，并通过双线性上采样的方式以从粗糙到精细的方式输出全分辨率光流。为了避免构建全对互相关体（All-pairs Correlation Volume）导致内存不足的问题[31]，本文采用对图像特征F_t^I和伪图像特征$\hat{F}^I_{t+dt}$之间进行局部特征互相关（Correlation）计算的方式[32]来构建代价体C，如下式所示：

$$\begin{aligned} C(x,\ d) &= F_t^I(x) \odot Warp\left(\hat{F}^I_{t+dt},\ f\right)(x+d) \\ &= F_t^I(x) \odot Warp\left(\hat{F}^I_{t+dt}(x+d),\ f(x+d)\right) \\ &= F_t^I(x) \odot \hat{F}^I_{t+dt}(x+f(x+d)+d) \end{aligned} \tag{7}$$

其中，$\odot$表示向量点积操作，$d \in D$表示x方向或者y方向的偏移量，$D=[-d_{\max},\ d_{\max}]^2$，$Warp$表示翘曲变换操作（Warping），$x \in X$表示像素位置，$X=[0,h)\times[0,w)$，$h$表示特征图的高度，$w$表示特征图的宽度。

在训练和推理阶段，本文首先使用1/16尺度的随机光流进行初始化，并使用Tanh函数对1/16尺度的上下文信息进行激活，生成初始的隐状态，然后将1/16尺度的隐状态H、上下文信息F_{con}、待更新的光流$\tilde{f}$和代价体C输入到光流递归更新器$\mathcal{D}$中，使其输出残差光流Δf，从而对光流进行迭代更新。在光流递归更新器中，第i个尺度第k次迭代的光流更新过程如下式所示：

$$m_k^i = \left[Conv\left(\theta^m, \left[\tilde{f}_{k-1}^i \| C_{k-1}^i\right]\right) \| F_{con}^i\right] \tag{8}$$

$$z_k^i = \sigma\left(Conv\left(\theta^z, [H_{k-1}^i \| m_k^i]\right)\right) \tag{9}$$

$$r_k^i = \sigma\left(Conv\left(\theta^r, [H_{k-1}^i \| m_k^i]\right)\right) \tag{10}$$

$$\tilde{H}_k^i = tanh\left(Conv\left(\theta^h, [r_k^i \odot H_{k-1}^i \| m_k^i]\right)\right) \tag{11}$$

$$H_k^i = (1-z_k^i) \odot H_{k-1}^i + z_k^i \odot \tilde{H}_k^i \tag{12}$$

$$\Delta f_k^i = Conv\left(\theta^{df},\ H_k^i\right) \tag{13}$$

$$f_k^i = f_{k-1}^i + \Delta f_k^i \tag{14}$$

式中，$Conv$表示卷积操作，θ^m、θ^z、θ^r、θ^h和θ^{df}表示这些卷积操作的网络参数，σ表示Sigmoid激活函数，$tanh$表示Tanh激活函数。在对单个尺度的光流进行多次更新后，本文通过2倍双线性上采样的方式将更新后的光流上采样到1/8尺度，该上采样后的光流将被作为1/8尺度迭代更新的初始光流，如下式所示：

$$f_0^{i+1} = \mathcal{S}_{\uparrow}(f_k^i) \tag{15}$$

式中$\mathcal{S}_{\uparrow}$表示双线性上采样操作。如此类推，通过这种方式，光流更新器可以对多种尺度的光流进行精化，并最终以从粗糙到精细的方式输出全分辨率的光流。在实际运行阶段，本文可以根据需要，增加光流迭代的尺度来提高光流估计的精度，或者减少网络的迭代次数来减少计算所需要的时间。

3 无监督训练

由于缺乏带有精确光流真值标注的事件光流数据集，本文采用基于图像的无监督学习的方式[17]来训练模型。本节主要介绍无监督训练过程中使用的损失函数和训练策略。除了使用了传统无监督光流估计方法常用的无监督损失函数外，本文还在训练过程中使用了动态损失过滤机制，该机制能够自适应地过滤不可靠的监督梯度信号，从而实现更加有效的网络训练。为了进一步提高光流的估计精度，本文还使用了特征相似性损失函数和双向光流训练策略。

3.1 无监督损失函数

传统的基于图像的无监督光流估计方法将光流网络的训练重新定义为图像重构问题，通过测量并最小化原图像和重构图像之间的差异，从而不依靠光流真值产生监督梯度信号来训练网络[33]。根据文献[17]，基于图像的无监督损失函数$\mathcal{L}_{un}$包含亮度损失函数$\mathcal{L}_{photo}$和平滑损失函数$\mathcal{L}_{smooth}$两部分。其中亮度损失函数被提出通过惩罚图像之间的亮度不一致性来对齐图

像，平滑损失函数用于增强空间一致性，并最小化邻域光流的差异。给定t时刻到$t+dt$时刻的前向预测光流f、开始图像帧I_t和结束图像帧I_{t+dt}，则相关的无监督损失如下式所示：

$$\mathcal{L}_{un}=\mathcal{L}_{photo}+\lambda\mathcal{L}_{smooth} \tag{16}$$

$$\mathcal{L}_{photo}=\sum_{x}\rho(I_t(x)-I_{t+dt}(x+f(x))) \tag{17}$$

$$\mathcal{L}_{smooth}=\frac{1}{HW}\sum_{x}|\nabla f^{u}(x)|+|\nabla f^{v}(x)| \tag{18}$$

式中，∇表示一阶差分算子；f^u和f^v分别表示光流的横向分量和纵向分量；H和W分别表示预测光流的高度和宽度，ρ表示鲁棒Charbonnier损失函数$\rho(s)=(s^2+\eta^2)^r$，与文献[17]和文献[22]一样，本文将参数设置为$r=0.45$以及$\eta=1e^{-3}$；权重因子λ被设置为10。

3.2 动态损失过滤机制

光流网络的无监督训练容易受到各种不利因素的影响，比如由稀疏的事件数据、事件噪声等，这些因素会增加光流网络输出任意像素运动的概率，也就是产生不准确的随机光流[26]，如果不准确的随机光流使得经过翘曲变换后重构的图像的像素位置与真实像素位置距离相差过大时，网络无法得到有意义的监督梯度信号[25]。为了解决这一问题，本文提出了动态损失过滤机制，这一机制包含了空间平滑过滤策略和自适应损失选择策略。

为了缓解稀疏的事件数据所带来的负面影响，由于存在图像编码的稠密空间上下文信息，本文假设在触发了事件的位置及其邻域的预测光流是相对可靠的。基于该假设，本文设计了空间平滑过滤策略。本文先获取事件有效掩膜M，该掩膜是二值掩膜，标记了当前像素位置是否有事件触发；然后对M进行空间平滑处理，得到空间平滑有效性掩膜$\hat{M}$；最后通过$\hat{M}$对亮度损失函数中不可靠的监督梯度信号进行过滤。事件有效掩膜M和空间平滑有效性掩膜$\hat{M}$的计算方式如下式所示：

$$M(x)=\begin{cases}1 & if\ E(x)>0\\ 0 & otherwise\end{cases} \tag{19}$$

$$\hat{M}=M\odot\mathcal{K} \tag{20}$$

式中，$\odot$表示零填充的卷积操作，$\mathcal{K}$表示尺寸为$m\times m$的均值卷积核，此处本文的设置为$m=3$。加入空间平滑有效性掩膜$\hat{M}$的亮度损失函数如下所示：

$$\mathcal{L}_{photo}=\sum_{x}\rho(I_t(x)-I_{t+dt}(x+f(x)))\cdot\hat{M}(x) \tag{21}$$

尽管大部分由稀疏的事件数据引起的不可靠监督信号会被空间平滑过滤策略给过滤掉，但是不可靠的监督梯度信号依旧存在，这些信号可能是大位移运动、像素遮挡、现实世界复杂的亮度变化等因素所导致的。为了缓解这一问题，本文引入了一种遵循小损失原则的自适应损失选择策略[34]。该策略会将损失数值过大的监督梯度信号去除，只保留特定比例的监督梯度信号，从而减少不可靠监督梯度信号对网络训练的影响。本文将空间平滑有效性掩膜$\hat{M}$中数值不为零的像素位置的集合表示为$\mathcal{D}=\{x|\hat{M}(x)>0\}$，则应用了空间平滑过滤策略的亮度损失函数可以表示为下式：

$$\mathcal{X}=\underset{\bar{\mathcal{D}}:|\bar{\mathcal{D}}|=round(RN_{\hat{M}})\wedge\bar{\mathcal{D}}\subseteq\mathcal{D}}{\operatorname{argmin}}\sum_{x\in\bar{\mathcal{D}}}\rho(I_t(x)-I_{t+dt}(x+f(x)))\cdot\hat{M}(x) \tag{22}$$

$$\mathcal{L}_{photo}=\sum_{x\in\mathcal{X}}\rho(I_t(x)-I_{t+dt}(x+f(x)))\cdot\hat{M}(x) \tag{23}$$

式中，R控制需要保留的监督梯度信号数量百分比，此处本文将参数设置为$R=0.8$；$round$表示向上取整操作；$N_{\hat{M}}$表示空间平滑有效性掩膜$\hat{M}$中数值不为零的像素的总数量；$\mathcal{X}$表示保留的监督梯度信号对应的像素位置集合。

3.3 特征相似性损失函数

在特征融合阶段，通过图像特征和事件特征来融合出另一个时刻的伪图像特征，以便于构建特征相关性。为了降低伪特征合成的误差，计算各个尺度原始图像特征和对应的伪图像特征的欧氏距离，将其作为特征相似性损失函数，如下式所示：

$$\mathcal{L}_{sim}=\left\|F_t^I-\hat{F}_t^I\right\|_2+\left\|F_{t+dt}^I-\hat{F}_{t+dt}^I\right\|_2 \tag{24}$$

3.4 双向光流训练策略

为了提高模型估计光流的性能，同时不增加额外的网络参数，引入并改进了传统无监督学习光流估计方法的双向光流训练策略[30]。在训练阶段，首先通过颠倒原始事件流数据的时间戳顺序和各个事件的极性，得到后向的事件流数据$E_{t+dt\to t}$。然后将图像I_t和前向事件流数据$E_{t\to t+dt}$输入到网络中，使其输出前向光流，与此同时将图像I_{t+dt}和后向的事件流数据

$E_{t+dt\to t}$输入到网络中，使其输出后向光流。根据公式（16），可以计算相应的前向无监督损失$\mathcal{L}_{un}^{f}$和后向无监督损失$\mathcal{L}_{un}^{b}$。总的来说，本文在无监督训练中所采用的总损失函数$\mathcal{L}_{total}$如下式所示：

$$\mathcal{L}_{un}=\frac{1}{2}(\mathcal{L}_{un}^{f}+\mathcal{L}_{un}^{b}) \tag{25}$$

$$\mathcal{L}_{total}=\mathcal{L}_{un}+\lambda_{sim}\mathcal{L}_{sim} \tag{26}$$

式中，λ_{sim}表示特征相似性损失函数的权重因子。

4 实验

在本节中，对针对模型进行一系列实验和分析。因此首先介绍了实验设置，包括所使用的数据集以及模型的训练参数设置，然后展示了一系列与现有方法对比的定量和定性结果并进行分析，最后通过消融实验验证了本文提出的各个策略的有效性。

4.1 实验数据集

本实验使用的MVSEC数据集[27]是一个包含事件流数据和连续灰度图像的室内室外场景数据集，这些数据主要是使用DAVIS346相机[19]记录的。在接下来的实验中，在两种不同的时间间隔（dt=1和dt=4图像帧间隔）下进行模型训练和评估，其中在3个室内的序列（indoor_flying1、indoor_flying2、indoor_flying3）和1个室外的序列（outdoor_day1）评估模型，将outdoor_day2序列用于模型训练。在模型训练时，以中心裁切的方式将灰度图像和事件体裁切到空间尺寸为256×256的大小，并以0.5的概率对它们进行随机水平和垂直翻转。

4.2 训练参数设置

本文方法是在PyTorch框架[35]下实现，并使用AdamW优化器[36]进行优化，优化器参数设置为$\beta_1=0.9$，$\beta_2=0.999$，$\epsilon=10^{-8}$。训练模型的batch设置为8，训练40个epoch。初始学习率设置为4×10^{-4}，并分别在第5、10、20个epoch时将学习率降低至原来的70%。在训练期间，光流回归器估计三个尺度的光流（1/16、1/8和1/4），各个尺度迭代次数为6。在推理期间，各个尺度的迭代次数被设置为12。此外，模型训练以及下文的实验均在NVIDIA RTX 4090上进行。

4.3 评估指标

在定量评估中，采用平均端点误差和异常值百分比两种量化评价指标对本文方法的光流计算结果进行量化评价。平均端点误差反映的是预测光流与真值光流之间的平均欧几里得距离，其计算公式如下：

$$AEE=\frac{1}{m}\sum_{m}\|f_{est}-f_{gt}\|_2 \tag{27}$$

式中，AEE表示平均端点误差，m表示像素数量，对于稠密光流评估，m表示具有有效的光流真值标注的像素的个数，对于稀疏光流评估，m是具有有效的光流真值标注并且至少触发了一个事件的像素个数。

异常值百分比反映的是平均端点误差大于3个像素的百分比，其计算公式如下：

$$Out=\frac{\sum P(AEE>\tau)}{m}\times100\% \tag{28}$$

式中，Out表示异常值百分比，$P(AEE>\tau)$表示平均端点误差大于τ的像素点，其中$\tau=3$。

表 1 各个使用了事件数据的无监督光流估计方法在 MVSEC 数据集上四个序列上不同时间间隔（dt=1 和 dt=4 图像帧间隔）的性能对比。注意表中参与对比的其它方法的结果来自对应的原始论文

Tab.1 Comparison of unsupervised optical flow estimation methods using event data on the four sequences from MVSEC dataset in different time intervals (dt=1 and dt=4 frame intervals). Note that the results of compared methods are extracted from original papers

输入	方法	评估类型	Indoor_flying1		Indoor_flying2		Indoor_flying3		Outdoor_day1	
			AEE	Out(%)	AEE	Out(%)	AEE	Out(%)	AEE	Out(%)
以下为时间间隔 dt=1 的性能对比结果										
E	EV-FlowNet[17]	稀疏光流	(1.03)	(2.2)	(1.72)	(15.1)	(1.53)	(11.9)	[0.49]	[0.2]
	文献[20]	稀疏光流	(0.58)	**(0.0)**	(1.02)	(4.0)	(0.87)	(3.0)	**[0.32]**	**[0.0]**
	Spike-FlowNet[22]	稀疏光流	[0.84]	–	[1.28]	–	[1.11]	–	[0.49]	–
	LIF-EV-FlowNet[23]	稀疏光流	0.71	1.41	1.44	12.75	1.16	9.11	0.53	0.33
	STE-FlowNet[15]	稀疏光流	[0.57]	[0.1]	[0.79]	[1.6]	[0.72]	[1.3]	[0.42]	**[0.0]**
	TMA[16]	稀疏光流	(1.06)	(3.63)	(1.81)	(27.29)	(1.58)	(23.26)	**(0.25)**	(0.07)
E+I	Fusion-FlowNet[11]	稠密光流	(0.62)	–	(0.89)	–	(0.85)	–	[1.02]	–
	Fusion-FlowNet[11]	稀疏光流	(0.56)	–	(0.95)	–	(0.76)	–	[0.59]	–
	DCEIFlow[13]	稠密光流	0.56	0.28	**0.64**	**0.16**	**0.57**	**0.12**	0.91	0.71
	DCEIFlow[13]	稀疏光流	0.57	0.30	0.70	**0.30**	0.58	**0.15**	0.74	0.29

	本文方法	稠密光流	**(0.51)**	(0.12)	**(0.62)**	(0.57)	**(0.58)**	(0.42)	(0.65)	(0.38)
	本文方法	稀疏光流	**(0.50)**	**(0.06)**	**(0.65)**	(0.40)	(0.61)	(0.57)	(0.50)	**(0.04)**
以下为时间间隔 dt=4 的性能对比结果										
E	EV-FlowNet[17]	稀疏光流	(2.25)	(24.7)	(4.05)	(45.3)	(3.45)	(39.7)	[1.23]	[7.3]
	文献[20]	稀疏光流	(2.18)	(24.2)	(3.85)	(46.8)	(3.18)	(47.8)	[1.30]	[9.7]
	Spike-FlowNet[22]	稀疏光流	[2.24]	–	[3.83]	–	[3.18]	–	[1.09]	–
	LIF-EV-FlowNet[23]	稀疏光流	2.63	29.55	4.93	51.10	3.88	41.49	2.02	18.91
	STE-FlowNet[15]	稀疏光流	[1.77]	[14.7]	[2.52]	[26.1]	[2.23]	[22.1]	**[0.99]**	**[3.9]**
	TMA[16]	稀疏光流	(2.43)	(29.91)	(4.32)	(52.74)	(3.60)	(42.02)	**(0.70)**	**(1.08)**
E+I	Fusion-FlowNet[11]	稠密光流	(1.81)	–	(2.90)	–	(2.46)	–	[3.06]	–
	Fusion-FlowNet[11]	稀疏光流	(1.68)	–	(3.24)	–	(2.43)	–	[1.17]	–
	DCEIFlow[13]	稠密光流	**1.49**	**8.14**	**1.97**	17.37	**1.69**	**12.34**	1.87	19.13
	DCEIFlow[13]	稀疏光流	1.52	8.79	2.21	22.13	1.74	13.33	1.37	8.54
	本文方法	稠密光流	**(1.43)**	**(7.54)**	**(1.87)**	**(14.36)**	**(1.68)**	**(11.46)**	(1.86)	(19.71)
	本文方法	稀疏光流	(1.51)	(8.64)	(2.01)	**(17.16)**	(1.78)	(13.21)	(1.34)	(9.90)

注：[]的结果是对在 outdoor_day1 和 outdoor_day2 序列上训练的模型的评估结果； ( )的结果是对在 outdoor_day2 序列上训练的模型的评估结果；“–”表示其对应的原始论文中并没有给出相应的评估结果；没有被括号包围结果是对并非在 MVSEC 数据集上训练的模型的评估结果；红色的结果表示最优值，蓝色的结果表示次优值；“E”表示该方法的输入仅有事件数据；“E+I”表示该方法的输入包含事件数据和图像数据。

表 2 本文方法与 DCEIFlow 在不同时间间隔下（dt=0.8、dt=1.8、dt=2、dt=3、dt=3.2、dt=4.4 和 dt=5 图像帧间隔）的稀疏光流性能对比

Tab.2 Comparison of sparse flow performance in different time intervals (dt=0.8, dt=1.8, dt=2, dt=3, dt=3.2, dt=4.4 and dt=5 frame intervals) between proposed method and DCEIFlow

方法	时间间隔	Indoor_flying1		Indoor_flying2		Indoor_flying3		Outdoor_day1	
		AEE	Out(%)	AEE	Out(%)	AEE	Out(%)	AEE	Out(%)
DCEIFlow[13]	dt=0.8	0.56	0.40	0.74	0.77	0.63	0.33	0.92	1.30
	dt=1.8	0.89	1.38	1.13	3.64	1.02	2.55	1.38	9.05
	dt=2	1.06	1.99	1.40	6.27	1.27	4.34	1.14	1.73
	dt=3	1.22	4.21	1.60	10.58	1.38	6.10	1.83	17.67
	dt=3.2	1.24	4.28	1.66	12.10	1.38	6.26	1.83	17.57
	dt=4.4	1.60	10.02	2.07	19.43	1.99	18.08	2.30	25.71
	dt=5	1.85	15.40	2.54	28.90	2.18	22.77	2.68	31.23
本文方法	dt=0.8	0.55	0.03	0.79	0.55	0.69	0.47	0.47	0.04
	dt=1.8	0.86	0.69	1.18	3.89	0.97	1.40	0.77	1.19
	dt=2	0.89	2.88	1.53	13.51	1.38	12.45	0.94	0.56
	dt=3	1.18	3.52	1.58	10.01	1.44	7.36	1.02	4.67
	dt=3.2	1.21	4.52	1.59	10.10	1.51	9.21	1.06	4.65
	dt=4.4	1.59	10.97	2.19	21.86	1.79	13.93	1.42	10.56
	dt=5	1.84	14.55	2.28	21.70	2.08	19.78	2.66	30.17

4.4 定量评估

为了验证本文提出的网络的有效性，本文选取了EV-FlowNet[17]、文献[20]、Spike-FlowNet[22]、LIF-EV-FlowNet[23]、STE-FlowNet[15]、TMA[16]、Fusion-FlowNet[11]和DCEIFlow[13]方法与本文方法进行比对分析。其中前六个方法都仅以事件流作为输入，而Fusion-FlowNet[11]和DCEIFlow[13]则是以图像帧以及对应的事件流数据作为输入。实验结果如表 1所示，其中对于仅使用了事件数据的光流估计方法，表格中只记录了稀疏光流的估计性能，而对于融合图像和事件的光流估计方法，表格记录了它们在稀疏光流和稠密光流估计方面的表现。

在与基于事件的光流估计方法的对比中，在dt=1的情况下，本文方法在三个室内序列上的平均端点误差和异常值百分比普遍要比其它方法更低，说明本文方法在整体性能上要优于其它对比算法。在dt=4的情况下，本文方法同样在三个室内序列上在平均端点误差和异常值百分比这两个指标中表现出了领先的性能，并且领先幅度相较于dt=1的情况更加大，这说明本文方法在大位移运动以及长时间间隔的场景中具有更加明显的性能优势。然而在室外场景中，无论是dt=1还是dt=4的情况下，本文方法的表现要略逊于STE-FlowNet[11]，这是因为室外场景的视觉条件相较于室内场景往往具有强烈的光照变化等更具挑

战性的视觉条件，而本文方法使用了图像数据，难免会受到不利的视觉条件的影响，因此本文方法在室外场景中仍存在一定的应用限制。在与融合图像和事件的光流估计方法的对比中，除了稀疏光流的估计精度外，本文还比对了各个方法的稠密光流估计精度。从结果中可知，相较于Fusion-FlowNet[11]和DCEIFlow[13]，本文在稀疏光流估计和稠密光流估计中，均具有更优的性能，这意味着相较于这两个方法，本文方法具有更强的多模态数据特征编码和融合能力，能够利用更少的信息估计出更加准确且稠密的光流。

此外，本文还与开源的基于图像和事件的光流估计方法DCEIFlow[13]进行多个时间间隔的稀疏光流估计性能对比分析，实验结果如表 2所示。从结果中可以看出，在长时间间隔光流估计上，本文方法在总体上性能要优于DCEIFlow[13]方法。

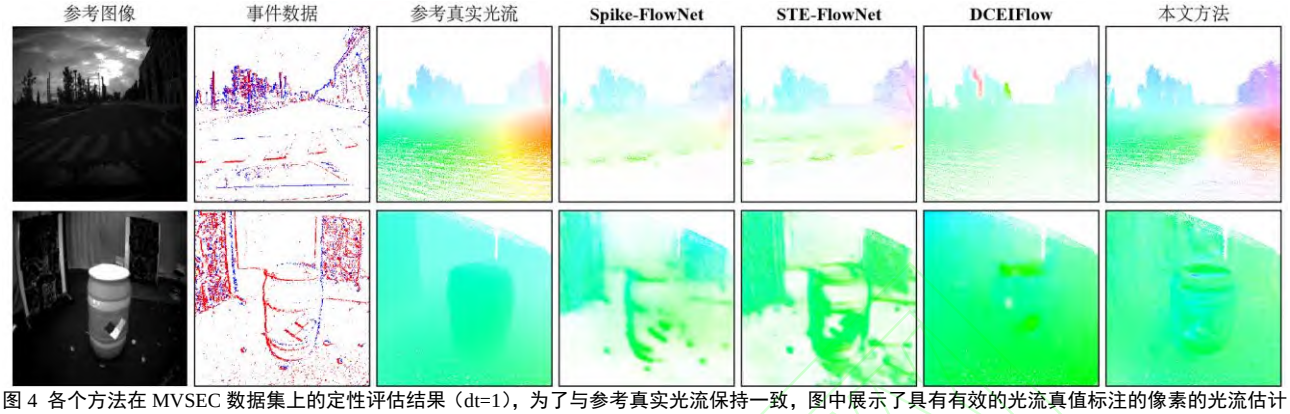


图 4 各个方法在 MVSEC 数据集上的定性评估结果（dt=1），为了与参考真实光流保持一致，图中展示了具有有效的光流真值标注的像素的光流估计结果。上半部分的样本来自 outdoor_day1 序列，下半部分的样本来自 indoor_flying1 序列

Fig.4 Qualitative evaluations on the MVSEC dataset for dt=1 case. To be consistent with ground truth, predicted flow maps demonstrate results only on the pixels with valid flow annotations. The samples are taken from outdoor_day1 (top) and indoor_flying1 (bottom) sequences

4.5 定性评估

图 4展示了各个方法在时间间隔dt=1情况下的光流估计结果可视化对比图。为了方便对比，各个预测光流图都与参考真值保持一致，即仅展示了具有有效的光流真值标注的像素的估计结果。从图中可以看出，基于纯事件数据的光流估计方法Spike-FlowNet[22]和STE-FlowNet[15]很难在没有事件触发的区域产生稠密的光流预测。在室外场景的outdoor_day1序列中，地面部分仅有斑马线边缘等有事件触发的区域存在光流估计结果。同样地，在室内场景的indoor_flying1序列中，也仅有水桶的边缘区域等有事件触发的区域存在较为准确的光流估计结果。相较之下，本文方法和同样基于图像和事件的DCEIFlow[13]方法则能够在没有事件触发的区域同样能够输出稠密的预测光流，但是本文方法所输出的光流更加接近参考真实光流。这说明本文方法相较于基于纯事件的光流估计方法，能够充分利用输入的图像信息，来补充事件数据难以提供的足够的上下文信息，从而输出更加稠密的光流，并且相较于其它基于图像和事件的方法，本文方法更加能够发挥快门相机和事件相机的优势，估计更加准确的光流。

本文方法使用了事件数据，能够估计在时间上具有连续性的光流，因此本文通过图 5展示了各个方法在连续光流估计方面的结果可视化对比图，其中本文展示了整数时间间隔（dt=1和dt=4）和非整数时间间隔（dt=0.8、dt=1.8、dt=3.2和dt=4.4）的光流可视化结果。从图中可以看出，相较于其它方法，本文方法获得的光流与参考真值更加接近，并且光流估计的时间连续性更好，这充分体现了本文方法在连续光流估计上优越性。

4.6 消融实验

为了进一步分析本文提出的网络模型以及训练策略对光流计算性能提升的影响，本文采用消融实验进行综合对比分析。实验采用MVSEC数据集，对各个消融模型在Indoor_flying1、Indoor_flying2、Indoor_flying3和Outdoor_day1四个序列上不同时间间隔（dt=1和dt=4图像帧间隔）的性能进行测试，采用平均端点误差和异常值百分比两个指标进行性能评估。各个消融模型的实验结果如表 3所示。从结果中可知，增加光流递归迭代精化的尺度以及增加递归迭代精化尺度的组合数量，能够有效地提升光流计算的精度，并且在两者的协同作用下，模型的性能提升更为显著，尤其是在dt=4的情况下，这说明多尺度光流迭代更新策略对光流计算精度的提升具有积极作用，并且能够提升模型在面对大位移运动以及长时间间隔的场景时的表现。在特征相似性损失的消融实验方面，选取了若干组权重（0.01、0.1、0.5和1.0）进行对比实验。从实验结果可以看出，过大或者过小的权重值均会对模型的性能造成不良的影响，而基于权重值0.5训练得到的消融模型在大多数情况下表现更好，因此选取了0.5作为模型训练的最佳权重。另外，在特征融合模块中加入梯度中断策略以及使用双向光流训练策略，同样能够有效地提高模型的光流估计性能。

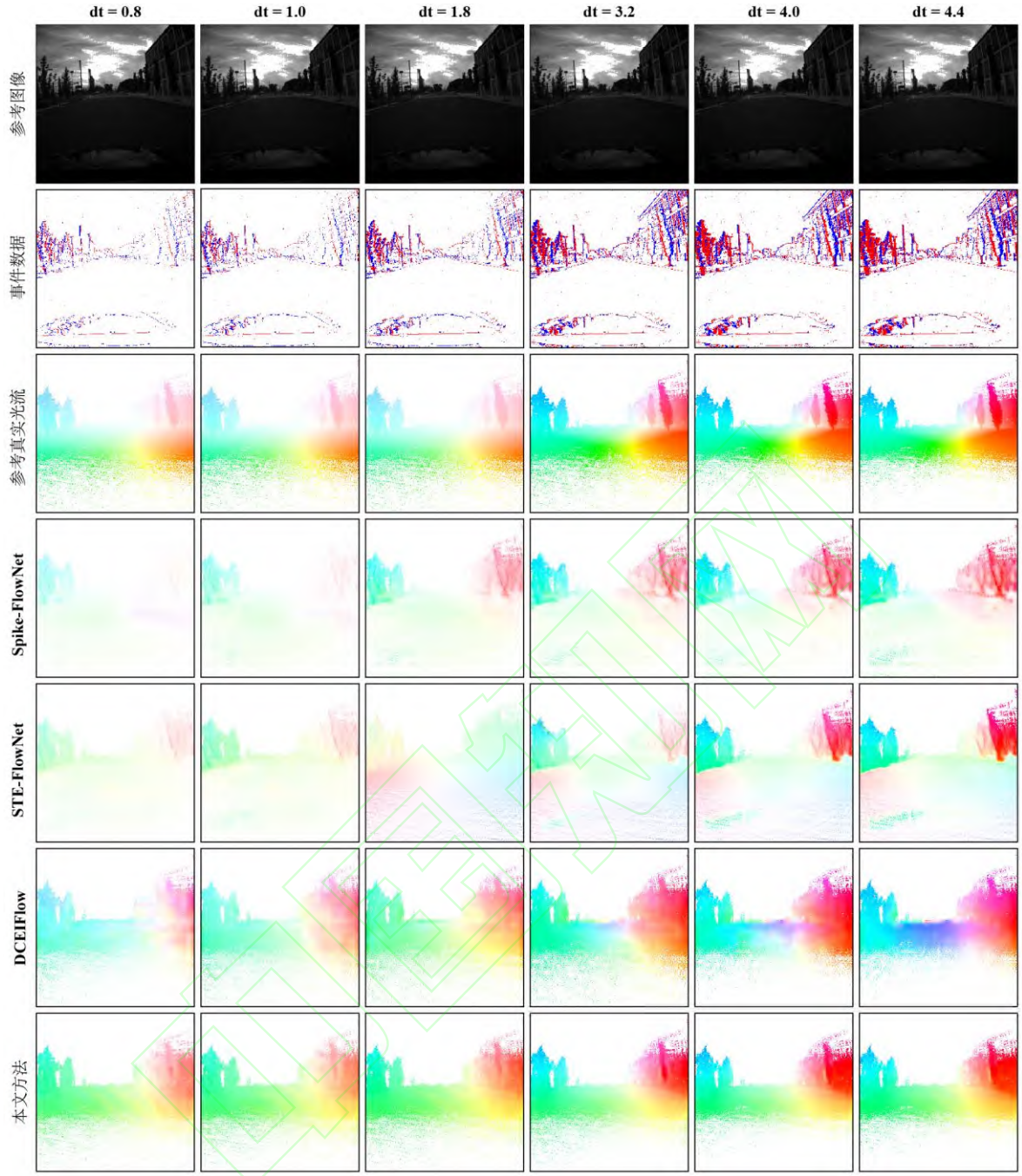


图 5 各个方法在 MVSEC 数据集的 outdoor_day1 序列上的定性评估结果（dt=0.8、dt=1.0、dt=1.8、dt=3.2、dt=4.0 和 dt=4.4 图像帧间隔），其中本文展示了具有有效的光流真值标注的像素的光流估计结果

Fig.5 Qualitative evaluations in the outdoor_day1 sequence of the MVSEC dataset for dt=0.8, dt=1.0, dt=1.8, dt=3.2, dt=4.0 and dt=4.4 case, where we show flow prediction on the pixels with valid flow annotations

为了测试本文方法在实际使用时所需要的计算资源，本文选出表 3中的三个尺度组合变种（1/4、1/8+1/4以及1/16+1/8+1/4），在模型性能、参数量、运行时的显存占用和推理时间方面，与开源的基于图像和事件的方法DCEIFlow[13]进行对比，结果如表 4所示。结果表明，在组合两个尺度以及三个尺度时，本文方法在性能上优于DCEIFlow[13]，但是所需要的显存和推理时间也随之增加。若需要实时运行本文方法，则需要下调组合尺度的数量，或者降低输入数据的空间分辨率。另外，本文方法虽然在参数量方面要高于DCEIFlow[13]，但是在显存占用方面比DCEIFlow[13]更加具有优势。

表 3 消融实验结果

Tab.3 The result of ablation studies

消融项目	设置	Indoor_flying1		Indoor_flying2		Indoor_flying3		Outdoor_day1	
		AEE	Out(%)	AEE	Out(%)	AEE	Out(%)	AEE	Out(%)
以下为时间间隔 dt=1 的实验结果									
尺度组合	1/8	0.71	0.14	0.89	0.98	0.77	0.70	0.78	0.10

	1/4	0.66	0.40	0.84	1.48	0.71	1.03	0.75	0.12
	1/16 + 1/8	0.70	0.28	0.84	0.93	0.74	0.71	0.69	0.27
	1/16 + 1/4	0.58	0.21	0.68	0.73	0.62	0.82	0.58	0.12
	1/8 + 1/4	0.55	0.12	0.66	0.49	0.61	0.67	0.56	**0.06**
	1/16 + 1/8 + 1/4	**0.53**	**0.09**	**0.64**	**0.46**	**0.58**	**0.62**	**0.56**	0.09
梯度停止	不启用	0.59	0.13	0.68	0.56	0.62	0.72	0.60	0.23
	启用	**0.53**	**0.09**	**0.64**	**0.46**	**0.58**	**0.62**	**0.56**	**0.09**
特征相似性损失权重	0.01	0.59	0.10	0.69	0.48	0.62	0.64	0.62	**0.08**
	0.1	0.56	0.17	0.67	**0.42**	0.60	0.64	**0.55**	0.09
	0.5	**0.53**	**0.09**	**0.64**	0.46	**0.58**	**0.62**	0.56	0.09
	1.0	0.59	0.13	0.75	0.66	0.65	0.70	0.61	0.24
双向训练	不启用	0.69	0.19	0.77	0.74	0.74	0.59	0.70	0.29
	启用	**0.53**	**0.09**	**0.64**	**0.46**	**0.58**	**0.62**	**0.56**	**0.09**
动态损失过滤机制	不启用	0.53	0.09	**0.64**	0.46	**0.58**	0.62	0.56	0.09
	启用	**0.50**	**0.06**	0.65	**0.40**	0.61	**0.57**	**0.50**	**0.04**
以下为时间间隔 dt=4 的实验结果									
尺度组合	1/8	2.71	31.50	4.40	43.37	3.26	33.52	1.97	18.10
	1/4	2.62	22.63	4.80	37.86	3.40	26.12	1.73	14.15
	1/16 + 1/8	2.18	21.25	3.61	34.13	2.68	24.93	2.05	16.54
	1/16 + 1/4	2.08	16.65	3.76	31.71	2.65	19.83	1.47	11.37
	1/8 + 1/4	1.76	12.08	3.10	25.56	2.20	16.63	1.50	11.44
	1/16 + 1/8 + 1/4	**1.58**	**9.43**	**2.24**	**18.82**	**2.01**	**14.86**	**1.46**	**11.32**
梯度停止	不启用	1.80	11.80	2.74	22.09	2.23	16.98	1.52	12.11
	启用	**1.58**	**9.43**	**2.24**	**18.82**	**2.01**	**14.86**	**1.46**	**11.32**
特征相似性损失权重	0.01	1.76	12.09	2.97	24.26	2.26	17.58	1.54	13.09
	0.1	1.68	11.36	2.92	24.23	2.19	16.51	**1.44**	**11.09**
	0.5	**1.58**	**9.43**	**2.24**	**18.82**	**2.01**	**14.86**	1.46	11.32
	1.0	1.76	11.98	2.78	23.27	2.23	17.36	1.54	12.20
双向训练	不启用	1.75	13.09	2.35	19.89	2.11	17.28	2.09	22.03
	启用	**1.58**	**9.43**	**2.24**	**18.82**	**2.01**	**14.86**	**1.46**	**11.32**
动态损失过滤机制	不启用	1.58	9.43	2.24	18.82	2.01	14.86	1.46	11.32
	启用	**1.51**	**8.64**	**2.01**	**17.16**	**1.78**	**13.21**	**1.34**	**9.90**

备注：“设置”栏中带有下划线的设置表示本文提出的网络进行实验所使用的默认设置，并且在“尺度组合”、“梯度停止”、“特征相似性损失权重”、“双向训练”这四项消融实验中，并没有启用“动态损失过滤机制”

表 4 本文方法与 DCEIFlow 在性能、参数量、运行时的显存占用和推理时间方面的对比结果

Tab.4 Comparison of performance, parameters, memory and inference time between proposed method and DCEIFlow

方法	Indoor_flying1		参数量	显存占用	平均推理时间
	AEE	Out(%)			
DCEIFlow[13]	0.57	0.30	**7.06M**	878MB	**22ms**
本文方法(1/4)	0.66	0.40	13.21M	**688MB**	31ms
本文方法(1/8 + 1/4)	0.55	0.12	13.21M	694MB	52ms
本文方法(1/16 + 1/8 + 1/4)	**0.53**	**0.09**	13.21M	852MB	74ms

5 结语

为了获得稠密且连续的光流，并实现长时间间隔光流估计，本文提出了一种基于图像和事件的多模态多尺度递归光流估计网络，它以单个图像和对应事件流作为输入，以从粗糙到精细和递归迭代精化的方式估计光流。为了避免使用标签数据并提高网络的模型性能，本文采用无监督学习的方式对网络进行训练，同时设计了动态损失过滤机制，通过过滤不可靠的监督

梯度信号来对网络进行更加有效的训练。本文在MVSEC数据集上进行了一系列实验，实验结果表明本文方法具有更高的光流估计精度，同时能够实现稠密且连续的光流估计，并且在估计长时间间隔光流方面具有更加突出的优势。然而本文方法的参数量偏大，并且运行速度较慢，应用时需要根据实际情况和需求进行精度和速度之间的平衡调整。针对这些问题，未来的研究重点是构建运行效率更高且更加轻量化的光流估计网络。

参考文献

[1] Tu Z, Xie W, Zhang D, et al. A survey of variational and CNN-based optical flow techniques[J]. Signal Processing: Image Communication, 2019, 72: 9-24.

[2] de Croon G C H E, De Wagter C, Seidl T. Enhancing optical-flow-based control by learning visual appearance cues for flying robots[J]. Nature Machine Intelligence, 2021, 3(1): 33-41.

[3] Wang C, Ji T, Nguyen T M, et al. Correlation flow: robust optical flow using kernel cross-correlators[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018: 836-841.

[4] Ho H W, De Wagter C, Remes B D W, et al. Optical flow for self-supervised learning of obstacle appearance[C]//2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2015: 3098-3104.

[5] Wang H, Cai P, Sun Y, et al. Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 13731-13737.

[6] Huang Y, Zhao B, Gao C, et al. Learning optical flow with R-CNN for visual odometry[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 14410-14416.

[7] HU Xuemin, ZHENG Hong, GUO Lin, XIONG Raorao. Crowd Motion Estimation Using a Fisheye Camera[J]. Geomatics and Information Science of Wuhan University, 2017, 42(4): 537-542. (胡学敏, 郑宏, 郭琳, 熊饶饶. 利用鱼眼相机对人群进行运动估计[J]. 武汉大学学报(信息科学版), 2017, 42(4): 537-542.)

[8] FU Jingyi, YU Lei, YANG Wen, LU Xin. Event-based Continuous Optical Flow Estimation. Acta Automatica Sinica, 2021, 47. (付婧祎, 余磊, 杨文, 卢昕. 基于事件相机的连续光流估计. 自动化学报, 2021, 47.)

[9] Bardow P, Davison A J, Leutenegger S. Simultaneous optical flow and intensity estimation from an event camera[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 884-892.

[10] Pan L, Liu M, Hartley R. Single image optical flow estimation with an event camera[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 1669-1678.

[11] Lee C, Kosta A K, Roy K. Fusion-FlowNet: Energy-efficient optical flow estimation using sensor fusion and deep fused spiking-analog network architectures[C]//2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022: 6504-6510.

[12] Gallego G, Delbrück T, Orchard G, et al. Event-based vision: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 44(1): 154-180.

[13] Wan Z, Dai Y, Mao Y. Learning dense and continuous optical flow from an event camera[J]. IEEE Transactions on Image Processing, 2022, 31: 7237-7251.

[14] Gehrig M, Millhäusler M, Gehrig D, et al. E-raft: Dense optical flow from event cameras[C]//2021 International Conference on 3D Vision (3DV). IEEE, 2021: 197-206.

[15] Ding Z, Zhao R, Zhang J, et al. Spatio-temporal recurrent networks for event-based optical flow estimation[C]//Proceedings of the AAAI conference on artificial intelligence. 2022, 36(1): 525-533.

[16] Liu H, Chen G, Qu S, et al. TMA: Temporal Motion Aggregation for Event-based Optical Flow[J]. arXiv preprint arXiv:2303.11629, 2023.

[17] Zhu A Z, Yuan L, Chaney K, et al. EV-FlowNet: Self-supervised optical flow estimation for event-based cameras[J]. arXiv preprint arXiv:1802.06898, 2018.

[18] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision[C]//IJCAI'81: 7th international joint conference on Artificial intelligence. 1981, 2: 674-679.

[19] Brandli C, Berner R, Yang M, et al. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor[J]. IEEE Journal of Solid-State Circuits, 2014, 49(10): 2333-2341.

[20] Zhu A Z, Yuan L, Chaney K, et al. Unsupervised event-based learning of optical flow, depth, and egomotion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 989-997.

[21] Amir A, Taba B, Berg D, et al. A low power, fully event-based gesture recognition system[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7243-7252.

[22] Lee C, Kosta A K, Zhu A Z, et al. Spike-flownet: event-based optical flow estimation with energy-efficient hybrid neural networks[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 366-382.

[23] Hagenaars J, Paredes-Vallés F, De Croon G. Self-supervised learning of event-based optical flow with spiking neural networks[J]. Advances in Neural Information Processing Systems, 2021, 34: 7167-7179.

[24] CHEN Zhen, ZHANG Daowen, ZHANG Congxuan, WANG Yang. Sparse-to-dense large displacement motion optical flow estimation based on deep matching. Acta Automatica Sinica, 2022, 48(9): 2316−2326. (陈震, 张道文, 张聪炫, 汪洋. 基于深度匹配的由稀疏到稠密大位移运动光流估计. 自动化学报, 2022, 48(9): 2316−2326.)

[25] Wang Y, Yang Y, Yang Z, et al. Occlusion aware unsupervised learning of optical flow[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4884-4893.

[26] Im W, Kim T K, Yoon S E. Unsupervised learning of optical flow with deep feature similarity[C]//Computer Vision

–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. Springer International Publishing, 2020: 172-188.
[27] Zhu A Z, Thakur D, Özaslan T, et al. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception[J]. IEEE Robotics and Automation Letters, 2018, 3(3): 2032-2039.
[28] Stoffregen T, Scheerlinck C, Scaramuzza D, et al. Reducing the sim-to-real gap for event cameras[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. Springer International Publishing, 2020: 534-549.
[29] Teed Z, Deng J. Raft: Recurrent all-pairs field transforms for optical flow[C]//European conference on computer vision. Springer, Cham, 2020: 402-419.
[30] Meister S, Hur J, Roth S. Unflow: Unsupervised learning of optical flow with a bidirectional census loss[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).
[31] Jiang S, Lu Y, Li H, et al. Learning optical flow from a few matches[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 16592-16600.
[32] Dosovitskiy A, Fischer P, Ilg E, et al. Flownet: Learning optical flow with convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2758-2766.
[33] Ren Z, Yan J, Ni B, et al. Unsupervised deep learning for optical flow estimation[C]//Proceedings of the AAAI conference on artificial intelligence. 2017, 31(1).
[34] Han B, Yao Q, Yu X, et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels[J]. Advances in neural information processing systems, 2018, 31.
[35] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library[J]. Advances in neural information processing systems, 2019, 32.
[36] Loshchilov I, Hutter F. Decoupled weight decay regularization[J]. arXiv preprint arXiv:1711.05101, 2017.

网络首发：
标题：基于图像和事件的无监督学习稠密连续光流估计
作者：胡建朗，郭迟，罗亚荣
收稿日期：2024-05-22
DOI:10.13203/j.whugis20230390

引用格式：
胡建朗，郭迟，罗亚荣. 基于图像和事件的无监督学习稠密连续光流估计[J].武汉大学学报(信息科学版),2024,DOI: 10.13203/j.whugis20230390（HU Jianlang，GUO Chi，LUO Yarong. Unsupervised Dense and Continuous Optical Flow Estimation Based on Image and Event Data[J].Geomatics and Information Science of Wuhan University,2024,DOI: 10.13203/j.whugis20230390）

网络首发文章内容和格式与正式出版会有细微差别，请以正式出版文件为准！
