EV-LayerSegNet：使用事件相机的自监督运动分割

Youssef Farah[1] Federico Paredes-Vallés[2] Guido C.H.E. De Croon[2]
Muhammad Ahmed Humais[1] Hussain Sajwani[1] Yahya Zweiri[1]
[1]Advanced Research and Innovation Center, Khalifa University [2]MAVLab, TU Delft
[1] { youssef.farah, 100061899, hussain.sajwani, yahya.zweiri } @ku.ac.ae
[2]fedeparedesv@gmail.com, [2]G.C.H.E.deCroon@tudelft.nl

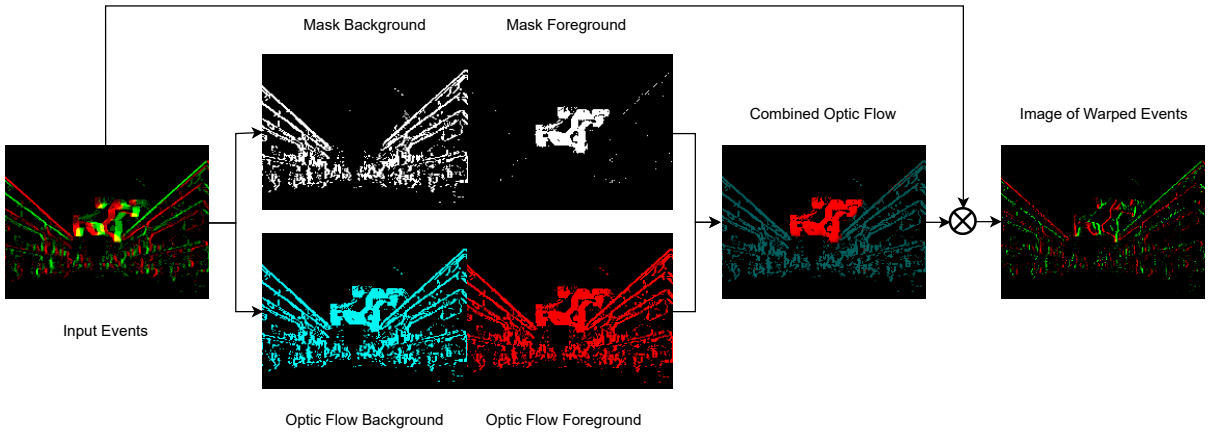

Figure 1. Our method takes as input an event volume that results in a blurry scene. It then attempts to generate two different masks to differentiate the background from the foreground. Next, it estimates the affine optical flow for both models, and combines the flow together using the masks. Finally, it warps the events according to the combined flow. A successful motion deblur leads to accurate segmentation.

Abstract

事件相机是一种新型的仿生传感器，与传统相机相比，它以更高的时间分辨率捕捉运动动态，因为像素对亮度变化做出异步反应。因此，它们更适合于运动分割等涉及运动的任务。然而，训练基于事件的网络仍然是一个困难的挑战，因为获取真实的标注代价高昂、容易出错并且频率有限。自我监督是发掘事件相机在运动分割上真正潜力的关键，但目前尚无存在这样的基于学习的方法。当前文献中的无监督方法利用事件特性并使用期望最大化来联合估计正确的聚类和相关的运动模型，而没有任何学习过程，从而阻碍了在真实世界中的实际应用。为此，本文介绍了 EV-LayerSegNet，这是一种基于事件的运动分割自我监督 CNN。受场景动态分层表示的启发，我们展示了可以分别学习仿射光流和分割掩码，并用它们来去模糊输入事件。然后测量去模糊质量并将其用作自我监督的学习损失。我们在只有仿射运动的模拟数据集上训练和测试该网络，分别实现了高达 71% 和 87% 的 IoU 和检测率。

1. 介绍

事件相机，例如动态视觉传感器（DVS），是一种新型的仿生传感器，通过检测亮度变化来感知运动，而不是在时间间隔内捕获图像中的强度变化 [5] 。换句话说，如果在时间 t ，像素坐标 x 和 y 的亮度变化超过了阈值 C ，一个事件就会通过记录亮度增加或减少，以及像素位置和事件产生的时间来生成。相比之下，传统相机以指定的帧速率记录所有像素中的色彩强度。

这种获取视觉信息的根本性转变带来了几个优点。高时间分辨率和低延迟（约为 μs ）使事件相机非常适合记录非常快速移动的场景，而不会像基于帧的相机那样遭受运动模糊的困扰。高动态范围（HDR）使相机能够在非常暗或非常亮的场景中捕捉细节，使其适合用于光线具有挑战性的环境，如海底、夜间驾驶或烟花表演显示 [34] 。

这些优势的结合开启了执行帧式相机无法实现的任务的可能性，例如低延迟光流估计、高速控制和跟踪以及同步定位与地图构建（SLAM） [5] 。然而，需要开发新一代的基于事件的视觉算法，以便事件相机能够展现其全部潜力，因为事件的数据结构与基于图像的

传统算法不兼容。

此外，由于缺乏具有微秒分辨率和 HDR 的真实世界数据集的真实值数据，带来了额外的挑战。在基于图像的数据集（如 COCO [15]，BD100K [32]）中，真实值通常是通过人类手动注释每一帧获得的。这个过程对于基于图像的数据集已经很昂贵且繁琐，而对于稀疏且低延迟的事件数据则变得不可行。为了避免人工注释，传统的基于事件的数据集如 DSEC [7] 和 MVSEC [36] 使用了额外的传感器和摄像机来生成真实值。然而，传感器和摄像机受限于其自然的视场（FOV）、空间和时间分辨率。因此，开发能够依靠事件的本质完成任务的算法，而不使用收集成本高昂且带有额外传感器误差测量的真实值，是极其重要的。

在运动估计及其在其他任务如监视、跟踪或障碍物规避中的应用中，将场景分割为独立运动的对象是基本的，并且通常被称为运动分割 [24]。尽管使用基于帧的相机取得了很大进展，但后者并不完全适合涉及运动的任务，因为它们受运动模糊的影响，并且即使在没有发生运动时也会跟踪所有像素值 [24]。在恒定的照明条件下，事件摄像机非常适合此任务，因为事件以与场景动态完全相同的速率采样，并且获取的信息仅与摄像机和场景中物体的运动相关。基于事件的运动分割可以被广泛地分类为基于模型的方法 [19, 24, 35] 和基于学习的方法 [17, 18, 22]。基于模型的方法利用事件数据的特性，以迭代的方式聚类由独立运动物体（IMO）生成的事件，因此它们是无监督的，不使用真实标签。然而，这种迭代方法需要较长的计算时间，限制了其在真实世界中的适用性。另一方面，基于学习的方法应用深度学习技术，同时也依赖于事件的特性。在某些情况下，它们结合其他任务如跟踪和光流进行运动分割，但它们在训练期间都需要真实标签或预训练权重。据我们所知，目前尚不存在使用无监督学习的基于学习的方法，即直接从输入事件中学习，而不是依赖于来自监督方法的真实标签或预训练权重。放眼事件领域之外，无监督图像分割已经取得了显著进展。这些方法将场景细分为前景中的主要移动物体，其余则为背景，学习则由图像或光流重建驱动。在大多数情况下，它们经常使用分层表示 [26]。像素根据运动相似性被分为不同层，并根据相关的运动进行移动，然后合成在一起以重建序列中的下一个图像。最近，Shrestha et al. [23] 将此过程整合到一个端到端可微分的 CNN 管道中，通过使用两个连续帧来分割视频中主要的移动物体。受他们工作的启发，我们提出了一种基于事件的无监督卷积神经网络用于运动分割。具体来说，我们将来自 [23] 的 CNN 分割方法转移到事件域，并将其与来自 [38] 的编码器-解码器结构结合。我们表明，在仿射运动和恒定亮度的假设下，可以使用对比度最大化损失来分割独立运动的物体。

我们将我们的主要贡献识别如下：

1. 我们提出了一种新颖的自监督网络，用于通过引入一个新颖的光流模块来学习基于事件的运动分割，该模块使仿射光流的自监督学习成为可能，并且有一个分割模块可单独学习与独立移动的物体相对应的掩码。
2. 我们贡献了一个新的基于事件的数据集，其中包含几个根据仿射运动移动的模拟背景和对象。
3. 我们通过实验展示了，与当前最先进的无监督运动分割方法相比，我们的方法表现出更优异的性能。

在计算机视觉中，运动分割被定义为获取运动物体形状的任务。图像差分，如 [33]，尝试通过找到视频帧间像素的强度差异来检测运动物体。基于层的技术 [6, 11] 根据统一运动的数量将帧分成多层。统计方法如期望最大化（EM）是最常用的方法之一。深度学习和光流估计的最新发展使这一运动任务有了显著改进，通常与统计和基于层的技术结合使用。然而，运动分割的最新方法主要的限制是依赖于真实标签，这限制了这些方法在注释数据集之外的场景的适用性。在这方面，开发自监督方法的尝试已经形成了在帧域中开发的框架，而对于事件相机的方法尚未出现。

在文献中，许多方法被认为是无监督或自监督的，但这仅限于推理和测试期间。实际上，它们的部分结构（即网络、掩码）是基于真实值预训练的。例如，COSNet 使用共同注意力来捕捉视频帧之间的丰富相关性，但掩码是基于真实值预训练的。类似地，Ye 进行利用全局精灵的运动分割，但它也需要预先计算的掩码。Li 提出使用实例嵌入根据运动显著性和对象性来查找运动物体。然而，密集的嵌入是通过一个在静态图像上预训练的网络进行的实例分割获得的。相比之下，我们只将那些在训练和推理过程中不需要任何类型的真实值，并且不依赖于有监督方法的预训练权重的方法视为自监督方法。早期的自监督或无监督学习的运动分割尝试始于 ConvNet。网络在单帧中学习高级特征如下。一组带有光流图的单帧传递给网络。然后，它通过将具有相似方向和速率的光流向量关联起来生成分割掩码。该网络使用这些掩码作为伪真实值，仅通过查看静态帧来尝试重现掩码。

Yang et al. [28] 提出了一种网络，该网络通过使用预先计算的光流作为输入来执行前景/背景运动分割。槽注意机制随后用于将具有视觉同质性的像素组合在一起，但结果对光流估计的准确性非常敏感。而在 [23, 28] 中，模型尝试通过使用对象自身的光流来分割移动对象时，Yang et al. [30] 提出使用生成模块（G）和修补模块（I）网络。生成模块对输入光流应用掩码，以隐藏前景的光流为目标。接下来，生成模块将被遮罩的光流传递给修补模块，其任务是重建被生成模块隐藏的光流。当光流重建效果不佳时，实现了准确的运动分割。它在公共数据集上取得了良好的结果，然而模型未能捕捉完整的对象或区分背景中的区域。这是因为它仅使用单尺度的时间信息，同时引入了摄像机运动的偏差。为了解决这个问题，Yang et al. [29] 使用 MASNet 扩展了这项工作，扩展了网络管道，添加了更多生成器和修补模块以处理多个输入光流图。最后，Shrestha et al. [23] 提出了 LayerSegNet，结合了场景的层次表示和仿射光流估计。网络以两张图像为输入，旨在分别估计运动模型（前景和背景）和掩码。随后，他们将掩码应用于第一幅图像，并根据相关的运动模

型扭曲这些掩码。然后，他们将这两个掩码组合在一起生成第二幅图像，并使用生成的第二幅图像与实际第二幅图像之间的差异作为学习参数。

1.1. 基于事件的运动分割

基于事件的运动分割是一个非常新的领域，试图利用事件摄像机的优异特性。

Stoffregen et al. [24] 使用期望最大化方法，并提出通过同时将事件分组到不同的簇中和估计与每个簇相关联的运动模型来分割移动物体的想法。然后，根据相关的运动模型对事件进行变形，生成变形事件图像(IWE)。接下来使用迭代过程来找到合适的运动模型和簇，以便最大化目标函数。该模型被证明对迭代过程中指定的簇数不敏感，并在真实世界数据集上表现良好，但无法估计场景中移动物体的数量。

Zhou et al. [35] 提出了通过引入两个空间正则化器来解决这个问题，这些正则化器最小化了簇的数量。其基本概念与 [24] 相同，但输入事件首先在时空事件图割中初始化，然后通过在目标函数中添加能量项来平滑锐化簇。

始终基于事件聚类，Chethan et al. [19] 将场景分解为多个运动并进行合并，同时支持特征跟踪。

尽管这些方法产生了有前景的结果，但由于迭代过程花费了大量时间和计算资源，使得它们不适合在现实世界中应用，从而阻碍了在无人机等移动平台上的潜在应用。因此，需要开发一种基于学习的方法。

维桑基特 et al. [22] 提出了一种名为 EVDodgeNet 的基于学习的流程，该流程同时解决运动分割、光流和三维运动，但是依赖于真实的遮罩。相反，米特罗欣 et al. [17] 使用图卷积网络从事件的三维表示中学习运动分割。而车坦 et al. [18] 提出了 SpikeMS，这是第一个基于事件的尖峰神经网络（SNN）用于运动分割。最近，乔尔吉斯 et al. [8] 首先通过自运动补偿事件，然后使用注意力模块以实现时间一致性来完成该任务。阿尔肯迪 et al. [1] 则使用图变换神经网络来去噪场景并执行分割。

尽管有这些努力，上述方法依然依赖于真实数据，导致自监督方法的可用性上存在差距。在这方面，Wang et al. [27] 从几何约束中生成伪标签，而 Arja et al. [2] 提出了使用自监督视觉变换器的方案。这两项工作都依赖于有监督网络的预训练权重。

受 LayerSegNet [23] 启发，我们提出了一种端到端的 CNN 架构，该架构通过联合估计仿射光流和分割掩模来学习运动分割，使用对比度最大化作为学习损失，并且不使用任何预训练权重。

2. 方法

分割任务的方法受到 [23] 的启发。然而，考虑到事件数据的独特性质，需要一种不同的输入方法。为此，我们使用来自 [38] 的输入事件表示。然后，我们联合估计分割掩码和仿射光流图。接下来，我们将分割掩码应用于相关的光流，并将这些图合并以获得单一的光流图。然后，我们使用该图对事件进行时间上的前向和后向变形，并按照 [9] 中的方法应用自监督损失。输入事件图像越被不清晰化，分割掩码的准确性就越高。

2.1. 输入事件表示

为给定任务选择正确的事件表示仍然是一个具有挑战性的问题。基于事件的方法可以处理单个事件 $e_k \doteq (x_k, t_k, p_k)$，但仅靠事件本身携带的信息非常有限，并且容易受到噪声的影响。通常更倾向于处理一组事件 $\mathcal{E} \doteq \{e_k\}_{k=1}^{N_e}$，以获得足够的信噪比，同时也为给定任务携带更多信息。最常见的方法要么是在不同的事件计数帧中离散化这一组事件 [12, 13, 20, 25, 38]，要么是每像素的平均/最近的事件时间戳 [14, 31, 37]。

由于我们开发了一个非递归的 CNN 流程，时间信息需要在输入中进行编码，因此我们采用来自 [38] 的事件表示。给定 N 个输入事件和 B 个 bin，我们首先将事件时间戳缩放到 $[0, B-1]$ 的范围内，并生成事件体积：

$$t_i^* = (B-1)\,(t_i - t_0)\,/\,(t_N - t_1) \tag{1}$$

$$V(x,y,t) = \sum_i p_i k_b\,(x - x_i)\,k_b\,(y - y_i)\,k_b\,(t - t_i^*) \tag{2}$$

$$k_b(a) = \max(0, 1 - |a|) \tag{3}$$

为了执行二维卷积，时间域被视为传统二维图像中的一个通道。

2.2. 运动模型

在 [23] 中，作者提出使用两层表示来对场景中的主要移动物体进行分割。前景层捕捉主要物体，而背景层则表示场景的其余部分。

在我们的情况中，假设光照是恒定的，事件是由相机和场景之间的相对运动产生的。由于事件是由亮度变化触发的，这些变化是由移动的边缘产生的。因此，如果相机在移动而场景是静止的，则事件是由相机的运动产生的。当物体在场景中也在运动时，事件的产生是由于运动物体的边缘以及由于相机自运动（egomotion）造成的静止物体的边缘。

让我们假设一个场景，其中物体和背景在明显移动。我们还假设运动是仿射变换，因此场景可能会经历平移、旋转、剪切和缩放。然后这两种运动可以用两个运动模型 A_1 和 A_2 来描述：

$$W_i(x,y) = \boldsymbol{A}_i \begin{bmatrix} 1 \\ x \\ y \end{bmatrix} = \begin{bmatrix} a_i^1 & a_i^2 & a_i^3 \\ a_i^4 & a_i^5 & a_i^6 \end{bmatrix} \begin{bmatrix} 1 \\ x \\ y \end{bmatrix}, \quad i \in \{1,2\} \tag{4}$$

其中 W_i 表示与仿射运动矩阵 A_i 相对应的密集流图。

对应于两个仿射运动，我们分别生成两个 alpha 蒙版，它们代表运动的物体。理想情况下，分割蒙版对于每个像素来说应该是一个 one-hot 向量，将像素分配给相应的物体。然而，这种操作不可微，阻碍了训练期间

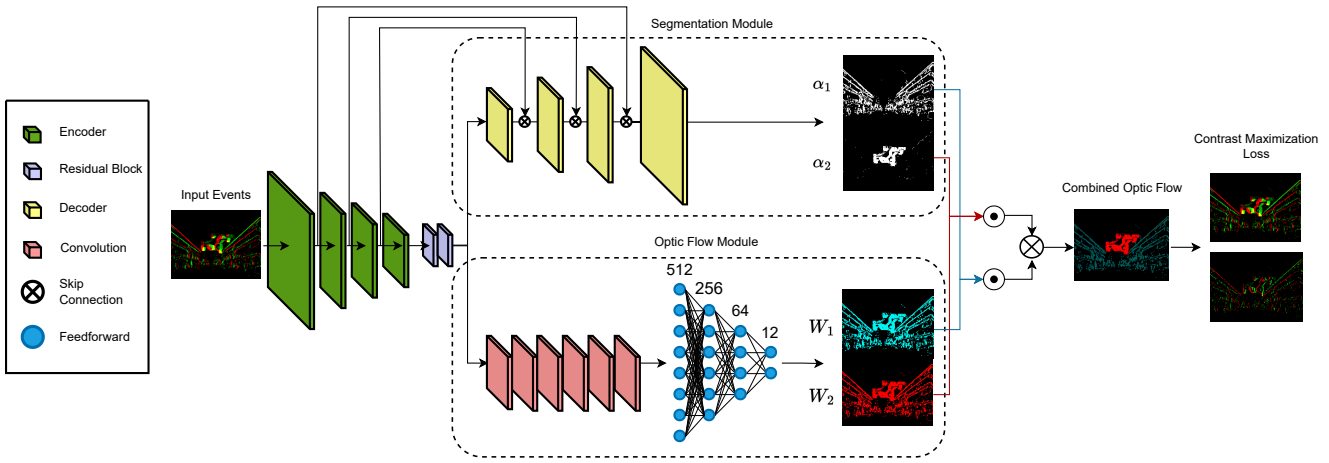

Figure 2. EV-LayerSegNet 架构。事件通过 4 个编码层进行降采样，然后传递到 2 个残差块。残差块的输出随后被传递到分割和光流模块。分割模块通过 4 个解码层进行上采样，并通过跳过连接与编码器连接。光流模块由 6 个卷积层和一个包含 4 层的前馈网络组成。

梯度的反向传播。我们通过为每个像素分配两个任意数值来克服这个问题，其中每个数值对应一个图层。然后我们应用 softmax 以确保这些数值在 [0,1] 范围内，并且这些数值的总和为 1。对于每个像素，我们保留两个图层中的最大值，并将另一个值设置为零。这种被称为 maxout 的操作使得分类可微分。它还确保一个像素不能同时属于两个物体。

接下来，我们通过将仿射流映射与对应的 alpha 映射进行元素级乘法来计算组合光流：

$$W_{\text{comb}} = \alpha_1 \odot W_1 + \alpha_2 \odot W_2 \tag{5}$$

Maxout 操作确保每个像素仅会有一个运动模型的流量分量，并根据与该像素关联的 alpha 图的相应值进行缩放。在训练过程中，我们期望网络能够自适应于流计算的缩放效果。

2.3. 通过对比度最大化进行自监督

自监督学习通过应用对比最大化来实现 [4]。正如在 [9] 中所描述的，由同一移动边缘生成的事件编码了精确的光流。由于这些事件在输入分区中存在错位（运动模糊），可以使用每像素光流 $\boldsymbol{u}(\boldsymbol{x}) = (u(\boldsymbol{x}), v(\boldsymbol{x}))^T$ 将事件传播到参考时间 t_{ref} 以重新对齐它们，并显示生成这些事件的初始边缘：

$$\boldsymbol{x}'_i = \boldsymbol{x}_i + (t_{\text{ref}} - t_i)\,\boldsymbol{u}(\boldsymbol{x}_i) \tag{6}$$

用来衡量去模糊质量的度量是扭曲事件图像 [16, 38] 的每像素和每极性平均时间戳。损失越低，去模糊效果越好，这也意味着估计的光流和 alpha 图更加准确。

我们最初使用双线性插值为每个极性生成每个像素处平均时间戳的图像，如 [9] 所示：

$$T_{p'}(\boldsymbol{x}; \boldsymbol{u} \mid t_{\text{ref}}) = \frac{\sum_j \kappa(x - x'_j)\kappa(y - y'_j)t_j}{\sum_j \kappa(x - x'_j)\kappa(y - y'_j) + \epsilon}$$
$$j = \{i \mid p_i = p'\}, \quad p' \in \{+, -\}, \quad \epsilon \approx 0 \tag{7}$$

损失是时间图像的平方和。这也通过具有至少一个事件的像素和进行缩放，以防止网络将具有大时间戳的事件保持在图像空间之外，使得它们不会对损失函数产生贡献。

$$\mathcal{L}_{\text{contrast}}(t_{\text{ref}}) = \frac{\sum_{\boldsymbol{x}} T_{\pm}(\boldsymbol{x}; \boldsymbol{u} \mid t_{\text{ref}})^2}{\sum_{\boldsymbol{x}} [n(\boldsymbol{x}') > 0] + \epsilon} \tag{8}$$

为了防止在反向传播期间出现时间尺度问题，扭曲过程在正向（t_{ref}^{fw}）和反向（t_{ref}^{bw}）都被执行。然后，总损失是反向和正向扭曲损失的总和，λ 是一个用来平衡 Charbonnier 平滑先验 $\mathcal{L}_{\text{smooth}}$ [3] 的标量。

$$\mathcal{L}_{\text{contrast}} = \mathcal{L}_{\text{contrast}}\left(t_{\text{ref}}^{\text{fw}}\right) + \mathcal{L}_{\text{contrast}}\left(t_{\text{ref}}^{\text{bw}}\right) \tag{9}$$

$$\mathcal{L}_{\text{flow}} = \mathcal{L}_{\text{contrast}} + \lambda \mathcal{L}_{\text{smooth}} \tag{10}$$

注意，只有当输入事件分区中有足够的模糊时，自监督损失才成为一个强有力的监督信号。检查输入分区中是否包含足够的事件以形成线性模糊是至关重要的。

3. 网络实现

我们的网络 EV-LayerSegNet 类似于编码器-解码器架构，并且受到 [23, 38] 的启发。

如同在 [38] 中一样，事件经过 4 个编码器层进行降采样，并传递到 2 个残差块。然后我们将残差块的输出进行堆叠，并传递到分割模块和光流模块。

3.1. 光流模块

光流模块包含 6 个卷积层，每个卷积层后跟一个 leaky ReLU 激活。然后我们将最后一个卷积层的输出展平，并将其传递给一个由 4 层（512, 256, 64 和 12 个输出单元）组成的前馈网络，网络中除最后一层外都使用 tanh 激活。然后我们使用前馈网络的输出并将其分成两个包含 6 个仿射运动参数的集合，接着我们计算两个流图 W_1 和 W_2 。

在分割部分，残差块的输出通过 4 个解码层进行双线性上采样。每个解码层通过跳跃连接与相应的编码层连接。我们应用了一个漏型双整流 ReLU 激活（漏型 DoReLU），这保证了分割性能的提高：

$$y(x) = \begin{cases} 1 + \frac{x-1}{\gamma} & , x > 1 \\ x & , \text{ if } 0 \leq x \leq 1 \\ \frac{x}{\gamma} & , x < 0 \end{cases} \tag{11}$$

在此模块中，每个解码层之后都跟随使用漏斗型 DoReLU 激活，其中 $\gamma = 100$。在最后一层，使用 softmax 以确保通道值被限制在 [0,1] 范围内，并且总和为 1。

4. 实验

4.1. 训练详情

我们在 Pytorch 中实现了我们的流程。我们使用 Adam 优化器 [10] 和学习率 $1 \cdot 10^{-5}$ 。在训练时，我们使用了 8 的批量大小，并训练了 400 个周期。输入由 N = 200,000 个事件组成，这是获得场景模糊的足够事件数。Charbonnier 损失由权重 $\lambda = 0.001$ 平衡。

4.2. 数据集

由于缺乏具有仿射运动的物体的大规模数据集，我们在使用 ESIM [21] 生成的数据集上训练了 EV-LayerSegNet。具体地说，我们利用 Multiple 2D Objects 渲染引擎来生成模拟事件。我们使用与 DVS 相机类似的设置，分辨率为 640x480 像素，阈值 C = 0.5。对于训练集，我们生成了 1000 个时长为 1 秒的序列，使用了 10 张背景图像和 10 个前景物体。同样，测试集由 25 个时长为 1 秒的序列组成，包含五张背景图像和五张前景图像。通过对光流设定阈值来生成真实标签，因为前景物体的移动速度明显快于背景图像。

4.3. 评估指标

我们使用交并比（IoU）和检测率（DR）作为定量评估的标准指标 [35] 。IoU 的公式如 Eq. (12) 所示，其中 S_D 代表预测的掩码，S_G 代表真实值掩码：

$$\text{IoU} = \frac{S_D \cap S_G}{S_D \cup S_G} \tag{12}$$

当满足以下条件时，检测率返回值为 1，其中 B_D 和 B_G 分别是预测掩码和真实值的边界框：

$$B_D \cap B_G > 0.5 \quad \text{and} \quad (B_D \cap B_G) > (S_D \cap \overline{B_G}) \tag{13}$$

5. 结果

5.1. 定性评价

我们在 Fig. 3 中提供了一个定性示例。该网络有效地消除了场景中的模糊，该场景由一只鸟组成，该鸟以高于相机的速度且与相机同方向飞行，并通过光流图的色彩进行编码。这表明我们的方法同样能够在背景和前景可能显现出类似运动的复杂场景中进行分割，这可能会导致网络将它们组合在一起。Fig. 4 展示了与真实情况和由 EMSMC [24] 和 EMSGC [35] 估计的相应运动分割的场景比较。我们的方法正确地将鸟类归入正确的类别，但受到背景稀疏事件的噪声影响。在这种情况下，EMSMC 从表面上显示出更好的分割效果，而 EMSGC 则将鸟分割成不同的簇。在定量评估中，虽然 EMSGC 错误地标记了前景在不同的簇中，我们将所有簇除背景外视为一个主要簇。

5.2. 定量评估

我们使用平均交并比（Mean IoU）和平均检测率（Mean DR）在测试数据集上对 EMSMC [24] 和 EMSGC [35] 进行方法基准测试。这个方法是相关的，因为这两种方法在测试时都不需要任何的真实值或预训练权重。与 EMSMC 的基准测试是公平的，因为这两种方法都需要一个初始聚类数，在我们的案例中为两个（背景和前景）。我们还与 EMSGC 进行了结果比较，同时牢记解决增加的聚类数量带来的挑战。由于测试数据集中存在大量运动，我们将所有方法的 N 设置为 50,000。这个选择也是出于考虑 EMSMC 和 EMSGC 在执行时所需的长时间计算。所有方法都发现测试数据集具有挑战性，在 Tab. 1 所示的 6 个序列中都表现出相关性能。在大多数情况下，我们的方法在场景分割中表现出更高的性能。此外，我们的方法执行即时运动分割，仅需几毫秒即可获得掩码，而由于迭代，EMSMC 和 EMSGC 需要几分钟。

6. 消融研究

6.1. Leaky ReLU 与 Leaky DoReLU 的比较

在 [23] 中，作者发现使用 leaky DoReLU 激活函数而不是 leaky ReLU 可以改善分割。这两个激活函数之间的主要区别在于 leaky DoReLU 对输入值大于 1 的斜率进行了限制。作者认为，使用 leaky DoReLU 改善分割的原因是梯度的稳定性增强，考虑到他们的 alpha 映射限制在 [0,1] 范围内。

受到他们分割方法的启发，我们选择实现 leaky DoReLU 而不是 leaky ReLU。然而，我们调查了这种修改后的激活函数在我们特定情况下的重要性。

基于这个原因，我们创建了一个备用的 EV-LayerSegNet 架构，其中使用了 leaky ReLU 激活函

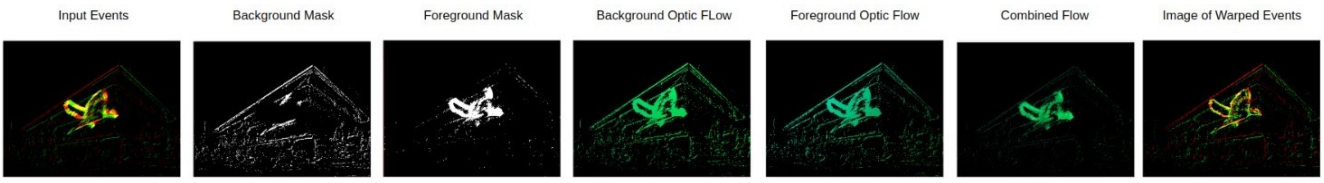

Figure 3. 合成数据集上的测试结果示例（” 建筑物前的鸟” 在 Tab. 1 中）。配色方案可以在 [9] 中找到。

	Mean IoU			Mean Detection Rate		
Sequence name	EMSMC	EMSGC	Ours	EMSMC	EMSGC	Ours
Drone above playground	0.20	0.05	0.71	0.00	0.00	0.87
Plane over city	0.06	0.03	0.67	0.00	0.00	0.77
Bird above playground	0.26	0.00	0.48	0.00	0.00	0.71
Second drone above playground	0.36	0.02	0.52	0.00	0.00	0.78
Bird in front of building	0.20	0.41	0.10	0.00	0.21	0.00
Helicopter over city	0.07	0.00	0.55	0.00	0.00	0.83

Table 1. 我们的方法、EMSMC [24] 和 EMSGC [35] 在使用恒等摄像机矩阵的 6 个模拟测试序列上的对比。

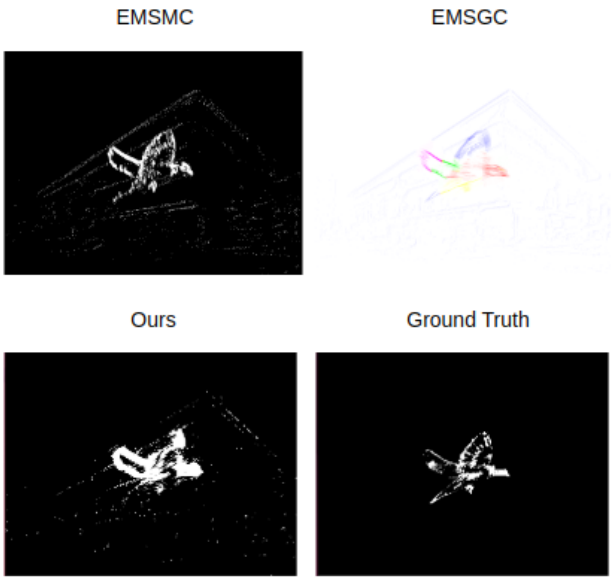

Figure 4. 对比 EMSMC [24] 、EMSGC [35] 和我们的方法与真实值（Tab. 1 中 “建筑前的鸟”）的定性测试结果。

	Mean IoU		Mean Detection Rate	
Sequence name	LeakyReLU	Ours	LeakyReLU	Ours
Drone above playground	0.38	0.71	0.00	0.87
Plane over city	0.48	0.67	0.65	0.77
Bird above playground	0.31	0.48	0.00	0.71
Second drone above playground	0.48	0.52	0.78	0.78
Helicopter over city	0.45	0.55	0.61	0.83

Table 2. 我们的方法与 Leaky ReLU 的测试结果对比基线。

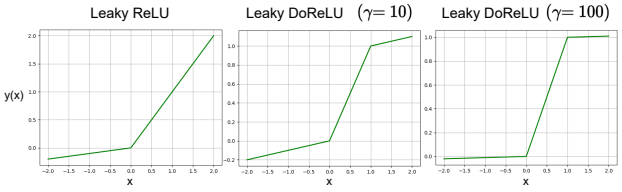

Figure 5. 漏泄 ReLU（左）与漏泄 DoReLU 的可视化，其中斜率系数 $\gamma = 10$（中）和 $\gamma = 100$（右）。x 轴是输入 x ，y 轴是输出 $y(x)$ 。

数而不是 leaky DoReLU 激活函数。我们在自己的数据集上对此进行了训练，并随后进行了测试。

Tab. 2 中的结果表明，与基线相比，网络遇到了更多的困难。这个观察结果表明，leaky DoReLU 激活函数提高了我们网络实现成功分割的能力。

通过在我们的分割模块中确定 leaky DoReLU 激活函数的重要性，我们转向探讨修改斜率系数 γ 的影响。我们研究了变化 γ 将如何影响我们网络的分割性能。

激活函数的特性如 Fig. 5 所示。可以观察到，泄漏的 DoReLU 在超过 $[0, 1]$ 范围时表现出比泄漏的 ReLU 更小的斜率。这种减小的斜率在稳定梯度流动中发挥了重要作用。我们怀疑这是因为该函数更具非线性，从而改善了我们的分割性能。为了验证这一假设，我们决定以更高的斜率（$\gamma = 10$）来训练网络，因此预期结果会较差。

相应的测试结果如 Tab. 3 所示，其中基线方法表现显著更好。这表明在 leaky DoReLU 函数中，范围 $[0, 1]$ 之外的近似平缓的斜率对于分割来说起着至关重要的作用。

7. 结论

我们提出了一种新颖的端到端卷积神经网络（CNN），利用事件相机进行自监督运动分割。我们借鉴了无监督运动分割领域的最新方法，并使用了最新发展的基于人工神经网络的事件光流估计算法。我们展示了从

	Mean IoU		Mean Detection Rate	
Sequence name	$\gamma = 10$	Ours	$\gamma = 10$	Ours
Drone above playground	0.06	0.71	0.00	0.87
Plane over city	0.03	0.67	0.00	0.77
Bird above playground	0.07	0.48	0.00	0.71
Second drone above playground	0.02	0.52	0.00	0.78
Helicopter over city	0.21	0.55	0.00	0.83

Table 3. 将我们的方法与基准在 $\gamma = 10$ 的测试结果进行对比。在所有场景中的分割性能显著下降。

模糊场景中分别学习运动模型和分割掩码并将其结合以对输入事件去模糊是可行的。为训练和测试我们的方法，我们生成了一个新的数据集，其中背景和物体进行仿射运动。我们展示了该网络能够正确分割事件流中的背景和独立移动的物体。我们还展示了我们的方法在无监督运动分割方面优于最新技术。不过，我们的方法仅限于仿射运动，并且对非常高的模糊度和噪声敏感。未来的工作中，我们计划扩展训练和测试数据集以获得更多的多样性，并将该方法与其他需要预训练权重的自监督方法以及最终的监督方法进行对比。在长期目标中，我们计划在真实的三维数据集上进行测试。额外的建议包括改进光流模块，以便也能估计非仿射运动，并使用脉冲神经网络开发该方法，从而允许在如无人机等对重量敏感的车辆上进行快速且低功耗的计算。我们希望我们的工作能够引起人们对事件相机自监督学习潜力的关注，从而允许基于学习的方法在没有昂贵的真实标注的情况下进行训练和使用。

8.

致谢

这项工作由 Sandooq Al Watan 在 SWARD-S22-015 资助协议下、STRATA Manufacturing PJSC 以及先进研究与创新中心（ARIC）支持。该中心是由 Mubadala 投资公司 PJSC 的全资子公司航空航天控股公司有限责任公司和哈利法大学科学与技术联合资助的。

References

[1] Yusra Alkendi, Rana Azzam, Sajid Javed, Lakmal Seneviratne, and Yahya Zweiri. Neuromorphic vision-based motion segmentation with graph transformer neural network. Trans. Multi., 27:385–400, 2025.

[2] Sami Arja, Alexandre Marcireau, Saeed Afshar, Bharath Ramesh, and Gregory Cohen. Motion segmentation for neuromorphic aerial surveillance, 2024.

[3] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In Proceedings of 1st International Conference on Image Processing, pages 168–172 vol.2, 1994.

[4] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12272–12281, 2019.

[5] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. IEEE Transactions on Pattern Analysis & Machine Intelligence, 44(01):154–180, 2022.

[6] Liyue Ge, Congxuan Zhang, Zhen Chen, and Ming Li. Optical flow estimation from layered nearest neighbor flow fields. In 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pages 1–6, 2018.

[7] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. IEEE Robotics and Automation Letters, 6(3):4947–4954, 2021.

[8] Stamatios Georgoulis, Weining Ren, Alfredo Bochicchio, Daniel Eckert, Yuanyou Li, and Abel Gawel. Out of the Room: Generalizing Event-Based Dynamic Motion Segmentation for Complex Scenes . In 2024 International Conference on 3D Vision (3DV), pages 442–452, Los Alamitos, CA, USA, 2024. IEEE Computer Society.

[9] Jesse Hagenaars, Federico Paredes-Valles, and Guido de Croon. Self-supervised learning of event-based optical flow with spiking neural networks. In Advances in Neural Information Processing Systems, pages 7167–7179. Curran Associates, Inc., 2021.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2014.

[11] M. Kumar, Philip Torr, and A. Zisserman. Learning layered motion segmentations of video. International Journal of Computer Vision, 76:301–319, 2008.

[12] Chankyu Lee, Adarsh Kosta, Alex Zihao Zhu, Kenneth Chaney, Kostas Daniilidis, and Kaushik Roy. Spike-flownet: Event-based optical flow estimation with energy-efficient hybrid neural networks. In European Conference on Computer Vision, 2020.

[13] Chankyu Lee, Adarsh Kumar Kosta, and Kaushik Roy. Fusion-flownet: Energy-efficient optical flow estimation using sensor fusion and deep fused spiking-analog network architectures. In 2022 International Conference on Robotics and Automation (ICRA), pages 6504–6510, 2022.

[14] Zhuoyan Li, Jiawei Shen, and Ruitao Liu. A lightweight network to learn optical flow from event data. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 1–7, 2021.

[15] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. CoRR, abs/1405.0312, 2014.

[16] Anton Mitrokhin, Cornelia Fermüller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object detection and tracking. In 2018

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1–9, 2018.

[17] Anton Mitrokhin, Zhiyuan Hua, Cornelia Fermüller, and Yiannis Aloimonos. Learning visual motion segmentation using event surfaces. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 14402–14411, 2020.

[18] Chethan M. Parameshwara, Simin Li, Cornelia Fermüller, Nitin J. Sanket, Matthew S. Evanusa, and Yiannis Aloimonos. Spikems: Deep spiking neural network for motion segmentation. In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 3414–3420, 2021.

[19] Chethan M. Parameshwara, Nitin J. Sanket, Chahat Deep Singh, Cornelia Fermüller, and Yiannis Aloimonos. 0-mms: Zero-shot multi-motion segmentation with a monocular event camera. In 2021 IEEE International Conference on Robotics and Automation (ICRA), pages 9594–9600, 2021.

[20] Federico Paredes-Vallés and Guido C. H. E. de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3445–3454, 2021.

[21] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. Conf. on Robotics Learning (CoRL), 2018.

[22] Nitin J. Sanket, Chethan M. Parameshwara, Chahat Deep Singh, Ashwin V. Kuruttukulam, Cornelia Fermüller, Davide Scaramuzza, and Yiannis Aloimonos. Evdodgenet: Deep dynamic obstacle dodging with event cameras. In 2020 IEEE International Conference on Robotics and Automation (ICRA), pages 10651–10657, 2020.

[23] Sahir Shrestha, Mohammad Ali Armin, Hongdong Li, and Nick Barnes. Learning to segment dominant object motion from watching videos. In 2021 Digital Image Computing: Techniques and Applications (DICTA), pages 01–08, 2021.

[24] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 7243–7252, 2019.

[25] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert E. Mahony. Reducing the sim-to-real gap for event cameras. In European Conference on Computer Vision, 2020.

[26] J.Y.A. Wang and E.H. Adelson. Representing moving images with layers. IEEE Transactions on Image Processing, 3(5):625–638, 1994.

[27] Ziyun Wang, Jinyuan Guo, and Kostas Daniilidis. Unevmoseg: Unsupervised event-based independent motion segmentation, 2023.

[28] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 7157–7168, 2021.

[29] Fan Yang, Srikrishna Karanam, Meng Zheng, Terrence Chen, Haibin Ling, and Ziyan Wu. Multi-motion and appearance self-supervised moving object detection. In 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 2101–2110, 2022.

[30] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 879–888, 2019.

[31] Chengxi Ye, Anton Mitrokhin, Cornelia Fermüller, James A. Yorke, and Yiannis Aloimonos. Unsupervised learning of dense optical flow, depth and egomotion with event-based sensors. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 5831–5838, 2020.

[32] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. ArXiv, abs/1805.04687, 2018.

[33] Luca Zappella, Xavier Llado, Edoardo Provenzi, and Joaquim Salvi. Enhanced local subspace affinity for feature-based motion segmentation. Pattern Recognition, 44:454–470, 2011.

[34] Bo Zhang, Jinli Suo, and Qionghai Dai. Event-enhanced snapshot compressive videography at 10k fps. IEEE Trans. Pattern Anal. Mach. Intell., 47(2): 1266–1278, 2025.

[35] Yi Zhou, Guillermo Gallego, Xiuyuan Lu, Siqi Liu, and Shaojie Shen. Event-based motion segmentation with spatio-temporal graph cuts. IEEE Transactions on Neural Networks and Learning Systems, pages 1–13, 2021.

[36] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters, 3(3):2032–2039, 2018.

[37] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. ArXiv, abs/1802.06898, 2018.

[38] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 989–997, 2019.