摘 要

事件相机（Event Camera）是一种新型的基于神经形态视觉原理的异步成像相机，又被称为“硅视网膜”。该相机模拟了人类视网膜的成像机制，成像时只对亮度发生变化的区域敏感，输出为包含变化的时空信息和极性信息的事件序列。相比于传统相机，事件相机具有功耗低、时延小、动态范围大等优点，在无人机视觉导航、无人驾驶、高速目标检测等方面具有很大的应用价值。但是，事件相机输出的原始事件序列信号质量较差，严重制约了其在实际场景中的应用性能。

针对这一问题，本文开展了基于深度网络的高质量事件序列重建研究，并主要完成了以下工作：首先，通过差分模拟和二次采集的方式完成了事件序列数据集的获取；其次，通过基于视觉暂留机制的映射算法实现了事件序列的可视化；最后，提出了两种基于深度网络的高质量事件序列重建方案——图像重构-卷积去噪自编码器方案和序列切分-循环神经网络方案，并通过不同的网络模型对两种方案进行了具体实现。经过测试，本文提出的两种方案均较好的完成了高质量事件序列的重建工作，并可以满足事件序列在不同场景下的应用需求。

关键词：深度网络　事件序列重建　卷积神经网络　自编码器　循环神经网络

ABSTRACT

Event cameras, also known as "silicon retina", are a new class of asynchronous imaging devices based on neuromorphic vision. Inspired by human being's retina, the camera is only sensitive to the region with brightness changes and outputs event sequences containing information of changing positions, time and polarities. Compared with traditional cameras, event cameras have many advantages, such as low power consumption, low latency and large dynamic range, which provide them with great potential applications in UAV's visual navigation, self-driving, high-speed target detection and so on. However, the poor quality of original event sequences seriously limits event cameras' application performance in actual scenarios.

To address this problem, we carry out the research of high-quality event sequence reconstruction based on deep neural networks(DNNs). Our main work is as follows. Firstly, we finish the acquisition of event sequence datasets through differential simulation and second shooting. Secondly, we realize the algorithm for event sequence visualization based on the persistence of vision. Finally, two high-quality event sequence reconstruction schemes — Convolutional Denoising AutoEncoder(ConvDAE) and Sequence segmentation Recurrent Neural Network(SeqRNN), based on DNNs are proposed and implemented through different network models. Through testing, we illustrate that this two schemes can successfully complete the reconstruction task and meet the requirements in corresponding scenarios.

Keywords: DNNs event sequence reconstruction CNN DAE RNN

目 录

第一章 绪论

1.1 视觉信息获取与成像技术简介

视觉是人类获取外界信息最主要的渠道之一，相关研究表明，人类在日常生活中获取的信息超过 80%来源于视觉。因此，视觉信息的获取、存储与处理也一直是信息领域中的重要课题之一。

世界上的第一台相机于 19 世纪 30 年代末由路易·达盖尔发明出来，在此后的数十年里，成像技术得到了迅速的发展。早期图像的获取和存储主要是通过化学原理将图像信息记录在含有卤化银的感光胶片上，然后通过在暗房进行冲洗来得到。通过这种方式获取的图像，实现了对视觉信息的采集和存储，是视觉信息获取技术的开拓性突破。但是，通过胶片存储的图像信息往往难以进行传输和处理，而且胶卷曝光后会使得信息丢失，因此也并不是一种可靠的长期保存方法。

传统成像技术存在的问题随着数字时代的到来得到了解决。1925 年前后，报纸业率先开创了数字图像技术应用的先河。此后，随着半导体器件、数字集成电路，特别是数字计算机的出现，数字图像的采集、存储及处理技术也得到了进一步的发展，并在医学影像、地理测绘、天文观测等许多领域得到了广泛应用。根据应用场景、采集谱段和获取信息维度等方面的不同，成像技术也多种多样，从普通的数码相机，到夜视红外成像的夜视仪，再到目前用于无人驾驶的激光雷达三维点云成像以及新兴的光场相机，成像技术在不断变革的同时也逐渐改变着生产生活的方方面面。近几年来，基于仿生学设计的类视网膜动态视觉传感器（Dynamic Vision Sensors，DVS）也崭露头角，并凭借低延迟、低功耗、低带宽和高动态范围等多方面的优势，在高速目标检测、无人驾驶、立体视觉等领域具有很大的应用潜力，成为新的研究热点。

1.2 动态视觉传感器与事件相机

动态视觉传感器是一种新兴的仿生视觉传感器，它基于神经形态视觉原理模拟了人眼视网膜的成像机制，又被称作“硅视网膜”（Silicon Retina）、神经形态视觉

传感器。

动态视觉传感器与传统图像传感器在成像机制上有着本质区别，最大的不同在于它没有“帧”的概念。工作时，传感器的各个像素探测器之间异步工作，并且各个像素探测器仅对亮度发生变化的区域敏感。每个检测到亮度变化超过给定阈值的像素探测器都会触发一个包含该像素的位置、变化时间和变化极性（即变化方向）的“事件”——$e(x, y, t, p)$，多个事件连续输出构成“事件序列”（Event Sequence）。这种成像信息的表示形式也被称作“地址事件表示”（Address Event Representation，AER），而基于这种芯片的相机则被称为“事件相机”（Event Camera）。

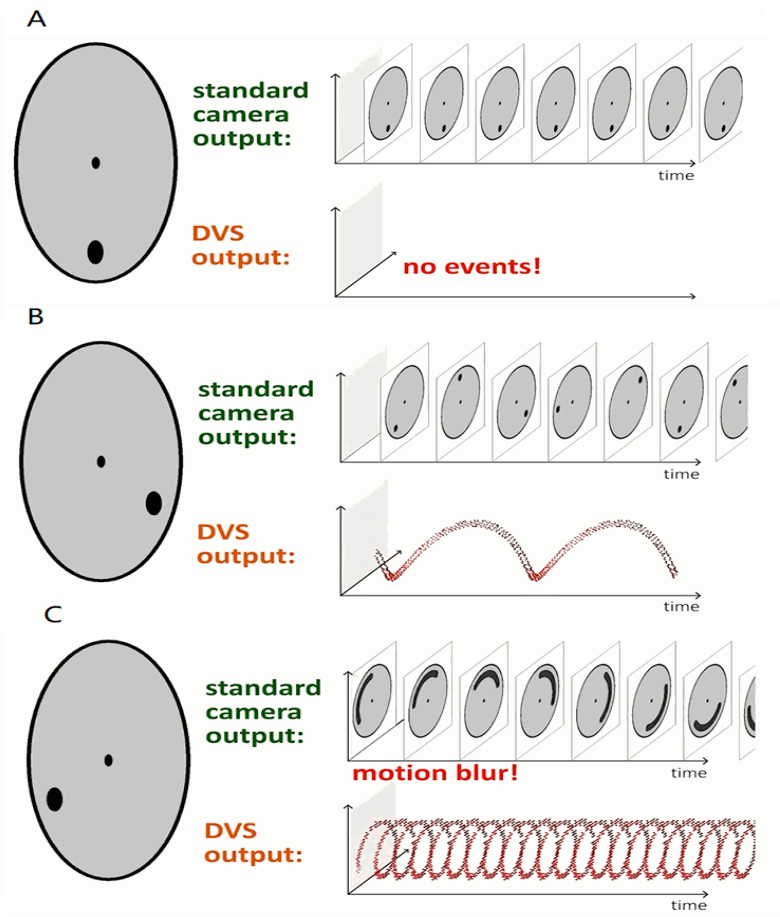

图 1.1 DVS 事件相机特性示意图（A：静止状态，B 低速运动，C：高速运动）

事件相机与普通相机的工作差异示意图如图 1.1 所示。事件相机工作时并不记录视野场景的全部信息，而是仅记录场景的变化，并且仅包含变化方向的二值信息，输出相应的 ON、OFF 两种事件，分别代表亮度增加或减小并超过给定阈值。当对静止无变化的场景进行拍摄时，事件相机没有事件触发，也不会产生输出信号。事

件相机的这种特性也大大降低了成像时的数据冗余，减少了数据传输所需带宽。由于事件相机的各个像素点异步工作，所以其时间分辨率很高，可以达到微秒量级，在对高速运动的物体进行拍摄时也不会存在运动模糊现象。此外，由于传感器对亮度变化的响应是对数的，所以相机的动态范围很大，对低光照环境下的目标也较为敏感。事件相机凭借在上述方面的优异性能，在低照度目标检测、高速目标跟踪识别、无人驾驶、立体视觉等方面具有很大的应用潜力。

虽然事件相机在很多方面的性能远远优于传统相机，但是其在空间分辨率和信号质量等方面的表现却还存在一定不足。由于事件相机的主要目标在于高速高动态成像，所以其在设计上采取了牺牲空间分辨率换取时间分辨率的折中方案。在应用中，通过结合传统相机和事件相机同步成像完成数据采集，可以充分扬长避短，补偿事件相机在空间分辨率方面的缺陷。在信号质量方面，由于事件相机灵敏度较高，所以受到电路热噪声和漏电流等方面的影响较大，目前已有事件相机的信噪比普遍不高。由于实际应用中信号质量对系统功能的实现具有重要影响，所以提高事件序列的信噪比，进行高质量事件序列重建具有重要的应用价值和工程意义。

1.3 事件序列去噪重建的研究现状及研究思路

由于事件相机的商业化发展刚刚起步，相关算法的适应性改进工作也刚刚开始，所以近几年事件相机相关的研究工作主要还是集中于芯片电路的设计优化以及底层特征检测等方面，而针对事件序列去噪重建和高层视觉任务实现等方面的研究还不是很多。特别是在去噪重建方面，目前已经发表的成果也主要是基于传统的时空最近邻或者稀疏表示等方法。

2015 年，苏黎世联邦理工学院的 Hongjie Liu 等人提出了一种基于时空相关性的事件序列滤波算法[1]，并通过 CMOS 工艺实现了功率 1mW，延迟 10ns 的混合信号滤波芯片设计。该芯片能够在保证系统最大数据通量不变的前提下，有效滤除动态视觉传感器输出事件序列中的不相关噪声。2016 年，南非开普敦大学的 Daniel Czech 和新加坡国立大学的 Garrick Orchard 对 8 种常见的基于事件的异步变化检测图像传感器滤波算法（最近邻法、极性法、不应期机制等）进行了实验评估[2]，并对各种滤波器的最优参数给出了参考建议。2018 年，西安电子科技大学的 Xuemei

Xie 等人提出了一种基于 K-SVD 稀疏表示和字典学习的去噪算法[3]，噪声滤除明显，但是其不足在于运算复杂度高，不能够实现在线实时处理，抑制了 DVS 在高速成像优势的体现。2018 年，南洋理工大学的 Vandana Padala 等人提出了一种基于神经网络的异步变化检测图像传感器滤波算法 NeuNN[4]，并将其部署于 IBM TrueNorth 神经形态芯片上。该算法不仅能够滤除采样信号中的噪声，同时还能产生新的信号事件，具有一定的缺失信号补全效果。

事件序列去噪的主要研究思路在于深度挖掘事件序列中信号和噪声所具有的不同时空分布特征以及这些特征所表现出来的稀疏特性等。通过直接或间接地利用信号和噪声在这些特征方面的差异，可以设计算法将其区分开来，并最终完成噪声的滤除。

在对时空分布特征的直接利用方面，基于时空最近邻和不应期机制的算法是最为常见的滤波方法。这类方法以噪声事件的随机杂散触发特性为依据，以信号事件发生的时间和空间间隔小于一定阈值为标准，完成信噪判别。具体来说，其滤波规则一般为：对于每个当前触发的事件，判断其邻域内是否也有事件触发过，如果有事件触发过，并且触发的时间相对于当前时间小于一定阈值，则判断当前的触发事件为信号。在一些改进算法中，往往引入不应期机制来进一步对时间相关性进行约束，该机制模拟了电生理学中动作电位触发特性，即在一个动作电位触发后，兴奋细胞会保持一段时间的休止期，不再响应刺激，在事件序列滤波中该机制表现为对于每个确定为信号的事件，其后一段时间（不应期）内触发的所有事件均判定为噪声。基于时空最近邻和不应期机制的算法原理简单，很容易在硬件电路上进行实现，并可以直接级联在传感器的输出端完成实时滤波。但是，该类算法往往需要人为地设定和调整滤波参数，对环境变化的自适应能力较差。此外，在噪声区域性集中或噪声信号与目标信号紧密耦合等特殊场景下，此类滤波方法往往很难取得良好效果。

在对时空特性的间接利用方面，根据事件序列重建出来的二维图像中噪声的不可稀疏特性，可以采用字典学习和稀疏表示的方法完成噪声滤除。这种方法首先需要通过离散余弦变换等方式构造一个过完备字典，然后通过映射从含噪事件序列中重构多张二维图片并拉平、归一化为图像块，接着使用 K-SVD 等方法利用含噪图像块训练字典，最后将含噪图像在训练好的字典上进行分解，加上去掉的均值，

重构出原始无噪声图像。由于在进行每段事件序列去噪时，都需要重复进行二维图像重构和 K-SVD 分解等操作来完成字典训练，所以基于字典学习的去噪算法往往需要耗费更多时间，且不能实现在线去噪，这在很大程度上限制了事件相机在低延迟方面优势的发挥。

1.4 本文研究内容与章节安排

高质量事件序列的获取对充分发挥事件相机的优异性能具有重要意义。但是现有的基于传统算法的事件序列去噪重建方案在自适应性、实时性等方面往往不能满足实际应用需求。目前，深度网络在图像处理、序列信号处理等方面均已取得了远超传统方法的效果。考虑到深度网络在表征能力、泛化能力和自适应性能等方面的优异表现，本文希望通过基于深度网络的方法来完成高质量事件序列的重建工作。

本文共分为六个章节，下面对各章节的主要内容进行简要概括。

第一章为绪论部分。本章首先对现代视觉信息获取和成像技术的发展进行了简要介绍，并引出了本文的主要研究对象——动态视觉传感器和事件相机；然后，总结了目前事件序列去噪重建的研究现状，并对研究思路进行了分析；最后，对本文的主要研究方法和研究目标进行了简单阐述。

第二章为事件相机和深度网络的背景简介。本章详细介绍了事件相机和深度网络的发展与应用情况。

第三章为事件序列数据集获取及事件序列二维可视化实现。本章简要介绍了目前事件序列数据集（神经形态视觉数据集）的发展状况和本文数据集的获取方法。然后，介绍了一种基于视觉暂留机制的事件序列二维可视化方法，为后续实验奠定了基础。

第四章和第五章分别对本文提出的两种高质量事件序列重建方案——图像重构-卷积去噪自编码器方案和序列切分-循环神经网络方案进行了详细介绍，并通过具体实验对两种方案的实现效果和应用性能进行了分析和比较。

第六章对本文的主要工作进行了总结，并对研究课题的未来发展进行了展望。

第二章 事件相机与深度网络背景简介

2.1 事件相机的出现、发展与应用

成像相机的核心器件是图像传感器，而图像传感器则是半导体器件、集成电路、光电信息技术等领域的综合发展所带来的产物，它通过光电转换器件实现了对二维光强分布信息的记录，形成了数字化、可存储、易处理的数字图像形式，是成像技术上的一个重要突破。1968 年前后，随着固态成像概念的提出，固体传感器技术便开始快速发展起来，并在许多领域中得到广泛应用。

2.1.1 传统图像传感器的发展与瓶颈

早期的图像传感器中，最主要的两种是互补金属氧化物半导体 CMOS（Complementary Metal-Oxide Semiconductor）图像传感器和电感耦合器件 CCD（Charge Coupled Device）图像传感器。二者的基本原理都是通过像素阵列对一段时间内接收到的光强的积累，实现对各像素点处灰度值的获取和记录。传感器的所有像素在固定曝光时间内同时曝光，获取的所有灰度值共同构成了对当前场景的一次全采样，即得到图像的一帧。按照预先给定的速率(即帧率)重复对场景进行“快照”，最终可以得到由一系列帧构成的视频。

受限于当时的半导体工艺水平， CMOS 图像传感器在分辨率、信噪比和灵敏度等方面的表现较差，发展也较为缓慢。相比之下，CCD 传感器凭借在光照灵敏度高、噪声小等方面的优势迅速发展起来，并占领了包括高端相机在内的绝大部分应用市场，而 CMOS 图像传感器则凭借着价格方面的优势，仅在中低端市场尚有一丝存活空间。但是近年来，随着半导体电路 CMOS 工艺的快速发展，CMOS 图像传感器在成像质量上也有很大提升，与 CCD 传感器之间的差距也越来越小。此外，凭借在低时延、低功耗、低成本和高集成度等方面的优势，CMOS 图像传感器日显后来居上之势。

虽然随着光电领域的快速发展，现在的 CCD 图像传感器和 CMOS 图像传感器的成像质量都有了很大提升，彩色图像、高清视频在生活中随处可见。但是，在一些特殊场景或者极端条件下，目前的成像技术仍然存在不足和瓶颈。其中，最突

出的两个场景是高动态范围成像和高速目标成像。动态范围的大小直接决定了图像传感器捕捉明暗场景的能力，具有大动态范围的传感器往往能够捕捉到场景更多的细节信息。受限于目前的技术水平，很难保证传感器对高亮度和低亮度场景同时具有细节捕捉能力，这也是我们日常拍摄明暗差异大的场景时出现亮处过曝、暗处模糊的原因。此外，受限于目前基于帧的成像体系的限制，传感器的帧率存在严重上限，普通相机的帧率往往仅在几十帧到几百帧每秒左右。虽然在日常场景中，40 帧每秒的拍摄帧率即可满足常规拍摄需要，但是在高清赛事直播、高速机动目标跟踪等方面，该帧率离实际需求仍相差甚远。此外，基于帧的成像体制带来的另一个明显劣势在于产生了大量的数据冗余，特别是在当前图像分辨率极高的情况下，即使经过压缩数据量仍比较大，对传输链路具有极大压力。当前，随着人工智能等领域的发展，无人驾驶、高速目标检测识别等技术对各种场景下的成像质量有了更高的要求，新型的成像体制和成像器件亟待开发来实现速度更快、冗余更小、动态范围更大的高质量成像。

2.1.2 仿生视觉传感器与事件相机的发展与应用

相比于现有的视觉传感器，人眼视觉系统这个“生物视觉传感器”在很多方面的表现都占据绝对优势。人眼视网膜中央凹具有很高的空间分辨率，可捕捉细节；视网膜外周具有很高的时间分辨率，可捕捉快速运动信息，提取并编码场景或物体的特征，如纹理、轮廓等。视网膜中央凹和视网膜外周协调工作，赋予了人眼在时空二维的优异性能。受到人眼视网膜外周视觉成像机制的启发，近些年来出现了一种新的仿生视觉传感器——神经形态视觉传感器，又被称作“硅视网膜”。

1991 年，Misha 和 Rodney 首次提出了“硅视网膜”的概念，并设计出第一个能够模拟视网膜外周光感受野、水平细胞和双极细胞的 VLSI 硅视网膜神经电路[5]。2006 年 Delbruck 研究组研制出首个分辨率 128×128 像素，动态范围 120 分贝的异步视觉传感器[6]，该传感器仅对场景中发生相对亮度变化的区域敏感，具有低冗余、低功耗和高动态范围的优点；2008 年，该研究组进一步将该异步视觉传感器的时间分辨率提高到了 15 微秒[7]，并首次推出商业化产品。2011 年，Posch 等人设计了一款基于时间的异步图像传感器 ATIS[8]，该传感器将动态检测电路和光强检测电路相结合，能够在保持较高时间分辨率的同时，获取事件触发处像素的多级灰度信息。 2013 年，Delbruck 研究组设计出了一种 DAVIS 视觉传感器[9]，该传感器融

合了传统灰度图像传感器和动态视觉传感器，从而同时具有传统相机和事件相机二者的优势（图 2.1 所示为 inivation 公司的两款商业 DAVIS 相机。）。2015 年，Chen 等人研制出了一款 Celex 事件相机[10]，该相机通过从对数光强电路取电压成像，解决了 DAVIS 传感器的时空域不一致性问题并在时间分辨率等性能上获得进一步提升。

图 2.1 inivation 公司的两款商业 DAVIS 相机

以仿生视觉传感器为核心的事件相机采用异步方式工作，仅关注有变化的像素点，而没有“帧”的概念，所以从根本上解决了传统相机数据高度冗余的问题，节约了数据传输所需带宽。在目前多媒体信息和大数据时代的背景下，人工智能发展的瓶颈之一仍在于对海量数据处理能力的算力的不足，而事件相机的低冗余特性显然为这一问题的解决提供了一条有效的途径。此外，异步刷新机制也使得事件相机摆脱了帧率的限制，具有低延迟的特性，能够实现对高速运动目标的准确捕捉跟踪，并且不会产生运动模糊。目前市场上可见的事件相机或神经形态视觉传感器（如 DAVIS240、ATIS 等）的时间分辨率已经可以达到亚微秒级别。在亮度检测上，事件相机的亮度采集采用了对数响应，极大的扩宽了其成像动态范围，从而提升了其在极端环境下的工作性能。

当然，事件相机的优势是以牺牲其在某些其他方面的性能为代价的。虽然事件相机具有极高的时间分辨率，但是相应的牺牲的是其空间分辨率。目前最新的神经形态视觉传感器 DAVIS 的分辨率也仅为 640×480＝307.2k 像素，远低于目前基于帧的图像分辨率。而且，事件相机并不记录变化像素的绝对亮度值，而仅仅记录其

亮度变化超过给定阈值时的变化方向（极性），输出 ON 事件（亮度增加并超过变化阈值上限）或者 OFF 事件（亮度降低并超过变化阈值下限），即仅有两级灰度值。另一方面，事件相机高度灵敏的检测性能也带来了噪声较大的问题。这些噪声一部分由作用于像素中浮动节点的开关上的漏电流引起，另一方面也来源于电路工作中的热噪声以及环境噪声。事件相机与普通相机的成像效果对比如图 2.2 所示。

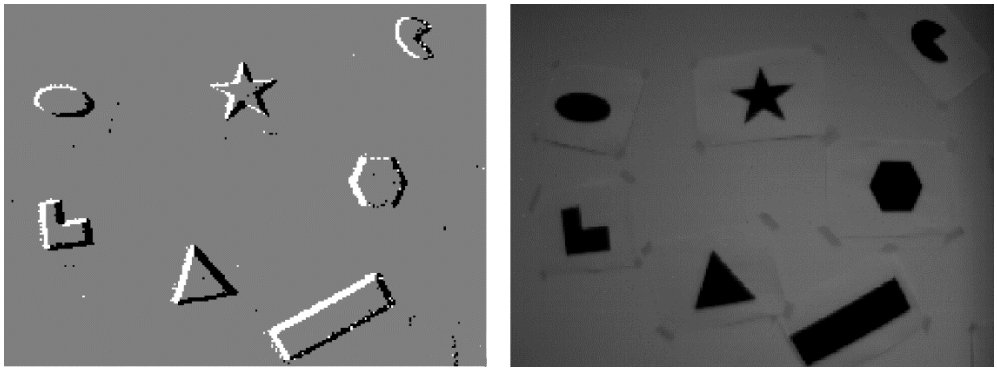

图 2.2 事件相机成像效果（左）和普通相机成像效果（右）对比

虽然事件相机在一些方面仍存在一定的不足，但是它在其他方面的优势已经足以为当前视觉成像领域开拓更广阔的应用市场。目前，在无人机视觉导航、无人驾驶、高速目标跟踪识别和立体视觉等方面，事件相机或者神经形态视觉传感器芯片已经得到初步应用。随着神经形态视觉（Neuromorphic Vision，NV）领域的进一步兴起和相关社区的进一步完善，基于事件的视觉数据集、算法和计算硬件也将得到进一步发展，一个独立于传统计算机视觉（Computer Vision，CV）的新兴领域正在逐渐形成。

2.2 深度网络的发展与应用

深度学习是目前人工智能领域最具影响力的技术之一，推动了包括无人驾驶、语音识别、智能医疗和智能图文信息处理等一大批新兴产业的发展。此外，深度学习作为一种数据驱动的有效工具，给通信、雷达、图像处理等许多传统信息领域也带来了新的活力，并逐渐形成了“深度学习+”的发展格局。然而，深度学习的发展和兴盛并非一日之功，而是伴随着数据、算力和算法的发展，积淀几十年的结果[11]。

2.2.1 深度网络的发展历史

深度学习是机器学习的一个子域，而深度学习中最重要，也是目前应用最广泛的部分则是深度神经网络。图 2.3 中展示了神经网络发展过程中的一些关键节点。早在 1943 年，美国数学家 W.Pitts 和心理学家 W.McCulloch 就已经提出了人工神经网络的概念[12]，并开始使用 MP 数学模型对人工神经网络中的神经元进行分析建模，这也被认为是人工神经网络研究的开端。20 世纪 50 年代末，美国科学家 Rosenblatt 基于 MP 模型提出了感知器（perceptron）算法[13]，用于完成分类任务。算法使用梯度下降策略从训练集中进行学习，更新网络参数，取得了很好的实践效果。此后该算法的收敛性也在理论上得到了严格证明，从而进一步推动了人工神经网络第一次浪潮的兴起。

虽然感知器的出现在神经网络的发展中具有里程碑式的意义，但是其存在的不足也导致了神经网络在随后数十年中的发展低谷。1969 年，人工智能之父 Marvin Lee Minsky 和 LOGO 语言创始人 Seymour Papert 指出，感知器本质上是一种只能处理线性分类问题的线性模型，并证明了单层感知机无法解决异或问题以及其他线性不可分问题[14]。这一结论极大限制了感知机的应用范围，也造成了人工神经网络在 20 世纪 70 年代和 80 年代期间将近 20 年的发展停滞。

人工神经网络的第一次低谷在 1980 年之后出现转机。1982 年，著名物理学家 John Hopfield 提出了 Hopfield 神经网络[15]，引入了能量函数的概念。随后 1985 年，Geoffrey Hinton 等人进一步提出了玻尔兹曼机[16]，它可以看作是带有隐藏单元的随机 Hopfield 网络，是能够学习内部特征表示的最早的神经网络之一，在表示并解决复杂的组合问题上有很好表现。1986 年 Hinton 又提出了适用于多层感知器的反向传播算法[17]，并引入了 Sigmoid 函数进行非线性映射。这些工作使得导致人工神经网络 20 年停滞不前的线性不可分问题得到了有效解决并产生了新的飞跃。此后，Kurt Hornik 和 G. Cybenko 等人又于 1989 年前后证明了多层感知器的万能逼近定理[18]，从理论上证明了人工神经网络对任意复杂度函数的强大表征能力，推动了神经网络的进一步发展。

神经网络发展中的第二次冰点在 1995 年到来。当时，在人们尝试训练大规模深度神经网络时，广泛应用的反向传播算法被发现存在梯度消失问题，这一问题造成了神经网络梯度学习速度极慢，并且容易收敛于局部最优解的现象。这一时期，以支持向量机为代表的浅层机器学习算法在应用性能上的优势凸显，进一步导致

了神经网络发展的停滞不前。

神经网络发展的第二次危机的解决仍然要归功于 Hinton、Bengio 和 Xavier 等人。2006 年，Hinton 针对深层网络训练中梯度消失的问题提出了“无监督预训练初始化权值+有监督训练微调权值” 的解决方案[19]，梯度消失的问题得到了一定的解决，深度神经网络的训练也重新成为可能。此后，Xavier 又于 2011 年提出使用修正线性单元作为神经元激活函数的有效方案[20]。该方案为网络引入了大量的特征稀疏性，加速了网络的特征学习速度，在很大程度上解决了深层网络中梯度消失的问题。至此，人工神经网络向着深度神经网络，或者说深度学习的发展迈出了关键一步。此后深度学习的发展便进入了爆发期，各种算法、模型如雨后春笋般在图像、语音等各个领域涌现出来，并在性能上远超其他传统算法。

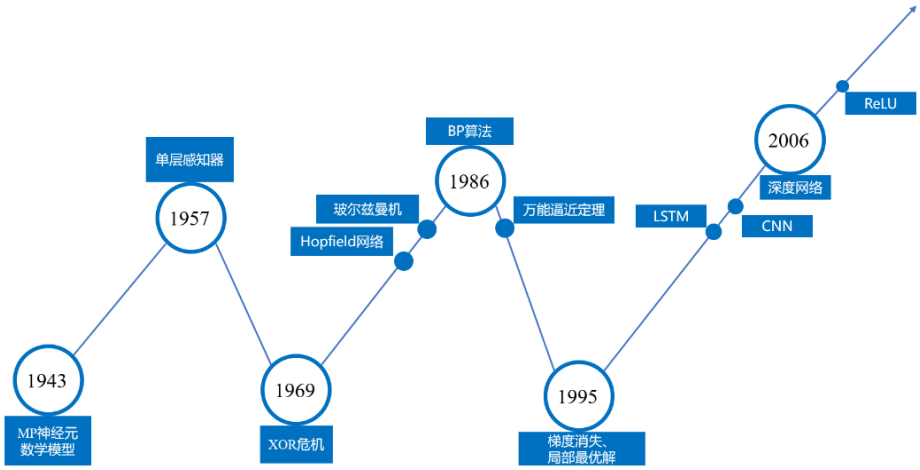

图 2.3 神经网络发展历程

2.2.2 深度网络的应用与现状

进入 21 世纪的第二个十年之后，许多优秀的深度学习模型，诸如 AlexNet、VGG、ResNet、BERT 等不断涌现，进一步完善了深度网络的应用基础，推动了深度网络在应用层面取得接连不断的突破。2011 年，微软公司首次将深度学习应用于语音识别领域，取得了很好的效果；2014 年，Facebook 公司开发的基于深度学习技术的人脸识别项目 DeepFace 准确率达到了 97%；2016 年，谷歌旗下 DeepMind 公司开发的 AlphaGo 战胜了围棋职业九段棋手李世石……深度学习现在已经进入了所未有的发展热潮之中。同时，伴随着互联网、多媒体的发展和 GPU 等硬件运算能力的提升，大规模深度学习所需的海量数据和强大运算处理能力也得以满足。

算法、数据、算力，这三个深度学习必不可少的组成要素一起，推动着深度学习向着更远的方向不断前进着。

目前深度学习的发展和应用中最主要的两个领域是在图像和语音的处理方面，涵盖了包括目标检测跟踪、图像识别分类、文字识别、图像去噪、三维重建、自然语言处理、语音合成等方方面面的多个子领域，催生了无人驾驶、智能医疗、推荐系统、合成主播等许多新兴技术与产业。现在，在学术界和工业界的共同推动下，深度学习仍在实践中高速发展着，并不断地重塑着人们的生活方式。相信在未来，深度学习及其衍生技术将会带来又一次划时代的变革。

2.3 本章小结

本章首先简述了传统图像传感器的发展历史和存在的不足，进而介绍了近几年来新出现的动态视觉传感器及事件相机。事件相机的出现改变了传统的成像体系，其在低冗余、低时延和高动态范围等方面的优势也为成像领域开拓了新的市场。接着，本章对深度网络的发展历史和应用现状进行了简要介绍。深度网络凭借强大的表征能力、学习能力和泛化能力在图像去噪、序列信号处理等方面取得了很大成功，这也为本文基于深度网络实现高质量的事件序列重建提供了良好借鉴。

第三章 事件序列数据集获取及事件序列二维可视化实现

3.1 事件序列数据集的获取

公开数据集的出现对计算机视觉领域的发展起到了重要的推动作用。一方面，可公开获取的数据集，如 MNIST、COCO、ImageNet 等为计算机视觉算法的开发提供了基本的“原料”，使得算法能够进行有效的训练和性能评估，并且省去了科研人员费时费力重复采集、标注数据的麻烦，从而能够节省更多的时间和精力专注于算法的开发。另一方面，使用被广泛接受的数据集进行基准测试可以在相同的条件下，对不同算法之间的性能进行定量评估和比较，并为算法开发者提供切实可信的最新性能指标以供挑战，能够有效促进社区良性竞争和技术进步[21]。

传统视觉数据集的素材主要来源于科研人员的合作采集以及网络图像的搜集标注。相比于传统计算机视觉领域，神经形态视觉领域的发展起步较晚，目前社区尚不完善。商业事件相机也是近几年才可以在市场上获取，并且成本较高。此外，事件相机获取的数据并不像传统图像那样可以直接呈现并在网络上分享，所以也进一步造成了现在事件序列数据集十分匮乏的局面。在这种状况下，将已有的传统图像数据集通过一定方法转化成为事件序列数据集成为一种短期可行的替代手段。本文中使用了两种类型的数据集，它们由常见的计算机视觉图像数据集分别通过差分模拟和二次采集得到，下面将对这两种数据集及其生成方式进行具体介绍。

3.1.1 差分模拟事件序列重构数据集

目前，基于事件的神经形态视觉算法的发展刚刚起步，与已经较为成熟的计算机视觉算法相比，仍存在许多问题需要解决。所以目前处理事件序列相关任务时，算法设计的主要思路之一仍是通过映射重构，将事件序列转换为适合于传统计算机视觉算法处理的二维图像，从而能够适配现有成熟算法框架解决特定任务。从这一思路出发，本文通过图 3.1 所示差分模拟的方式转换 MNIST 数据集获取了二维 MNIST-DIFF 数据集，用以完成基于二维映射的事件序列去噪重构算法的训练及验证。

MNIST 手写数字数据集[22]是机器学习和计算机视觉领域最常用的数据集之一，

被广泛应用于数据训练和算法基准测试。它由美国国家标准与技术研究所搜集的原始数据集 NIST 部分混合重组而来。其中，训练集部分包含了 60 000 张手写数字图片，测试集部分包含了 10 000 张手写数字图片。训练集和测试集图片均由 50%的美国人口普查局雇员手写数字和 50%的美国高中生手写数字组成。数据集中的图片均已被预处理为大小 28×28、灰度值 0-255 且数字基本位于图片中央的灰度图像。目前，基于卷积神经网络的识别算法在 MNIST 数据集上已经达到了超过 99.7%的识别准确率。

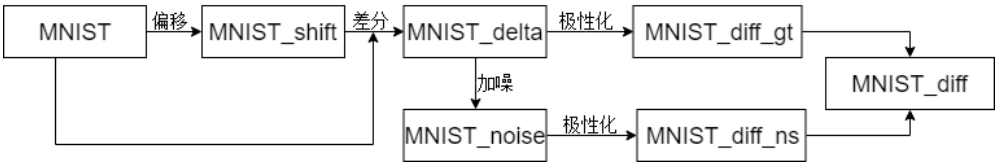

图 3.1 差分模拟事件序列重构数据集的生成原理

根据事件相机的基本原理和特性可知，事件相机仅对亮度变化超过给定阈值的像素敏感。当对移动物体进行成像时，在物体存在亮度梯度的地方将会出现亮度变化，产生事件输出。对一定的时间间隔内的事件序列进行二维映射后，其结果相当于对物体运动前后的灰度图像进行差分和极性化处理。基于此原理，本文首先获取了 MNIST 数据集，并将 MNIST 中的图片分别进行上、下、左、右，左上，左下，右上，右下共 8 个方向的平移，然后将平移后的图像与原图像进行差分和极性化处理（即将大于一定阈值的像素值置 1，小于一定阈值的像素值置-1，两阈值之间的像素值置 0），获得像素值为-1,0 或+1 的模拟事件序列二维映射图像，其中-1 和+1 分别对应于事件序列中的 OFF 事件和 ON 事件，0 代表未激发状态，即表示该像素处无事件输出。此外，本文还通过对差分后的图像进行加噪处理来模拟实际事件序列重构中的噪声。含噪图像与无噪图像分别作为样本和真值构成了 MNIST-DIFF 差分模拟事件序列重构数据集，如图 3.2 所示，其中 A 为原始 MNIST 数据集的图像，B 为重构数据集的样本图像，C 为重构数据集的真值图像。

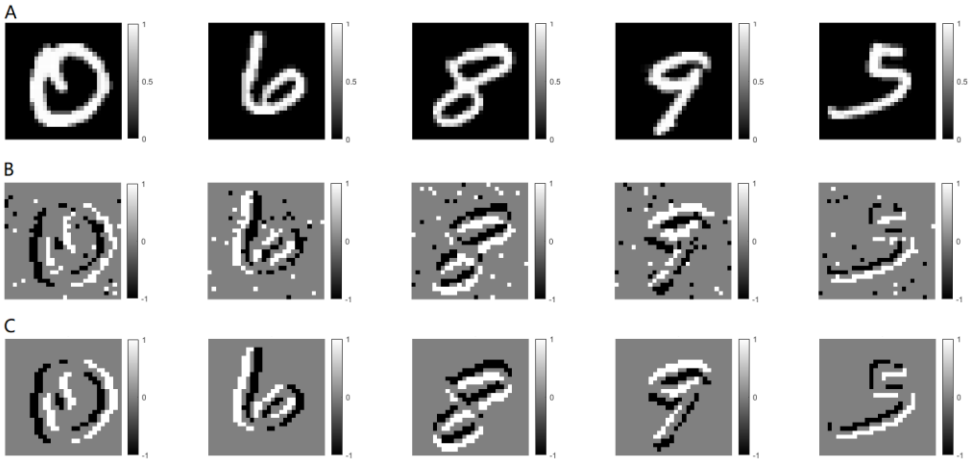

图 3.2 差分模拟事件序列重构数据集图像示例

3.1.2 二次采集事件序列数据集

虽然通过差分模拟的方式能够获得与事件序列二维重构结果相似的图像数据集，并可用于基于重构的相关算法的训练和验证工作，但是它也存在很大的局限性。一方面，由于此类数据集是通过模拟的方式获得的，并非实际采集，因此它和实际场景中的数据特征，特别是噪声特征，可能会存在较大差异，从而导致训练的模型在实际应用场景中可能失效。另一方面，此类数据集只能模拟事件序列二维重构之后的图像结果，不能模拟原始事件序列中的时间戳信息，从而导致了信息丢失。并且，该数据集的二维图像表示方式也无法用于基于原始事件序列（AER 数据表示方式）的相关算法的开发。鉴于此，一些研究团队利用现有的计算机视觉图像数据集，通过二次采集的方式获取了真正事件相机输出的事件序列，在一定程度上解决了通过模拟方式获取的事件序列重构数据集的局限性。

目前已经公开的此类事件序列数据集主要有 N-MNIST、N-Caltech101 和 CIFAR10-DVS 等。其中，N-MNIST 和 N-Caltech101 是由 Garrick Orchard 等人于 2015 年基于 MNIST 图像数据集和 Caltech101 图像数据集利用 ATIS 动态视觉传感器通过二次采集和后处理制作得到[23]。其具体操作为：将现有计算机视觉图像数据集 N-MNIST 或 Caltech101 中的图片按照固定时间间隔依次在显示器上进行放映，同时利用装在云台机械装置上的动态视觉传感器 ATIS 对显示器进行拍摄。为了产生相对运动触发事件输出，拍摄期间，图片保持静止，而搭载传感器的云台按照设定的闭环等边三角形轨迹进行变速运动。每一张图像采集时长约 300ms，其中

在 50ms、150ms 和 250ms 处相机的速度达到最大值，此时的数据率也是最高的。整个采集装置的运行和同步由一块 FPGA 进行编程控制，并且采集过程中进行了外部遮光处理，防止环境光污染。采集完成后，通过后处理对采集数据进行对齐、规则化，得到和原始图片像素范围一致，且图片内容位于图片中央的事件序列，最终结果如图 3.4 所示。与之相似，CIFAR10-DVS 数据集是由 Hongmin Li 等人于 2017 年基于 CIFAR10 图像数据集通过 DVS 事件相机二次采集处理得到[24]，采集装置如图 3.3 所示。与 N-MNIST 采集方式的不同之处在于，其采集过程中，传感器保持不动，而由显示器中移动的图片来产生相对位移，从而触发事件输出。

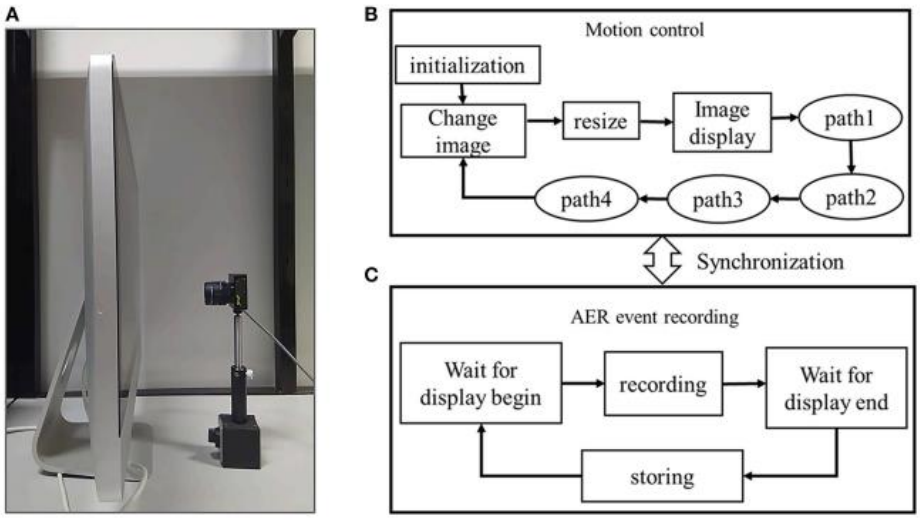

图 3.3 CIFAR10-DVS 数据集的二次采集示意图

通过二次采集的方式来获取事件序列数据集的优势在于，一方面重复利用了现有素材资源，省去了重新收集大规模数据并进行标注的麻烦。另一方面，由于采用了和传统计算机视觉领域相同的素材内容，更方便进行基于事件的神经形态视觉算法和传统的基于帧的计算机视觉算法之间的性能比较，有利于两个领域在公平可量化的竞争中共同发展。

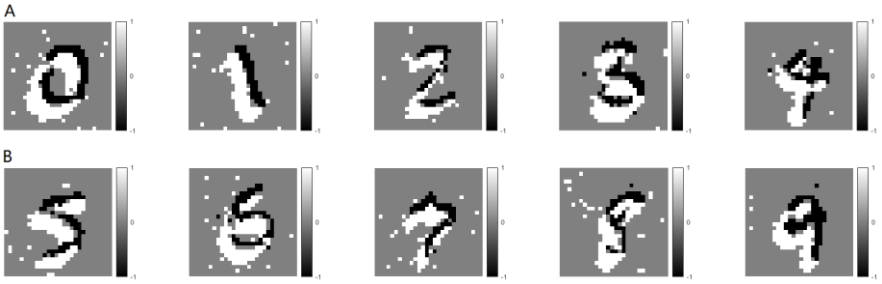

图 3.4 N-MNIST 事件序列数据集重构图像示例

3.2 事件序列的二维可视化实现

由于事件相机输出的原始信号为地址事件表示类型的事件序列，无法通过直接观察获得场景信息，所以需要对其进行可视化处理（如图 3.5 所示），以转换为便于直观观察的二维图像形式。

事件序列中的事件包含两维坐标信息 x、y，一维时间信息 t 和一维极性信息 p，因此，截取固定时间间隔 δ_t 内的事件序列，并通过二维映射的方式将其中包含的事件点的极性值 p 赋予其所在的坐标$(x，y)$，即可重构出一帧二维图像。重复累积截取多段事件序列并进行映射重构，即可生成二维可视化视频帧序列。由于事件相机具有微秒级的时间分辨率，因此在对高速运动目标进行拍摄时，可以达到数百帧每秒的重构帧率，足以满足一般高速场景下运动成像无模糊的需求。

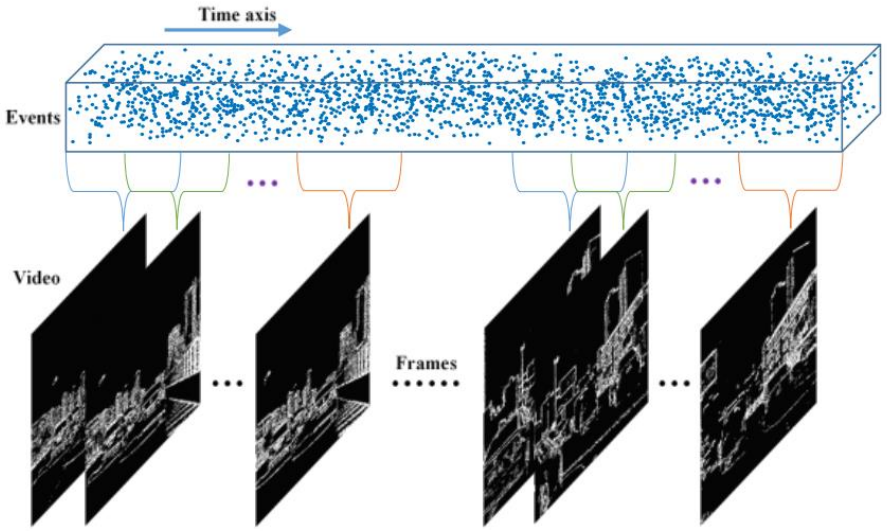

图 3.5 事件序列可视化示意图

3.2.1 “多对多”映射和“多对一”映射

实际操作中，在进行每帧二维图像的映射重构时，可能会出现在同一时间间隔内的同一坐标处产生多次事件触发的现象。考虑到这种情况，具体实现时主要有两种映射方案：一种是进行“多对多”映射，即对同一坐标处触发的多次事件进行带极性的积累加和，最终得到一幅具有多级灰度值的灰度图像；另一种方案是进行“多对一”映射，即对同一坐标处触发的多次事件，仅保留最新一次的触发事件，并将其极性值+1 或者-1 作为该坐标处的像素值，最终得到一幅仅包含+1、-1 和 0 三级

灰度值的灰度图像。由事件相机的动态成像特性可知，同一坐标处触发的多次事件和该坐标处的绝对亮度值并没有直接关系，而仅代表拍摄对象在该坐标处可能存在连续亮度梯度变化，因此“多对多”的映射方案缺乏一定的合理性，本文主要采用“多对一”的映射方案。

3.2.2 基于视觉暂留机制的“重叠”可视化

在对事件序列进行二维视频可视化时，如何选取合适的时间间隔需要综合考量重构速率和重构质量。如图 3.6 所示，当选取的事件序列时间间隔较大时，每秒输出的帧数也会相应的减少，造成帧率下降；而如果选取的时间间隔较小，则该时段内包含的事件较少，重构出的图像也会比较稀疏。为了解决上述矛盾，Xuemei Xie 等人与 2017 年提出了一种具有“重叠”的事件序列二维视频可视化方法[3]，该方法模拟了人眼视觉的暂留机制，在保证重构速率的同时提升了每帧图像包含的信息。

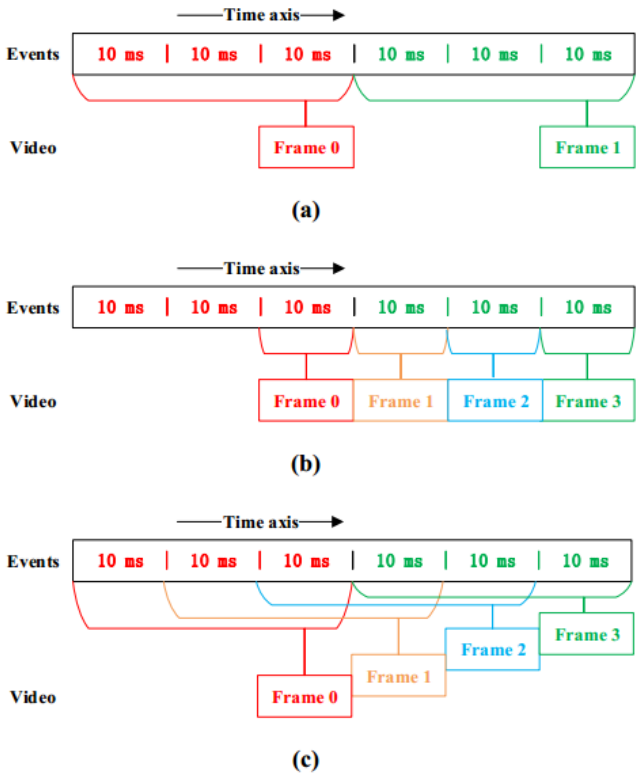

图 3.6 事件序列可视化中帧间间隔和帧率的关系

视觉暂留机制又称“余晖效应”，它是指当人眼观察外部场景时，所形成的视觉信息会在人的大脑中保持 0.1s-0.4s 的短暂时间，这也是我们能够观看 24 帧/秒的

“离散”视频而不会觉得卡顿的原因。受此机制启发，在对事件序列进行二维视频可视化时，可以人为地在相邻两段事件序列切片之间保持一定比例的“重叠”，这样相当于每一个新的视频帧仅包含部分新的事件和部分原有事件。在总的事件数不变的情况下，由于每帧图像仅包含较少的新事件，所以平均帧率仍然较高。同时，由于每帧图像也包含了部分先前事件，所以其内容也更加丰富，视觉效果更好。图 3.7 中展示了有“重叠”和无“重叠”时的二维图像重构差异。

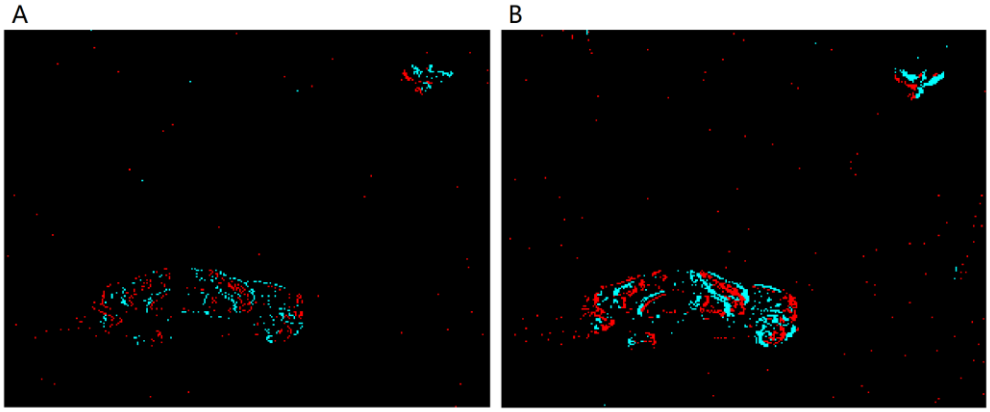

图 3.7 事件序列无“重叠”重构图像（左）和有“重叠”重构图像（右）

3.3 本章小结

目前，事件相机的商业化应用刚刚开始，神经形态视觉领域的发展也还不完善。在这种背景下，利用现有图像资源完成数据集的转换具有重要的实际意义。本章分别通过差分模拟和二次采集的方式完成了从传统图像数据集到事件序列及其重构数据集的转换，为后续深度网络的训练提供了数据资源。此外，本文又在“多对一”的映射原则下实现了基于视觉暂留机制的事件序列可视化算法，该算法在保证帧率的同时具有较高的图像质量，为事件序列重建效果的可视化评价打下了基础。

第四章 基于图像重构-卷积去噪自编码器的高质量事件序列重建方法

目前，事件驱动的神经形态视觉算法的发展刚刚起步，该类算法在很多方面的表现和传统计算机视觉算法相比仍存在较大不足。所以目前在事件相机和事件序列相关的许多问题中，一个主要的研究思路仍是首先将事件序列重构为传统的二维图像，然后利用现有计算机视觉框架完成后续处理。在这种思路的指导下，本章提出了一种基于图像重构-卷积去噪自编码器的高质量事件序列重建方法——ConvDAE。该方法充分挖掘了事件序列所包含的二维场景结构信息，并与现有的计算机视觉框架和高层视觉任务解决方案相匹配，为基于事件相机开发端到端的视觉任务解决方案提供了可能。

4.1 图像重构-卷积去噪自编码器方案的基本原理

4.1.1 卷积去噪自编码器

去噪自编码器是深度网络中最基本的网络架构之一，早在 2008 年 Vincent 等人就已经在国际机器学习会议中首次提出了用于抽取和重构鲁棒特征信息的去噪自编码器[25]，并通过无监督的方式在图像去噪任务中取得了良好表现。去噪自编码器的基本结构与普通自编码器基本相同，即都是通过编码网络首先对输入数据进行编码，得到特定的空间表征，然后再通过“对称”网络将编码所得空间表征恢复到原始输入形状后输出，其训练目标为得到与输入尽可能相同的输出，从而完成“自编码”任务。通过对编码获得的空间表征附加维度、稀疏性等方面的约束，可以获得具有相应性质的空间表征，从而使得自编码器在数据降维、图像压缩、稀疏表示等方面具有很大应用价值。去噪自编码器的基本结构如图 4.1 所示，与普通自编码器不同，去噪自编码器首先通过加噪、随机丢弃等方式对输入数据进行了“破损”（corrupt）操作，然后才将其输入后级自编码器结构进行训练。通过训练，网络能够学习到从“破损”数据中重建出原始输入的能力，即本质上为学习到了对输入中所包含的鲁棒特征信息的表征，从而相应地能够对噪声进行滤除。

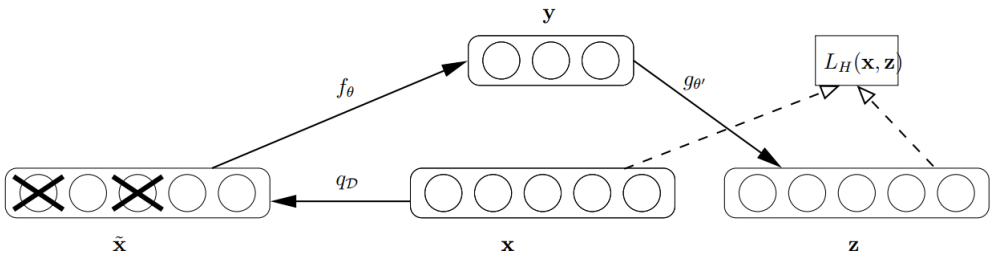

图 4.1 去噪自编码器的基本结构

近些年来，随着便携成像设备的发展和互联网的普及，我们生活中出现了越来越多的图像、视频等视觉数据，视觉信息处理也因此成为机器学习的主要研究与应用方向之一，并催生出了“计算机视觉”这一相对独立的学科分支。作为目前计算机视觉领域最主流的框架之一，卷积神经网络的应用对该领域的发展起到了巨大的推动作用。卷积神经网络最早在 20 世纪 80 年代就已经初现雏形，但是限于当时算力和训练样本的不足，卷积神经网络在当时并没有发挥太大作用。1998 年，LeCun 等人提出基于卷积神经网络的 LeNet-5（如图 4.2 所示）并在手写数字识别问题上取得成功后，卷积神经网络逐渐进入全盛期，此后基于卷积神经网络的 AlexNet、VGGNet、ResNet 等相继出现，并在图像分类、识别等领域取得惊人表现[26]。与全连接网络相比，卷积神经网络通过“局域卷积”和“层级递进”等策略完成了对图像特征从低层到高层的提取和抽象，更好地模拟了人类视觉系统处理视觉信息的机制。此外，卷积神经网络也大大减少了处理大量图像数据所需参数，节省了算力资源，加快了网络模型的训练速度。

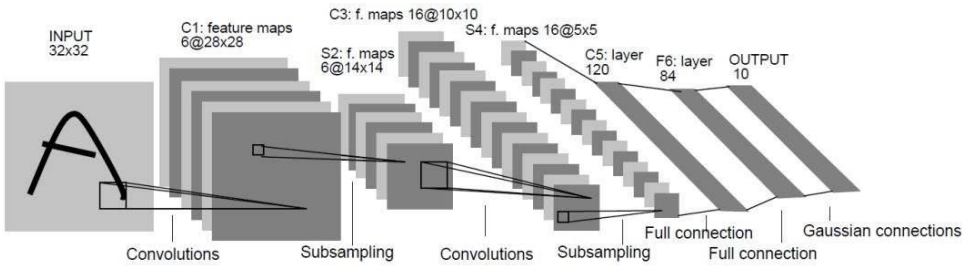

图 4.2 基于卷积神经网络的 LeNet-5 网络结构

考虑到卷积神经网络在图像特征抽象方面的优异性能和去噪自编码器在鲁棒特征信息提取和重构方面的良好表现，将二者相结合构成“卷积去噪自编码器”——ConvDAE 来完成图像去噪任务成为一种自然而然的研究思路[27]。目前，卷积去噪自编码器在医学影像去噪、光学字符识别预处理等方面已经取得良好效果，本章提

出的方案也将以卷积去噪自编码器为基础，来完成高质量事件序列去噪重建网络的搭建。

4.1.2 “skip connection”——跳跃连接

随着硬件运算能力越来越高，需要解决的问题越来越复杂，现在卷积神经网络的层数也越来越多。从最初的五层 LeNet 到现在的几百层甚至上千层的极深网络，网络深度的增加不仅带来了更强的表征能力，同时也带来了梯度消失、信息传递损耗等挑战。而“skip connection”——跳跃连接的出现则有效地解决了极深网络所存在的这些问题。

跳跃连接策略首次出现在 Kaiming He 等人于 2015 年提出的残差网络中[28]，如图 4.3 所示，其思路是在网络基本模块设计时引入一条新的数据流支路，该支路不经过中间层的卷积加权操作，而是直接完成从模块输入到输出的数据传递。跳跃连接的引入使得网络前级的信息能够更有效地传递到网络后级，而且当前网络模块也只需学习上一个模块输出的残差，减轻了网络的负担，这也是残差网络名称的由来。引入跳跃连接策略的残差网络以更快的速度、更少的参数在当年的 ILSVRC 2015 比赛中大放异彩、取得冠军，也使得跳跃连接成为后续极深网络设计的必选组件。此后，DenseNet、U-Net 等网络的设计纷纷效仿，通过在网络结构中加入类似的跳跃连接结构，取得了很好的效果。

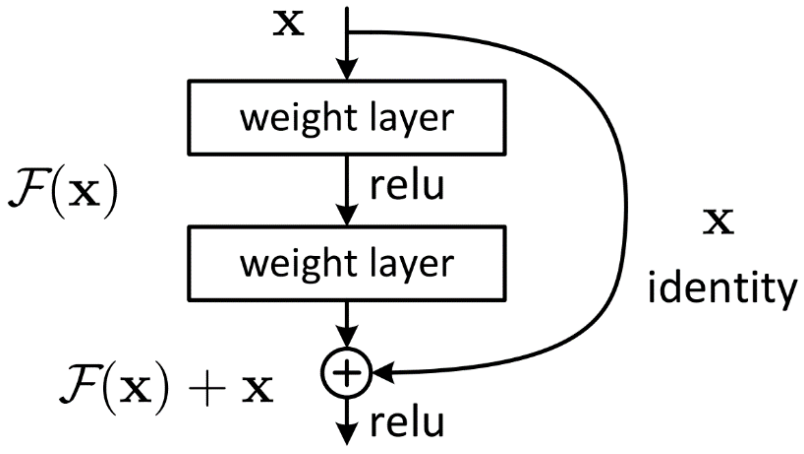

图 4.3 残差网络基本模块设计中所采用的跳跃连接

4.2 图像重构-卷积去噪自编码器方案的设计与实现

4.2.1 数据集预处理与二维图像重构

本章所提出的图像重构-卷积去噪自编码器方案分别在差分模拟事件序列重构数据集 MNIST-DIFF 和二次采集事件序列数据集 N-MNIST 上进行了训练和测试。差分模拟事件序列重构数据集和二次采集事件序列数据集的生成方式已经在本文 3.1 节中进行了说明。下面分别对两种数据集的具体格式、参数和预处理方法进行介绍。

MNIST-DIFF 数据集通过对 MNIST 数据集进行偏移、差分、加噪和极性化等处理来模拟事件序列重构的二维结果。其中偏移操作的尺度为 2 个像素，加噪处理中所加噪声为零均值标准差 0.25 的高斯白噪声，极性化操作的阈值为±0.5。数据集中包含了训练集和测试集两部分，其中训练集含有样本 60 000 个、测试集含有样本 10 000 个。由于该数据集本身即为二维图像数据集，所以不需要进行二维重构，该数据集的意义在于为算法的可行性提供一个简单验证并为实际采集数据集做一个对比参照。

N-MNIST 数据集是通过对 MNIST 数据集进行二次采集获得的，由于二次采集数据集的原始数据格式为事件序列，所以需要首先通过预处理将其重构为二维图像。重构时，仅在每条事件序列记录具有较高数据率的 150ms 附近截取时长 1/24s 的一段事件序列，利用 3.2 节提出的可视化方法映射为一帧图像，作为该条事件序列记录重构出的样本。此外，预处理时还利用 1.3 节中提到的时空最近邻滤波方法获取了该段事件序列预去噪之后的序列（滤波阈值为人工精心挑选），并用与上述相同的方法重构出一帧图像，作为模型训练和测试的参考。最终，通过预处理同样可以重构出一个包含 120 000 个训练样本和 20 000 个测试样本的重构数据集 N-MNIST-PIC，其中原始序列重构样本和预去噪序列重构样本各占一半。

4.2.2 基于卷积去噪自编码器的网络架构设计

本节以卷积去噪自编码器为基础构建了两种形式的深度网络模型——线性串联型卷积去噪自编码器 L-ConvDAE 和带有跳跃连接的 U 型卷积去噪自编码器 U-ConvDAE，下面分别对两种模型进行具体介绍。

（1）线性串联型卷积去噪自编码器 L-ConvDAE

L-ConvDAE 是在传统去噪自编码器的基础上，通过将普通的全连接网络结构替换为全卷积网络结构来实现的，其主体部分为一个包含 8 个串联相接卷积层的卷积去噪自编码器，具体配置如表 4.1 所示。

表 4.1 L-ConvDAE 的网络层级及参数配置

layers		patch size	stride	output size
inut				28×28×1
encoder	conv_1	3×3	1	28×28×32
	max pool_1	2×2	2	14×14×32
	conv_2	3×3	1	14×14×64
	max pool_2	2×2	2	7×7×64
	conv_3	3×3	1	7×7×64
	max pool_3	2×2	2	4×4×64
mid-layer	conv_4	3×3	1	4×4×128
decoder	conv_5	3×3	1	4×4×64
	deconv_1	2×2	2	7×7×64
	conv_6	3×3	1	7×7×64
	deconv_2	2×2	2	14×14×64
	conv_7	3×3	1	14×14×32
	deconv_3	2×2	2	28×28×32
output	conv_8	3×3	1	28×28×1

网络的输入为一个 28×28 的单通道事件序列重构图像，该图像具有三级灰度值，分别为+1，0，-1，代表着各个像素的极性状态。网络编码器和解码器具有对称结构，各包含了 3 个卷积层，卷积层采用 tanh 函数作为激活函数。为了防止训练中出现过拟合现象，每个卷积层之后进行了 dropout 操作，丢弃比例为 30%。在编码器中，通过最大池化的方式进行下采样来降低特征图的尺寸，在解码器中通过转置卷积的方式进行上采样来提高特征图尺寸，最终在输出层采用一个单通道的卷积核获得一个与输入尺寸相同的输出图像。

与传统去噪自编码器不同，L-ConvDAE 并非直接使用原始输入作为真值来和

网络输出求取损失函数，而是先将原始输入通过 1.3.2 节中介绍的时空最近邻滤波算法进行预去噪之后再作为真值和网络输出求取损失函数。通过引入时空最近邻滤波模块，可以给原有的无监督网络引入了监督特性，有利于网络更好的学习到去噪特性。此外，由于时空最近邻滤波运算复杂度很低，所以该模块并不会对网络运行速度产生明显影响。L-ConvDAE 网络的整体模型示意图如图 4.4 所示。

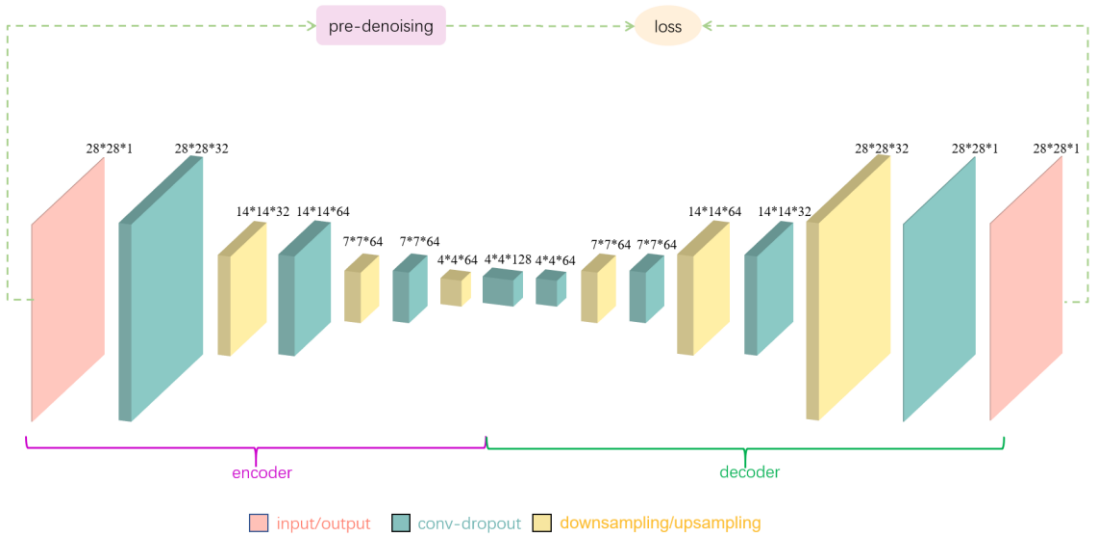

图 4.4 L-ConvDAE 网络模型示意图

（2）U 型卷积去噪自编码器 U-ConvDAE

U 型卷积去噪自编码器是在线性串联型卷积去噪自编码器的基础上，借鉴 U-Net 的设计思路，通过在编码器和解码器的相同层级之间引入跳跃连接支路来提高前后级信息传递效率的一种网络结构。

U-ConvDAE 的网络层级和参数配置与 L-ConvDAE 基本相同（见表 4.1），二者的不同之处在于，U-ConvDAE 网络模型在编码器和解码器的对应层级之间引入了跳跃连接，即直接将编码器各卷积层的输出传递到解码器对应层上并与之拼接起来，然后再完成解码器的后续卷积操作；此外，U-ConvDAE 的 dropout 比例设置为 10%。在损失函数计算方面，与 L-ConvDAE 类似，U-ConvDAE 同样先引入了基于时空最近邻滤波的预去噪模块再去计算损失函数，从而为网络引入监督特性。U-ConvDAE 网络的整体模型示意图如图 4.5 所示。

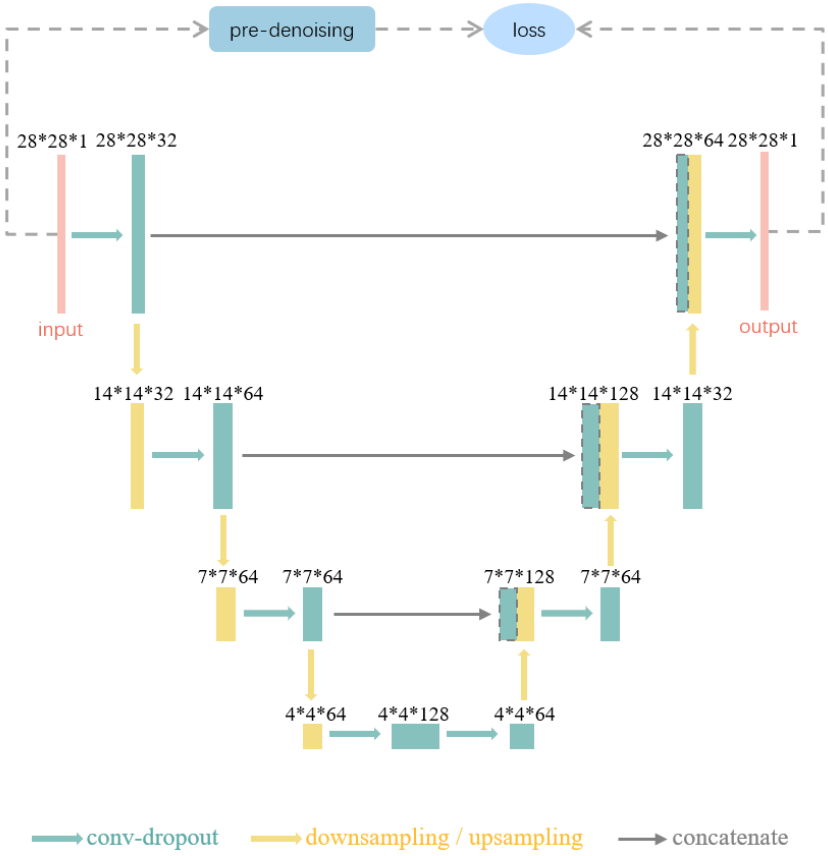

图 4.5 U-ConvDAE 网络模型示意图

4.3 图像重构-卷积去噪自编码器方案的实验与分析

网络训练阶段选取的损失函数为均方误差函数，选用的优化器为 AdamOptimizer，采用随机梯度下降策略进行训练，批量大小为 128，学习率为 0.0005。将两种网络模型分别在差分模拟事件序列重构数据集 MNIST-DIFF 和二次采集事件序列重构数据集 N-MNIST-PIC 上进行训练和测试，训练与测试平台配置为 Core i5-7300HQ CPU，2.5GHz 主频; NVIDIA GeForce GTX 1050Ti GPU，4G 显存；操作系统为 Windows 10，网络架构基于 TensorFlow 编写。网络训练过程和测试结果如图 4.6-4.8 所示。

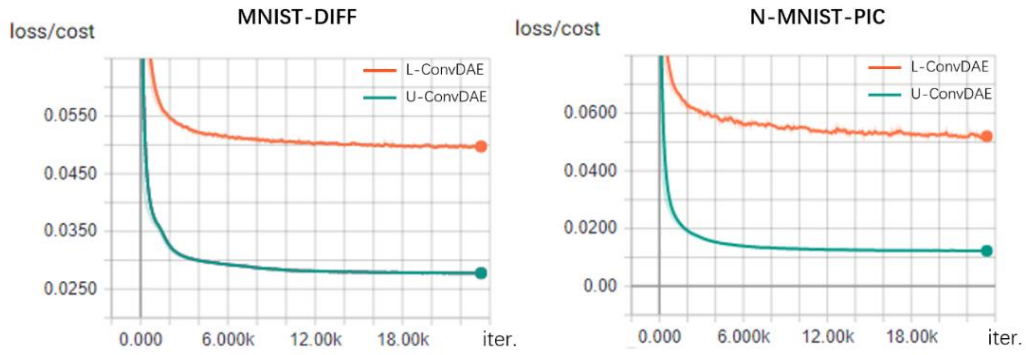

图 4.6 MNIST-DIFF 和 N-MNIST-PIC 数据集上两种模型的测试误差曲线

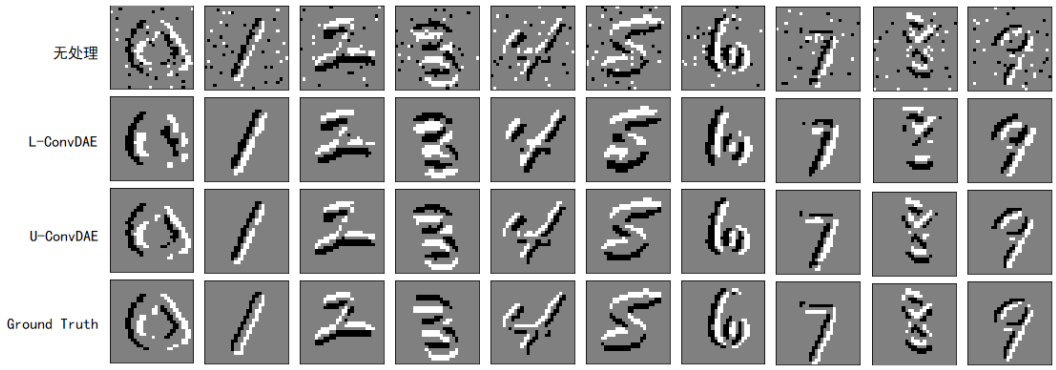

图 4.7 MNIST-DIFF 数据集上两种模型的处理结果

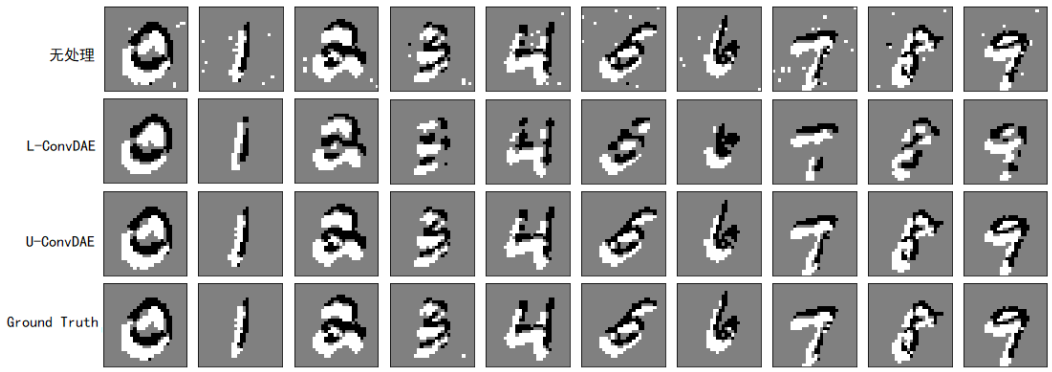

图 4.8 N-MNIST-PIC 数据集上两种模型的处理结果

通过训练过程中的测试误差曲线和训练完成后的测试集处理结果可以直观看出，本章所提出的两种网络模型——L-ConvDAE 和 U-ConvDAE 均实现了较好的去噪重建效果。相比之下， 在两种数据集上 U-ConvDAE 的测试误差均更小；在测试集处理结果上， L-ConvDAE 的少量重建结果细节部分有断裂（如图 4.7 中的数字“6”、“8”及图 4.8 中的数字“3”、“7”）且整体较为粗糙，而 U-ConvDAE 的重建结果则更加细腻，贴近真值。

为了进一步定量衡量网络对事件序列重构图像的去噪重建效果，本文选取了均方误差 MSE，峰值信噪比 PSNR 和结构化相似度 SSIM 作为评价指标，对去噪前后图像的质量进行了分析对比，同时对单幅图像重构的平均耗时进行了统计，结果如表 4.2 所示。

表 4.2 网络去噪重建效果测试与对比

网络	数据集	MSE	PSNR	SSIM	time(ms)
MNIST-DIFF	无处理	0.072	17.457	0.599	
	L-ConvDAE	0.050	17.890	0.674	0.143
	U-ConvDAE	0.028	20.554	0.756	0.148
N-MNIST-PIC	无处理	0.052	18.452	0.752	
	L-ConvDAE	0.022	22.664	0.774	0.139
	U-ConvDAE	0.010	26.294	0.950	0.149

由表 4.2 可以看出，两种网络模型重建得到的图像的各项指标均有一定改善，特别是经过 U-ConvDAE 重建得到的图像，在均方误差、峰值信噪比和结构化相似度等方面改善十分明显，这也与上面图 4.6-4.8 中的直观结果相吻合，表明了引入跳跃连接的 U-ConvDAE 模型能够更好的实现编解码器相同层级之间的信号传递，从而有效的改善网络性能。重建时间方面可以看到，对于本实验中的 28×28 的手写数字图像，在笔记本平台上以 100 个样本为一个批次完成共 10 000 个样本的测试，单张图像的网络平均处理时间仅为不到 0.15ms，完全能够满足实时性的需求。

4.4 本章小结

本章提出了两种基于图像重构-卷积去噪自编码器方案的高质量事件序列重建模型——L-ConvDAE 和 U-ConvDAE，并对两种网络模型的实现效果进行了定性对比和定量分析。通过对比分析可以发现，本章提出的两种网络模型均达到了较好的事件序列重建效果，相比之下， U-ConvDAE 由于引入了跳跃连接，能够更有效地实现了深层网络前后级的信息传递，所以表现更好。

在本方案的基础上，可以将 L-ConvDAE 或者 U-ConvDAE 作为事件相机的“转接器”与“滤波器”，直接完成事件序列到图像信号的转换及去噪。网络模块的输出结果也可以方便地与现有的基于图像的后级高层网络相级联，从而实现事件相机端到端的高层视觉任务解决方案。

第五章 基于序列切分-循环神经网络的高质量事件序列重建方法

基于图像重构-卷积去噪自编码器的高质量事件序列重建方法通过二维重构的方式显式地利用了事件序列所包含的二维场景结构化信息，在当前计算机视觉框架下取得了较好的重建效果。但是，其不足在于在重构的过程中丢失了原始事件序列所包含的时间信息，导致相邻事件之间的时间相关性没能得到充分利用。随着事件驱动的神经形态视觉算法和第三代人工神经网络——脉冲神经网络的发展，基于事件的脉冲信号处理框架也在逐渐兴起，并已经在一些领域取得较好效果。为了更好地挖掘事件序列之间的时间相关性并适配正在兴起的事件驱动算法，本章提出了一种基于序列切分-循环神经网络的高质量事件序列重建方法——SeqRNN。该方法将原始事件序列按固定步长进行切分后直接作为输入，并输出具有相同数据类型的去噪重建序列，是一种“从序列到序列”的高效重建算法。

5.1 序列切分-循环神经网络方案的基本原理

循环神经网络（Recurrent Neural Network, RNN）是用于序列信号处理的重要网络架构之一，目前在机器翻译、语音识别等领域已经取得了很好的效果。循环神经网络的发展雏形最早可以追溯到 1982 年由美国学者 John Hopfield 提出的 Hopfield 神经网络[15]，该网络的特点是具有结合存储能力并且包含了外部记忆，其设计思想为后续循环神经网络的发展奠定了基础。此后几年，用于序列信号处理的简单循环网络——Jordan 网络和 Elman 网络相继出现，同时循环神经网络的学习理论也快速发展起来，沿用至今的“实时循环学习”和“随时间反向传播学习”的网络学习方法就是在这一时期出现的。20 世纪 90 年代初，Sepp Hochreiter 发现了循环神经网络的长期依赖问题[29]，该问题会导致循环神经网络在对序列进行学习时出现梯度消失和梯度爆炸的现象，极大限制了循环神经网络对长时间跨度的序列信号的处理能力。为解决这一问题，许多研究者对相关优化理论进行了大量研究，并提出了许多行之有效的网络结构，其中就包括现在广泛应用的长短期记忆网络 LSTM[30]、门控循环单元网络 GRU[31]及许多其他网络。此后，随着深度网络的发展，循环神经网络的层数和表征泛化能力也在逐渐增加，并在语言识别、机器翻译

等领域取得了很好的效果。

目前基于 RNN 架构的网络模型已经有上万种，其中长短期记忆网络 LSTM 及其变体（包括门控循环单元 GRU、双向长短期记忆网络 BiLSTM 等）是应用最为广泛的网络类型之一。本章以 LSTM 网络及其变体为基础，完成了基于序列切分-循环神经网络方案的模型构建，并使用普通 RNN 网络作为参照进行了对比。下面分别对普通 RNN 和 LSTM 及其变体的基本原理[32]进行简要介绍。

（1）普通循环神经网络 RNN

与普通的全连接网络相比，循环神经网络 RNN 通过在网络结构中引入隐藏状态，实现了信息的记忆功能，从而更适合于含有前后相关性的序列信号的处理。普通 RNN 的原理示意图如图 5.1 所示，其中左侧是 RNN 的实际网络结构，右侧是将该网络结构按隐藏信息随时间步的传递展开后的结果。

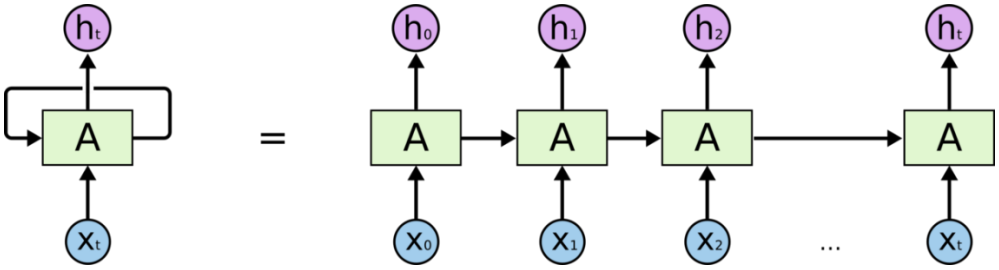

图 5.1 普通 RNN 的原理示意图

如图 5.2 所示，单个时间步的 RNN 单元结构和使用 tanh 作为激活函数的普通全连接神经网络的结构基本相同。不同在于，网络单元的输入不仅包含了当前时间步的序列信号输入，同时也包含了上一时间步的隐藏状态，而隐藏状态中则包含了序列信号的历史信息，是 RNN“记忆能力”的根本来源。RNN 的“记忆能力”与网络所包含的步长数有着密切关系，更长的输入步长往往包含着更加丰富的序列信号相关信息。但是对于普通 RNN 网络结构来说，随着步长的增加，训练中往往会出现梯度消失或者梯度爆炸的现象，也就是所谓的长期依赖问题。LSTM 和 GRU 等网络结构的出现使得长期依赖问题得到了有效解决。

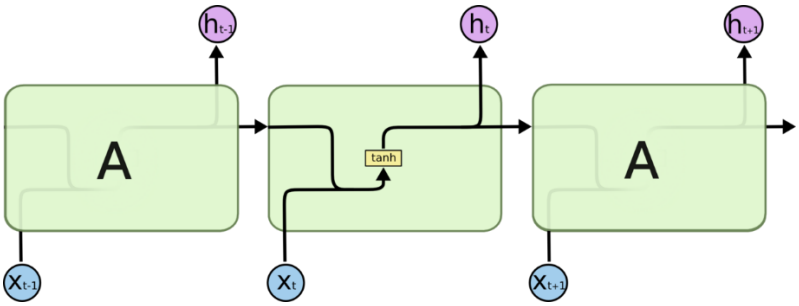

图 5.2 RNN 网络的内部结构

（2）长短期记忆网络 LSTM

LSTM 全称为 Long Short-Term Memory，即长短期记忆网络，它是循环神经网络中的一种。普通 RNN 内部仅含有一层用 tanh 激活的全连接层，且前一时间步的隐藏状态只是简单的和当前时间步的序列信号拼接后即作为当前输入，这些因素限制了隐藏信息的长期传递效率。为了解决这一问题，LSTM 对普通 RNN 的网络结构进行了优化，通过引入“遗忘门”、“输入门”和“输出门”，使得网络能够自主学习到对信息的遗忘状态，从而提高了隐藏信息的传递效率。

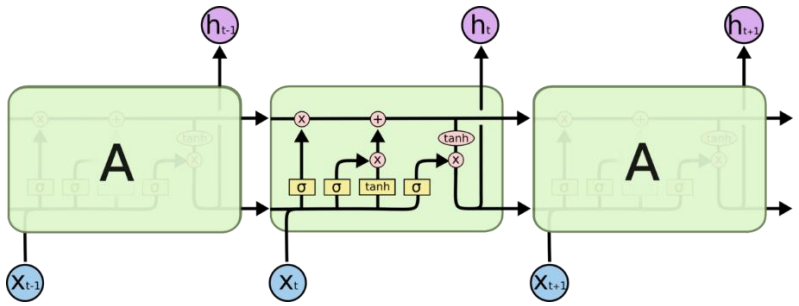

图 5.3 LSTM 网络的内部结构

如图 5.3 所示，LSTM 内部包含着两条贯穿全部时间步的横向“信息流”，其中下面一条为 “隐藏状态流”，当前时间步的隐藏状态和下一时间步的序列信号一起构成了下一时间步的输入；上面一条为“细胞状态流”，它是“细胞”——即每个时间步的网络单元对输入信息和隐藏状态进行进一步处理后得到的结果。“细胞”的详细结构如图 5.3 中的中间时间步所示。

左侧的向上分支为“遗忘门”，它决定着网络对历史信息的遗忘程度。“遗忘门”的数学表达为：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (5\text{-}1)$$

$$F_t = f_t * C_{t-1} \quad (5\text{-}2)$$

其中 W_f 和 b_f 为网络的线性参数，h_{t-1} 和 x_t 分别为上一时间步的隐藏状态和当前时间步的序列信号输入，方括号表示拼接操作，点号和星号分别表示元素相乘和矩阵相乘，σ 为 sigma 激活函数，式（5-1）的结果 f_t 为一个介于 0-1 之间的比例门控，它在式（5-2）中与上一“细胞状态”C_{t-1} 相乘，控制了对历史信息的遗忘程度。

中间的向上分支为“输入门”，它控制着对当前输入信息的传递比例。输入门的数学表达为：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5\text{-}3)$$

$$I_t = i_t * [tanh\,(W_i \cdot [h_{t-1}, x_t] + b_i)] \quad (5\text{-}4)$$

与“遗忘门”中类似，式（5-3）中上一隐藏状态和当前序列信号共同决定了输入的比例门控项 i_t，而在式（5-4）中二者则在进行线性变换后又通过 tanh 函数进行激活得到了当前的输入状态，该输入状态与比例门控相乘，控制了输入信号的传递比例。

此后，新的“细胞状态”C_t 由比例门控后的历史信息和输入信息进行加和来得到更新：

$$C_t = F_t + I_t \quad (5\text{-}5)$$

右侧的向下分支为“输出门”，它通过对当前的“细胞状态”进行进一步处理得到当前时间步下的输出，也即为当前时间步下的“隐藏状态”h_t。“输出门”的数学表达为：

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5\text{-}6)$$

$$h_t = o_t * \tanh(C_t) \quad (5\text{-}7)$$

“输出门”中输出的新的隐藏状态携带着历史信息再次回到“隐藏状态流”中，继续向下一时间步传递。

LSTM 通过引入三个可学习门控对历史信息、输入信息和输出信息进行控制，使得网络能够很好的适应不同长度序列所包含的信息，从而具有较好的长短期信息综合提取能力。

（3）门控循环单元 GRU 和双向 LSTM

GRU 全称为 Gated Recurrent Unit，即门控循环单元，它是 LSTM 的一种常用变体，其基本原理与 LSTM 类似，只是在结构上进行了进一步的简化。如图 5.4 所示，GRU 内部合并了 LSTM 的遗忘门和输入门得到了“重置门”，而原来的“输出门”则变为“更新门”。此外，GRU 也对“隐藏状态流”和“细胞状态流”进行合并得到新的“隐藏状态流”，贯穿所有时间步。

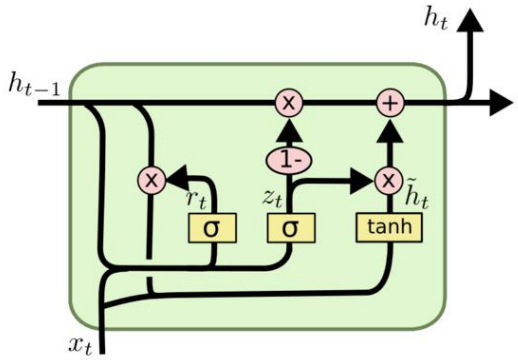

图 5.4 GRU 网络的内部结构

GRU 的数学表达如下：

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{5-8}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{5-9}$$

$$\hat{h}_t = tanh\,(W_s \cdot [r_t * h_{t-1}, x_t]) \tag{5-10}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \hat{h}_t \tag{5-11}$$

其中，W_z、W_r、W_s 均为网络参数，z_t 和 r_t 为“更新门”和“重置门”的门控比例项，x_t 为当前时间步的序列信号输入，h_t 为当前时间步的隐藏状态。从 GRU 的网络结构和数学表达可以看出，GRU 与 LSTM 相比，结构更加紧凑，从而节省了网络参数。

双向 LSTM（Bidirection LSTM, BiLSTM）是 LSTM 的另一种变体，如图 5.5 所示，它通过堆叠两个逆向的 LSTM 将原始的单向结构扩展为双向结构，从而更加有助于提取序列所包含的两个方向的信息。BiLSTM 中每个方向的网络单元与 LSTM 中相同，此处不再赘述。两个方向的 LSTM 的输出一般沿批量维度拼接作为网络的最终输出。

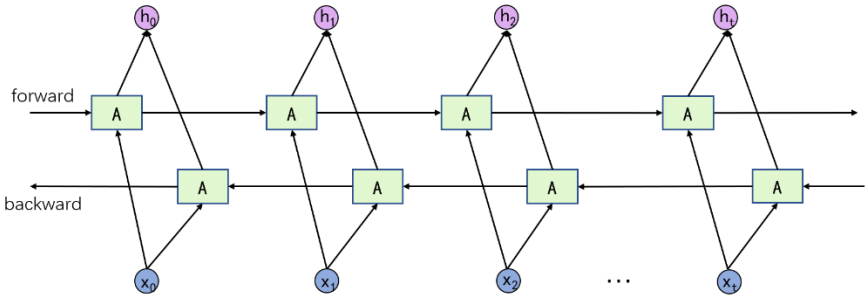

图 5.5 BiLSTM 的网络结构

5.2 序列切分-循环神经网络方案的设计与实现

5.2.1 数据集预处理与事件序列切分

本章所提出的序列切分-循环神经网络方案在 N-MNIST 事件序列数据集上进行了训练和测试，数据集的具体预处理过程如下：首先，对 N-MNIST 中的每个事件记录进行切分，以 135ms 为起点截取 512 个事件点，得到包含 60 000 个样本的训练集和包含 10 000 个样本的测试集。接着，通过时空最近邻滤波对获取的训练集和测试集进行去噪处理，去噪时，判定为信号的事件极性不变，判定为噪声的事件极性取零，将去噪得到的极性作为训练真值。然后，对训练集和测试集中每段事件序列的时间和空间坐标进行线性归一化，使其均位于 0 到 1 之间。最后，按照设定的输入时间步长，对数据集的尺寸进行变换。经过调参测试，取步长为 64 时能够满足循环神经网络对时间相关性的需求，同时具有较快的训练速度。所以，尺寸变换后，得到新的训练集尺寸为 480 000×64×4，测试集的尺寸为 80 000×64×4，其对应真值的尺寸分别为 480 000×64 和 80 000×64。

5.2.2 基于循环神经网络的网络架构设计

本章提出的序列切分-循环神经网络方案 SeqRNN 以 RNN、LSTM 及其变体为

基础，结合全连接网络实现了对输入序列的极性预测，从而能够达到去噪重建效果，网络模型的基本结构如图 5.6 所示。

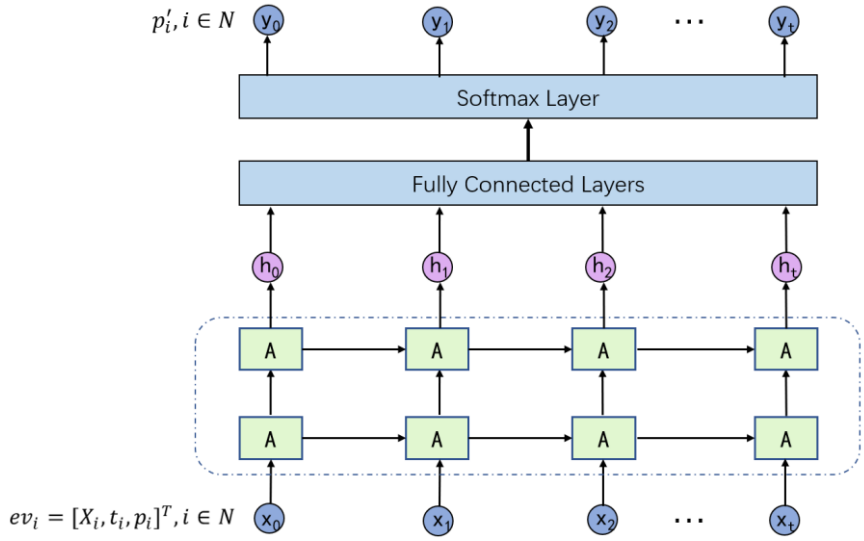

图 5.6 SeqRNN 方案网络模型示意图（以双层 RNN 为例）

图示网络模型以事件序列为输入，以极性预测为输出。其底层模块（虚线框内）为一个两层的 RNN 网络，该模块可以替换为具有相同输入输出的 LSTM、GRU 或 BiLSTM 网络。底层模块的输出连接到了一个全连接网络层进行后续处理，最后通过一个 softmax 层对输出极性进行预测。本章以该网络模型为基本框架，分别实现并测试了 RNN、LSTM、GRU、BiLSTM 的重建效果，网络配置如表 5.1 所示。

表 5.1 SeqRNN 方案的四种具体网络模型实现

layers		input size	output size
inut		64×4	64×4
bottom（4 选 1）	RNN_1	64×4	64×64
	RNN_2	64×64	64×32
	LSTM_1	64×4	64×64
	LSTM_2	64×64	64×32
	GRU_1	64×4	64×64
	GRU_2	64×64	64×32
	BiLSTM_f	64×4	64×32
	BiLSTM_b	64×4	64×32
fully connected		64×32	64×3
output		64×3	64×1

5.3 序列切分-循环神经网络方案的实验与分析

网络训练阶段选用的损失函数为交叉熵损失函数，选用的优化器为AdamOptimizer，采用随机梯度下降策略进行训练，批量大小为 128，学习率为 0.001。将基于普通 RNN、LSTM、GRU 和 BiLSTM 的四种网络模型分别在预处理后的 N-MNIST 数据集上进行训练和测试，训练平台配置为 NVIDIA GeForce GTX 1080Ti GPU，11G 显存；测试平台配置为 Core i5-7300HQ CPU，2.5GHz 主频、NVIDIA GeForce GTX 1050Ti GPU，4G 显存、网络架构基于 TensorFlow 编写。网络训练过程中的准确率变化曲线如图 5.7 所示。

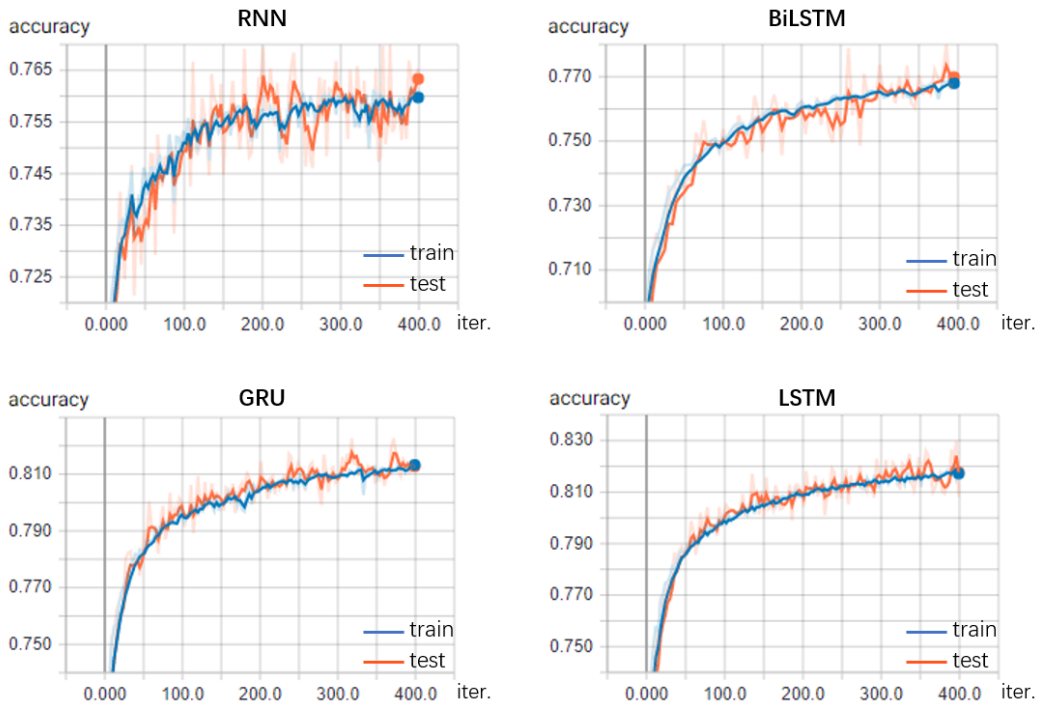

图 5.7 基于 SeqRNN 方案的四种网络模型的训练过程准确率曲线

图中的横坐标为训练迭代次数，纵坐标为事件极性预测准确率。从图中曲线可以看出，基于 SeqRNN 方案的四种网络模型经过训练之后极性预测准确率均有明显提升。训练过程中，普通 RNN 网络的准确率曲线波动较大，网络鲁棒性较差，且经过 400 个迭代后其在测试集上的预测准确率最低，为 76.3%左右；BiLSTM 网络的准确率曲线相对平坦，在测试集上的事件极性预测准确率为 77.0%左右；GRU 和 LSTM 网络的准确率曲线收敛过程最为平坦，在测试集上的预测准确率分别为 81.4%和 81.7%。综上可见，四种网络模型中 LSTM 和 GRU 模型表现较好。

利用训练完成的四种网络模型分别对测试集的事件序列进行去噪重建测试，并将去噪重建后的结果重构为二维图像序列，每个图像序列取 4 帧（纵列），结果如图 5.8 所示。

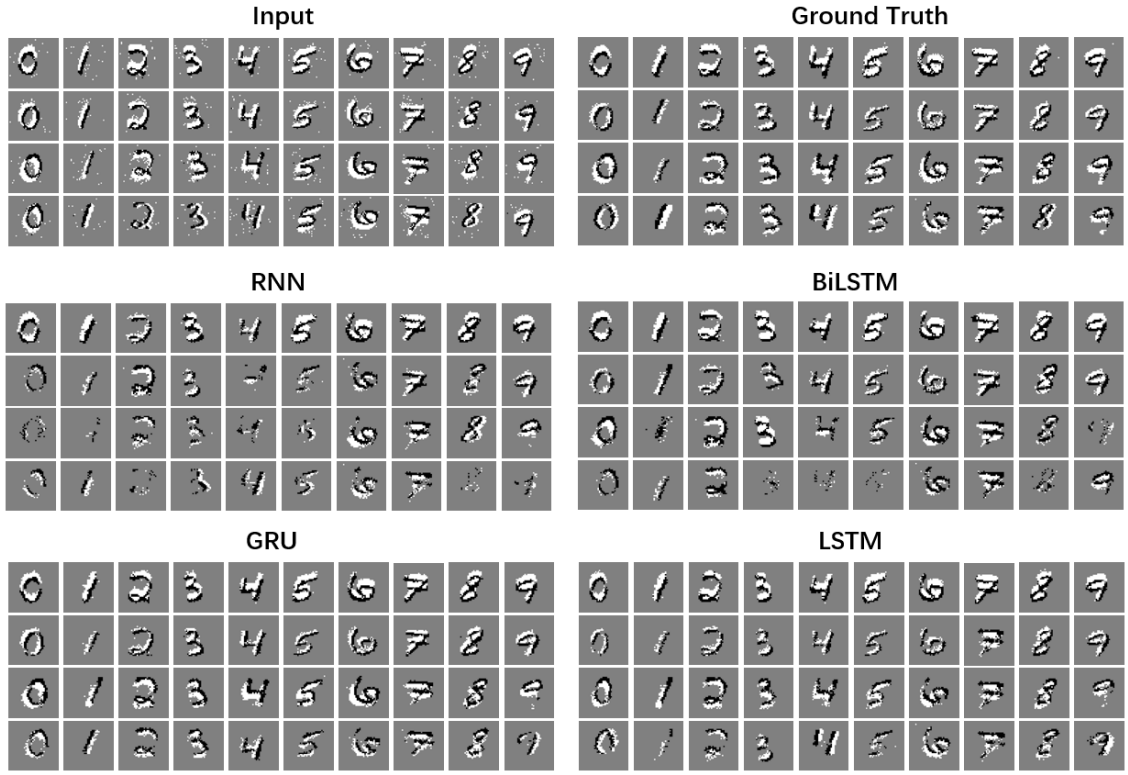

图 5.8 基于 SeqRNN 方案的四种网络模型去噪重建效果

由图 5.8 可以看出，四种网络模型对事件序列的背景噪声均具有明显的去除效果。其中，RNN 和 LSTM 在去除噪声的同时也会损失部分正常信号，断点较多。相比之下，GRU 和 LSTM 在噪声去除和信号保留方面表现更好，这也与训练结果的准确率数据相匹配。本实验的最终测试集预测准确率数据及平均事件处理速率数据如表 5.2 中所示。

表 5.2 基于 SeqRNN 方案的四种网络模型的预测准确率及平均处理速度（兆事件每秒）

网络	RNN	LSTM	BiLSTM	GRU
测试集准确率	76.3%	81.7%	77.0%	81.4%
处理速度	3.78	2.21	2.63	2.53

由于四种网络模型的任务本质为像素（事件）级的三分类（正极性、负极性和非触发）问题，且模型所用的数据集真值是通过最近邻滤波方法模拟产生的，并非实际真值，所以其整体预测准确率比常见的图像级分类预测准确率略低，但是通过重建实验可以发现，该准确率下的重建图像视觉效果已经较好，能够满足实际场景下的应用需求。处理速度方面，四种网络模型在笔记本平台上测试的平均事件处理速度均大于 2 兆事件每秒，考虑到目前的事件相机的时间分辨率为微秒量级，可以确定所提出的网络模型能够满足事件序列处理的实时性要求。

5.4 本章小结

本章以序列切分-循环神经网络方案为基础，分别基于普通 RNN、LSTM、GRU 和 BiLSTM 完成了网络模型的构建。通过测试，四种模型均实现了一定的去噪重建效果并且能够满足事件序列的实时处理需求，其中基于 LSTM 和 GRU 的网络模型达到了 80%以上的事件极性预测准确率，背景噪声去除效果明显。

本方案的主要意义在于能够保持网络对于事件序列的“透明性”，由于网络不改变原始事件序列的数据结构，而是仅对事件序列中的噪点进行去除，所以该网络模型的输出可以直接与更加贴近生物神经机制的新兴脉冲神经网络或事件驱动算法相级联，从而完成基于事件相机的端到端的高层视觉任务解决方案。

第六章 总结与展望

6.1 论文总结

事件相机是一种新型的基于神经形态学原理的异步成像相机，在无人驾驶、三维立体视觉、SLAM 等领域具有巨大的应用潜力。由于事件相机输出的原始事件序列噪声较大，限制了其在很多场景下的应用性能，因此，通过算法设计来实现事件序列的高速高质量去噪重建对发挥事件相机的独特优势具有重要价值。

现有的基于最近邻机制的事件序列滤波算法对不同环境噪声的自适应能力较差，而基于稀疏表示的去噪算法则不能够实现事件序列的实时处理。为了解决上述问题，实现事件相机效率和优势的最大化，本文开展了基于深度网络的高速高质量事件序列去噪重建研究。主要工作包括以下几个方面：

（1）通过调研和分析，详细阐述了课题背景与研究意义，具体分析了课题研究现状、研究思路及其存在的难点与不足。

（2）通过差分模拟等方式获取了事件序列数据集，并实现了基于视觉暂留机制的事件序列可视化方法，为后续工作奠定了基础。

（3）设计并实现了一种基于图像重构-卷积去噪自编码器的高质量事件序列重建方案。该方案深入挖掘了事件序列所包含的二维场景结构信息，能够较好地完成了高质量事件序列去噪重建任务，并与现有计算机视觉框架相适配。

（4）设计并实现了一种基于序列切分-循环神经网络的高质量事件序列重建方案。该方案利用了事件序列中所包含的时空相关性，能够在保持原有事件序列数据结构不变的前提下，较好地完成高质量事件序列重建任务，为后续基于脉冲神经网络和事件驱动算法的高层任务实现打下了基础。

通过利用深度网络的高效学习能力，本文提出的事件序列去噪重建方案能够自适应地调整网络参数，达到了理想、实时的去噪效果。而且，通过利用深度网络进行前端事件序列的预处理，我们很容易与后端执行高层视觉任务的神经网络级联匹配、协同工作，从而实现端到端的高效视觉任务解决方案。

6.2 工作展望

深度网络凭借在特征抽取、数据拟合、自适应学习等方面的优势，在图像处理、语音识别等许多领域取得了显著效果，而决定深度网络最终应用性能的重要因素则在于数据集质量的好坏和模型表征能力的高低。目前，神经形态视觉领域的发展刚刚起步，数据资源，特别是有标注或真值的数据资源十分稀缺，这也是限制本文网络模型训练效果的一个主要因素。另一方面，目前现有的计算机视觉网络架构主要是基于传统成像体系进行设计的，其对脉冲事件信号的处理和表征能力的不足也从理论上限制着本文事件序列去噪重建方案的表现。

近几年来，随着神经、认知科学与类脑智能领域的发展，更贴近人类视觉与认知机理的神经形态视觉及第三代人工神经网络——脉冲神经网络也取得了很大进步。相信随着该领域的进一步发展，越来越多的优质数据资源和适合于脉冲事件信号处理的新兴网络架构也将会出现，并将进一步促进现有高质量事件序列去噪重建方案效果的提升。

致谢

本论文是在清华大学自动化系索津莉副教授和西安电子科技大学电子工程学院李洁教授的悉心指导下完成的。从论文的最初选题到最终定稿，两位老师都给予了我极大的帮助。索老师平易近人、治学严谨，在软硬件资源等方面给予了我很大支持，并在课题上对我进行了耐心指导，使得本论文得以最终完成。李老师学识渊博、循循善诱，不仅在毕业设计期间给予了我很多的帮助，而且也在本科阶段的学习中帮我打下了良好的基础。在此向索老师和李老师表示衷心的感谢。

毕业设计期间，清华大学自动化系 BBNC 实验室的师兄师姐也给予了我许多耐心的指导和帮助，让我取得了很大的进步，在此向他们表示感谢。同时，也要感谢我的任课老师、辅导员和同学们四年来对我的关心和支持。

最后，特别要感谢我的室友，感谢他们四年的陪伴，让我度过了大学中许多值得回忆的时光；感谢我的父母、家人和朋友，他们无私的支持是我坚强的后盾和不懈努力的动力。研究生期间，我也会继续努力，期待能够在科研和学术的道路上走得更远！

参考文献

[1] Liu H, Brandli C, Li C, et al. Design of a spatiotemporal correlation filter for event-based sensors[C]//2015 IEEE International Symposium on Circuits and Systems (ISCAS). Lisbon, Portugal: IEEE, 2015: 722–725.

[2] Czech D, Orchard G. Evaluating noise filtering for event-based asynchronous change detection image sensors[C]//2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob). Singapore, Singapore: IEEE, 2016: 19–24.

[3] Xie X, Du J, Shi G, et al. An Improved Approach for Visualizing Dynamic Vision Sensor and its Video Denoising[C]//Proceedings of the International Conference on Video and Image Processing - ICVIP 2017. Singapore, Singapore: ACM Press, 2017: 176–180.

[4] Padala V, Basu A, Orchard G. A Noise Filtering Algorithm for Event-Based Asynchronous Change Detection Image Sensors on TrueNorth and Its Implementation on TrueNorth[J]. Frontiers in Neuroscience, 2018, 12: 14.

[5] Mahowald Misha, Douglas Rodney. A silicon neuron[J]. Nature, 1991, 354(6354): 515–518.

[6] Kleinfelder S, Lim S, Liu X Q, et al. A 128×128 120dB 30mW asynchronous vision sensor that responds to relative intensity change[C]//2006: 10.

[7] Lichtsteiner P, Posch C, Delbruck T. A 128×128 120 dB 15us Latency Asynchronous Temporal Contrast Vision Sensor[J]. IEEE Journal of Solid-State Circuits, 2008, 43(2): 566–576.

[8] Posch C, Matolin D, Wohlgenannt R. A QVGA 143 dB Dynamic Range Frame-Free PWM Image Sensor With Lossless Pixel-Level Video Compression and Time-Domain CDS[J]. IEEE Journal of Solid-State Circuits, 2011, 46(1): 259–275.

[9] Berner Raphael, Brandli Christian, Yang Minhao, et al. A 240x180 120dB 10mW 12us-latency sparse output vision sensor for mobile applications[R]. Intl. Image Sensors Workshop, Snowbird Resort, UT, USA: 2013.

[10] Guo M, Ding R, Chen S. Live demonstration: A dynamic vision sensor with direct logarithmic output and full-frame picture-on-demand[C]//2016 IEEE International Symposium on Circuits and Systems (ISCAS). Montréal, QC, Canada: IEEE, 2016: 456–456.

[11] Schmidhuber J. Deep learning in neural networks: An overview[J]. Neural Networks, 2015, 61: 85–117.

[12] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. The Bulletin of Mathematical Biophysics, 1943, 5(4): 115–133.

[13] Rosenblatt F. The perceptron: A probabilistic model for information storage and

organization in the brain.[J]. Psychological Review, 1958, 65(6): 386–408.

[14] Minsky M, Papert S A. Perceptrons: An Introduction to Computational Geometry[M]. MIT Press, 2017.

[15] Hopfield J J. Neural networks and physical systems with emergent collective computational abilities.[J]. Proceedings of the National Academy of Sciences, 1982, 79(8): 2554–2558.

[16] Hinton G E, Sejnowski T J. Learning and relearning in Boltzmann machines, Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations[M]. MIT Press, Cambridge, MA, 1986.

[17] Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation[R]. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[18] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(5): 359–366.

[19] Hinton G E, Osindero S, Teh Y-W. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006, 18(7): 1527–1554.

[20] Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks[M]. 2010, 15.

[21] Tan C, Lallee S, Orchard G. Benchmarking neuromorphic vision: lessons learnt from computer vision[J]. Frontiers in Neuroscience, 2015, 9: 374.

[22] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324.

[23] Orchard G, Jayawant A, Cohen G K, et al. Converting Static Image Datasets to Spiking Neuromorphic Datasets Using Saccades[J]. Frontiers in Neuroscience, 2015, 9.

[24] Li H M, Liu H C, Ji X Y, et al. CIFAR10-DVS: An Event-Stream Dataset for Object Classification[J]. Frontiers In Neuroscience, 2017, 11.

[25] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th international conference on Machine learning - ICML '08. Helsinki, Finland: ACM Press, 2008: 1096–1103.

[26] Sakib S, Ahmed, Jawad A, et al. An Overview of Convolutional Neural Network: Its Architecture and Applications[M]. 2018.

[27] Masci J, Meier U, Cireşan D, et al. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction[G]//Honkela T, Duch W, Girolami M, et al. Artificial Neural Networks and Machine Learning – ICANN 2011. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, 6791: 52–59.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern

Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770–778.

[29] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 1994, 5(2): 157–166.

[30] Sepp Hochreiter, J urgen Schmidhuber. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735–1780.

[31] Cho K, van Merrienboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1724–1734.

[32] Christopher Olah. Understanding LSTM Networks[EB/OL]. colah's blog, 2015-08-27. (2015-08-27). http://colah.github.io/posts/2015-08-Understanding-LSTMs/.