\_欧阳玉宋worklog

# 9.23-9.30

## 当前任务

了解snn的基本原理，结构，设计。

了解田永红老师的团队工作，

了解ann based的snn加购，

了解以qkformer为代表的snn transformer等论文。

田老师有做行为识别的工作 要额外注意。

## 执行

### SNN基本原理

模拟大脑神经细胞，包含神经元、突触和网格结构三部分

胞体积累脉冲信号，当积累的脉冲信号超过阈值之后发出脉冲（脉冲与脉冲之前并无明显差别，因此信息主要依靠脉冲频率和时间来记录。这里描述并不准确），另外胞体所累积的脉冲信号还会逐渐下降，叫做泄漏

不应期：发放一次脉冲后会有一段时间无法再发放脉冲，可以简单理解为冷却cd

对于脉冲信号累计的过程，就是简单的突触连接权重相加

网格结构暂不清楚（好像有传统的层级结构，还有一些注重生物合理性即仿大脑网格结构的以模块为单元的复杂网格结构）

SNN有强化学习、无监督学习和自监督学习

其中无监督学习主要通过STDP来实现，STDP简单来说在一个神经细胞发生脉冲信号前有突出传入脉冲信号，则认为该突触连接是有效的，则突触权重加强（LTP）；相反如果在发生脉冲信号后有突触传入脉冲信号则认为该突触连接是无效的，则减弱该突触连接（LTD）。 并且这个权重加强减弱的程度是与时间差密切相关的。

### 结构

传统层状结构（工程化任务）和类脑拓扑结构（生物合理性）

**设计（不太理解这个设计是指什么）**

神经元模型：LIF、EIF、SRM、Hodgkin-Huxley（HH）

### ANN based SNN：

一种是先训练好ANN再转换成SNN，会损失一些SNN特性，感觉单纯为了低功耗

另外一种是直接训练，但是会基于一些ANN的训练方法甚至结构

### 田永鸿老师团队相关工作

Brain-like and Deep Computing:

* Neural inversion computing
* Deep learning for video analysis
* Biologically simulating for human vision system

Multimedia Big Data

### 田老师动作识别相关论文：

Yemin Shi, Wei Zeng, Tiejun Huang, Yaowei Wang∗, [Learning Deep Trajectory Descriptor for Action Recognition in Videos using Deep Neural Networks](http://mlg.idm.pku.edu.cn/wp-content/uploads/2013/07/2015_ICME_DTD.pdf), ***Proc.******Int’l Conf. Multimedia and Expo******(ICME 2015)***, Torino, Italy.

Zhengying Chen, Yonghong Tian\*, Wei Zeng and Tiejun Huang, [Detecting Abnormal Behaviors in Surveillance Videos Based on Fuzzy Clustering and Multiple Auto-Encoders](http://mlg.idm.pku.edu.cn/wp-content/uploads/2013/07/2015_ICME_FMAE.pdf), ***Proc.******Int’l Conf. Multimedia and Expo******(ICME 2015)***, Torino, Italy.

都是ANN而不是SNN，大概看了一下处理逻辑，没有细看

### 一些数据集

数据形式：RGB、keleton , depth , infrared sequences , point clouds 、event streams.

DVS：

HARDVS：<https://github.com/Event-AHU/HARDVS?tab=readme-ov-file>

UCF-Crime-DVS：[GitHub - YBQian-Roy/UCF-Crime-DVS](https://github.com/YBQian-Roy/UCF-Crime-DVS)

**Celex-HAR：**<https://github.com/Event-AHU/CeleX-HAR>

**DailyDVS-200：**[**GitHub - QiWang233/DailyDVS-200: [ECCV-2024] DailyDVS-200: A Comprehensive Benchmark Dataset for Event-Based Action Recognition**](https://github.com/QiWang233/DailyDVS-200)

RGB：

**UCF101：**[UCF101 – Action Recognition Data Set – Center for Research in Computer Vision](https://www.crcv.ucf.edu/research/data-sets/ucf101/)

**HMDB51** ：[数据集-OpenDataLab](https://opendatalab.org.cn/OpenDataLab/HMDB51)

**Kinetics** ：[GitHub - cvdfoundation/kinetics-dataset](https://github.com/cvdfoundation/kinetics-dataset)

**AVA** ：[GitHub - cvdfoundation/ava-dataset: The AVA dataset densely annotates 80 atomic visual actions in 351k movie clips with actions localized in space and time, resulting in 1.65M action labels with multiple labels per human occurring frequently.](https://github.com/cvdfoundation/ava-dataset)

**KTH** ：[GitHub - vkhoi/KTH-Action-Recognition: Action Recognition on the KTH Dataset](https://github.com/vkhoi/KTH-Action-Recognition)

RGB+depth：

**NTU RGB+D** ：[GitHub - shahroudy/NTURGB-D: Info and sample codes for "NTU RGB+D Action Recognition Dataset"](https://github.com/shahroudy/NTURGB-D)

SL-Animals-DVS, N-LSA64, DVS128 Gesture

and DailyAction-DVS,

### snn transformer

qkformer：能大概理解处理过程，不过还是有很多不理解的地方

Spikformer

# 10.1-10.8

## 当前任务

广泛调研

## 模型网络架构调研

1. **Spike-HAR++: an energy-efficient and lightweight parallel spiking transformer for event-based human action recognition**

输入数据：dvs

网络架构：

图示

AI 生成的内容可能不正确。

原始 DVS 事件帧 → 输入预处理 → PE 块（Patch Embedding）→ 并行脉冲 Transformer 块（多组堆叠）→ 注意力分支（动态掩码）→ SNN 分类头 → 行为类别输出

各模块详细说明：

**1. 输入预处理模块**

**（1）模块组成**

* DVS 事件帧格式转换：将 DVS 相机输出的 “离散事件流” 整合为**2D 事件帧**
* 随机裁剪与标准化：对事件帧进行空间裁剪（减少冗余），并标准化特征分布。

**（2）核心作用**

* 适配 SNN 输入：DVS 原始输出为 “离散事件”，需转化为 SNN 可处理的 “帧结构”；
* 降低数据复杂度：通过裁剪和标准化，减少后续模块的计算量，符合 SNN “低能耗” 设计目标。

**（3）Problem**

* DVS输出为离散事件，不正适合SNN处理吗？为什么还要转换帧？我记得没看到论文中提到，估计ai瞎编的吧？？

**2. PE 块（Patch Embedding 块）—— 基础时空特征提取与脉冲转换**

**（1）模块组成**

4 个 Conv2D（2D 卷积层） + 3 个 BN（批归一化层） + 2 个 MP（最大池化层） + 3 个 SNN 层（基于 LIF 神经元） + 膜电位残差连接。

**（2）核心作用**

* **空间特征提取**：通过 Conv2D 提取 DVS 帧中的 “局部空间特征”（如动作的边缘、手部轮廓）
* **空间维度压缩**：2 个 MP 层将空间尺寸下采样，大幅减少计算量；
* **连续特征→脉冲转换**：3 个 SNN 层（核心为 LIF 神经元）将 Conv2D/BN 输出的 “连续特征” 转化为 SNN 专属的 “脉冲信号”，最终输出**时空脉冲令牌（SPE）**；
* **避免膜电位衰减**：膜电位残差连接（将 SNN 层输出（SPE）与初始卷积层输出IPE相加），解决 SNN 深层传递中 “脉冲信号衰减” 问题。

**（3）Problem**

* 这里的卷积层、归一化层和池化层是SNN版本的，还是单纯的ANN处理方式？？
* 架构图中归一化层处理后要接由SNN层再由池化层处理，这里ai说SNN层的作用是将连续特征转换为脉冲信号但是后续的池化层如何处理脉冲信号？？？
* PE块中模块设计成这个样子的原因是什么？？不断的实验尝试出来的吗？？？为什么要按照这样一个顺序去安排各个模块？？？
* 膜电位残差连接的作用还是缺乏理解

**3. 并行脉冲 Transformer 块 —— 核心特征精细化（模型效率关键）**

**（1）模块组成**

LayerNorm（层归一化） + 并行的 “CB-S3A 块（卷积基简化脉冲自注意力）” 与 “MLP 块（多层感知机）” + 膜电位残差连接。

**（2）核心作用**

* **简化注意力计算，适配 SNN 低能耗需求**：CB-S3A 块用 “卷积操作替代传统 Transformer 的自注意力矩阵乘法”，将计算复杂度从\(O(ND^2)\)降至\(O(N)\)（N为特征数量，D为通道数），同时基于 “脉冲信号” 建模帧间 “局部 - 全局关联”（如挥手时手臂的连贯运动）；
* **局部特征强化**：MLP 块通过 “卷积 + 激活” 进一步加工特征，且会用 “注意力分支生成的脉冲发放率图” 做 “掩码操作”（仅保留动作核心区域特征，过滤背景噪声）；
* **并行计算提效**：CB-S3A 与 MLP 块 “同时接收 LayerNorm 处理后的特征” 并独立计算，避免串行耗时，符合 SNN “实时性” 需求；
* **特征传递稳定**：最终将 CB-S3A 与 MLP 的输出求和，再通过膜电位残差连接与模块输入相加，防止深层 SNN 特征丢失。

**（3）Problem**

* Transformer做了哪些优化？？

**4. 注意力分支 —— 动态聚焦动作核心区域（Spike-HAR 与 Spike-HAR++ 核心差异）**

**（1）模块组成（分两种实现）**

* **Spike-HAR 的注意力分支**：原始 DVS 帧 → 第一次 SAR 操作（时间求和→平均→重复）→ 两轮 Conv + 下采样 → 第二次 SAR 操作 → 脉冲发放率图（注意力掩码）；
* **Spike-HAR++ 的注意力分支**：PE 块输出的\(S\_{PE}\)（时空脉冲令牌）→ 第一次 SAR 操作 → 两轮 LIF-Conv-BN（SNN 专属特征加工）→ 第二次 SAR 操作 → 多通道注意力掩码。

**（2）核心作用**

* **降低背景噪声干扰**：通过 “时间维度聚合（SAR 操作）” 生成 “脉冲发放率图”，为 “脉冲信号密集的动作区域” 分配高权重，过滤背景噪声（如灯光波动）；
* **适配长时间动作识别**：
  + Spike-HAR 局限：依赖 “原始 DVS 帧求和”，长时间动作的 “时序细节（如挥手的‘抬手 - 摆臂 - 落手’）” 会被平均模糊，无法适配长动作；
  + Spike-HAR++ 改进：用\(S\_{PE}\)（已提取深层时空特征的脉冲令牌）替代原始帧，且新增 “LIF-Conv-BN” 加工，保留长动作的 “动态关联特征”，大幅提升长动作识别精度。

**（3）Problem**

* 短时间注意力机制很简单，能够理解。对于长时间动作识别，为什么这样更改就可以，不太理解
* 二次SAR操作没看懂

**5. SNN 分类头 —— 最终行为类别输出**

**（1）模块组成**

全局平均池化（GAP） + 全连接 SNN 层（基于 LIF 神经元）。

**（2）核心作用**

* **特征降维**：全局平均池化将 “多通道脉冲特征” 压缩为 “单通道全局特征”，减少最终计算量；
* **类别映射**：全连接 SNN 层将 “全局脉冲特征” 映射到 “行为类别空间”，输出每个类别的概率，完成识别任务。

**（3）Problem**

* 不了解具体细节

**Problem：**

* 对于很多内容的理解缺乏细节和经验支撑
* 另外不太懂这些架构设计的依据
* 不理解SNN在其中起到的作用

**2. SVFormer: A Direct Training Spiking Transformer for Efficient Video Action Recognition**

图示

AI 生成的内容可能不正确。

**1. 输入预处理模块**

**（1）模块组成**

* 帧维度重组：将原始 B×T×3×H\_in×W\_in重组为 T×B×3×H\_in×W\_in，按 “时间步（帧）” 逐帧输入 SNN（帧 - by-frame 处理）；
* 数据增强：随机裁剪、水平翻转、随机擦除（提升泛化性）；
* 空间标准化：将帧尺寸统一为 224×224（RGB 数据）或 128×128（DVS 数据），避免后续模块尺寸不匹配。

**（2）核心作用**

* 适配 SNN “时间步驱动” 逻辑：SNN 需按时间步（T）处理信号，重组后每一时间步输入 1 帧，T 同时对应 SNN 的 “模拟时间步”（ANN 无时间步概念）；
* 降低数据冗余：通过裁剪和标准化减少无效计算，契合 SNN “低能耗” 设计目标。

**（3）Problem**

* 无

**2. 骨干网络（4 阶段分层结构）**

骨干网络是 SVFormer 的核心，分 4 个阶段，每阶段由 “Patch Embedding（PE）模块”+“特征提取模块（LFE/GSA）” 组成，且通过**膜电位残差连接**避免深层 SNN 特征衰减。

**2.1 Patch Embedding（PE）模块 —— 特征降维与脉冲转换**

**（1）模块组成**

* 第 1 个 PE：Conv2D（卷积）+ BN（批归一化）；
* 第 2-4 个 PE：SN（Spiking Neuron，脉冲神经元）+ Conv2D + BN（简写为 \(X\_{PE\_{out}} = ConvBN(SN(X\_{PE\_{in}}))\)）。

**（2）核心作用**

* **空间下采样**：通过 Conv2D（步长 = 2）将特征图空间尺寸逐步缩小（如从 224×224→112×112→56×56→28×28），减少后续计算量；
* **通道维度扩展**：将输入 3 通道（RGB）逐步提升至 128→256→384→512 通道（\(C=[128,256,384,512]\)），增强特征表达能力；
* **脉冲转换（第 2-4 个 PE）**：SN 层将 Conv2D 输出的 “连续特征” 转化为 “脉冲信号”，确保后续模块均处理 SNN 专属的脉冲数据（第 1 个 PE 无 SN，因输入为原始帧，需先通过 Conv 提取基础特征）。

**（3）Problem**

* **其它都好理解，就是这个SN**层的作用看不懂

**2.2 局部特征提取器（LFE）—— 前两阶段的局部空间特征提取**

**（1）模块组成**

3 个级联 Conv 层（PWConv-DWConv-PWConv，类似 MobileNet）+ MLP 模块 + 膜电位残差连接；

* Conv 层：前加 1 个 SN 层（将输入转为脉冲），PWConv（点卷积）负责通道融合，DWConv（深度卷积，5×5 核）负责局部空间特征提取；
* MLP 模块：2 个 “SN-Linear-BN” 单元（先扩通道 2 倍，再缩回原尺寸）。

**（2）核心作用**

* **提取局部空间特征**：通过 DWConv 捕捉视频帧内的局部细节（如动作的边缘、肢体局部形态），弥补 Transformer 自注意力 “局部感知弱” 的缺陷；
* **降参增效**：PWConv+DWConv 比普通 Conv 参数减少 80% 以上，契合 SNN “低能耗” 需求；
* **残差稳定训练**：膜电位残差连接将 LFE 输入与输出相加，避免深层 SNN 膜电位衰减（ANN 残差连接处理连续值，SNN 残差处理膜电位 / 脉冲信号）。

**（3）Problem**

* **无**

**2.3 全局自注意力（GSA）模块 —— 后两阶段的全局时空关联捕捉**

**（1）模块组成**

SSA（Spiking Self-Attention，脉冲自注意力）模块 + MLP 模块（同 LFE 的 MLP） + 膜电位残差连接；

**（2）核心作用**

* **捕捉全局时空关联**：通过 SSA 建模帧间长距离依赖（如挥手动作中 “抬手→摆臂→落手” 的时序关联），解决 CNN “局部感知局限”；
* **脉冲特征强化**：MLP 模块进一步加工注意力输出的脉冲特征，提升动作区分度。

**（3）Problem**

* **缺乏对SSA的深入理解**

**3. 局部路径（LP）—— 多尺度特征融合**

**（1）模块组成**

级联 “SN-DWConv-PWConv-BN” 单元，输出与最后一个 GSA 模块的输出尺寸一致（通道数相同，空间尺寸匹配）。

**（2）核心作用**

* **补充局部细节**：从第 3 个 PE 的输出提取局部特征，弥补后两阶段 GSA 模块 “重全局、轻局部” 的缺陷；
* **多尺度融合**：将 LP 输出与最后 GSA 输出沿通道维度拼接，让分类头同时利用 “全局时空关联” 和 “局部细节”。

**（3）Problem**

* **No Problem**

**4. 分类头（CH）—— 动作类别预测**

**（1）模块组成**

SN 层 + 3D Depthwise Conv + BN + Linear 层；

**（2）核心作用**

* **时空特征整合**：3D Depthwise Conv（仅作用于时空维度）捕捉 “空间特征 + 时间动态” 的联合信息（如动作在不同帧的空间位置变化），避免简单平均丢失时序细节；
* **脉冲到类别映射**：Linear 层将整合后的脉冲特征映射到类别空间，输出每个动作的概率。

**（3）Problem**

* **不懂分类头内部细节**