基于动态原型在视频异常检测的元学习方法

摘要

基于自编码器 (AE) 的当前帧重构或未来帧预测是一类流行的视频异常检测方法。使用正常数据训练的模型，异常场景的重建误差通常比正常场景的重建误差大得多。以前的方法将记忆库引入 AE，用于在训练视频中编码不同的正常模式。然而，以前的方法会消耗较大的内存存储特征向量，并且无法处理测试数据中未见过的新场景。在这项工作中，本文提出了一个动态原型单元 (DPU) 来实时地将正常的动态行为编码为原型，而不需要额外的内存开销。此外，本文将元学习的方法引入到 DPU 中，形成了一个新的少量正常学习器，即元原型单元 (MPU)。它通过只消耗几次更新迭代来实现对新场景的快速适应能力。在各种基准上进行了广泛的实验。优于最先进技术的性能证明了本文提出的方法的有效性。在原有的模型基础上，本项工作还改进了原型的生成机制与帧的深度预测方案，使得方法的性能得到提升。

关键词： 元学习；动态原型；视频异常检测；注意力机制

1 引言

视频异常检测是指检测目标的动作或行为模式是否与当前场景的状态相符合 [3] [15] [2]。在过去的几年中，监控摄像头在各个场景下的重要程度日渐提高，这使得越来越多人关注视频异常检测的方法。这些监控场景下的应用涉及到公共安全的问题，因此必须考虑到视频异常检测方法的性能和适应多个场景的能力。然而，异常这个定义是模糊的，由于在不同场景下产生的异常的情况各不相同，因此想要将所有可能出现的异常情况作为数据提取出来是不可能完成的任务。因此，异常检测通常被表述为一个无监督学习问题，旨在学习一个仅使用正常数据学习规则模式的模型。在推理过程中，与编码规则不一致的模式被视为异常。

深度自编码器（AE）方法是视频异常检测中流行的方法，其方法是将正常模式下的一系列帧作为输入，并对当前帧进行重构 [11] [17] 或者对未来帧进行预测 [9]。由于输入的异常是在正常场景下没用出现过的行为，因此会产生较大的误差，通过误差的大小来判断是否发生异常。一方面，现有的方法依赖于大量的正常训练数据来建模共同的正常模式。由于卷积神经网络 (CNN) 强大的表示能力，这些模型会面临 “过度泛化” 的问题，即无论是正常的视频帧还是异常的视频帧都可以很好地预测。为解决上述问题，以前的方法提出使用记忆库对普通训练视频中的共同的正常模式进行建模 [6] [2]，以此促进正常区域的检测能力并抑制异常区域的检测。但是，这类方法非常消耗内存空间存储正常模式的特征向量。

为了解决这一限制，把本文提出对正常动态进行编码作为一种注意力机制。本文提出的动态原型单元 (Dynamic Prototype Unit, DPU) 是一种针对正常模式的学习器，可以方便地

集成到自编码器的主干中。DPU 以连续帧的编码作为输入，学习各种不同正常模式的行为作为动态原型。具体的操作为，在自编码器编码形成编码图的过程中添加了一个较为新颖的注意力机制模块，该操作为图像中的每一个像素都赋予一个正常权值，在这些正常权值的引导下形成的局部编码向量即为学习到的原型。最终，将自编码器映射的特征与原型重构的编码特征进行聚合，用于后期的帧预测。

另一方面，在各个场景中可能出现的正常模式也各不相同，并且当前场景下的正常模式处于另外一个场景可能会转换为异常模式 [14]。例如，一辆在马路上行驶的汽车是正常的，而在人行道上行驶的汽车就是异常。以前的方法假设训练视频中的正常模式与 VAD 无监督设置下的测试场景的正常模式一致。但是这种假设是不可靠的，特别是在实际应用中，监控摄像头安装在各种不同场景的地方，无法将之前训练好的视频异常检测模型直接运用到新的场景。因此，迫切需要开发一种具有自适应能力的异常检测器。

基于上述问题，本文从一个新的角度来解决这个问题，即视频异常检测的小样本学习。在小样本学习中，在训练过程中可以使用多个场景的视频帧，在推理过程中可以使用目标场景的视频帧 [11]。这类问题的解决方案是使用元学习技术。在此元学习的训练阶段，通过少量帧数和参数更新迭代，训练少量视频帧的目标模型以适应新的场景。使用来自不同场景的视频数据重复该过程以获得模型初始化，这是快速适应新场景的良好起点。因此，本文将 DPU 模块制定为一个少量的常态学习器，即元原型单元 (Meta Prototype Unit, MPU)，其目的是学习目标场景中的常态。与其通过调整整个网络粗略地转移到新的场景，这可能会导致“过度泛化”问题，本文提出冻结预训练的自编码器的思想，只更新 MPU 的参数。本文提出的元学习模型仅消耗少量参数和更新迭代，具有快速有效地适应未知场景常态的能力。

2 相关工作

2.1 视频异常检测方法

由于视频中异常数据数量较少且昂贵的标注成本，视频异常检测被分为几种类型的训练问题。在无监督的设置下仅有正常数据作为数据 [13] [10]，在弱监督的设置下会有少量的带标签的视频数据作为输入 [15] [22]。由于在实际运用场景中无监督方法更加贴切，因此本文主要关注无监督学习下的方法。早期的方法有基于稀疏编码 [4] [10]，混合动态纹理 [20]，马尔科夫随机场 [7] 等方法。由于早期的手工特征提取的方法需要较多的先验知识，且难以迁移到其他场景使用，深度学习方法的流行特别是 CNN 方法，取代了早期的传统手工特征提取方法。在 [13] 中，Luo 等人提出的时间稀疏编码的方法能运用到 RNN 框架中。

最近，许多方法利用深度自动编码器 (deep Auto-Encoder, AE) 来建模规则模式并重建视频帧 [1] [6]。为了结合时空信息进行视频异常检测，许多作者开发了多种 AE 变体。在 [12] 中，作者研究了长短期记忆 (LSTM) 在视觉中的时间序列模式进行建模。Liu 等人 [9] 提出用 AE 和生成对抗网络 (GAN) 预测未来帧。这是基于一个假设，即视频序列中的异常帧是不可预测的。与基于重建的方法相比，该方法取得了更好的性能。然而，这种方法存在“过度泛化”的问题，即有时异常帧也可以像正常帧一样被很好地预测 (即预测误差小)。

Gong 等人 (MemAE) [11] 和 Park 等人 (LMN) [17] 在模型中引入了一个内存记忆库来进行异常检测。他们将训练视频中的正常模式的特征向量进行存储，这带来了额外的内存开销。与 [6] [17] 中使用预先定义的规则查询并更新记忆内存库来记录训练数据中的粗略模式相比，

本文提出的注意机制来衡量正常模式类别的原型，这个学习过程的消耗是非常小的，并且学习原型的过程是动态学习的，能够适应当前场景的空间和时间的优点。此外，在推理过程中，原型是基于实时视频数据自动导出的，而不是训练阶段收集的记忆的特征项 [6] [17]。为了适应测试场景，Park 等 [17] 进一步扩展了记忆内存库的更新规则，使用阈值来区分异常帧并记录正常模式。然而，在各种场景下不可能找到一个统一的、最优的阈值来区分正常帧和异常帧。相反，本文在 DPU 模块中引入了元学习技术，使其能够快速适应新环境。

2.2 注意力机制

注意机制在许多计算机视觉任务中被广泛采用。目前的方法大致可以分为两类，即通道型注意 [21] [18] 和空间型注意 [5] [8]。Wang 等人 [19] 在 CNN 的中间特征提取阶段提出 trunk-mask 注意力。然而，大多数先前关注的模块都集中在优化主干以此进行特征学习和增强。本文提出利用注意机制对空间局部向量编码并生成对正常模式的编码的原型。

2.3 小样本学习和元学习

在小样本学习任务中，研究目标是模仿人类的快速灵活学习能力，即只需要少量的数据样本就能快速适应新的场景任务 [11]。元学习的发展就是为了解决这个问题。这类方法可以通过多任务间的元更新机制快速适应新任务。

3 本文方法

3.1 本文方法概述

带有动态原型单元的自编码器模型如图 1所示，DPU 学习并压缩实时序列中信息生成多个正常模式的动态原型，并用动态信息丰富输入序列的编码。首先，将连续的视频帧作为自编码器的输入 $(I_{\mathrm{k-T+1}}, I_{\mathrm{k-T+2}}, \cdots, I_{\mathrm{k}})$，简单记为 x_k。接着选择自编码器中的隐藏编码作为 DPU 模块的前馈输入。最终将 DPU 的结果应用在自编码器剩余的层中并对帧进行预测，预测的结果为 $y_k = I_{\mathrm{k+1}}$。因此，本文将 (x_k, y_k) 记为 k 时刻的输入输出序列对。

DPU 的前向传递是通过完全可微分的注意力方式生成动态原型池，然后通过检索原型重建正常编码，最终将输入编码与正常编码聚合为输出来实现的。整个过程可分为注意、融合和检索三个子过程，整体的模型框架如图 1所示：

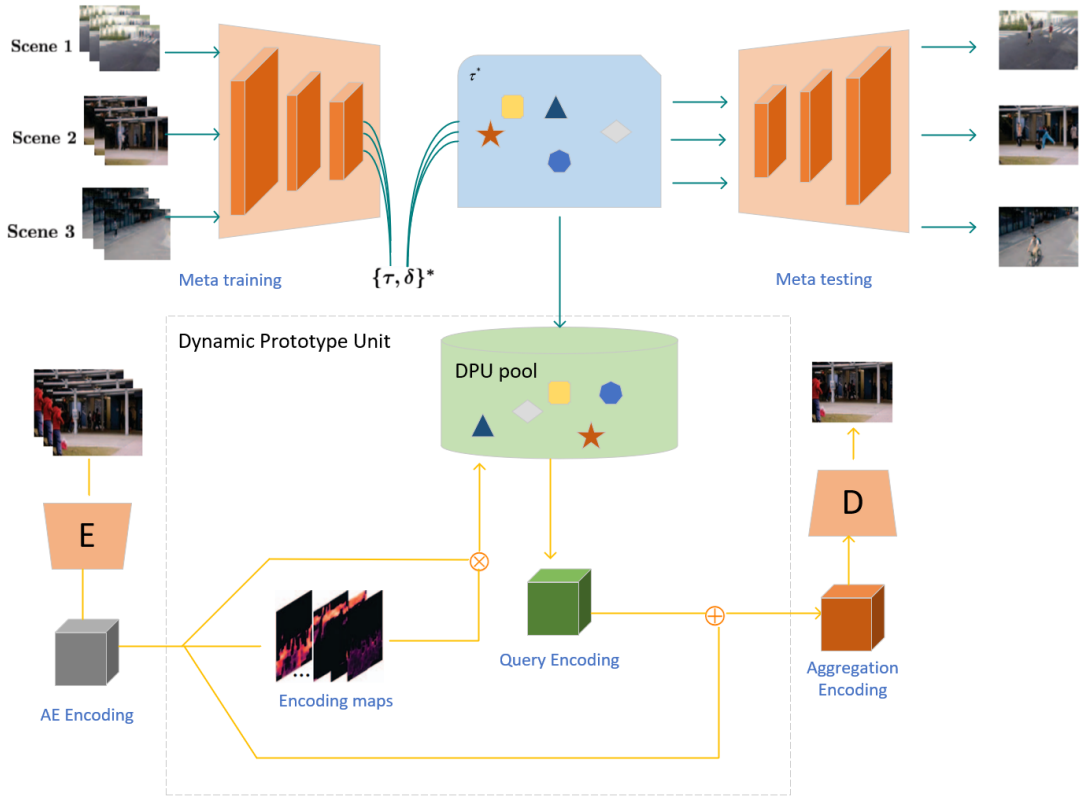

图 1. 模型整体框架，上半部分为训练原型部分可以用于元学习，下半部分是具体训练模式包括注意力机制、聚合、检索三部分。

详细的过程如下：将输入的第 t 个序列，从 AE 中的编码器中提取生成编码图 $X_t = f_\theta(x_t)$。在注意力机制的子过程中，通过注意力图生成函数生成 M 个正常模式的编码向量，即 $\omega_t^{n,m} \in W_t^m = \psi_m(X_t)$。这里的 W_t^m 代表通过第 m 个注意力生成函数生成的第 m 个正常模式的编码向量图。通过此过程就可以得到 M 个独特的动态原型，其意义就是为了表示编码图中每一个像素点 $\{x_t^N, N = w * c\}$ 对于该原型的重要程度，这个步骤就是融合的过程：

$$p_t^m = \sum_{n=1}^{N} \frac{\omega_t^{n,m}}{\sum_{n'=1}^{N} \omega_t^{n',m}} x_t^n$$

通过多头注意力函数生成的 M 个正常模式的动态原型形成一个动态原型池，即 $P_t = \{p_t^m\}_{m=1}^M$。

最终，将编码器生成的特征向量与原型池中的向量进行检索，通过输入的初始编码特征图中每一个像素点的特征向量与动态原型池中的正常模式的原型对特征向量进行重构，其操作为：

$$\tilde{X}_t^n = \sum_{m=1}^{M} \beta_t^{n,m} p_t^m$$

$$\beta_t^{n,m} = \frac{x_t^n p_t^m}{\sum_{m'=1}^{M} x_t^n p_t^{m'}}$$

其中，$\beta_t^{n,m}$ 表示特征图中像素点对应的特征向量与动态原型池中的原型的相似程度。最终得到的正常样本的特征图为原始编码器得到的特征图与上述检索操作得到的权重特征的通

道和。其核心思想就是为了丰富自编码器中编码器对正常模式的表示并抑制异常模式的部分。从编码器输出的结果用于后续的帧预测。

本文为整体原型的学习中对于正常模式的表示、对正常模式编码的增强进行的重构以及最终帧预测结果构建损失函数。全局的损失函数

$$\mathcal{L}$$

由特征重构损失部分 $\mathcal{L}_{fea}$ 与帧预测损失部分 $\mathcal{L}_{fra}$ 两部分组成，这两部分权重值由 λ_1 进行调整。其形式如下：

$$\mathcal{L} = \mathcal{L}_{fra} + \lambda_1 \mathcal{L}_{fea}$$

对于帧预测损失，本文使用预测帧与真实帧之间的 L2 距离：

$$\mathcal{L}_{fra} = \|\widehat{y}_t - y_t\|_2$$

特征重构损失是使得学习到的原型之间与正常模式的编码特征紧密且互不相同。因此描述这一特征需要两部分组成，分别为正常模式编码特征与原型 $\mathcal{L}_c$、原型与原型之间的差异 $\mathcal{L}_d$，并通过 λ_2 调整两者之间的权重，其形式为：

$$\mathcal{L}_{fea} = \mathcal{L}_c + \lambda_2 \mathcal{L}_d$$

由于 $\mathcal{L}_c$ 表示的含义为正常模式的编码特征需要与原型紧密相关，因此用平均 L2 距离进行衡量：

$$\begin{aligned} &\mathcal{L}_c = \tfrac{1}{N} \sum_{n=1}^{N} \|x_t^n - p_t^*\|_2 \\ &s.t., * = \arg\max \beta_t^{n,m}, m \in [1, M] \end{aligned}$$

计算原型与特征向量中最相关的一项之间的距离作为损失，使得特征向量与原型之间联系更加紧凑。同时，原型之间应该满足互不相同，这样才能尽可能表示更多的正常模式，因此计算原型之间的差异如下：

$$\mathcal{L}_d = \frac{2}{M(M-1)} \sum_{m=1}^{M} \sum_{m=1}^{M} [- \|p_m - p_{m'}\|_2 + \gamma]_+$$

通过 γ 调整原型之间的边缘值。

3.2 视频异常检测的小样本元学习方法

首先本文考虑一个 VAD 的整体框架，即 $f_\theta(E_\eta(x)) = D_\delta(P_\tau(E_\eta(x)))$，其中 η 和 δ 代表自编码器中编码器和解码器函数 E、D 的参数。通过输入的一系列连续帧 x，经过编码器生成编码特征图 $X = E_\eta(x)$，接着将该特征图作为 DPU 模块的输入，经过参数 τ 调整，对输入连续帧的特征图中的正常模式的信息进行编码。本文的小样本训练方法的目标函数为 $f_\theta(X)$，通过由 DPU 模块和解码器组成的元动态学习单元（Meta-Prototype Unit, MPU）进行训练，并由参数 $\theta = \tau \cup \delta$ 进行调整。

在推理期间，少量的测试视频片段用于调整小样本学习模型的参数。为了模仿这一过程，获得一个较好的初始化参数 θ，首先随机初始化一个 θ_0，接着经过更新函数的少量次数迭代，使得模型在少量的数据下也能能快速适应新场景。本文采取梯度下降的方法，通过参数 α 进行优化。更新函数如下：

$$U\left(\theta, \nabla_\theta \mathcal{L}; \alpha\right) = \theta - \alpha \odot \nabla_\theta \mathcal{L}$$

为了保证场景自适应的鲁棒性，在元训练过程中，对同一场景中不同输入输出的误差信号对目标模型进行更新。关键思想是目标模型推广到同一场景中的其他帧，而不仅仅是模型训练的几个帧。目标模型的更新过程为：

$$\theta_0^{i+1} = U\left(\theta_0^i, \nabla_{\theta_0^i} \mathcal{L}\left(y_k, f_{\theta_0}\left(E_\eta\left(x_k\right)\right)\right)\right)$$

在经过 T 步的更新迭代得到较优的场景优化参数 $\widehat{\theta}$。本文评估模型参数在随机输入输出对 (x_j, y_j) 中最小化场景误差信号。整体的优化更新过程为：

$$\theta_0^*, \alpha^* = \arg\min_{\theta_0, \alpha} \mathrm{E}\left[\mathcal{L}\left(y_j, f_{\widehat{\theta}}\left(E_\eta\left(x_j\right)\right)\right)\right]$$

3.3 改进的原型机制和帧深度预测方案

由于原模型中使用的原型都来自于正常模式，这使得模型在实际场景中的鲁棒性较差，本文通过添加少量的异常模式让其也作为原型的一部分，以部分噪声作为影响原型的生成。这样使得生成的原型更具有鲁棒性，能够在场景中预测效果更好，并且通过对比损失调整原型之间的距离，即：

$$\mathcal{L}_p = \frac{1}{T} \sum_{t=1}^{T} \left\| p_i^a - p_i^m \right\| - \left\| p_i^a - q_i^m \right\| + \alpha$$

其中 p_i^a 表示为该序列生成动态原型的特征，p_i^m 表示为该序列中为正常动态原型的特征，q_i^m 表示为该序列中为异常动态原型的特征，通过 α 参数进行调整。

同时，本文还提出了帧深度预测的方法，经过 DPU 模块查询操作并融合得到的预测帧将其与真实帧同时输入至预训练好的 VGG-16 网络中，并提取 K 层特征进行比较，与原来帧之间像素层面的比较相比，本文提出的方法将其中的图像深层信息，如纹理等细节也进行比较，修改原来的帧预测损失，即：

$$\mathcal{L}_{fra} = \sum_{k=1}^{K} \left\| \widehat{y}_t^k - y_t^k \right\|_2$$

整体的改进模型如图 2

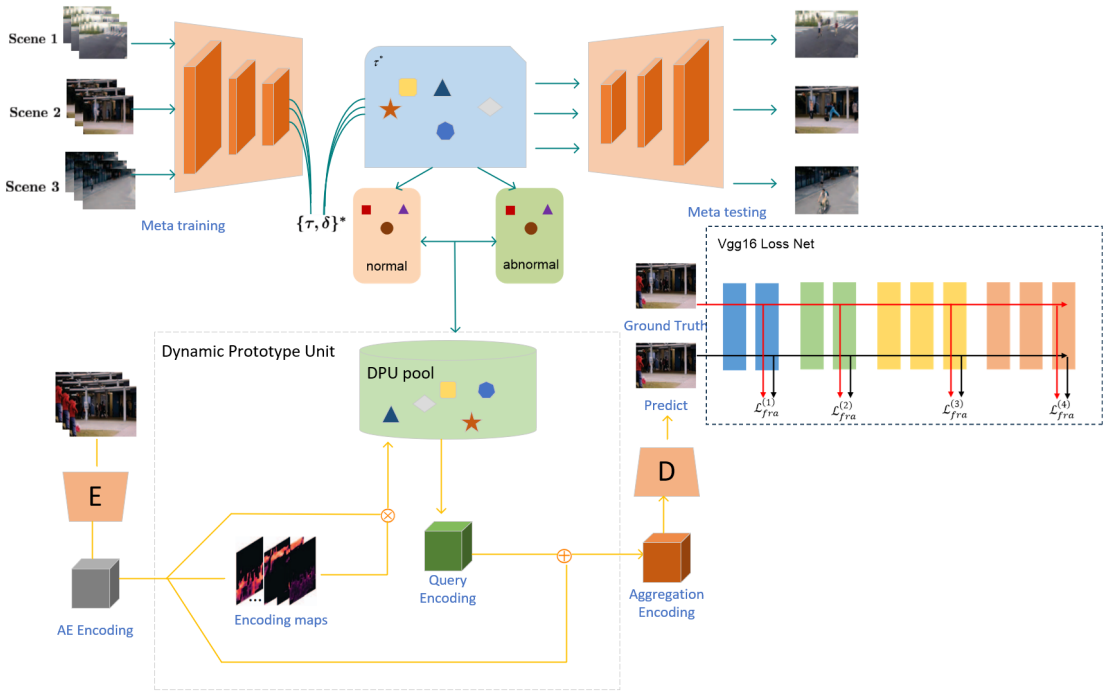

图 2. 模型改进后的整体框架添加了正常异常原型对比和 VGG-16 感知损失网络。

4 复现细节

4.1 与已有开源代码对比

使用的源代码基础上对原型重建的部分进行修改，同时在损失函数的设置上新增了一个预训练好的 VGG-16 网络重新计算预测损失，修改其中的损失函数。

4.2 实验环境搭建

4.3 界面分析与使用说明

Train 和 Test 代码文件为无监督学习的训练方法，而 Train_meta 和 Test_meta 则为元学习训练部分，数据集放在./data 路径下，根据数据集分享者原本的路径对一些细节进行修改，如 ped2 下的图片格式为.tif，而 Avenue 和 Shanghai 的图片格式为.jpg，需要在对应的 util 进行修改。最终生成的模型放在./exp 文件下，测试的模型需要放在./data/exp_900_exp 中。

5 实验结果分析

对于问题的设置，为了更好地评估把本文提出方法的有效性，本文遵循两种视频异常检测问题设置，分别为无监督视频异常检测和小样本视频异常检测。第一种是对无监督训练方法的评估，在训练过程中只有正常的视频帧，测试视频的场景是在此训练期间出现的相同场景。第二种是元学习评估，基于来自不同数据集的训练和测试视频，确保训练和测试过程中场景的多样性。这种设置也被称为“跨数据集”测试。第一种设置测试了在固定摄像机下的表现，而后一种设置测试在给定新摄像机时的适应能力。本文认为上述设置对于评估一个鲁棒

和实用的异常检测方法是必不可少的。本文选择了四个流行的异常检测数据集，在不同的问题设置下评估本文提出的方法。1) UCSD Pedl & Ped2 数据集，分别包含 34 个和 16 个训练视频，36 个和 12 个测试视频，其中有 12 个不规则事件，包括骑自行车和驾驶汽车。2)Avenue 数据集由 16 个训练视频和 21 个测试视频组成，其中包含跑步、扔东西等 47 个异常事件。3) ShanghaiTech 数据集包含 330 个训练视频和 107 个测试视频，共 13 个场景。4) UCF-Crime 数据集包含从大量真实世界的监控摄像头收集的正常视频和犯罪视频，每个视频来自不同的场景。本文使用来自该数据集的普通视频进行元训练，然后在跨数据集测试中在其他数据集上测试模型。本文使用 ROC 曲线下面积 (AUC) 来评估性能。ROC 曲线是通过改变每帧预测异常评分的阈值来获得的。对于无监督训练的测试结果如下表 1：

表 1. **与最先进的异常检测方法的定量比较。并将原模型和改进模型进行比较。在无监督环境下的平均 AUC(%)。加粗的数字表示最好的性能，下划线表示第二好的性能。DPU 是原文方法，DPU-c 是改进后的方法。**

Methods	Ped1	Ped2	Avenue	Shanghai
MDT [20]	81.8	82.9	-	-
TSC [13]	-	91.0	80.6	60.9
Frame-Pred [9]	83.1	95.4	85.1	72.8
AMC [16]	-	96.2	86.9	-
rGAN [11]	86.3	96.2	85.8	77.9
MemAE [11]	-	94.1	83.3	71.2
LMN [17]	-	97.0	88.5	70.5
Ours DPU	85.1	94.5	89.0	71.13
Ours DPU-c	85.7	96.4	89.6	71.41

表中在几个异常检测数据集上使用标准无监督异常检测。其他方法之间与本文提出结构之间进行比较。其中 MemAE 和 LMN 是与本文提出的方法最相关的方法。这两个方法都是通过构成一个大的记忆库，用于存储并训练视频帧中的正常模式。本文提出的方法学习一些以输入数据为条件的动态标准原型，更节省内存，提高速度。优异的性能也证明了本文提出的 DPU 模块有效性。在 ped1 和 Shanghai Tech 上，本文使用的方法的 AUC 低于 rGAN 方法，这是合理的，因为 rGAN 的模型架构更复杂。rGAN 使用 ConvLSTM 通过多次叠加 AE 来保留历史信息。然而，本文提出的模型结构只应用一个 AE。

对于小样本模型方面的评估，为了证明本文提出的方法场景适应能力，本文在 Shanghai Tech 数据集和 UCF-Crime 数据集的正常视频上进行了跨数据集的元训练测试，然后使用其他数据集 (UCSD Ped1, UCSD Ped2，Avenue) 进行验证。比较结果如表 2 所示。在大多数情况下，预训练的 DPU 模型比 rGAN 更具泛化性。基于原型的特征重构极大地提高了基于帧预测的异常检测的鲁棒性。

表 2. **跨数据集测试下 K-shot (K = 1,5,10) 场景自适应异常检测对比。**

Shanghai Tech

Target	Methods	1-shot	5-shot	10-shot
Ped1	rGAN(Meta) [11]	80.6	81.42	82.38
	Ours(Meta)	78.54	79.35	80.20
	Ours(Meta)-c	78.84	79.87	80.53
Ped2	rGAN(Meta) [11]	91.19	91.8	92.8
	Ours(Meta)	94.46	94.67	95.75
	Ours(Meta)-c	94.51	95.05	95.93
Avenue	rGAN(Meta) [11]	76.58	77.1	78.79
	Ours(Meta)	78.92	80.25	95.75
	Ours(Meta)-c	79.05	80.78	81.94

UCF crime

Target	Methods	1-shot	5-shot	10-shot
Ped1	rGAN(Meta) [11]	78.44	80.43	81.62
	Ours(Meta)	77.19	78.33	79.53
	Ours(Meta)-c	77.36	78.64	79.65
Ped2	rGAN(Meta) [11]	83.08	86.41	90.21
	Ours(Meta)	88.43	87.83	89.89
	Ours(Meta)-c	89.02	89.56	89.75
Avenue	rGAN(Meta) [11]	72.62	74.68	79.02
	Ours(Meta)	85.62	85.66	85.91
	Ours(Meta)-c	85.58	85.72	85.86

本文还分析了两部分损失函数对实验效果的影响，因此设置了消融实验，具体结果如表 3 所示，并且还分析了 DPU 的有效性。本文设置 M = 10 作为 DPU 中注意映射函数的默认数。结果列在表 4 中。DPU 提高了各种基准测试的总体性能。

表 3. **设计的 DPU 模块 AUC 分析。表中 FR 和 FP 分别代表特征重构 Feature Reconstruction 和帧预测 Frame Prediction 得到的异常分数。**

Setting	Shanghai	Avenue	Ped1	Ped2
AE baseline(FP)	66.7	83.9	83.2	93.2
AE with DPU(FP)	69.2	68.8	83.5	93.7
AE with DPU(FR)	70.4	87.1	84.1	94.2
AE with DPU(FP&FR)	71.1	89.0	85.1	94.5
AE with DPU(FP&FR)-c	71.4	89.6	85.7	96.4

表 4. **DPU 中原型数量的 AUC 分析，在 Ped2 数据集下进行分析。**

Number	1	5	10	20
Overall	92.89	93.15	94.52	94.20

6 总结与展望

本文引入了一个原型学习模块，通过无监督异常检测的注意机制显式地对视频序列中的正常动态进行建模。原型模块是完全可微的，并以端到端方式进行训练。在没有额外内存消耗的情况下，本文的方法在无监督设置的各种异常检测基准上性能较优。此外，本文利用元学习技术将原型模块改进为少量正常学习器。大量的实验评估证明了场景自适应方法的有效性。其中，对于原型的设置学习的都是正常模式的原型，实际情况可以结合一些异常模式进行对比学习，并且学习到的原型表示的区域较为随机，可以尝试结合一些特征提取对视频帧中出现的目标进行提取，将这些特征作为特定的原型。

参考文献

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 481–490, 2019.

[2] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560, 2008.

[3] Yannick Benezeth, P-M Jodoin, Venkatesh Saligrama, and Christophe Rosenberger. Abnormal events detection based on spatio-temporal co-occurences. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2458–2465. IEEE, 2009.

[4] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456. IEEE, 2011.

[5] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

[6] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.

[7] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE conference on computer vision and pattern recognition*, pages 2921–2928. IEEE, 2009.

[8] Idan Kligvasser, Tamar Rott Shaham, and Tomer Michaeli. xunit: Learning a spatial activation function for efficient image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2433–2442, 2018.

[9] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018.

[10] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[11] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 125–141. Springer, 2020.

[12] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International conference on multimedia and expo (ICME)*, pages 439–444. IEEE, 2017.

[13] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, pages 341–349, 2017.

[14] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15425–15434, 2021.

[15] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing*, 30:4505–4515, 2021.

[16] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1273–1283, 2019.

[17] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020.

[18] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. Multi-person pose estimation with enhanced channel-wise and spatial information. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5674–5682, 2019.

[19] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.

[20] Shu Wang and Zhenjiang Miao. Anomaly detection in crowd scene. In *IEEE 10th International Conference on Signal Processing Proceedings*, pages 1220–1223. IEEE, 2010.

[21] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[22] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246, 2019.