Latent Space Autoregression for Novelty Detection

Davide Abati Angelo Porrello Simone Calderara Rita Cucchiara

University of Modena and Reggio Emilia
{name.surname}@unimore.it

Abstract

Novelty detection is commonly referred to as the discrimination of observations that do not conform to a learned model of regularity. Despite its importance in different application settings, designing a novelty detector is utterly complex due to the unpredictable nature of novelties and its inaccessibility during the training procedure, factors which expose the unsupervised nature of the problem. In our proposal, we design a general framework where we equip a deep autoencoder with a parametric density estimator that learns the probability distribution underlying its latent representations through an autoregressive procedure. We show that a maximum likelihood objective, optimized in conjunction with the reconstruction of normal samples, effectively acts as a regularizer for the task at hand, by minimizing the differential entropy of the distribution spanned by latent vectors. In addition to providing a very general formulation, extensive experiments of our model on publicly available datasets deliver on-par or superior performances if compared to state-of-the-art methods in one-class and video anomaly detection settings. Differently from prior works, our proposal does not make any assumption about the nature of the novelties, making our work readily applicable to diverse contexts.

1. Introduction

Novelty detection is defined as the identification of samples which exhibit significantly different traits with respect to an underlying model of regularity, built from a collection of normal samples. The awareness of an autonomous system to recognize unknown events enables applications in several domains, ranging from video surveillance [7, 11], to defect detection [22] to medical imaging [38]. Moreover, the surprise inducted by unseen events is emerging as a crucial aspect in reinforcement learning settings, as an enabling factor in curiosity-driven exploration [34].

However, in this setting, the definition and labeling of novel examples are not possible. Accordingly, the literature agrees on approximating the ideal shape of the boundary separating normal and novel samples by modeling the intrinsic characteristics of the former. Therefore, prior works tackle such problem by following principles derived from the unsupervised learning paradigm [9, 37, 11, 26, 30]. Due to the lack of a supervision signal, the process of feature extraction and the rule for their normality assessment can only be guided by a proxy objective, assuming the latter will define an appropriate boundary for the application at hand.

According to cognitive psychology [4], novelty can be expressed either in terms of capabilities to *remember* an event or as a degree of *surprisal* [42] aroused by its observation. The latter is mathematically modeled in terms of low probability to occur under an expected model, or by lowering a variational free energy [16]. In this framework, prior models take advantage of either parametric [49] or non-parametric [14] density estimators. Differently, remembering an event implies the adoption of a memory represented either by a dictionary of normal prototypes - as in sparse coding approaches [9] - or by a low dimensional representation of the input space, as in the self-organizing maps [20] or, more recently, in deep autoencoders. Thus, in novelty detection, the remembering capability for a given sample is evaluated either by measuring reconstruction errors [11, 26] or by performing discriminative in-distribution tests [37].

Our proposal contributes to the field by merging remembering and surprisal aspects into a unique framework: we design a generative unsupervised model (i.e., an autoencoder, represented in Fig. 1i) that exploits end-to-end training in order to maximize remembering effectiveness for normal samples whilst minimizing the surprisal of their latent representation. This latter point is enabled by the maximization of the likelihood of latent representations through an autoregressive density estimator, which is performed in conjunction with the reconstruction error minimization. We show that, by optimizing both terms jointly, the model implicitly seeks for minimum entropy representations maintaining its remembering/reconstructive power. While entropy minimization approaches have been adopted in deep neural compression [3], to our knowledge this is the first proposal

潜在空间自回归用于新颖性检测

达维德 · 阿巴蒂 安杰洛 · 波雷洛 西蒙内 · 卡尔德拉拉 丽塔 · 库基亚拉

摩德纳和雷焦艾米利亚大学
{名 . 姓 }@unimore.it

摘要

异常检测通常指对不符合已学习常态模型的观测数据进行区分的任务。尽管在不同应用场景中具有重要意义，但由于异常现象的不可预测性及其在训练过程中的不可触及性 —— 这些因素揭示了该问题的无监督本质 —— 设计异常检测器变得极为复杂。在我们的方案中，我们设计了一个通用框架：通过自回归过程为深度自编码器配备参数化密度估计器，从而学习其潜在表征所遵循的概率分布。我们证明，在正常样本重构过程中联合优化的最大似然目标函数，可通过最小化潜在向量所构成分布的微分熵，有效充当当前任务的正则化器。除了提供高度通用的 formulation，我们在公开数据集上的大量实验表明：与单类别异常检测和视频异常检测领域的先进方法相比，本模型实现了相当或更优的性能。与先前研究不同，我们的方案不对异常性质做任何假设，使其能够直接适用于多样化场景。

1. 引言

新颖性检测的定义是，识别那些相对于基于正常样本集合建立的规律性基础模型表现出显著不同特征的样本。自主系统识别未知事件的能力使其在多个领域得到应用，范围从视频监控 [7, 11]，到缺陷检测 [22] 再到医学成像 [38]。此外，由未见事件引发的惊奇感正逐渐成为强化学习环境中的关键要素，作为驱动好奇心探索的赋能因素 [34]。

然而，在此设定下，无法对新型示例进行定义和标记。因此，文献在通过建模正常样本的内在特性来逼近区分正常样本与新颖样本的理想边界形状方面达成共识。因此，先前的研究遵循无监督学习范式 [9, 37, 11, 26, 30] 衍生的原则来解决这一问题。由于缺乏监督信号，特征提取过程及其正常性评估规则只能通过代理目标进行引导，并假设后者能为当前应用定义合适的边界。

根据认知心理学 [4], , 新奇性既可以通过记忆事件的能力来表征，也可通过观测事件时引发的惊异程度 [42] 来衡量。后者在数学上通过低概率事件发生于预期模型之下的概率来建模，或通过降低变分自由能 [16] 来实现。在此框架中，先验模型采用参数化 [49] 或非参数化 [14] 密度估计器。与之不同，记忆事件意味着采用某种记忆表征：或通过正常原型字典（如稀疏编码方法 [9] ），或通过输入空间的低维表征（如自组织映射 [20]，或近年兴起的深度自编码器）。因此在新奇性检测中，特定样本的记忆能力可通过测量重构误差 [11, 26] 或执行判别性分布内测试 [37] 来评估。

我们的提案通过将记忆与惊奇度方面融合到一个独特框架中，为该领域做出贡献：我们设计了一种生成式无监督模型（即自编码器，如图 1i 所示），该模型利用端到端训练来最大化正常样本的记忆效率，同时最小化其潜在表示的惊奇度。后一点是通过自回归密度估计器最大化潜在表示的可能性实现的，该过程与重构误差最小化同步进行。我们表明，通过联合优化这两个项，模型会隐式地寻求保持其记忆 / 重构能力的最小熵表示。虽然熵最小化方法已在深度神经压缩中得到应用 [3], , 但据我们所知，这是首个将 ...

tailored for novelty detection. In memory terms, our procedure resembles the concept of prototyping the normality using as few templates as possible. Moreover, evaluating the output of the estimator enables the assessment of the surprisal aroused by a given sample.

2. Related work

Reconstruction-based methods. On the one hand, many works lean toward learning a parametric projection and reconstruction of normal data, assuming outliers will yield higher residuals. Traditional sparse-coding algorithms [48, 9, 27] adhere to such framework, and represent normal patterns as a linear combination of a few basis components, under the hypotheses that novel examples would exhibit a non-sparse representation in the learned subspace. In recent works, the projection step is typically drawn from deep autoencoders [11]. In [30] the authors recover sparse coding principles by imposing a sparsity regularization over the learned representations, while a recurrent neural network enforces their smoothness along the time dimension. In [37], instead, the authors take advantage of an adversarial framework in which a discriminator network is employed as the actual novelty detector, spotting anomalies by performing a discrete in-distribution test. Oppositely, future frame prediction [26] maximizes the expectation of the next frame exploiting its knowledge of the past ones; at test time, observed deviations against the predicted content advise for abnormality. Differently from the above-mentioned works, our proposal relies on modeling the prior distribution of latent representations. This choice is coherent with recent works from the density estimation community [41, 6]. However, to the best of our knowledge, our work is the first advocating for the importance of such a design choice for novelty detection.

Probabilistic methods. A complementary line of research investigates different strategies to approximate the density function of normal appearance and motion features. The primary issue raising in this field concerns how to estimate such densities in a high-dimensional and complex feature space. In this respect, prior works involve handcrafted features such as optical flow or trajectory analysis and, on top of that, employ both non-parametric [1] and parametric [5, 31, 25] estimators, as well as graphical modeling [17, 23]. Modern approaches rely on deep representations (e.g., captured by autoencoders), as in Gaussian classifiers [36] and Gaussian Mixtures [49]. In [14] the authors involve a Kernel Density Estimator (KDE) modeling activations from an auxiliary object detection network. A recent research trend considers training Generative Adversarial Networks (GANs) on normal samples. However, as such models approximate an implicit density function, they can be queried for new samples but not for likelihood values. Therefore, GAN-based models employ different heuristics for the evaluation of novelty. For instance, in [38] a guided latent space search is exploited to infer it, whereas [35] directly queries the discriminator for a normality score.

3. Proposed model

Maximizing the probability of latent representations is analogous to lowering the surprisal of the model for a normal configuration, defined as the negative log-density of a latent variable instance [42]. Conversely, remembering capabilities can be evaluated by the reconstruction accuracy of a given sample under its latent representation.

We model the aforementioned aspects in a latent variable model setting, where the density function of training samples $p(\mathbf{x})$ is modeled through an auxiliary random variable $\mathbf{z}$, describing the set of causal factors underlying all observations. By factorizing

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}, \tag{1}$$

where $p(\mathbf{x}|\mathbf{z})$ is the conditional likelihood of the observation given a latent representation $\mathbf{z}$ with prior distribution $p(\mathbf{z})$, we can explicit both the memory and surprisal contribution to novelty. We approximate the marginalization by means of an inference model responsible for the identification of latent space vector for which the contribution of $p(\mathbf{x}|\mathbf{z})$ is maximal. Formally, we employ a deep autoencoder, in which the reconstruction error plays the role of the negative logarithm of $p(\mathbf{x}|\mathbf{z})$, under the hypothesis that $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\tilde{\mathbf{x}}, I)$ where $\tilde{\mathbf{x}}$ denotes the output reconstruction. Additionally, surprisal is injected in the process by equipping the autoencoder with an auxiliary deep parametric estimator learning the prior distribution $p(\mathbf{z})$ of latent vectors, and training it by means of Maximum Likelihood Estimation (MLE). Our architecture is therefore composed of three building blocks (Fig. 1i): an encoder $f(\mathbf{x};\theta_f)$, a decoder $g(\mathbf{z};\theta_g)$ and a probabilistic model $h(\mathbf{z};\theta_h)$:

$$\begin{gathered} f(\mathbf{x};\theta_f) : \mathbb{R}^m \to \mathbb{R}^d, \qquad g(\mathbf{z};\theta_g) : \mathbb{R}^d \to \mathbb{R}^m, \\ h(\mathbf{z};\theta_h) : \mathbb{R}^d \to [0,1]. \end{gathered} \tag{2}$$

The encoder processes input $\mathbf{x}$ and maps it into a compressed representation $\mathbf{z} = f(\mathbf{x};\theta_f)$, whereas the decoder provides a reconstructed version of the input $\tilde{\mathbf{x}} = g(\mathbf{z};\theta_g)$. The probabilistic model $h(\mathbf{z};\theta_h)$ estimates the density in $\mathbf{z}$ via an autoregressive process, allowing to avoid the adoption of a specific family of distributions (i.e., Gaussian), potentially unrewarding for the task at hand. On this latter point, please refer to supplementary materials for comparison w.r.t. variational autoencoders [19].

With such modules, at test time, we can assess the two sources of novelty: elements whose observation is poorly

专为新颖性检测而设计。在内存方面，我们的程序类似于使用尽可能少的模板来构建正常性原型的概念。此外，通过评估估计器的输出，能够衡量给定样本所引起的惊讶程度。

2. 相关工作

基于重构的方法。一方面，许多研究倾向于学习正常数据的参数化投影与重构，假定异常值会产生更高残差。传统稀疏编码算法 [48, 9, 27] 遵循此框架，在假设新样本会在学习子空间中呈现非稀疏表示的前提下，将正常模式表示为若干基组分的线性组合。近期研究中，投影步骤通常源于深度自编码器 [11]。[30] 作者通过对学习表征施加稀疏正则化来恢复稀疏编码原理，同时利用循环神经网络增强其在时间维度上的平滑性。[37], 则另辟蹊径，采用对抗框架 —— 将判别器网络作为实际的新颖性检测器，通过执行离散分布内测试来识别异常。与之相反，未来帧预测 [26] 通过利用历史帧知识来最大化下一帧的期望值；测试时，观测内容与预测内容之间的显著偏差即提示异常。与上述研究不同，我们的方案依赖于对潜在表征先验分布的建模。这一选择与密度估计领域的最新研究 [41, 6] 相契合。但据我们所知，本研究是首个论证该设计选择对新奇检测重要性的工作。

概率化方法。另一条互补的研究路线致力于探索近似正常外观与运动特征密度函数的不同策略。该领域的核心问题在于如何在高维复杂特征空间中估算此类密度函数。对此，早期研究采用光流或轨迹分析等手工设计特征，并在此基础上结合了非参数 [1] 与参数 [5, 31, 25] 估计器，以及图模型 [17, 23]。现代方法则依托深度表征（如自编码器提取的特征），例如高斯分类器 [36] 和高斯混合模型 [49] 的应用。[14] 研究者引入核密度估计器（ KDE ）对辅助物体检测网络的激活值进行建模。近期研究趋势关注在正常样本上训练生成对抗网络（ GANs ），但因此类模型逼近的是隐式密度函数，可通过查询生成新样本。

但不适用于似然值。因此，基于 GAN 的模型采用不同的启发式方法来评估新颖性。例如，在 [38] 中利用引导潜空间搜索进行推断，而 [35] 则直接向判别器查询正态性得分。

3. 提出的模型

最大化潜在表示的概率类似于降低模型对正常配置的惊奇度，其定义为潜在变量实例 [42] 的负对数密度。反之，记忆能力可以通过给定样本在其潜在表示下的重构精度来评估。

我们在潜变量模型设置中对上述方面进行建模，其中训练样本 $p(\mathbf{x})$ 的概率密度函数通过辅助随机变量 $\mathbf{z}$ 进行建模，该变量描述了所有观测值背后的因果因子集合。通过因子分解

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}, \tag{1}$$

其中 $p(\mathbf{x}|\mathbf{z})$ 是给定潜在表示 $\mathbf{z}$ 及先验分布 $p(\mathbf{z})$ 的观测条件似然函数，我们可以显式分解记忆与信息惊奇对新奇性的贡献。通过采用负责识别潜在空间向量的推理模型，我们近似边缘化处理，该模型旨在找到使 $p(\mathbf{x}|\mathbf{z})$ 贡献最大的潜向量。形式上，我们采用深度自编码器，在假设 $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\tilde{\mathbf{x}}\ I)$ 的前提下，其重构误差扮演 $p(\mathbf{x}|\mathbf{z})$ 负对数的角色（此处 $\tilde{\mathbf{x}}$ 表示输出重构）。此外，通过为自编码器配备辅助深度参数估计器来学习潜在向量的先验分布 $p(\mathbf{z})$，并采用最大似然估计进行训练，将信息惊奇注入流程。因此我们的架构包含三个核心模块（图 1i ）： 编码器 $f(\mathbf{x};\theta_f)$ 、解码器 $g(\mathbf{z};\theta_g)$ 以及概率模型 $h(\mathbf{z};\theta_h)$：

$$\begin{gathered} f(\mathbf{x};\theta_f) : \mathbb{R}^m \to \mathbb{R}^d, \qquad g(\mathbf{z};\theta_g) : \mathbb{R}^d \to \mathbb{R}^m, \\ h(\mathbf{z};\theta_h) : \mathbb{R}^d \to [0,1]. \end{gathered} \tag{2}$$

编码器处理输入 $\mathbf{x}$ 并将其映射为压缩表示 $\mathbf{z} = f(\mathbf{x};\theta_f)$，而解码器则提供输入的重构版本 $\tilde{\mathbf{x}} = g(\mathbf{z};\theta_g)$。概率模型 $h(\mathbf{z};\theta_h)$ 通过自回归过程估计 $\mathbf{z}$ 中的密度，从而避免采用可能对当前任务无益的特定分布族（例如高斯分布）。关于后一点，请参阅补充材料以与变分自编码器 [19] 进行比较。

借助此类模块，在测试时我们可以评估两种新颖性来源：观测效果较差的元素

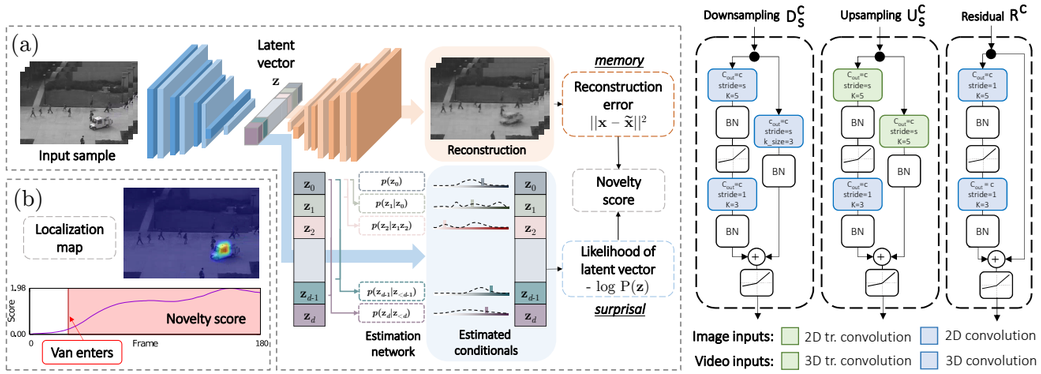


Figure 1: (i) The proposed novelty detection framework. The overall architecture, depicted in (a), consists of a deep autoencoder and an autoregressive estimation network operating on its latent space. The joint minimization of their respective objective leads to a measure of novelty - (b) - obtained by assessing the remembrance of the model when looking to a new sample, combined with its surprise aroused by causal factors. (ii) Building blocks employed in the autoencoder's architecture.

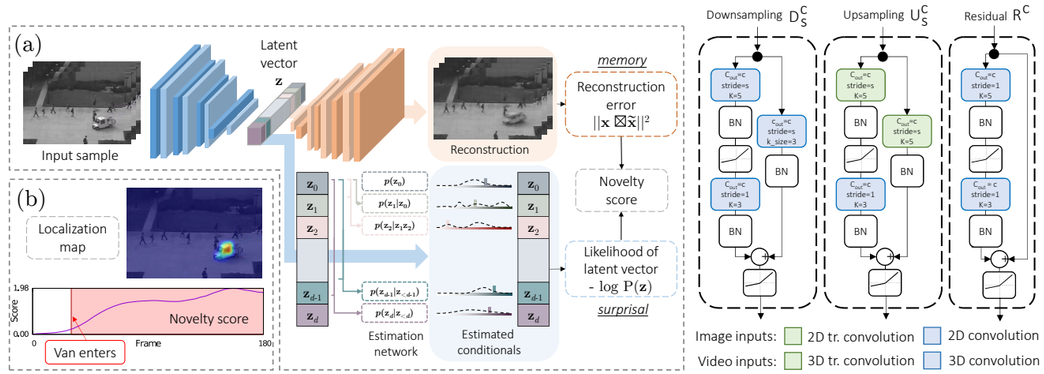


图 1：(i) 提出的新颖性检测框架。整体架构如 (a) 所示，包含一个深度自编码器及其潜在空间上运行的自回归估计网络。通过联合最小化各自目标函数，可获得新颖性度量 ——(b)—— 该度量通过评估模型观察新样本时的记忆强度，并结合因果因素引发的惊异值来实现。(ii) 自编码器架构中采用的构建模块。

explained by the causal factors inducted by normal samples (i.e., high reconstruction error); elements exhibiting good reconstructions whilst showing surprising underlying representations under the learned prior.

Autoregressive density estimation. Autoregressive models provide a general formulation for tasks involving sequential predictions, in which each output depends on previous observations [28, 32]. We adopt such a technique to factorize a joint distribution, thus avoiding to define its landscape a priori [24, 43]. Formally, $p(\mathbf{z})$ is factorized as

$$p(\mathbf{z}) = \prod_{i=1}^{d} p(z_i|\mathbf{z}_{<i}), \tag{3}$$

so that estimating $p(\mathbf{z})$ reduces to the estimation of each single Conditional Probability Density (CPD) expressed as $p(z_i|\mathbf{z}_{<i})$, where the symbol $<$ implies an order over random variables. Some prior models obey handcrafted orderings [46, 45], whereas others rely on order agnostic training [44, 10]. Nevertheless, it is still not clear how to estimate the proper order for a given set of variables. In our model, this issue is directly tackled by the optimization. Indeed, since we perform autoregression on learned latent representations, the MLE objective encourages the autoencoder to impose over them a pre-defined causal structure. Empirical evidence of this phenomenon is given in the supplementary material.

From a technical perspective, the estimator $h(\mathbf{z}; \theta_h)$ outputs parameters for d distributions $p(z_i|\mathbf{z}_{<i})$. In our implementation, each CPD is modeled as a multinomial over B=100 quantization bins. To ensure a conditional estimate of each underlying density, we design proper layers guaranteeing that the CPD of each symbol z_i is computed from inputs $\{z_1, \ldots, z_{i-1}\}$ only.

Objective and connection with differential entropy. The three components f, g and h are jointly trained to minimize $\mathcal{L} \equiv \mathcal{L}(\theta_f, \theta_g, \theta_h)$ as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{REC}}(\theta_f, \theta_g) + \lambda \mathcal{L}_{\text{LLK}}(\theta_f, \theta_h) \\ &= \mathbb{E}_{\mathbf{x}}\Big[\underbrace{||\mathbf{x} - \tilde{\mathbf{x}}||^2}_{\text{reconstruction term}} - \lambda \underbrace{\log(h(\mathbf{z}; \theta_h))}_{\text{log-likelihood term}} \Big],\end{aligned} \tag{4}$$

where λ is a hyper-parameter controlling the weight of the $\mathcal{L}_{\text{LLK}}$ term. It is worth noting that it is possible to express the log-likelihood term as

$$\begin{aligned}&\mathbb{E}_{\mathbf{z}\sim p^*(\mathbf{z};\theta_f)}\big[-\log h(\mathbf{z}; \theta_h)\big] \\ &= \mathbb{E}_{\mathbf{z}\sim p^*(\mathbf{z};\theta_f)}\big[-\log h(\mathbf{z}; \theta_h) + \log p^*(\mathbf{z}; \theta_f) - \log p^*(\mathbf{z}; \theta_f)\big] \\ &= D_{\text{KL}}(p^*(\mathbf{z}; \theta_f) \parallel h(\mathbf{z}; \theta_h)) + \mathbb{H}[p^*(\mathbf{z}; \theta_f)],\end{aligned} \tag{5}$$

where $p^*(\mathbf{z}; \theta_f)$ denotes the true distribution of the codes produced by the encoder, and is therefore parametrized by θ_f. This reformulation of the MLE objective yields meaningful insights about the entities involved in the optimization. On the one hand, the Kullback-Leibler divergence ensures that the information gap between our parametric model h and the true distribution p^* is small. On the other hand, this framework leads to the minimization of the differential entropy of the distribution underlying the codes produced by the encoder f. Such constraint constitutes a crucial point when learning normality. Intuitively, if we think

由正常样本引入的因果因素（即高重构误差）所解释；在学习的先验条件下，表现出良好重构效果，同时显示出令人惊讶的潜在表征的元素。

自回归密度估计。自回归模型为涉及序列预测的任务提供了通用框架，其中每个输出都依赖于先前的观测结果 [28, 32]。我们采用这种技术来分解联合分布，从而避免先验地定义其分布形态 [24, 43]。形式上， $p(\mathbf{z})$ 被分解为

$$p(\mathbf{z}) = \prod_{i=1}^{d} p(z_i|\mathbf{z}_{<i}), \tag{3}$$

因此，估计 $p(\mathbf{z})$ 可简化为对每个条件概率密度 (CPD) $p(z_i|\mathbf{z}_{<i})$ 的单独估计，其中符号 $<$ 表示随机变量间的顺序关系。部分先验模型遵循人工设计的顺序 [46, 45], ，而其他模型则采用顺序无关的训练方法 [44, 10]。然而，如何确定给定变量集的最优顺序仍不明确。在我们的模型中，该问题通过优化过程直接解决。由于我们对学习到的潜在表示进行自回归操作，最大似然估计目标会驱动自编码器在这些表示上施加预定义的因果结构。补充材料中提供了这一现象的实验证据。

从技术角度来看，估计器 $h(\mathbf{z}; \theta_h)$ 会输出 d 分布 $p(z_i|\mathbf{z}_{<i})$ 的参数。在我们的实现中，每个条件概率分布被建模为 B=100 量化箱上的多项式分布。为确保对每个变量的条件估计在底层密度的基础上，我们设计了合适的层，确保每个符号 z_i 的条件概率分布仅从输入 $\{z_1, \ldots, z_{i-1}\}$ 计算得出。

目标及其与微分熵的关联。三个组件 f、 g 和 h 通过联合训练最小化 $\mathcal{L} \equiv \mathcal{L}(\theta_f, \theta_g, \theta_h)$，具体实现方式如下：

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{REC}}(\theta_f, \theta_g) + \lambda \mathcal{L}_{\text{LLK}}(\theta_f, \theta_h) \\ &= \mathbb{E}_{\mathbf{x}}\Big[\underbrace{||\mathbf{x} - \tilde{\mathbf{x}}||^2}_{\text{reconstruction term}} - \lambda \underbrace{\log(h(\mathbf{z}; \theta_h))}_{\text{log-likelihood term}} \Big],\end{aligned} \tag{4}$$

其中 λ 是控制 $\mathcal{L}_{\text{LLK}}$ 项权重的超参数。值得注意的是，可以将对数似然项表示为

$$\begin{aligned}&\mathbb{E}_{\mathbf{z}\sim p^*(\mathbf{z};\theta_f)}\big[-\log h(\mathbf{z}; \theta_h)\big] \\ &= \mathbb{E}_{\mathbf{z}\sim p^*(\mathbf{z};\theta_f)}\big[-\log h(\mathbf{z}; \theta_h) + \log p^*(\mathbf{z}; \theta_f) - \log p^*(\mathbf{z}; \theta_f)\big] \\ &= D_{\text{KL}}(p^*(\mathbf{z}; \theta_f) \parallel h(\mathbf{z}; \theta_h)) + \mathbb{H}[p^*(\mathbf{z}; \theta_f)],\end{aligned} \tag{5}$$

其中 $p^*(\mathbf{z}; \theta_f)$ 表示编码器生成代码的真实分布，因此由 θ_f 参数化。这种对 MLE 目标的重构为优化中涉及的实体提供了有意义的见解。一方面，Kullback-Leibler 散度确保参数模型 h 与真实分布 p^* 之间的信息差距很小。另一方面，该框架导致编码器 f 生成的代码所基于分布的微分熵最小化。在学习正态性时，这种约束构成了一个关键点。直观地说，如果我们考虑

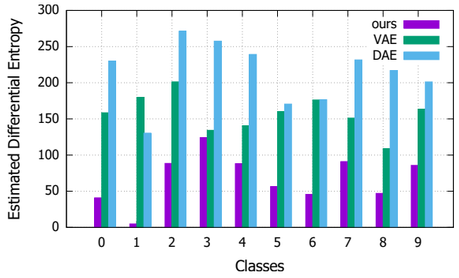


Figure 2: Estimated differential entropies delivered on each MNIST class in the presence of different regularization strategies: our, divergence w.r.t a Gaussian prior (VAE) and input perturbation (DAE). For each class, the estimate is computed on the training samples' hidden representations, whose distribution are fit utilizing a Gaussian KDE in a 3D-space. All models being equal, ours exhibits lower entropies on all classes.

about the encoder as a source emitting symbols (namely, the latent representations), its desired behavior, when modeling normal aspects in the data, should converge to a 'boring' process characterized by an intrinsic low entropy, since surprising and novel events are unlikely to arise during the training phase. Accordingly, among all the possible settings of the hidden representations, the objective begs the encoder to exhibit a low differential entropy, leading to the extraction of features that are easily predictable, therefore common and recurrent within the training set. This kind of features is indeed the most useful to distinguish novel samples from the normal ones, making our proposal a suitable regularizer in the anomaly detection setting.
We report empirical evidence of the decreasing differential entropy in Fig. 2, that compares the behavior of the same model under different regularization strategies.

3.1. Architectural Components

Autoencoder blocks. Encoder and decoder are respectively composed by downsampling and upsampling residual blocks depicted in Fig. 1ii. The encoder ends with fully connected (FC) layers. When dealing with video inputs, we employ *causal* 3D convolutions [2] within the encoder (i.e., only accessing information from previous time-steps). Moreover, at the end of the encoder, we employ a temporally-shared full connection (TFC, namely a linear projection sharing parameters across the time axis on the input feature maps) resulting in a temporal series of feature vectors. This way, the encoding procedure does not shuffle information across time-steps, ensuring temporal ordering.

Autoregressive layers. To guarantee the autoregressive nature of each output CPD, we need to ensure proper connectivity patterns in each layer of the estimator h. Moreover, since latent representations exhibit different shapes depending on the input nature (image or video), we propose two different solutions.
When dealing with images, the encoder provides feature vectors with dimensionality d. The autoregressive estimator is composed by stacking multiple Masked Fully Connections (MFC, Fig. 3-(a)). Formally, it computes output feature map $\mathbf{o} \in \mathbb{R}^{d \times co}$ (where co is the number of output channels) given the input $\mathbf{h} \in \mathbb{R}^{d \times ci}$ (assuming $ci = 1$ at the input layer). The connection between the input element $\mathbf{h}_i^k$ in position i, channel k and the output element $\mathbf{o}_j^l$ is parametrized by

$$\begin{cases} w_{i,j}^{k,l} & \text{if } i < j \\ \begin{cases} w_{i,j}^{k,l} & \text{if type = B} \\ 0 & \text{if type = A} \end{cases} & \text{if } i = j \\ 0 & \text{if } i > j. \end{cases} \tag{6}$$

Type A forces a strict dependence on previous elements (and is employed only as the first estimator layer), whereas type B masks only succeeding elements. Assuming each CPD modeled as a multinomial, the output of the last autoregressive layer (in $\mathbb{R}^{d \times B}$) provides probability estimates for the B bins that compose the space quantization.
On the other hand, the compressed representation of video clips has dimensionality $t \times d$, being t the number of temporal time-steps and d the length of the code. Accordingly, the estimation network is designed to capture two-dimensional patterns within observed elements of the code. However, naively plugging 2D convolutional layers would assume translation invariance on both axes of the input map, whereas, due to the way the compressed representation is built, this assumption is only correct along the temporal axis. To cope with this, we apply d different convolutional kernels along the code axis, allowing the observation of the whole feature vector in the previous time-step as well as a portion of the current one. Every convolution is free to stride along the time axis and captures temporal patterns. In such operation, named Masked Stacked Convolution (MSC, Fig. 3-(b)), the i-th convolution is equipped with a kernel $\mathbf{w}^{(i)} \in \mathbb{R}^{3 \times d}$ kernel, that gets multiplied by the binary mask $\mathbf{M}^{(i)}$, defined as

$$m_{j,k}^{(i)} \in \mathbf{M}^{(i)} = \begin{cases} 1 & \text{if } j = 0 \\ 1 & \text{if } j = 1 \text{ and } k < i \text{ and type=A} \\ 1 & \text{if } j = 1 \text{ and } k \leq i \text{ and type=B} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where j indexes the temporal axis and k the code axis. Every single convolution yields a column vector, as a result of its stride along time. The set of column vectors resulting

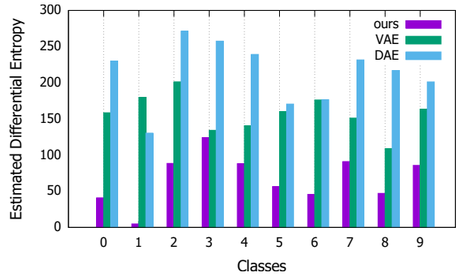


图 2：在不同正则化策略下，各 MNIST 类别估计得到的微分熵：我们的方法、高斯先验散度（VAE）和输入扰动（DAE）。针对每个类别，该估计值基于训练样本的隐层表示计算得出，其分布通过三维空间中的高斯核密度估计进行拟合。在模型结构相同的情况下，我们的方法在所有类别上均展现出更低的熵值。

关于编码器作为发射符号（即潜在表示）的源，其在建模数据中的正常方面时，期望行为应收敛于一种以内在低熵为特征的 " 乏味 " 过程 —— 因为训练阶段不太可能出现令人惊讶的新颖事件。相应地，在所有可能的隐表示设置中，该目标要求编码器展现出低微分熵，从而提取易于预测的特征，这些特征因此成为训练集中常见且重复出现的模式。这类特征确实最能有效区分新颖样本与正常样本，使得我们的提案成为异常检测场景中合适的正则化器。

我们在图 2 中报告了微分熵递减的经验证据，该图比较了同一模型在不同正则化策略下的行为。

3.1. 架构组件

自编码器模块。编码器和解码器分别由图 1ii 所示的下采样和上采样残差块构成。编码器末端为全连接层。处理视频输入时，我们在编码器中采用因果三维卷积 [2]（即仅访问先前时间步的信息）。此外，在编码器末端采用时序共享全连接层（TFC，即在输入特征图上沿时间轴共享参数的线性投影），生成特征向量的时间序列。这种编码方式不会跨时间步混淆信息，从而确保时序顺序。

自回归层。为保证每个输出条件概率分布的自回归特性，我们需要确保

估计器 h 每一层的连接模式。此外，由于潜在表征根据输入性质（图像或视频）会呈现不同形态，我们提出了两种不同的解决方案。
处理图像时，编码器会生成维度为 d 的特征向量。自回归估计器通过堆叠多个掩码全连接层（MFC，图 3-(a）构成。其形式化计算可表示为：给定输入特征图 $\mathbf{h} \in \mathbb{R}^{d \times ci}$（假定输入层维度为 $ci = 1$），输出特征图 $\mathbf{o} \in \mathbb{R}^{d \times co}$（其中 co 表示输出通道数）。位于位置 i、通道 k 的输入元素 $\mathbf{h}_i^k$ 与输出元素 $\mathbf{o}_j^l$ 之间的连接关系由以下参数定义：

$$\begin{cases} w_{i,j}^{k,l} & \text{if } i < j \\ \begin{cases} w_{i,j}^{k,l} & \text{if type = B} \\ 0 & \text{if type = A} \end{cases} & \text{if } i = j \\ 0 & \text{if } i > j. \end{cases} \tag{6}$$

类型 A 强制对先前元素进行严格依赖（仅作为第一个估计器层使用），而类型 B 仅遮蔽后续元素。假设每个条件概率分布被建模为多项式分布，最后一个自回归层（位于 $\mathbb{R}^{d \times B}$ 中）的输出为构成空间量化的 B 区间提供概率估计。
另一方面，视频片段的压缩表示维度为 $t \times d$，其中 t 表示时间步数，d 表示代码长度。相应地，估计网络被设计用于捕捉代码观测元素中的二维模式。然而，直接使用二维卷积层会假设输入映射的两个轴都具有平移不变性，而由于压缩表示的构建方式，这种假设仅沿时间轴成立。为解决此问题，我们沿代码轴应用 d 个不同的卷积核，使其既能观测前一时刻的完整特征向量，也能捕捉当前时刻的部分特征。每次卷积可沿时间轴自由滑动以捕获时序模式。在这种称为掩码堆叠卷积（MSC，图 3-(b)）的操作中，第 i 个卷积配备的卷积核 $\mathbf{w}^{(i)} \in \mathbb{R}^{3 \times d}$ 会与二进制掩码 $\mathbf{M}^{(i)}$ 相乘，该掩码定义为

$$m_{j,k}^{(i)} \in \mathbf{M}^{(i)} = \begin{cases} 1 & \text{if } j = 0 \\ 1 & \text{if } j = 1 \text{ and } k < i \text{ and type=A} \\ 1 & \text{if } j = 1 \text{ and } k \leq i \text{ and type=B} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

其中 j 索引时间轴，k 索引代码轴。
每次卷积都会产生一个列向量，这是其在时间维度上滑动步长的结果。由这些卷积操作产生的一系列列向量

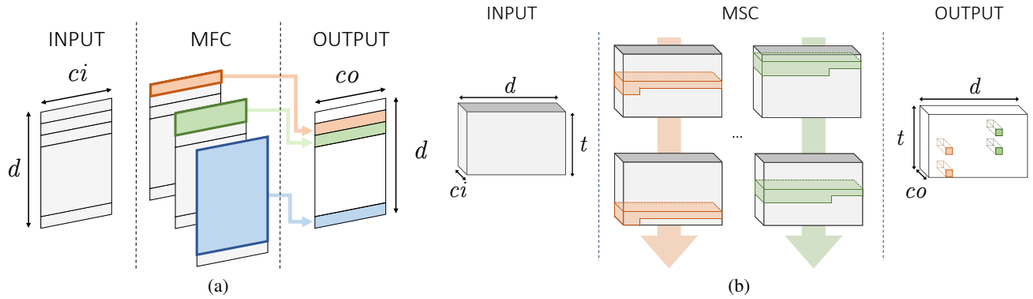


Figure 3: Proposed autoregressive layers, namely the Masked Fully Connection (a, Eq. 6) and the Masked Stacked Convolution (b, Eq. 7). For both layers, we represent type A structure. Different kernel colors represent different parametrizations.

from the application of the d convolutions to the input tensor $\mathbf{h} \in \mathbb{R}^{t \times d \times ci}$ are horizontally stacked to build the output tensor $\mathbf{o} \in \mathbb{R}^{t \times d \times co}$, as follows:

$$\mathbf{o} = \Big|\Big|_{i=1}^{d} [(\mathbf{M}^{(i)} \odot \mathbf{w}^{(i)}) * \mathbf{h}], \tag{8}$$

where $||$ represents the horizontal concatenation operation.

4. Experiments[1]

We test our solution in three different settings: images, videos, and cognitive data. In all experiments the novelty assessment on the i-th example is carried out by summing the reconstruction term (REC_i) and the log-likelihood term (LLK_i) in Eq. 4 in a single novelty score NS_i:

$$NS_i = norm_S(REC_i) + norm_S(LLK_i). \tag{9}$$

Individual scores are normalized using a reference set of examples S (different for every experiment),

$$norm_S(L_i) = \frac{L_i - \max_{j \in S} L_j}{\max_{j \in S} L_j - \min_{j \in S} L_j}. \tag{10}$$

Further implementation details and architectural hyperparameters are in the supplementary material.

4.1. One-class novelty detection on images

To assess the model's performances in one class settings, we train it on each class of either MNIST or CIFAR-10 separately. In the test phase, we present the corresponding test set, which is composed of 10000 examples of all classes, and expect our model to assign a lower novelty score to images sharing the label with training samples. We use standard train/test splits, and isolate 10% of training samples for validation purposes, and employ it as the normalization set (S in Eq. 9) for the computation of the novelty score.
As for the baselines, we consider the following:

- standard methods such as OC-SVM [39] and Kernel Density Estimator (KDE), employed out of features extracted by PCA-whitening;
- a denoising autoencoder (DAE) sharing the same architecture as our proposal, but defective of the density estimation module. The reconstruction error is employed as a measure of normality vs. novelty;
- a variational autoencoder (VAE) [19], also sharing the same architecture as our model, in which the Evidence Lower Bound (ELBO) is employed as the score;
- Pix-CNN [45], modeling the density by applying autoregression directly in the image space;
- the GAN-based approach illustrated in [38].

We report the comparison in Tab. 1 in which performances are measured by the Area Under Receiver Operating Characteristic (AUROC), which is the standard metric for the task. As the table shows, our proposal outperforms all baselines in both settings.

Considering MNIST, most methods perform favorably. Notably, Pix-CNN fails in modeling distributions for all digits but one, possibly due to the complexity of modeling densities directly on pixel space and following a fixed autoregression order. Such poor test performances are registered despite good quality samples that we observed during training: indeed, the weak correlation between sample quality and test log-likelihood of the model has been motivated in [40]. Surprisingly, OC-SVM outperforms most deep learning based models in this setting.
On the contrary, CIFAR10 represents a much more significant challenge, as testified by the low performances of most models, possibly due to the poor image resolution and visual clutter between classes. Specifically, we observe

[1]Code to reproduce results in this section is released at `https://github.com/aimagelab/novelty-detection`.

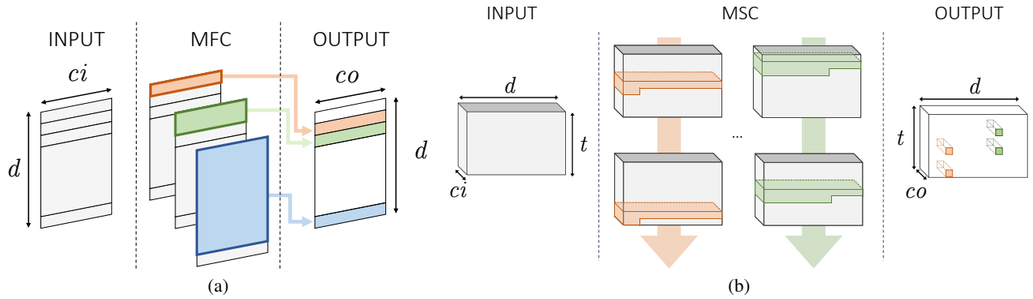


图 3：提出的自回归层，即掩码全连接层（a，公式 6）与掩码堆叠卷积层（b，公式 7）。两种层结构均采用 A 型架构，不同颜色的卷积核代表不同的参数化形式。

将 d 卷积应用于输入张量 $\mathbf{h} \in \mathbb{R}^{t \times d \times ci}$ 后，通过水平堆叠构建输出张量 $\mathbf{o} \in \mathbb{R}^{t \times d \times co}$，具体如下：

$$\mathbf{o} = \Big|\Big|_{i=1}^{d} [(\mathbf{M}^{(i)} \odot \mathbf{w}^{(i)}) * \mathbf{h}], \tag{8}$$

其中 $||$ 表示水平拼接操作。

4. 实验 [1]

我们在三种不同设置下测试我们的解决方案：图像、视频和认知数据。所有实验中，对第 i 个样本的新颖性评估均通过将重构项（REC_i）和似然对数项（LLK_i）在公式 4 中合并为单一新颖性分数 NS_i 来实现：

$$NS_i = norm_S(REC_i) + norm_S(LLK_i). \tag{9}$$

个体分数通过参考示例集 S 进行归一化（每个实验有所不同），

$$norm_S(L_i) = \frac{L_i - \max_{j \in S} L_j}{\max_{j \in S} L_j - \min_{j \in S} L_j}. \tag{10}$$

进一步的实现细节和架构超参数详见补充材料。

4.1. 图像单类新颖性检测

为了评估模型在单类设置中的性能，我们分别在 MNIST 或 CIFAR-10 的每个类别上单独训练模型。在测试阶段，我们使用由所有类别共 10000 个样本组成的对应测试集，并期望模型能为与训练样本标签相同的图像分配较低的新颖性分数。我们采用标准的训练 / 测试分割，并隔离 10% 的训练样本用于

[1]本节结果重现代码发布于 https://github.com/aimagelab/novelty-detection。

出于验证目的，我们将其作为归一化集（公式 9 中的 S）用于新颖性得分的计算。关于基线方法，我们考虑以下方案：

- 采用 PCA 白化提取特征的标准方法，如 OC-SVM [39] 和核密度估计器 (KDE)；- 使用与我们提案相同架构的去噪自编码器 (DAE)，但缺少密度估计模块。其重构误差被用作正常性与新颖性的衡量指标；- 变分自编码器 (VAE)[19]，同样采用与我们模型相同的架构，其中使用证据下界 (ELBO) 作为评分标准；- 在图像空间直接应用自回归进行密度建模的 Pix-CNN [45]，；- 基于 GAN 的方法如文献 [38] 所述。

我们在表 1 中报告了比较结果，其中性能通过接收者操作特征曲线下面积（AUROC）进行衡量，这是该任务的标准评估指标。如表所示，我们的方法在两种设置下均优于所有基线模型。

考虑到 MNIST 数据集，大多数方法都表现良好。值得注意的是，Pix-CNN 在建模除一个数字外的所有数字分布时均告失败，这可能是由于直接在像素空间建模密度并遵循固定自回归顺序的复杂性。尽管在训练过程中我们观察到了高质量的样本，但测试性能仍然很差：事实上，样本质量与模型测试对数似然之间的弱相关性已在 {v1} 中得到解释。令人惊讶的是，在这种设置下 OC-SVM 的表现优于大多数基于深度学习的模型。
相反，CIFAR10 则代表了更为严峻的挑战，大多数模型的低性能便是明证，这可能是由于图像分辨率低以及类别间视觉杂乱造成的。具体而言，我们观察到

	MNIST							CIFAR10						
	OC SVM	KDE	DAE	VAE	Pix CNN	GAN	ours	OC SVM	KDE	DAE	VAE	Pix CNN	GAN	ours
0	0.988	0.885	0.991	0.998	0.531	0.926	0.993	0.630	0.658	0.718	0.688	0.788	0.708	0.735
1	0.999	0.996	0.999	0.999	0.995	0.995	0.999	0.440	0.520	0.401	0.403	0.428	0.458	0.580
2	0.902	0.710	0.891	0.962	0.476	0.805	0.959	0.649	0.657	0.685	0.679	0.617	0.664	0.690
3	0.950	0.693	0.935	0.947	0.517	0.818	0.966	0.487	0.497	0.556	0.528	0.574	0.510	0.542
4	0.955	0.844	0.921	0.965	0.739	0.823	0.956	0.735	0.727	0.740	0.748	0.511	0.722	0.761
5	0.968	0.776	0.937	0.963	0.542	0.803	0.964	0.500	0.496	0.547	0.519	0.571	0.505	0.546
6	0.978	0.861	0.981	0.995	0.592	0.890	0.994	0.725	0.758	0.642	0.695	0.422	0.707	0.751
7	0.965	0.884	0.964	0.974	0.789	0.898	0.980	0.533	0.564	0.497	0.500	0.454	0.471	0.535
8	0.853	0.669	0.841	0.905	0.340	0.817	0.953	0.649	0.680	0.724	0.700	0.715	0.713	0.717
9	0.955	0.825	0.960	0.978	0.662	0.887	0.981	0.508	0.540	0.389	0.398	0.426	0.458	0.548
avg	0.951	0.814	0.942	0.969	0.618	0.866	**0.975**	0.586	0.610	0.590	0.586	0.551	0.592	**0.641**

Table 1: AUROC results for novelty detection on MNIST and CIFAR10. Each row represents a different class on which baselines and our model are trained.

that our proposal is the only model outperforming a simple KDE baseline; however, this finding should be put into perspective by considering the nature of non-parametric estimators. Indeed, non-parametric models are allowed to access the whole training set for the evaluation of each sample. Consequently, despite they benefit large sample sets in terms of density modeling, they lead into an unfeasible inference as the dataset grows in size.

The possible reasons behind the difference in performance w.r.t. DAE are twofold. Firstly, DAE can recognize novel samples solely based on the reconstruction error, hence relying on its memorization capabilities, whereas our proposal also considers the likelihood of their representations under the learned prior, thus exploiting surprisal as well. Secondly, by minimizing the differential entropy of the latent distribution, our proposal increases the discriminative capability of the reconstruction. Intuitively, this last statement can be motivated observing that novelty samples are forced to reside in a high probability region of the latent space, the latter bounded to solely capture unsurprising factors of variation arising from the training set. On the other hand, the gap w.r.t. VAE suggests that, for the task at hand, a more flexible autoregressive prior should be preferred over the isotropic multivariate Gaussian. On this last point, VAE seeks representations whose average surprisal converges to a fixed and expected value (i.e., the differential entropy of its prior), whereas our solution minimizes such quantity within its MLE objective. This flexibility allows modulating the richness of the latent representation vs. the reconstructing capability of the model. On the contrary, in VAEs, the fixed prior acts as a blind regularizer, potentially leading to over-smooth representations; this aspect is also appreciable when sampling from the model as shown in the supplementary material.

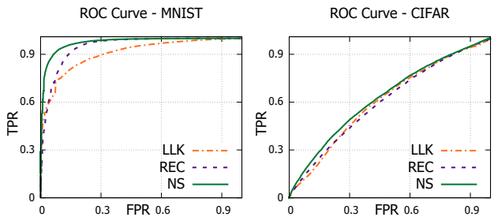


Figure 4: ROC curves delivered by different scoring strategies on MNIST and CIFAR-10 test sets. Each curve is an interpolation over the ten classes.

Fig. 4 reports an ablation study questioning the loss functions aggregation presented in Eq. 9. The figure illustrates ROC curves under three different novelty scores: i) the log-likelihood term, ii) the reconstruction term, and iii) the proposed scheme that accounts for both. As highlighted in the picture, accounting for both memorization and surprisal aspects is advantageous in each dataset. Please refer to the supplementary material for additional evidence.

4.2. Video anomaly detection

In video surveillance contexts, novelty is often considered in terms of abnormal human behavior. Thus, we evaluate our proposal against state-of-the-art anomaly detection models. For this purpose, we considered two standard benchmarks in literature, namely UCSD Ped2 [8] and ShanghaiTech [30]. Despite the differences in the number of videos and their resolution, they both contain anomalies that typically arise in surveillance scenarios (e.g., vehicles in pedestrian walkways, pick-pocketing, brawling). For UCSD Ped, we preprocessed input clips of 16 frames to extract smaller patches (we refer to supplementary materials for details) and perturbed such inputs with random Gaussian noise with $\sigma = 0.025$. We compute the novelty score of each input clip as the mean novelty score among all patches. Concerning ShanghaiTech, we removed the dependency on

	MNIST							CIFAR10						
	OC SVM	KDE	DAE	VAE	Pix CNN	GAN	ours	OC SVM	KDE	DAE	VAE	Pix CNN	GAN	ours
0	0.988	0.885	0.991	0.998	0.531	0.926	0.993	0.630	0.658	0.718	0.688	0.788	0.708	0.735
1	0.999	0.996	0.999	0.999	0.995	0.995	0.999	0.440	0.520	0.401	0.403	0.428	0.458	0.580
2	0.902	0.710	0.891	0.962	0.476	0.805	0.959	0.649	0.657	0.685	0.679	0.617	0.664	0.690
3	0.950	0.693	0.935	0.947	0.517	0.818	0.966	0.487	0.497	0.556	0.528	0.574	0.510	0.542
4	0.955	0.844	0.921	0.965	0.739	0.823	0.956	0.735	0.727	0.740	0.748	0.511	0.722	0.761
5	0.968	0.776	0.937	0.963	0.542	0.803	0.964	0.500	0.496	0.547	0.519	0.571	0.505	0.546
6	0.978	0.861	0.981	0.995	0.592	0.890	0.994	0.725	0.758	0.642	0.695	0.422	0.707	0.751
7	0.965	0.884	0.964	0.974	0.789	0.898	0.980	0.533	0.564	0.497	0.500	0.454	0.471	0.535
8	0.853	0.669	0.841	0.905	0.340	0.817	0.953	0.649	0.680	0.724	0.700	0.715	0.713	0.717
9	0.955	0.825	0.960	0.978	0.662	0.887	0.981	0.508	0.540	0.389	0.398	0.426	0.458	0.548
avg	0.951	0.814	0.942	0.969	0.618	0.866	**0.975**	0.586	0.610	0.590	0.586	0.551	0.592	**0.641**

表 1：MNIST 和 CIFAR10 上新奇检测的 AUROC 结果。每行代表基线模型与我们的模型所训练的不同类别。

我们的提案是唯一优于简单 KDE 基线的模型；但这一发现需结合非参数估计器的特性进行辩证看待。事实上，非参数模型被允许访问整个训练集以评估每个样本。因此，尽管它们在大样本集的密度建模方面具有优势，但随着数据集规模增大，会导致推理过程难以实现。

关于与 DAE 性能差异的可能原因有两点。首先，DAE 仅能基于重构误差识别新样本，因此依赖其记忆能力；而我们的方案还考虑了这些表征在所学先验下的似然度，从而同时利用了惊奇度。其次，通过最小化潜在分布的微分熵，我们的方案增强了重构的判别能力。直观而言，最后这一论断的合理性在于：新颖样本被迫位于潜在空间的高概率区域，而该区域仅能捕捉训练集中出现的非惊奇变异因素。另一方面，与 VAE 的差距表明，对于当前任务，应采用更灵活的自回归先验 —— 与各向同性的多元高斯分布相比，我们的方法更具优势。关键区别在于：VAE 追求的是平均信息量收敛于固定期望值的表征（即先验分布的微分熵），而我们的解决方案则在其最大似然估计目标中最小化该数值。这种灵活性能够调节潜在表征的丰富度与模型重建能力之间的平衡。相反在 VAE 中，固定先验如同盲目正则化器，可能导致过度平滑的表征；这一特点在补充材料所示的模型采样过程中也得到印证。

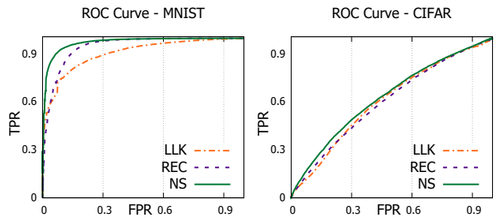


图 4：不同评分策略在 MNIST 和 CIFAR-10 测试集上呈现的 ROC 曲线。每条曲线均为对十个类别的插值结果。

图 4 展示了针对公式 9 中损失函数聚合方式的消融研究。该图呈现了三种不同新颖性评分下的 ROC 曲线：i) 对数似然项，ii) 重构项，以及 iii) 同时考虑两者的 proposed 方案。如图所示，在每项数据集中同时考虑记忆性与意外性两方面因素均能带来性能提升。更多佐证请参阅补充材料。

4.2. 视频异常检测

在视频监控场景中，异常通常指异常的人类行为。为此，我们将提出的方法与最先进的异常检测模型进行比较评估。我们采用文献中两个标准基准测试集 —— UCSD Ped2 [8] 与 ShanghaiTech[30]。尽管这两个数据集在视频数量和分辨率上存在差异，但都包含了监控场景中典型的异常行为（例如机动车驶入人行道、扒窃、斗殴）。针对 UCSD Ped 数据集，我们对 16 帧输入片段进行预处理以提取更小的图像块（详见补充材料），并使用 $\sigma = 0.025$ 的高斯噪声对这些输入进行随机扰动。每个输入片段的新颖性得分通过计算所有图像块的平均新颖值得出。对于 ShanghaiTech 数据集，我们消除了

	UCSD Ped2	ShanghaiTech
MPPCA [17]	0.693	-
MPPC+SFA [31]	0.613	-
MDT [31]	0.829	-
ConvAE [11]	0.850	0.609
ConvLSTM-AE [29]	0.881	-
Unmasking [15]	0.822	-
Hinami *et al.* [14]	0.922	-
TSC [30]	0.910	0.679
Stacked RNN [30]	0.922	0.680
FFP [26]	0.935	-
FFP+MC [26]	**0.954**	**0.728**
Ours	**0.954**	**0.725**

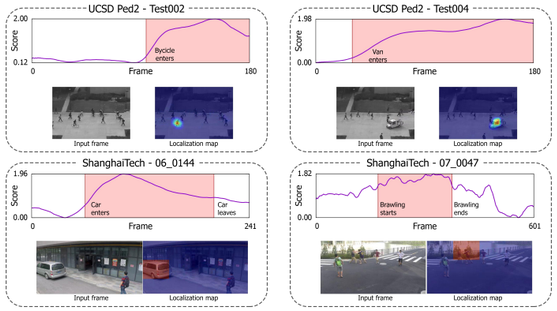


Figure 5: On the left, AUROC performances of our model w.r.t. state-of-the-art competitors. On the right, novelty scores and localizations maps for samples drawn from UCSD Ped2 and ShanghaiTech. For each example, we report the trend of the assessed score, highlighting with a different color the time range in which an anomalous subject comes into the scene.

the scenario by estimating the foreground for each frame of a clip with a standard MOG-based approach and removing the background. We fed the model with 16-frames clips, but ground-truth anomalies are labeled at frame level. In order to recover the novelty score of each frame, we compute the mean score of all clips in which it appears. We then merge the two terms of the loss function following the same strategy illustrated in Eq. 9, computing however normalization coefficients in a per-sequence basis, following the standard approach in the anomaly detection literature. The scores for each sequence are then concatenated to compute the overall AUROC of the model. Additionally, we envision localization strategies for both datasets. To this aim, for UCSD, we denote a patch exhibiting the highest novelty score in a frame as anomalous. Differently, in ShanghaiTech, we adopt a sliding-window approach [47]: as expected, when occluding the source of the anomaly with a rectangular patch, the novelty score drops significantly.
Fig. 5 reports results in comparison with prior works, along with qualitative assessments regarding the novelty score and localization capabilities. Despite a more general formulation, our proposal scores on-par with the current state-of-the-art solutions specifically designed for video applications and taking advantage of optical flow estimation and motion constraints. Indeed, in the absence of such hypotheses (FFP entry in Fig. 5), our method outperforms future frame prediction on UCSD Ped2.

4.3. Model Analysis

CIFAR-10 with semantic features. We investigate the behavior of our model in the presence of different assumptions regarding the expected nature of novel samples. We expect that, as the correctness of such assumptions increases, novelty detection performances will scale accordingly. Such a trait is particularly desirable for applications in which prior beliefs about novel examples can be envisioned. To this end, we leverage the CIFAR-10 benchmark described in Sec. 4.1 and change the type of information provided as input. Specifically, instead of raw images, we feed our model with semantic representations extracted by ResNet-50 [12], either pre-trained on Imagenet (i.e., assume semantic novelty) or CIFAR-10 itself (i.e., assume data-specific novelty). The two models achieved respectively 79.26 and 95.4 top-1 classification accuracies on the respective test sets. Even though this procedure is to be considered unfair in novelty detection, it serves as a sanity check delivering the upper-bound performances our model can achieve when applied to even better features. To deal with dense inputs, we employ a fully connected autoencoder and MFC layers within the estimation network.
Fig. 6-(a) illustrates the resulting ROC curves, where semantic descriptors improve AUROC w.r.t. raw image inputs (entry "Unsupervised"). Such results suggest that our model profitably takes advantage of the separation between normal and abnormal input representations and scales accordingly, even up to optimal performances for the task under consideration. Nevertheless, it is interesting to note how different degrees of supervision deliver significantly different performances. As expected, dataset-specific supervision increases the AUROC from 0.64 up to 0.99 (a perfect score). Surprisingly, semantic feature vectors trained on Imagenet (which contains all CIFAR classes) provide a much lower boost, yielding an AUROC of 0.72. Such result suggests that, even in the rare cases where the semantic of novelty can be known in advance, its contribution has a limited impact in modeling the normality, mostly because novelty can depend on other cues (e.g., low-level statistics).

Autoregression via recurrent layers. To measure the contribution of the proposed MFC and MSC layers described in Sec. 3, we test on CIFAR-10 and UCSD

	UCSD Ped2	ShanghaiTech
MPPCA [17]	0.693	-
MPPC+SFA [31]	0.613	-
MDT [31]	0.829	-
ConvAE [11]	0.850	0.609
ConvLSTM-AE [29]	0.881	-
Unmasking [15]	0.822	-
Hinami *et al.* [14]	0.922	-
TSC [30]	0.910	0.679
Stacked RNN [30]	0.922	0.680
FFP [26]	0.935	-
FFP+MC [26]	**0.954**	**0.728**
Ours	**0.954**	**0.725**

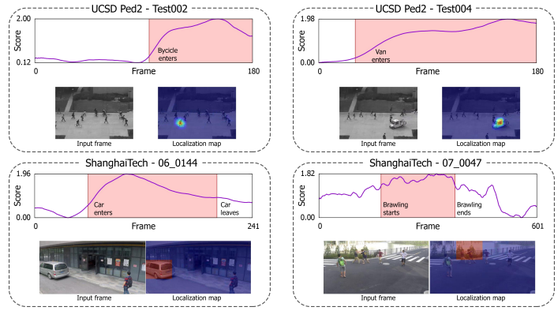


图 5：左侧为本模型相较于顶尖竞争者的 AUROC 性能表现；右侧为从 UCSD Ped2 和 ShanghaiTech 数据集中抽取样本的新颖性评分与定位图谱。每个示例均展示评估得分的走势，并使用不同颜色高亮异常目标进入场景的时间区间。

我们通过使用基于标准 MOG 的方法估计视频片段每一帧的前景并移除背景来构建场景。模型输入为 16 帧片段，但真实异常标注是在帧级别进行的。为恢复每帧的新颖性评分，我们计算该帧所在所有片段的平均得分。随后按照公式 9 所示的相同策略合并损失函数的两项，但遵循异常检测文献的标准方法，以逐序列方式计算归一化系数。各序列的评分最终被拼接起来计算模型的整体 AUROC。此外，我们为两个数据集设计了定位策略：对于 UCSD 数据集，将帧中具有最高新颖性评分的图像块标记为异常；而
在 ShanghaiTech 数据集中，我们采用滑动窗口方法 [47]—— 当用矩形遮挡块覆盖异常源时，新颖性评分会如预期般显著下降。

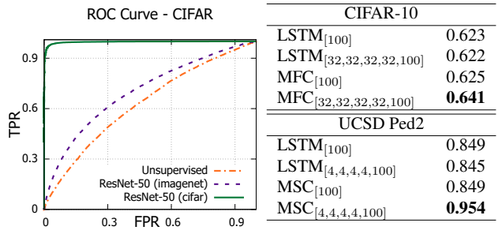
图 5 展示了与现有研究的对比结果，同时提供了关于新颖性评分和定位能力的定性评估。尽管采用了更通用的公式化方法，我们的方案在性能上与当前专门为视频应用设计的先进解决方案相当 —— 这些方案利用了光流估计和运动约束。实际上，在不依赖这些假设的情况下（图 5 中 FFP 条目所示），我们的方法
在 UCSD Ped2 数据集上超越了未来帧预测的表现。

4.3. 模型分析

带有语义特征的 CIFAR-10 数据集。我们研究了在不同假设条件下模型的行为，这些假设涉及新样本的预期性质。我们预期，随着这些假设正确性的提高，novelty 检测性能也会相应提升。这一特性对于事先对新样本存在认知的应用场景尤为重要。

可以设想。为此，我们利用第 4.1 节中描述的 CIFAR-10 基准，并改变输入信息的类型。具体而言，我们不再输入原始图像，而是向模型提供通过 ResNet-50 [12]，提取的语义表示 —— 该网络或在 Imagenet 上预训练（即假设存在语义新颖性），或在 CIFAR-10 本身上训练（即假设存在数据特定新颖性）。两个模型在各自测试集上分别达到了 79.26% 和 95.4% 的 top-1 分类准确率。尽管这种方法在新颖性检测中应被视为不公平，但它可作为验证基准，展示我们的模型在应用于更优质特征时能达到的性能上限。为处理密集输入，我们在估计网络中采用了全连接自动编码器和 MFC 层。
图 6-(a) 展示了最终的 ROC 曲线，其中语义描述符相较于原始图像输入（" 无监督 " 条目）提升了 AUROC 指标。这些结果表明，我们的模型有效利用了正常与异常输入表征之间的分离性并进行相应扩展，甚至在该任务中达到接近最优的性能。但值得注意的是，不同监督程度会带来显著差异的性能表现。正如预期，数据集特定监督将 AUROC 从 0.64 提升至 0.99（满分）。令人惊讶的是，基于 Imagenet（包含所有 CIFAR 类别）训练的语义特征向量带来的提升较为有限，仅产生 0.72 的 AUROC。这一现象表明，即使在某些罕见情况下可以预知新颖性的语义信息，其对正常性建模的贡献仍然有限，这主要是因为新颖性可能依赖于其他线索（例如低层统计特征）。

通过循环层进行自回归。为衡量第 3 节所述 MFC 与 MSC 层的贡献，我们在 CIFAR-10 和 UCSD 数据集上进行测试



(a)

CIFAR-10	
$\text{LSTM}_{[100]}$	0.623
$\text{LSTM}_{[32,32,32,32,100]}$	0.622
$\text{MFC}_{[100]}$	0.625
$\text{MFC}_{[32,32,32,32,100]}$	**0.641**
UCSD Ped2	
$\text{LSTM}_{[100]}$	0.849
$\text{LSTM}_{[4,4,4,4,100]}$	0.845
$\text{MSC}_{[100]}$	0.849
$\text{MSC}_{[4,4,4,4,100]}$	**0.954**

(b)

Figure 6: (a) CIFAR-10 ROC curves with semantic input vectors. Each curve is an interpolation among the ten classes. (b) Comparison of different architectures for the autoregressive density estimation in feature space. We indicate with $\text{LSTM}_{[F_1,F_2,\dots,F_N]}$ - same goes for MFC and MSC - the output shape for each of the N layers composing the estimator. Results are reported in terms of test AUROC.

Ped2, alternative solutions for the autoregressive density estimator. Specifically, we investigate recurrent networks, as they represent the most natural alternative featuring autoregressive properties. We benchmark the proposed building blocks against an estimator composed of LSTM layers, which is designed to sequentially observe latent symbols $\mathbf{z}_{<i}$ and output the CPD of z_i as the hidden state of the last layer. We test MFC, MSC and LSTM in single-layer and multi-layer settings, and report all outcomes in Fig. 6-(b).

It emerges that, even though our solutions perform similarly to the recurrent baseline when employed in a shallow setting, they significantly take advantage of their depth when stacked in consecutive layers. MFC and MSC, indeed, employ disentangled parametrizations for each output CPD. This property is equivalent to the adoption of a specialized estimator network for each z_i, thus increasing the proficiency in modeling the density of its designated CPD. On the contrary, LSTM networks embed all the history (i.e., the observed symbols) in their memory cells, but manipulate each input of the sequence through the same weight matrices. In such a regime, the recurrent module needs to learn parameters shared among symbols, losing specialization and eroding its modeling capabilities.

4.4. Novelty in cognitive temporal processes

As a potential application of our proposal, we investigate its capability in modeling human attentional behavior. To this end, we employ the DR(eye)VE dataset [33], introduced for the prediction of focus of attention in driving contexts. It features 74 driving videos where frame-wise fixation maps are provided, highlighting the region of the scene attended by the driver. In order to capture the dynamics of attentional patterns, we purposely discard the visual content of

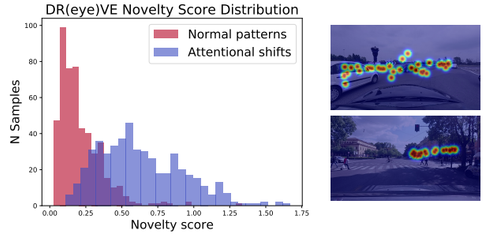


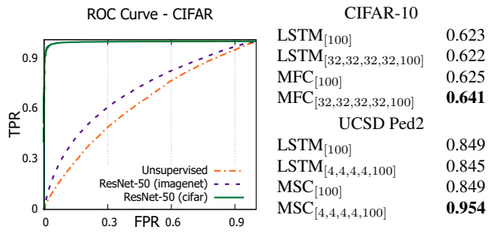
Figure 7: Left, the distribution of novelty scores assigned to normal patterns against attentional shifts labeled within the DR(eye)VE dataset. Right, DR(eye)VE clips yielding the highest novelty score (i.e., clips in which the attentional pattern shifts from the expected behavior). Interestingly, they depict some peculiar situations such as waiting for the traffic light or approaching a roundabout.

the scene and optimize our model on clips of fixation maps, randomly extracted from the training set. After training, we rely on the novelty score of each clip as a proxy for the uncommonness of an attentional pattern. Moreover, since the dataset features annotations of peculiar and unfrequent patterns (such as distractions, recording errors), we can measure the correlation of the captured novelty w.r.t. those. In terms of AUROC, our model scores 0.926, highlighting that novelty can arise from unexpected behaviors of the driver, such as distractions or other shifts in attention. Fig. 7 reports the different distribution of novelty scores for ordinary and peculiar events.

5. Conclusions

We propose a comprehensive framework for novelty detection. We formalize our model to capture the twofold nature of novelties, which concerns the incapability to remember unseen data and the surprisal aroused by the observation of their latent representations. From a technical perspective, both terms are modeled by a deep generative autoencoder, paired with an additional autoregressive density estimator learning the distribution of latent vectors by maximum likelihood principles. To this aim, we introduce two different masked layers suitable for image and video data. We show that the introduction of such an auxiliary module, operating in latent space, leads to the minimization of the encoder's differential entropy, which proves to be a suitable regularizer for the task at hand. Experimental results show state-of-the-art performances in one-class and anomaly detection settings, fostering the flexibility of our framework for different tasks without making any data-related assumption.

Acknowledgements. We gratefully acknowledge Facebook Artificial Intelligence Research and Panasonic Silicon Valley Lab for the donation of GPUs used for this research.



(a)

CIFAR-10	
$\text{LSTM}_{[100]}$	0.623
$\text{LSTM}_{[32,32,32,32,100]}$	0.622
$\text{MFC}_{[100]}$	0.625
$\text{MFC}_{[32,32,32,32,100]}$	**0.641**
UCSD Ped2	
$\text{LSTM}_{[100]}$	0.849
$\text{LSTM}_{[4,4,4,4,100]}$	0.845
$\text{MSC}_{[100]}$	0.849
$\text{MSC}_{[4,4,4,4,100]}$	**0.954**

(b)

图 6：(a) 使用语义输入向量的 CIFAR-10 ROC 曲线。每条曲线均为十个类别间的插值结果。(b) 特征空间中自回归密度估计的不同架构对比。$\text{LSTM}_{[F_1,F_2,\dots,F_N]}$ 表示估计器各 N 层的输出形状（MFC 与 MSC 标注方式相同）。实验结果以测试 AUROC 值呈现。

Ped2，自回归密度估计器的替代解决方案。具体而言，我们研究了循环网络，因为它们代表了最具自回归特性的自然替代方案。我们将提出的构建模块与由 LSTM 层组成的估计器进行基准测试，该估计器被设计为顺序观察潜在符号 $\mathbf{z}_{<i}$ 并输出 z_i 的条件概率分布作为最后一层的隐藏状态。我们在单层和多层设置中测试了 MFC 、 MSC 和 LSTM，并在图 6-(b) 中报告了所有结果。

结果表明，尽管我们的解决方案在浅层设置中表现与循环基线相似，但在连续堆叠多层时能显著发挥其深度优势。MFC 和 MSC 确实为每个输出 CPD 采用解缠结的参数化方法，这一特性等同于为每个 z_i 配置专用估计器网络，从而提升对其指定 CPD 密度建模的专业能力。相比之下，LSTM 网络虽将所有历史信息（即已观测符号）嵌入记忆单元，但通过相同的权重矩阵处理序列中的每个输入。在这种机制下，循环模块需要学习符号间共享的参数，从而丧失 specialization 并削弱其建模能力。

4.4. 认知时间过程中的新颖性

作为我们方案的一个潜在应用，我们研究了其在模拟人类注意力行为方面的能力。为此，我们采用 DR(eye)VE 数据集 [33], —— 该数据集专为驾驶场景中的注意力焦点预测而开发。该数据集包含 74 段驾驶视频，其中逐帧提供了注视点分布图，用以标示驾驶员关注的场景区域。为了捕捉注意力模式的动态特性，我们特意舍弃了视觉内容中的

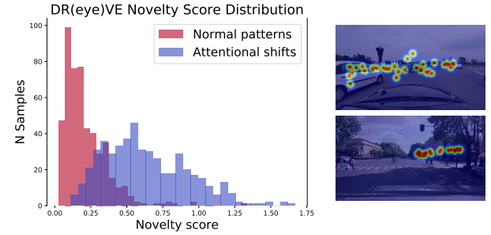


图 7：左图展示了在 DR(eye)VE 数据集中标注的注意力转移情况下，正常模式所获得的新颖性分数分布。右图呈现了 DR(eye)VE 数据集中产生最高新颖性分数的视频片段（即注意力模式偏离预期行为的片段）。值得注意的是，这些片段描绘了某些特殊场景，例如等待交通灯或接近环岛时的情形。

场景并基于从训练集中随机提取的注视点地图片段优化我们的模型。训练完成后，我们依据每个片段的新颖度评分作为注意力模式非普遍性的代理指标。此外，由于数据集包含特殊及罕见模式（如分心、录制错误）的标注，我们可以衡量所捕获新颖度与这些标注之间的相关性。在 AUROC 指标上，我们的模型达到 0.926 分，表明新颖度可能源于驾驶员的意外行为，例如分心或其他注意力转移。图 7 展示了普通事件与特殊事件在新颖度评分上的差异分布。

5. 结论

我们提出了一个用于新颖性检测的综合框架。我们将模型形式化以捕捉新颖性的双重本质：其既涉及对未见数据记忆的不可行性，也涉及观察其潜在表征所引发的惊奇感。从技术角度来看，这两个方面均通过深度生成自编码器进行建模，并辅以一个额外的自回归密度估计器，该估计器通过最大似然原则学习潜在向量的分布。为此，我们引入了两种适用于图像和视频数据的掩码层。我们证明，在潜在空间中引入此类辅助模块可导致编码器微分熵的最小化，这被证明是适用于当前任务的正则化器。实验结果显示，在单类别和异常检测设置中均实现了最先进的性能，从而证明了我们的框架在不同任务中的灵活性，且无需做出任何与数据相关的假设。

致谢。我们衷心感谢 Facebook 人工智能研究院和松下硅谷实验室为本研究捐赠 GPU 设备。

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. 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. 2

[2] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*, 2018. 4

[3] J. Ballé, V. Laparra, and E. P. Simoncelli. End-to-end optimized image compression. *International Conference on Learning Representations*, 2017. 1

[4] A. Barto, M. Mirolli, and G. Baldassarre. Novelty or surprise? *Frontiers in psychology*, 4:907, 2013. 1

[5] A. Basharat, A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[6] M. Bauer and A. Mnih. Resampled priors for variational autoencoders. *International Conference on Artificial Intelligence and Statistics*, 2019. 2

[7] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara, and N. Tishby. Detecting anomalies in peoples trajectories using spectral graph analysis. *Computer Vision and Image Understanding*, 115(8):1099–1111, 2011. 1

[8] A. Chan and N. Vasconcelos. Ucsd pedestrian database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008. 6

[9] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3449–3456. IEEE, 2011. 1, 2

[10] M. Germain, K. Gregor, I. Murray, and H. Larochelle. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889, 2015. 3

[11] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 733–742. IEEE, 2016. 1, 2, 7

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 7

[13] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, pages 6626–6637, 2017. 12

[14] R. Hinami, T. Mei, and S. Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *IEEE International Conference on Computer Vision*, pages 3639–3647, 2017. 1, 2, 7

[15] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu. Unmasking the abnormal events in video. *IEEE International Conference on Computer Vision*, 2017. 7

[16] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009. 1

[17] J. Kim and K. Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2921–2928. IEEE, 2009. 2, 7

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. 11

[19] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014. 2, 5, 11

[20] T. Kohonen. *Self-organization and associative memory*, volume 8. Springer Science & Business Media, 2012. 1

[21] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. 13

[22] A. Kumar. Computer-vision-based fabric defect detection: A survey. *IEEE Transactions on Industrial Electronics*, 55(1):348–363, 2008. 1

[23] J. Kwon and K. M. Lee. A unified framework for event summarization and rare event detection from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1737–1750, 2015. 2

[24] H. Larochelle and I. Murray. The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 29–37, 2011. 3

[25] W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):18–32, 2014. 2

[26] W. Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection – a new baseline. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 7

[27] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *IEEE International Conference on Computer Vision*, pages 2720–2727. IEEE, 2013. 2

[28] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun. Predicting deeper into the future of semantic segmentation. In *IEEE International Conference on Computer Vision*, pages 648–657, 2017. 3

[29] W. Luo, W. Liu, and S. Gao. Remembering history with convolutional lstm for anomaly detection. In *IEEE International Conference on Multimedia and Expo*, pages 439–444. IEEE, 2017. 7

[30] W. Luo, W. Liu, and S. Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. *IEEE International Conference on Computer Vision*, 2017. 1, 2, 6, 7

[31] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. IEEE, 2010. 2, 7

[32] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016. 3

参考文献

[1] A. Adam, E. Rivlin, I. Shimshoni 和 D. Reinitz。使用多固定位置监控器的鲁棒实时异常事件检测。《IEEE 模式分析与机器智能汇刊》，30(3):555–560，2008 年。2[2] S. Bai, J. Z. Kolter 和 V. Koltun。通用卷积与循环网络在序列建模中的实证评估。《*arXiv:1803.01271*》，2018 年。4[3] J. Ball´e, V. Laparra 和 E. P. Simoncelli。端到端优化的图像压缩。《国际学习表征大会》，2017 年。1[4] A. Barto, M. Mirolli 和 G. Baldassarre。新颖性或惊奇度?《心理学前沿》，4:907，2013 年。1[5] A. Basharat, A. Gritai 和 M. Shah。通过学习物体运动模式实现异常检测与改进的物体检测。见《IEEE 国际计算机视觉与模式识别会议》，第 1–8 页。IEEE，2008 年。2[6] M. Bauer 和 A. Mnih。变分自编码器的重采样先验。《国际人工智能与统计大会》，2019 年。2[7] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara 和 N. Tishby。使用谱图分析检测人员轨迹中的异常。《计算机视觉与图像理解》，115(8):1099–1111，2011 年。1[8] A. Chan 和 N. Vasconcelos。UCSD 行人数据库。《IEEE 模式分析与机器智能汇刊》，2008 年。6[9] Y. Cong, J. Yuan 和 J. Liu。基于稀疏重构成本的异常事件检测。见《IEEE 国际计算机视觉与模式识别会议》，第 3449–3456 页。IEEE，2011 年。1, 2[10] M. Germain, K. Gregor, I. Murray 和 H. Larochelle。MADE：用于分布估计的掩码自编码器。见《国际机器学习大会》，第 881–889 页，2015 年。3[11] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury 和 L. S. Davis。学习视频序列中的时间规律性。见《IEEE 国际计算机视觉与模式识别会议》，第 733–742 页。IEEE，2016 年。1, 2, 7[12] K. He, X. Zhang, S. Ren 和 J. Sun。用于图像识别的深度残差学习。见《IEEE 国际计算机视觉与模式识别会议》，第 770–778 页，2016 年。7[13] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler 和 S. Hochreiter。通过双时间尺度更新规则训练的 GAN 收敛至局部纳什均衡。见《神经信息处理系统》，第 6626–6637 页，2017 年。12[14] R. Hinami, T. Mei 和 S. Satoh。通过学习深度通用知识实现异常事件的联合检测与叙述。见《IEEE 国际计算机视觉大会》，第 3639–3647 页，2017 年。1, 2, 7[15] R. T. Ionescu, S. Smeureanu, B. Alexe 和 M. Popescu。视频中异常事件的解蔽。《IEEE 国际计算机视觉大会》，2017 年。7[16] L. Itti 和 P. Baldi。贝叶斯惊奇度吸引人类注意力。《视觉研究》，49(10):1295–1306，2009 年。1

[17] J. Kim 与 K. Grauman。《局部观察，全局推断：支持增量更新的异常活动检测时空马尔可夫随机场》。发表于 IEEE 国际计算机视觉与模式识别会议，第 2921–2928 页。IEEE 出版社，2009 年。2, 7[18] D. P. Kingma 与 J. Ba。《Adam：一种随机优化方法》。发表于国际学习表征会议，2015 年。11[19] D. P. Kingma 与 M. Welling。《变分贝叶斯自编码器》。发表于国际学习表征会议，2014 年。2, 5, 11[20] T. Kohonen。《自组织与关联记忆》第 8 卷。Springer 科学与商业媒体，2012 年。1[21] D. Koller 与 N. Friedman。《概率图模型：原理与技术 - 自适应计算与机器学习》。MIT 出版社，2009 年。13[22] A. Kumar。《基于计算机视觉的织物缺陷检测综述》。载于 IEEE 工业电子学汇刊，第 55 卷第 1 期，第 348–363 页，2008 年。1[23] J. Kwon 与 K. M. Lee。《多视角事件摘要与稀有事件检测的统一框架》。载于 IEEE 模式分析与机器智能汇刊，第 37 卷第 9 期，第 1737–1750 页，2015 年。2[24] H. Larochelle 与 I. Murray。《神经自回归分布估计器》。发表于第十四届人工智能与统计国际会议论文集，第 29–37 页，2011 年。3[25] W. Li 、 V. Mahadevan 与 N. Vasconcelos。《拥挤场景中的异常检测与定位》。载于 IEEE 模式分析与机器智能汇刊，第 36 卷第 1 期，第 18–32 页，2014 年。2[26] W. Liu 、 W. Luo 、 D. Lian 与 S. Gao。《基于未来帧预测的异常检测 —— 新基线》。发表于 IEEE 国际计算机视觉与模式识别会议，2018 年。1, 2, 7[27] C. Lu 、 J. Shi 与 J. Jia。《在 MATLAB 中以 150 帧 / 秒检测异常事件》。发表于 IEEE 国际计算机视觉会议，第 2720–2727 页。IEEE 出版社，2013 年。2[28] P. Luc 、 N. Neverova 、 C. Couprie 、 J. Verbeek 与 Y. Le-Cun。《语义分割的未来深度预测》。发表于 IEEE 国际计算机视觉会议，第 648–657 页，2017 年。3[29] W. Luo 、 W. Liu 与 S. Gao。《利用卷积 LSTM 记忆历史实现异常检测》。发表于 IEEE 国际多媒体与博览会会议，第 439–444 页。IEEE 出版社，2017 年。7[30] W. Luo 、 W. Liu 与 S. Gao。《基于稀疏编码的堆叠 RNN 异常检测框架再探索》。发表于 IEEE 国际计算机视觉会议，2017 年。1, 2, 6, 7[31] V. Mahadevan 、 W. Li 、 V. Bhalodia 与 N. Vasconcelos。《拥挤场景中的异常检测》。发表于 IEEE 国际计算机视觉与模式识别会议，第 1975–1981 页。IEEE 出版社，2010 年。2, 7[32] A. v. d. Oord 、 S. Dieleman 、 H. Zen 、 K. Simonyan 、 O. Vinyals 、 A. Graves 、 N. Kalchbrenner 、 A. Senior 与 K. Kavukcuoglu。《WaveNet：原始音频的生成模型》。arXiv 预印本 arXiv:1609.03499，2016 年。3

[33] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara. Predicting the driver's focus of attention: the dr(eye)ve project. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 8

[34] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, volume 2017, 2017. 1

[35] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe. Training adversarial discriminators for cross-channel abnormal event detection in crowds. *arXiv preprint arXiv:1706.07680*, 2017. 2

[36] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, 2018. 2

[37] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli. Adversarially learned one-class classifier for novelty detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3379–3388, 2018. 1, 2

[38] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017. 1, 2, 5

[39] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt. Support vector method for novelty detection. In *Neural Information Processing Systems*, 2000. 5

[40] L. Theis, A. v. d. Oord, and M. Bethge. A note on the evaluation of generative models. *International Conference on Learning Representations*, 2016. 5

[41] J. M. Tomczak and M. Welling. Vae with a vamp prior. *International Conference on Artificial Intelligence and Statistics*, 2018. 2

[42] M. Tribus. *Thermostatics and thermodynamics: an introduction to energy, information and states of matter, with engineering applications*. van Nostrand, CS7, 1961. 1, 2

[43] B. Uria, I. Murray, and H. Larochelle. Rnade: The real-valued neural autoregressive density-estimator. In *Advances in Neural Information Processing Systems*, pages 2175–2183, 2013. 3

[44] B. Uria, I. Murray, and H. Larochelle. A deep and tractable density estimator. In *International Conference on Machine Learning*, pages 467–475, 2014. 3

[45] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, and A. Graves. Conditional image generation with pixelcnn decoders. In *Neural Information Processing Systems*, 2016. 3, 5

[46] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *International Conference on Machine Learning*, 2016. 3, 11

[47] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014. 7

[48] B. Zhao, L. Fei-Fei, and E. P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3313–3320. IEEE, 2011. 2

[49] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018. 1, 2

[33] A. Palazzi, D. Abati, S. Calderara, F. Solera 与 R. Cucchiara。预测驾驶员注意力焦点：dr(eye)ve 项目。IEEE 模式分析与机器智能汇刊，2018 年。8[34] D. Pathak, P. Agrawal, A. A. Efros 与 T. Darrell。通过自监督预测实现好奇心驱动探索。发表于国际机器学习大会，2017 年第 1 卷。1[35] M. Ravanbakhsh, E. Sangineto, M. Nabi 与 N. Sebe。训练对抗判别器用于跨通道人群异常事件检测。arXiv 预印本 arXiv:1706.07680，2017 年。2[36]M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed 与 R. Klette。Deep-anomaly：面向拥挤场景快速异常检测的全卷积神经网络。计算机视觉与图像理解，2018 年。2[37] M. Sabokrou, M. Khalooei, M. Fathy 与 E. Adeli。基于对抗学习的单类分类器用于新颖性检测。发表于 IEEE 国际计算机视觉与模式识别会议，第 3379-3388 页，2018 年。1, 2 [38] T. Schlegl, P. Seeb¨ock, S. M. Waldstein, U. Schmidt-Erfurth 与 G. Langs。利用生成对抗网络进行无监督异常检测以指导标记发现。发表于医学影像信息处理国际会议，第 146-157 页。Springer，2017 年。1, 2, 5[39] B. Sch¨olkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor 与 J. C. Platt。新颖性检测的支持向量方法。发表于神经信息处理系统大会，2000 年。5[40] L. Theis, A. v. d. Oord 与 M. Bethge。关于生成模型评估的说明。国际学习表征大会，2016 年。5[41] J. M. Tomczak 与 M. Welling。采用 VAMP 先验的变分自编码器。国际人工智能与统计会议，2018 年。2[42] M. Tribus。热静力学与热力学：能量、信息与物态导论及工程应用。van Nostrand 出版社，CS7 系列，1961 年。1, 2[43] B. Uria, I. Murray 与 H. Larochelle。RNNADE：实值神经自回归密度估计器。发表于神经信息处理系统进展，第 2175-2183 页，2013 年。3 [44] B. Uria, I. Murray 与 H. Larochelle。一种深度可处理的密度估计器。发表于国际机器学习大会，第 467-475 页，2014 年。3[45] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals 与 A. Graves。基于 PixelCNN 解码器的条件图像生成。发表于神经信息处理系统大会，2016 年。3, 5[46] A. van den Oord, N. Kalchbrenner 与 K. Kavukcuoglu。像素循环神经网络。国际机器学习大会，2016 年。3, 11[47] M. D. Zeiler 与 R. Fergus。卷积网络的可视化与理解。发表于欧洲计算机视觉大会，第 818-833 页。Springer，2014 年。7[48] B. Zhao, L. Fei-Fei 与 E. P. Xing。通过动态稀疏编码实现视频异常事件在线检测。发表于 *IEEE* 国际计算机视觉与模式识别会议，第 3313–3320 页。IEEE，2011 年。2[49] B. Zong、 Q. Song、 M. R. Min、 W. Cheng、 C. Lumezanu、 D. Cho 和 H. Chen。基于深度自编码高斯混合模型的无监督异常检测。见国际学习表征大会，2018 年。1, 2

Supplementary material

6. On the implementation details

Architectures and hyperparameters employed for each experiment are reported in Tab. 2, in terms of the type of blocks, autoregressive layers, mini-batch size, learning rate and weight of the log-likelihood objective. All intermediate layers are Leaky ReLU activated. The objective function is optimized using Adam [18]. All hyperparameters are tuned on a held-out validation set, by minimizing the raw objective (Eq. 4 with $\lambda = 1$).

7. On the log-likelihood objective

In this section, we detail how the log-likelihood term (Eq. 4 in the main paper) has been computed and optimized. Importantly, as mentioned in the main paper, we model each CPD through a multinomial. To this aim, we firstly need that the encoder acts as a bounded function. To achieve such desideratum, we simply employ a sigmoidal activation, ensuring that latent representations $\mathbf{z} = f(\mathbf{x}; \theta_f)$ reside in $[0,1]^d$. Therefore, for each z_j with $j = 1, 2, \ldots, d$, we perform a linear quantization of the space $[0,1]$ in B bins (where B is a hyperparameter). This latter step provides for z_j a B-dimensional categorical distribution $\phi(z_j)$, highlighting the correct bin to which z_j belongs. For each CPD, such distribution will serve as ground truth for the estimator $h(\mathbf{z}; \theta_h)$, the latter coherently predicting d distributions $p(z_j|\mathbf{z}_{<j})$ across the B bins, employing a softmax activation. This way, as shown in Eq. 11, the $\mathcal{L}_{\text{LLK}}$ loss turns out to be a valid likelihood term, defined as the cross-entropy loss between each one of the estimated CPD and their categorical counterparts:

$$\mathcal{L}_{\text{LLK}}(\theta_f, \theta_h) = \mathbb{E}_{\mathbf{x}\sim P}\left[-\sum_{j=1}^{d}\sum_{k=1}^{B}\phi(z_j)_k \log(p(z_j|\mathbf{z}_{<j})_k)\right]. \tag{11}$$

It is worth noting that multinomials are just one of the plausible models for the CPDs. Indeed, if we replace them with Gaussians, the overall framework would leave standing. However, as we observed in different trials, this choice does not yield considerable improvements but rather numerical instabilities, as described in prior works [46].

8. On the relations to Variational Autoencoders

Our framework yields some similarities with the Variational Autoencoder (VAE) [19]. Indeed, they both approximate the integral of Eq. 1 in the main paper through the minimization of the reconstruction error under a regularization constraint involving a prior distribution on latent vectors. However, it is worth noting several fundamental distinctions. Firstly, our model does not provide an explicit strategy to sample from the posterior distribution, thus resulting in a deterministic mapping from the input to the hidden representation. Secondly, while VAE specifies an explicit and adamant form for modeling the prior $p(\mathbf{z})$, in our formulation its landscape is free from any assumption and directly learnable as a result of the estimator's autoregressive nature. On this point, our proposal leads to two beneficial aspects. First, as the VAE forces the codes' distribution to match the prior, their differential entropy converges to be the same as the prior. This behavior results in approximately stationary entropies across different settings (appreciable in Fig. 2 in the main paper, where we discuss the intuition behind the entropy minimization within a novelty detection task). Secondly, the employment of a too simplistic prior may lead to over-regularized representations, whereas our proposal is less prone to such risk. Empirical evidence of

	MNIST	CIFAR-10	UCSD Ped2	ShanghaiTech	DR(eye)VE
Input Shape	1,28,28	3,32,32	1,8,32,32*	3,16,256,512	1,16,160,256
Encoder Network	$\mathrm{D}^{32}_{2,2}$ $\mathrm{D}^{64}_{2,2}$ FC^{64} FC^{64}	2D $\mathrm{Conv}^{32}_{3\times3}$ R^{32} $\mathrm{D}^{64}_{2,2}$ $\mathrm{D}^{128}_{2,2}$ $\mathrm{D}^{256}_{2,2}$ FC^{256} FC^{64}	$\mathrm{D}^{8}_{1,2,2}$ $\mathrm{D}^{16}_{2,1,1}$ $\mathrm{D}^{32}_{1,2,2}$ $\mathrm{D}^{64}_{2,1,1}$ $\mathrm{D}^{40}_{1,2,2}$ TFC^{64}	$\mathrm{D}^{8}_{1,2,2}$ $\mathrm{D}^{16}_{1,2,2}$ $\mathrm{D}^{32}_{2,2,2}$ $\mathrm{D}^{64}_{1,2,2}$ $\mathrm{D}^{64}_{2,2,2}$ TFC^{512} TFC^{64}	$\mathrm{D}^{8}_{1,2,2}$ $\mathrm{D}^{16}_{1,2,2}$ $\mathrm{D}^{32}_{2,2,2}$ $\mathrm{D}^{64}_{1,2,2}$ $\mathrm{D}^{64}_{2,2,2}$ TFC^{512} TFC^{64}
Decoder Network	FC^{64} FC^{64} $\mathrm{U}^{32}_{2,2}$ $\mathrm{U}^{16}_{2,2}$ 2D $\mathrm{Conv}^{1}_{1\times1}$	FC^{256} FC^{256} $\mathrm{U}^{128}_{2,2}$ $\mathrm{U}^{64}_{2,2}$ $\mathrm{U}^{32}_{2,2}$ R^{32} 2D $\mathrm{Conv}^{3}_{1\times1}$	TFC^{64} $\mathrm{U}^{40}_{1,2,2}$ $\mathrm{U}^{27}_{1,2,2}$ $\mathrm{U}^{18}_{2,1,1}$ $\mathrm{U}^{12}_{1,2,2}$ $\mathrm{U}^{8}_{2,1,1}$ 3D $\mathrm{Conv}^{1}_{1\times1}$	TFC^{64} TFC^{512} $\mathrm{U}^{64}_{2,2,2}$ $\mathrm{U}^{32}_{1,2,2}$ $\mathrm{U}^{16}_{2,2,2}$ $\mathrm{U}^{8}_{1,2,2}$ 3D $\mathrm{Conv}^{3}_{1\times1}$	TFC^{64} TFC^{512} $\mathrm{U}^{64}_{2,2,2}$ $\mathrm{U}^{32}_{1,2,2}$ $\mathrm{U}^{16}_{2,2,2}$ $\mathrm{U}^{8}_{1,2,2}$ 3D $\mathrm{Conv}^{1}_{1\times1}$
Estimator Network	MFC^{32} MFC^{32} MFC^{32} MFC^{32} MFC^{100}	MFC^{32} MFC^{32} MFC^{32} MFC^{32} MFC^{100}	MSC^{4} MSC^{4} MSC^{4} MSC^{4} MSC^{100}	MSC^{4} MSC^{4} MSC^{100}	MSC^{4} MSC^{4} MSC^{4} MSC^{4} MSC^{100}
Mini Batch	256	256	2760	8	16
Learning Rate	10^{-4}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
λ	1	0.1	0.1	1	1

*Patches extracted from input clips having shape 1,16,256,384.

Table 2: Architectural and optimization hyperparameters of each setting. We denote with D^C_S (downsampling), U^C_S (upsampling) and R^C (residual) the parametrizations for the employed building blocks (see Fig. 1ii in the main paper). On the one hand, C is the number of output channels, whereas S is the stride of the first convolution in the block. Additionally, FC^C and TFC^C denote dense layers and temporally-shared full connections respectively (in this case, C is the number of output features). Finally, we refer to MFC^C and MSC^C for the proposed autoregressive layers, illustrated in Fig. 3 in the manuscript. For a comprehensive description of each type of layer, please refer to Sec. 3.1 of the main paper.

补充材料

6. 关于实现细节

各实验采用的架构与超参数详见表 2，包括模块类型、自回归层数、小批量规模、学习率以及对数似然目标的权重。所有中间层均采用 Leaky ReLU 激活函数。使用 Adam [18] 优化器对目标函数进行优化，所有超参数通过在保留验证集上最小化原始目标函数（公式 4 结合 $\lambda = 1$）进行调优。

7. 关于对数似然目标函数

在本节中，我们将详细说明如何计算和优化对数似然项（主论文中的公式 4）。重要的是，如主论文所述，我们通过多项分布对每个条件概率分布进行建模。为此，我们首先需要编码器充当有界函数。为实现这一目标，我们简单地采用 S 形激活函数，确保潜在表示 $\mathbf{z} = f(\mathbf{x}; \theta_f)$ 位于 $[0,1]^d$ 范围内。因此，对于每个 z_j（其中 $j = 1$、2、$\ldots$、d），我们对空间 $[0,1]$ 执行线性量化，将其划分为 B 个区间（其中 B 为超参数）。此步骤为 z_j 生成一个 B 维分类分布 $\phi(z_j)$，标示出 z_j 所属的正确区间。对于每个条件概率分布，该分布将作为估计器 $h(\mathbf{z}; \theta_h)$的真实标签，该估计器通过 softmax 激活函数一致地预测 d 个分布 $p(z_j|\mathbf{z}_{<j})$，覆盖全部 B 个区间。如此，如公式 11 所示，$\mathcal{L}_{\text{LLK}}$ 损失转化为有效的似然项，定义为每个估计条件概率分布与其对应分类分布之间的交叉熵损失：

$$\mathcal{L}_{\text{LLK}}(\theta_f, \theta_h) = \mathbb{E}_{\mathbf{x}\sim P}\left[-\sum_{j=1}^{d}\sum_{k=1}^{B}\phi(z_j)_k \log(p(z_j|\mathbf{z}_{<j})_k)\right]. \tag{11}$$

值得注意的是，多项式只是 CPD 的合理模型之一。实际上，若将其替换为高斯模型，整体框架仍能成立。但正如我们在不同试验中观察到的，这种选择并未带来显著改进，反而会引发数值不稳定问题 —— 正如先前研究 [46] 所述。

8. 关于与变分自编码器的关系

我们的框架与变分自编码器（VAE） [19] 存在一些相似之处。事实上，它们都通过最小化重构误差来近似主论文中公式 1 的积分，同时受到涉及隐向量先验分布的正则化约束。然而，有几个根本区别值得注意。首先，我们的模型没有提供从后验分布采样的显式策略，因此形成了从输入到隐层表示的确定性映射。其次，VAE 为建模先验 $p(\mathbf{z})$ 指定了显式且固定的形式，而在我们的公式中，其分布形态不受任何假设约束，并可直接通过估计器的自回归特性进行学习。这一点使我们的方案具有两个优势：第一，由于 VAE 强制编码分布与先验匹配，其微分熵会收敛至与先验相同。这种行为导致不同设置下的熵值近似恒定（可参见主论文图 2，我们在其中讨论了新颖性检测任务中熵最小化的直观意义）；第二，使用过于简单的先验可能导致过度正则化的表示，而我们的方案不易出现此类风险。相关实证证据

	MNIST	CIFAR-10	UCSD Ped2	ShanghaiTech	DR(eye)VE
Input Shape	1,28,28	3,32,32	1,8,32,32*	3,16,256,512	1,16,160,256
Encoder Network	$\mathrm{D}^{32}_{2,2}$ $\mathrm{D}^{64}_{2,2}$ FC^{64} FC^{64}	2D $\mathrm{Conv}^{32}_{3\times3}$ R^{32} $\mathrm{D}^{64}_{2,2}$ $\mathrm{D}^{128}_{2,2}$ $\mathrm{D}^{256}_{2,2}$ FC^{256} FC^{64}	$\mathrm{D}^{8}_{1,2,2}$ $\mathrm{D}^{16}_{2,1,1}$ $\mathrm{D}^{32}_{1,2,2}$ $\mathrm{D}^{64}_{2,1,1}$ $\mathrm{D}^{40}_{1,2,2}$ TFC^{64}	$\mathrm{D}^{8}_{1,2,2}$ $\mathrm{D}^{16}_{1,2,2}$ $\mathrm{D}^{32}_{2,2,2}$ $\mathrm{D}^{64}_{1,2,2}$ $\mathrm{D}^{64}_{2,2,2}$ TFC^{512} TFC^{64}	$\mathrm{D}^{8}_{1,2,2}$ $\mathrm{D}^{16}_{1,2,2}$ $\mathrm{D}^{32}_{2,2,2}$ $\mathrm{D}^{64}_{1,2,2}$ $\mathrm{D}^{64}_{2,2,2}$ TFC^{512} TFC^{64}
Decoder Network	FC^{64} FC^{64} $\mathrm{U}^{32}_{2,2}$ $\mathrm{U}^{16}_{2,2}$ 2D $\mathrm{Conv}^{1}_{1\times1}$	FC^{256} FC^{256} $\mathrm{U}^{128}_{2,2}$ $\mathrm{U}^{64}_{2,2}$ $\mathrm{U}^{32}_{2,2}$ R^{32} 2D $\mathrm{Conv}^{3}_{1\times1}$	TFC^{64} $\mathrm{U}^{40}_{1,2,2}$ $\mathrm{U}^{27}_{1,2,2}$ $\mathrm{U}^{18}_{2,1,1}$ $\mathrm{U}^{12}_{1,2,2}$ $\mathrm{U}^{8}_{2,1,1}$ 3D $\mathrm{Conv}^{1}_{1\times1}$	TFC^{64} TFC^{512} $\mathrm{U}^{64}_{2,2,2}$ $\mathrm{U}^{32}_{1,2,2}$ $\mathrm{U}^{16}_{2,2,2}$ $\mathrm{U}^{8}_{1,2,2}$ 3D $\mathrm{Conv}^{3}_{1\times1}$	TFC^{64} TFC^{512} $\mathrm{U}^{64}_{2,2,2}$ $\mathrm{U}^{32}_{1,2,2}$ $\mathrm{U}^{16}_{2,2,2}$ $\mathrm{U}^{8}_{1,2,2}$ 3D $\mathrm{Conv}^{1}_{1\times1}$
Estimator Network	MFC^{32} MFC^{32} MFC^{32} MFC^{32} MFC^{100}	MFC^{32} MFC^{32} MFC^{32} MFC^{32} MFC^{100}	MSC^{4} MSC^{4} MSC^{4} MSC^{4} MSC^{100}	MSC^{4} MSC^{4} MSC^{100}	MSC^{4} MSC^{4} MSC^{4} MSC^{4} MSC^{100}
Mini Batch	256	256	2760	8	16
Learning Rate	10^{-4}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
λ	1	0.1	0.1	1	1

*Patches extracted from input clips having shape 1,16,256,384.

表 2：各配置的架构与优化超参数。我们以 D^C_S（下采样）、U^C_S（上采样）和 R^C（残差）表示所采用构建模块的参数化配置（详见主论文图 1ii）。其中 C 表示输出通道数，S 表示模块中首个卷积层的步长。全连接层 FC^C 与时序共享全连接层 TFC^C 分别表示稠密层和时序共享的全连接层（此时 C 指输出特征数）。最后，MFC^C 和 MSC^C 对应本文提出的自回归层（详见稿件图 3）。关于各类层的完整说明，请参阅主论文第 3.1 节。

Figure 8: For all CIFAR-10 classes (organized in different rows), we report images sampled from VAEs (left) and the proposed autoencoders with autoregressive priors. As can be seen, our samples visually exhibit fine-grained details and sharpness, differently from the heavily blurred ones coming from VAEs. Finally, the over-regularization arising from VAE is confirmed when looking at FID scores (at the extremes of the figure, the lower, the better).

such behavior can also be appreciated in Fig. 8, where we draw new samples from VAE and our model, both of which has been trained on CIFAR-10. All settings being equal, our hallucinations are visually much more realistic than the ones coming from VAEs, the latter leading to over-smooth shapes and lacking any details, as further confirmed by the substantial differences in Fréchet Inception Distance (FID) scores [13].

9. On the dual nature of novelty

In this section, we stress how significant is the presence of both terms for obtaining a highly discriminative novelty score (NS, Eq. 9 in the main paper): namely the reconstruction error (REC), modeling the memory capabilities, and the log-likelihood term (LLK), capturing the surprisal inducted from latent representations. Aiming to reinforce this latter point, just briefly illustrated in Fig. 4 of the manuscript, we report in Tab. 3 performances - expressed in AUROC - delivered by different scoring strategies on each setting mentioned in the main paper. Except for ShanghaiTech, we systematically observe a reward in accounting for both aspects. Furthermore, for MNIST and CIFAR-10, we find particularly interesting the gap in performance arising from our reconstruction error w.r.t. the one arising from the denoising autoencoder (DAE) variants (0.942 and 0.590 for the two datasets respectively, as reported in Tab. 1 of the main paper). In this respect, we gather new evidence supporting that surprisal minimization acts as a novelty-oriented

	LLK	REC	NS
MNIST	0.926	0.949	**0.975**
CIFAR-10	0.627	0.603	**0.641**
UCSD Ped2	0.933	0.909	**0.954**
ShanghaiTech	0.695	**0.726**	0.725
DR(eye)VE	0.917	0.863	**0.926**

Table 3: For each setting, AUROC performances under three different novelty scores: i) the log-likelihood term (LLK), ii) the reconstruction term (REC), and iii) the proposed scheme accounting for both (NS).

图 8：针对所有 CIFAR-10 类别（按不同行排列），我们展示了从 VAE （左侧）和采用自回归先验的提议自编码器中采样的图像。可见，我们的样本在视觉上呈现出细腻的细节和清晰度，这与 VAE 产生的严重模糊图像形成鲜明对比。最后，通过观察 FID 分数（位于图表两端，数值越低越好）可以证实 VAE 存在的过度正则化问题。

此类行为亦可在图 8 中得到印证 —— 我们分别从 VAE 与本文模型中抽取新样本，二者均在 CIFAR-10 数据集上完成训练。在同等条件下，本模型生成的幻象视觉上远比 VAE 产生的更为真实，后者会导致过度平滑的形态且缺乏细节特征，这一结论进一步得到了两者在 Fr{v1} echet Inception Distance （ FID ）分数 {v2} 上的显著差异所证实。

9. 论新颖性的双重本质

在本节中，我们重点强调了同时保留两项要素对于获得高区分度新颖性评分（ NS，主论文公式 9 ）的重要性：即建模记忆能力的重构误差（ REC ），以及从潜在表示中提取意外信息的对数似然项（ LLK ）。为强化这一在稿件图 4 中仅简要说明的观点，我们在表 3 中汇报了主论文所述各场景下不同评分策略的表现（以 AUROC 表示）。除 ShanghaiTech 数据集外，我们系统性地观察到兼顾两项要素的评估方式能持续带来性能提升 ——

此外，对于 MNIST 和 CIFAR-10 数据集，我们发现由我们的重构误差产生的性能差距与去噪自编码器（ DAE ）变体产生的性能差距尤为引人关注（根据主论文表 1 报告，两个数据集上的数值分别为 0.942 和 0.590 ）。在这方面，我们收集到的新证据表明，惊奇最小化起到了面向新颖性的 {v*} 作用。

	LLK	REC	NS
MNIST	0.926	0.949	**0.975**
CIFAR-10	0.627	0.603	**0.641**
UCSD Ped2	0.933	0.909	**0.954**
ShanghaiTech	0.695	**0.726**	0.725
DR(eye)VE	0.917	0.863	**0.926**

表 3：针对每种设置，三种不同新颖性评分下的 AUROC 性能表现：i) 对数似然项（ LLK ），ii) 重构项（ REC ），以及 iii) 同时考虑二者的 proposed 方案（ NS ）。

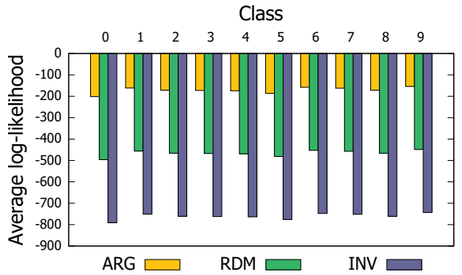


Figure 9: Sample training log-likelihood of a Bayesian Network modeling the distribution of latent codes produced by the encoder of our model trained on MNIST digits. When the BN structure resembles the autoregressive order imposed during training, a much higher likelihood is achieved. This behavior is consistent in all classes and supports the capability of the encoder to produce codes that respect a pre-imposed autoregressive structure.

regularizer for the overall architecture, as it improves the discriminative capability of the reconstruction (as already conjectured in Sec. 4.1 of the main paper).

10. On the causal structure of representations

We now investigate the capability of our encoder to produce representations that respect the autoregressive causal structure imposed by the LLK loss (mentioned in Sec. 3 of the main paper). To this aim, we extract representations out of the ten models trained on MNIST digits and fit their distribution using a structured density estimator. Specifically, we employ Bayesian Networks (BNs) with different autoregressive structures. In this respect, each BN is modeled with Linear Gaussians [21], s.t. each CPD $p(z_i|Pa(z_i))$ with $i = 1, 2, \ldots, d$ is given by:

$$p(z_i|Pa(z_i)) = \mathcal{N}(z_i \mid w_0^{(i)} + \sum_{z_j \in Pa(z_i)} w_j^{(i)} z_j, \sigma_i^2), \quad (12)$$

where each $w_j^{(i)}$, σ_i^2 are learnable parameters. We indicate with $Pa(z_i)$ the parent variables of z_i in the BN. The previous equation holds for all nodes, except for the root one, which is modeled through a Gaussian distribution. Concerning the BN structure, we test:

- Autoregressive order: the BN structure follows the autoregressive order imposed during training, namely $Pa(z_i) = \{z_j \mid j = 1, 2, \ldots, i-1\}$
- Random order: the BN structure follows a random autoregressive order.
- Inverse order: the BN structure follows an autoregressive order which is the inverse with respect to the one imposed during training, namely $Pa(z_i) = \{z_j \mid j = i+1, i+2, \ldots, d\}$

It is worth noting that, as the three structures exhibit the same number of edges and independent parameters, the difference in their fitting capabilities is only due to the causal order imposed over variables.
Fig. 9 reports the sample training log-likelihood of all BN models. Remarkably, the autoregressive order delivers a better fit, supporting the capability of the encoder network to extract features with learned autoregressive properties. Moreover, to show that this result is not due to overfitting or other lurking behaviors, we report in Tab. 4 log-likelihoods for training, validation and test set.

11. On the entropy minimization

To provide an additional grasp about the role of the representation's entropy minimization, we focus on a single MNIST digit (class 7) and report in Fig. 10 some randomly sampled reconstructions from the training set. Such reconstructions are learned under three different regularization regimes, represented by different weights on the log-likelihood objective (λ, Eq. 4 in the main paper). As shown in Fig. 10, higher degrees of regularization (i.e., stricter constraints on entropy) deliver near mode-collapsed reconstructions, losing sharp variations in favor of capturing fewer prototypes for the input distribution.

12. On the complexity of autoregressive layers

In this section, we briefly discuss the complexity of Masked Fully Connected (MFC) and Masked Stacked Convolution

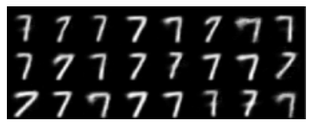
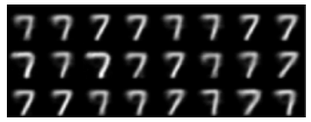
Loss weight	Reconstructions
$\lambda = 0.01$	
$\lambda = 1$	
$\lambda = 100$	

Figure 10: MNIST reconstructions delivered by different values of λ, the latter controlling the impact of the differential entropy minimization.

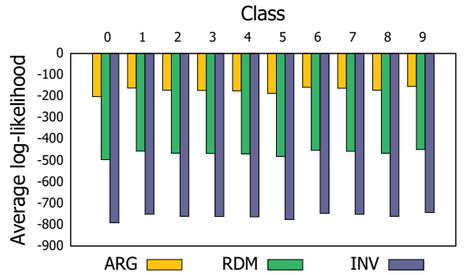


图 9：贝叶斯网络建模我们模型在 MNIST 数字数据集上训练时编码器所生成隐码分布的样本训练对数似然。当 BN 结构接近训练时施加的自回归顺序，可获得显著更高的似然值。该现象在所有类别中保持一致，证实了编码器能够生成遵循预设自回归结构的编码。

整体架构的正则化器，因为它提升了重建的判别能力（正如主论文第 4.1 节中已推测的那样）。

10. 关于表征的因果结构

我们现在研究编码器生成表征的能力，这些表征遵循主论文第 3 节中提到的 LLK 损失所施加的自回归因果结构。为此，我们从在 MNIST 数字上训练的十个模型中提取表征，并使用结构化密度估计器拟合它们的分布。具体来说，我们采用具有不同自回归结构的贝叶斯网络（ BNs ）。在这方面，每个 BN 都使用线性高斯模型 [21]，进行建模，使得每个条件概率分布 $p(z_i|Pa(z_i))$ 在 $i = 1$、 2、 ...、 d 条件下由以下公式给出：

$$p(z_i|Pa(z_i)) = \mathcal{N}(z_i \mid w_0^{(i)} + \sum_{z_j \in Pa(z_i)} w_j^{(i)} z_j, \sigma_i^2), \quad (12)$$

其中每个 $w_j^{(i)}$、 σ_i^2 都是可学习参数。我们用 $Pa(z_i)$ 表示贝叶斯网络中 z_i 的父变量。上述等式对所有节点都成立，除根节点外 —— 根节点通过高斯分布进行建模。关于贝叶斯网络结构，我们测试了以下方案：

- 自回归顺序：BN 结构遵循训练期间施加的自回归顺序，即 $Pa(z_i) = \{z_j \mid j = 1, 2, \ldots, i-1\}$
- 随机顺序：BN 结构遵循随机的自回归顺序。
- 逆序：BN 结构遵循自回归顺序，该顺序与在训练过程中施加的顺序即 $Pa(z_i) = \{z_j \mid j = i+1, i+2, \ldots, d\}$

值得注意的是，由于这三种结构具有相同数量的边和独立参数，它们拟合能力的差异仅源于变量间施加的因果顺序。
图 9 展示了所有 BN 模型的训练对数似然样本。值得注意的是，自回归排序实现了更好的拟合效果，这印证了编码器网络能够提取具有学习自回归特性的特征。此外，为证明该结果并非由过拟合或其他潜在行为导致，我们在表 4 中同步呈现了训练集、验证集与测试集的对数似然值。

11. 关于熵最小化

为了进一步理解表示熵最小化的作用，我们聚焦于单个 MNIST 数字（类别 7 ），并在图 10 中展示了从训练集中随机抽取的部分重建样本。这些重建结果是在三种不同正则化机制下学习得到的，通过对数似然目标函数（ λ，主论文公式 4 ）的不同权重来体现。如图 10 所示，更高强度的正则化（即对熵的更严格约束）会生成近乎模式坍塌的重建结果，牺牲了鲜明特征变化以捕捉输入分布中更少的原型。

12. 论自回归层的复杂性

在本节中，我们简要讨论掩码全连接（ MFC ）和掩码堆叠卷积的复杂度。

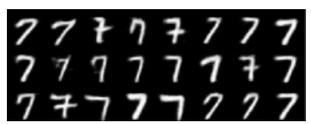
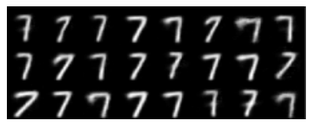
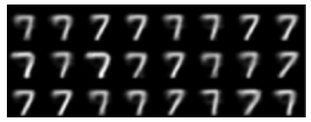
Loss weight	Reconstructions
$\lambda = 0.01$	
$\lambda = 1$	
$\lambda = 100$	

图 10：由不同 λ 值提供的 MNIST 重建结果，后者控制着微分熵最小化的影响程度。

		Classes									
		0	1	2	3	4	5	6	7	8	9
ARG	Train	-201.60	-161.60	-171.43	-172.73	-174.17	-186.48	-158.22	-162.37	-171.65	-154.11
	Val	-200.96	-160.38	-170.10	-172.29	-173.85	-185.25	-157.22	-162.20	-171.42	-154.02
	Test	-200.89	-159.73	-169.64	-170.75	-172.40	-184.27	-157.74	-161.65	-170.10	-152.70
RDM	Train	-496.33	-456.34	-466.16	-467.47	-468.90	-481.21	-452.95	-457.10	-466.39	-448.84
	Val	-495.69	-455.11	-464.83	-467.02	-468.58	-479.98	-451.95	-456.93	-466.15	-448.75
	Test	-495.62	-454.47	-464.37	-465.48	-467.13	-479.00	-452.48	-456.38	-464.83	-447.43
INV	Train	-791.06	-751.07	-760.89	-762.20	-763.63	-775.94	-747.68	-751.83	-761.12	-743.57
	Val	-790.42	-749.84	-759.56	-761.75	-763.31	-774.71	-746.68	-751.66	-760.88	-743.48
	Test	-790.35	-749.20	-759.11	-760.22	-761.86	-773.73	-747.21	-751.12	-759.56	-742.16

Table 4: Sample log-likelihood obtained by different BN structures when fitting MNIST representations. Each BN is trained on latent codes computed from the training set of a single class, following either the autoregression order (ARG), a random order (RDM) or the order inverse to autoregression (INV). We report the log-likelihood also on the validation and test set. For train-val-test split, see Sec 4.1 of the paper. Only "normal" test samples are used in this evaluation.

(MSC) layers (Fig. 3 of the main paper)[2]: adhering to the notation introduced in Sec. 3 from the main paper, MFC exhibits $\frac{d^2+d}{2} \cdot ci \cdot co + d \cdot co$ trainable parameters and a computational complexity $\mathcal{O}(d^2 \cdot ci \cdot co)$. MSC, instead, features $\frac{3d^2+d}{2} ci \cdot co + d \cdot c_o$ free parameters and a time complexity $\mathcal{O}(d^2 \cdot ci \cdot co \cdot t)$.

13. On the localizations and novelty scores in video anomaly detection

We show in Fig. 11 other qualitative evidence of the behavior of our model in video anomaly detection settings, namely UCSD Ped2 and ShanghaiTech.

[2]We refer to the type 'B' of both layers, since it is an upper bound to the type 'A'

		Classes									
		0	1	2	3	4	5	6	7	8	9
ARG	Train	-201.60	-161.60	-171.43	-172.73	-174.17	-186.48	-158.22	-162.37	-171.65	-154.11
	Val	-200.96	-160.38	-170.10	-172.29	-173.85	-185.25	-157.22	-162.20	-171.42	-154.02
	Test	-200.89	-159.73	-169.64	-170.75	-172.40	-184.27	-157.74	-161.65	-170.10	-152.70
RDM	Train	-496.33	-456.34	-466.16	-467.47	-468.90	-481.21	-452.95	-457.10	-466.39	-448.84
	Val	-495.69	-455.11	-464.83	-467.02	-468.58	-479.98	-451.95	-456.93	-466.15	-448.75
	Test	-495.62	-454.47	-464.37	-465.48	-467.13	-479.00	-452.48	-456.38	-464.83	-447.43
INV	Train	-791.06	-751.07	-760.89	-762.20	-763.63	-775.94	-747.68	-751.83	-761.12	-743.57
	Val	-790.42	-749.84	-759.56	-761.75	-763.31	-774.71	-746.68	-751.66	-760.88	-743.48
	Test	-790.35	-749.20	-759.11	-760.22	-761.86	-773.73	-747.21	-751.12	-759.56	-742.16

表 4：不同 BN 结构在拟合 MNIST 表征时获得的样本对数似然。每个 BN 均在单类别训练集的潜码上训练，遵循自回归序（ARG）、随机序（RDM）或逆自回归序（INV）。我们同时报告了验证集和测试集上的对数似然。关于训练 - 验证 - 测试集划分方法，请参阅论文第 4.1 节。本评估仅使用 " 正常 " 测试样本。

(MSC) 层（主论文图 3）[2]：遵循主论文第 3 节引入的符号表示，MFC 具有 $\frac{d^2+d}{2} \cdot ci \cdot co + d \cdot co$ 个可训练参数，计算复杂度为 $\mathcal{O}(d^2 \cdot ci \cdot co)$。而 MSC 则具有 $\frac{3d^2+d}{2} ci \cdot co + d \cdot c_o$ 个自由参数，时间复杂度为 $\mathcal{O}(d^2 \cdot ci \cdot co \cdot t)$。

13. 视频异常检测中的定位与新颖性评分

我们在图 11 中展示了模型在视频异常检测场景下行为的其他定性证据，即 UCSD Ped2 和上海科技大学数据集。

[2]我们将这两层的类型称为 "B"，因为它是类型 "A" 的上界

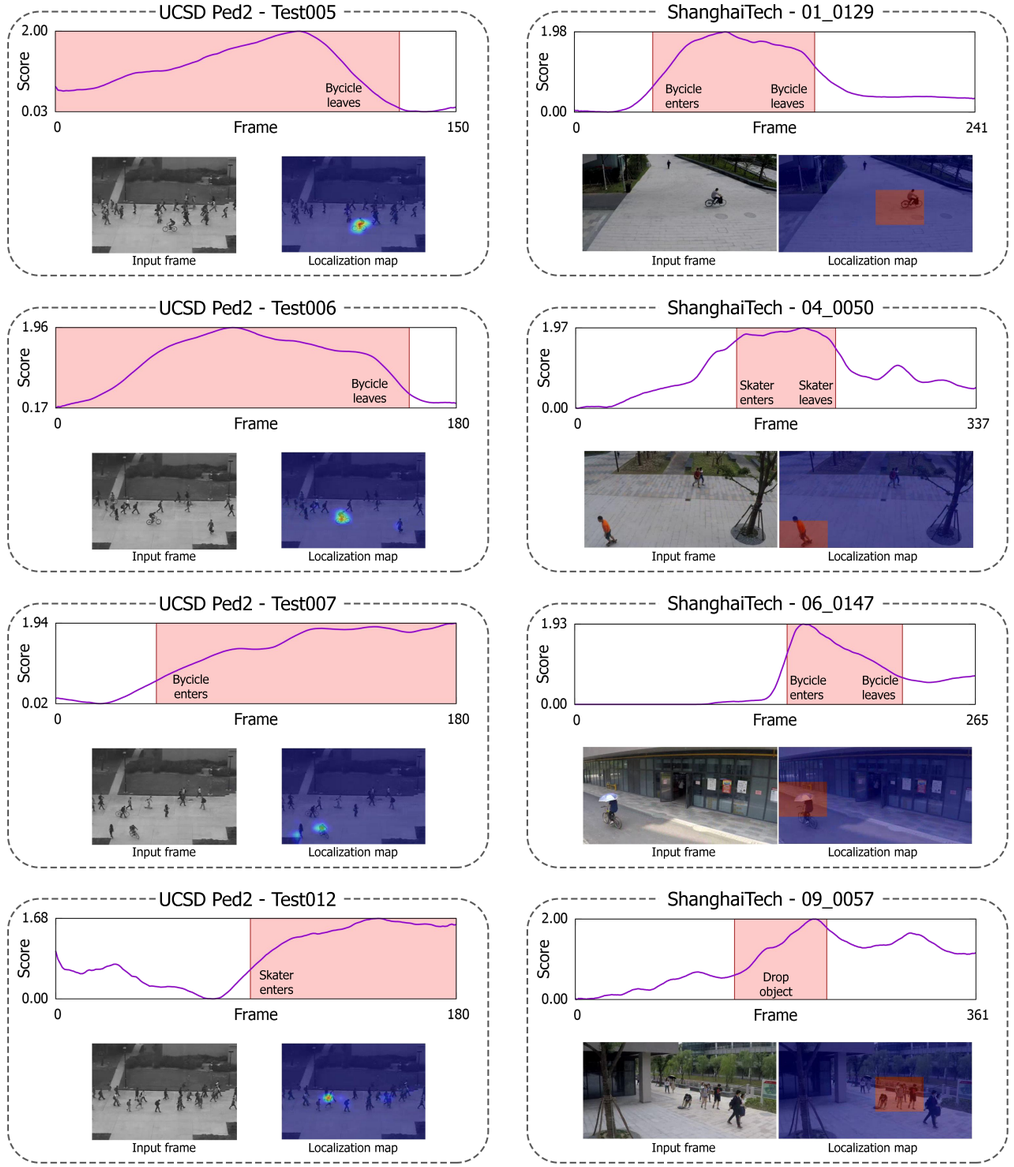


Figure 11: Novelty scores and localizations maps for several test clips from UCSD Ped2 (left) and ShanghaiTech (right).

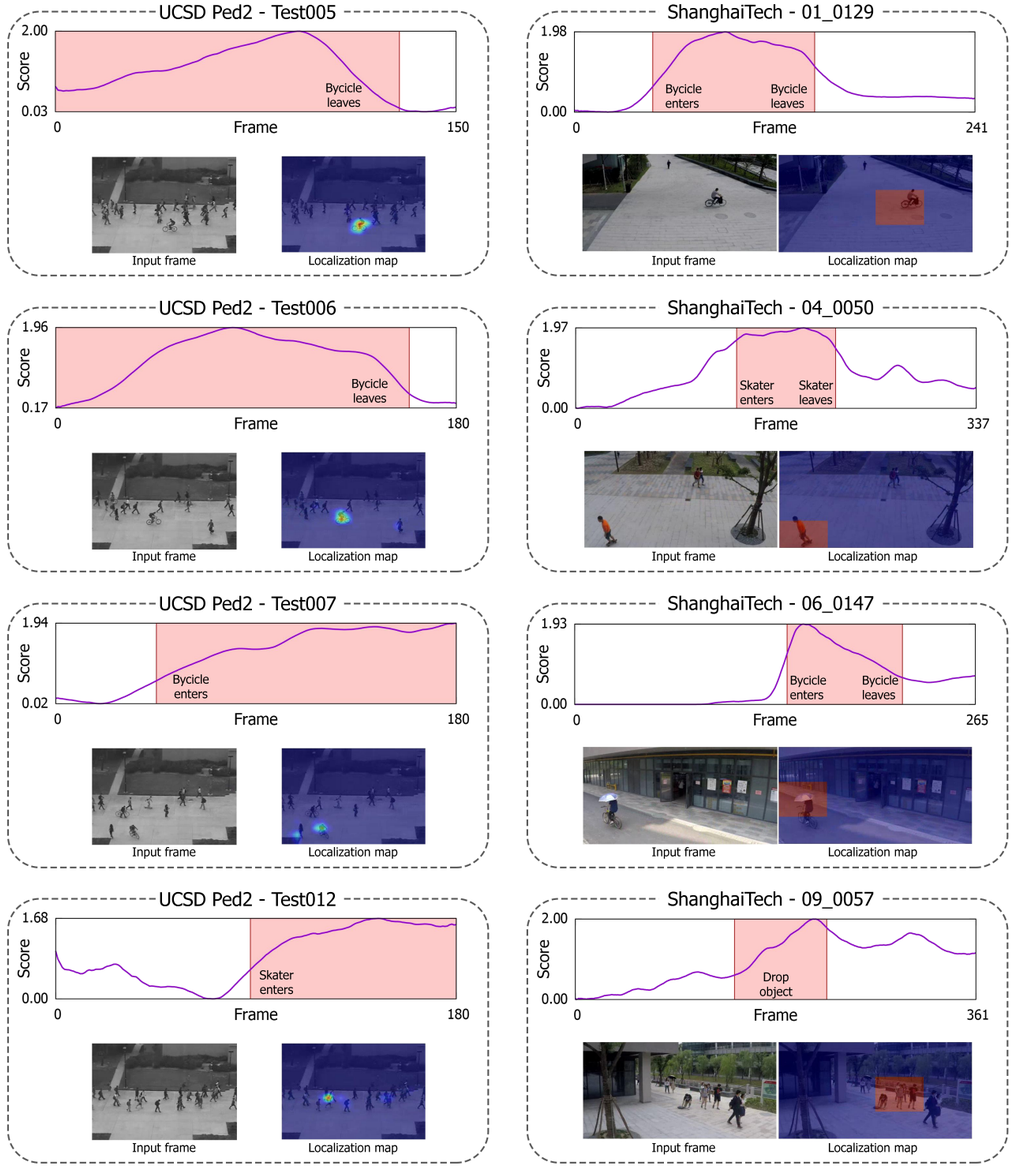


图 11：来自 UCSD Ped2 （左图）和 ShanghaiTech （右图）若干测试片段的新颖性评分及定位图谱。