arXiv:2304.02221v1 [cs.LG] 5 Apr 2023

面向半监督异常检测的异常样本零样本领域自适应

Tomoya Nishida、Takashi Endo、Yohei Kawaguchi

日立制作所研究与开发集团，日本东京国分寺市东恋洼 1-280，
邮编 185-8601

摘要

半监督异常检测（SSAD）是一项在训练中可获得正常数据和少量异常数据的任务。在实际情况下，由于训练阶段难以获得目标域的异常数据，半监督异常检测方法难以适应领域偏移。为解决这一问题，我们提出了一种半监督异常检测的领域自适应方法，该方法无需目标域的异常数据。首先，我们在基于变分自编码器的半监督异常检测模型中引入域对抗网络，以获得领域不变潜变量。由于解码器仅凭领域不变潜变量无法重构原始数据，因此我们以域标签作为解码器的条件输入。为弥补目标域缺失的异常数据，我们引入了一种基于重要性采样的加权损失函数，用于近似理想损失函数。实验结果表明，在目标域没有异常数据可用的情况下，所提方法有助于使半监督异常检测模型适应目标域。

索引词——异常检测、领域自适应、重要性加权

1. 引言

异常检测是工厂自动化 [1, 2] 和自动化监控 [3]等各类业务领域中的一项关键技术。在这些领域中，该技术可用于自动检测工厂设备故障或异常事件的发生（例如尖叫或打斗）。通常，异常检测被视为一项无监督任务，即训练过程中仅提供正常数据。这是因为异常数据通常发生在罕见的情况下，且具有高度多样性，因此难以收集。在这种无监督场景下，已有许多成熟的方法，例如基于支持向量机 [4], 、自编码器 [5],或变分自编码器（VAEs） [6]的方法。

在某些情况下，训练数据 [7] 中可能包含少量异常数据。例如，目标设备可能意外发生故障，或者在数据收集过程中可能发生异常事件。在这种情况下，我们可以利用已有的异常数据来提升异常检测性能。这种情况的一个显著特点是，由于异常数据具有多样性，所收集的异常数据不一定涵盖所有可能的异常数据。例如，工厂设备可能以多种不同的方式发生故障，每种故障都会产生不同类型的异常数据。在这里，我们需要检测两种类型的异常数据：所收集的数据已涵盖的异常数据（已见异常数据）以及未被其涵盖的异常数据（未见异常数据）。因此，尽管异常检测旨在将每条数据分类为正常样本或异常数据，但它无法作为标准的二元分类任务来解决。在本文中，我们将这一任务称为半监督异常检测（SSAD）任务。并专注于此。半监督异常检测旨在比无监督异常检测方法更好地检测已见异常数据，同时保持对未见异常数据的检测性能。文献中已有基于变分自编码器 [8, 9], 、回归模型 [10], 或分类器 [11] 的半监督异常检测方法。 [12, 13], 中也研究了类似的任务，其中假设除了带标签的正常样本和异常数据外，还有大量无标签数据可用。

在实际应用中，用于执行异常检测的数据分布，即目标域数据分布，可能与最初收集的训练数据分布，即源域数据分布，不同。例如，目标设备的测量条件或运行状态存在差异，都可能导致这种分布差异。在这种情况下，我们必须将训练好的模型适配到目标域。如果目标域同时具有正常数据和异常数据，则可以通过典型的域适应方法 [14, 15] 或重新训练模型轻松实现这一过程。然而，由于异常数据很少出现，目标域的异常数据往往无法获取。在这种情况下，由于目标域仅有正常数据可用，半监督异常检测模型的适配将会失败。

本文提出了一种用于半监督异常检测的领域自适应方法：源域中有正常数据和已见异常数据，而目标域中只有正常数据。在这种情况下，我们旨在比无监督异常检测方法更准确地检测目标域中的已见异常数据，同时保持对未见异常数据的检测性能。首先，我们在基于变分自编码器的半监督异常检测模型中加入域对抗网络，以获得领域不变潜变量。由于变分自编码器的解码器无法仅凭领域不变潜变量重建原始数据，因此我们将域标签作为解码器的条件。为弥补目标域异常数据的缺失，我们使用基于重要性采样的加权损失函数 [16] 近似理想损失函数，以处理现有数据。通过这种方法，即使目标域中没有这些数据，训练后的模型也能比无监督异常检测模型更好地检测已见异常数据。

2. 相关工作

文献中已有多项关于异常检测领域自适应的研究。例如， [17], 将归一化流与自适应批归一化相结合，以实现无监督异常检测模型的领域自适应。 [11], 也探讨了半监督异常检测的领域自适应，其中半监督异常检测模型由一个有监督异常检测模型和一个无监督异常检测模型组成。然而，在该研究中，只有无监督异常检测模型适应了目标域。因此，这种自适应无助于检测目标域中的已见异常数据。

If w我们考察异常检测以外的任务，领域... n 适应-

表1：我们的问题设定中的数据可用性。

	正常样本	已见异常	未见异常
源域	Yes	Yes	No
目标域	Yes	No	No

领域自适应已在分类任务中得到广泛研究。通常，域适应方法旨在匹配源域数据与目标域数据的嵌入特征之间的分布 [14, 18]。研究人员还解决了一些复杂任务，例如部分领域自适应 [19, 20]。在这种情况下，目标域数据中将缺失特定子集的类别，这与我们的问题设定类似。然而，在该任务中，缺失类别不会出现在测试数据中，因此无需对其进行正确区分。相比之下，在我们的问题设定中，来自缺失类别的数据，即已见异常数据，将出现在测试数据中，并且需要被正确检测。与我们最接近的任务由 [16], 研究，该研究假设目标训练数据中缺失若干类别，而所有类别都会出现在测试数据中。与该任务相比，我们进一步假设，同时从源域和目标域中缺失的数据，即未见异常数据，也会出现在测试数据中。

3. 问题陈述

在本文中，我们研究半监督异常检测的领域自适应问题。表 1 展示了训练阶段可用的数据类型。对于源域，可获得正常数据和部分异常数据。我们将可用的异常数据称为“已见”异常数据。对于目标域，仅可获得正常数据。因此，整体训练数据集总结如下。

$$\mathcal{D} = \{(\boldsymbol{x}, c, d) \mid \boldsymbol{x} \in \mathbb{R}^s, (c, d) \in \{(\mathrm{N}, \mathrm{S}), (\mathrm{A}, \mathrm{S}), (\mathrm{N}, \mathrm{T})\}\}, \quad (1)$$

其中，$\boldsymbol{x} \in R^s$ 表示输入数据，c 表示输入数据是否属于正常数据（$c = \mathrm{N}$）或已见异常数据（$c = \mathrm{A}$），而 d 表示输入数据所属的领域；$d = \mathrm{S}$表示源域数据，$d = \mathrm{T}$表示目标域数据。我们还假设测试数据中也存在与已见异常数据不同的异常数据。我们将其称为“未见异常数据”。

在本文中，我们旨在构建一种异常检测模型，使其能够比 UAD 方法更准确地检测已见异常数据，同时保持对目标域中未见异常数据的检测性能。需要注意的是，目标域中不存在可用的已见异常数据。因此，我们无法直接训练半监督异常检测模型。此外，采用标准领域自适应技术将源域中训练好的模型迁移到目标域也会失败。

4. 所提方法

4.1. 用于半监督异常检测的双类别变分自编码器

作为基线方法，我们引入了 [8, 9] 中提出的半监督异常检测框架。在这里，使用潜变量 $\boldsymbol{z}$ 的两个先验分布 $p(\boldsymbol{z} \mid c = \mathrm{N})$ 和 $p(\boldsymbol{z} \mid c = \mathrm{A})$ 训练一个变分自编码器。每个先验分布分别对应正常数据和异常数据的分布。通过将两个先验分布设置得相距较远，我们可以通过观察潜变量，轻松地区分已见异常数据与正常数据。我们将此模型称为双类别变分自编码器（2C-VAE）。2C-VAE 的损失函数定义为

$$l_{\mathrm{VAE}}(\boldsymbol{x}, c; \boldsymbol{\theta}_{\mathrm{Enc}}, \boldsymbol{\theta}_{\mathrm{Dec}}) = \mathbb{E}_{q_{\boldsymbol{\theta}_{\mathrm{Enc}}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}_{\mathrm{Dec}}}(\boldsymbol{x} \mid \boldsymbol{z})] - \mathrm{KL}[q_{\boldsymbol{\theta}_{\mathrm{Enc}}}(\boldsymbol{z} \mid \boldsymbol{x}) \| p(\boldsymbol{z} \mid c)], \quad (2)$$

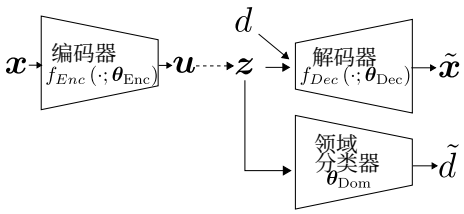

图 1：所提方法的模型架构

这就是 $\log\ p(\boldsymbol{x})$ 的证据下界（ELBO）。其中，$q_{\boldsymbol{\theta}_{\mathrm{Enc}}}(\boldsymbol{z}|\boldsymbol{x})$ 和 $p_{\boldsymbol{\theta}_{\mathrm{Dec}}}(\boldsymbol{x} \mid \boldsymbol{z})$ 由神经网络构建。$q_{\boldsymbol{\theta}_{\mathrm{Enc}}}(\boldsymbol{z}|\boldsymbol{x})$ 由编码器 f_{Enc} 构建。该编码器输出概率密度函数 $p_{\mathrm{Enc}}(\cdot; \boldsymbol{u})$ 的参数 $\boldsymbol{u}$，该函数将用作 $q_{\boldsymbol{\theta}_{\mathrm{Enc}}}(\boldsymbol{z}|\boldsymbol{x})$，作为

$$\boldsymbol{u} = f_{\mathrm{Enc}}(\boldsymbol{x}; \boldsymbol{\theta}_{\mathrm{Enc}}), \quad (3)$$

$$q_{\boldsymbol{\theta}_{\mathrm{Enc}}}(\boldsymbol{z} \mid \boldsymbol{x}) = p_{\mathrm{Enc}}(\boldsymbol{z}; \boldsymbol{u}). \quad (4)$$

对于 p_{Enc}，采用高斯分布 [8] 或 von Mises-Fisher（vMF）分布 [9]。$p_{\boldsymbol{\theta}_{\mathrm{Dec}}}(\boldsymbol{x} \mid \boldsymbol{z})$ 由解码器f_{Dec}构建。该解码器通常输出 x的期望值，并在其周围使用高斯分布来构成$p_{\boldsymbol{\theta}_{\mathrm{Dec}}}(\boldsymbol{x} \mid \boldsymbol{z})$，如

$$\tilde{\boldsymbol{x}} = f_{\mathrm{Dec}}(\boldsymbol{z}; \boldsymbol{\theta}_{\mathrm{Dec}}), \quad (5)$$

$$p_{\boldsymbol{\theta}_{\mathrm{Dec}}}(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \tilde{\boldsymbol{x}}, \boldsymbol{I}). \quad (6)$$

因此，$\log p_{\boldsymbol{\theta}_{\mathrm{Dec}}}(\boldsymbol{x} \mid \boldsymbol{z})$ 可以计算为 $\tilde{\boldsymbol{x}}$ 与 $\boldsymbol{x}$ 之间的均方误差。

4.2. 2C-VAE 的域对抗网络

为了使 2C-VAE 适应目标域，我们引入一个域分类器 $p_{\boldsymbol{\theta}_{\mathrm{Dom}}}(d \mid \boldsymbol{z})$ 来构建一个对抗网络 [14]。域分类损失表示为

$$l_{\mathrm{dom}}(\boldsymbol{x}, d; \boldsymbol{\theta}_{\mathrm{Enc}}, \boldsymbol{\theta}_{\mathrm{Dom}}) = \mathbb{E}_{q_{\boldsymbol{\theta}_{\mathrm{Enc}}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}_{\mathrm{Dom}}}(d \mid \boldsymbol{z})]. \quad (7)$$

通过对（7）进行对抗训练，我们旨在匹配 $\boldsymbol{z}$ 源域数据与目标域数据之间的分布，即获得一个域不变潜变量。然而，这一目标与（2）的优化相矛盾，因为如果没有任何领域信息，解码器就无法重构原始数据。为了解决这一问题，除了潜变量之外，我们还将域标签输入解码器。通过使用域标签，即使 $\boldsymbol{z}$ 变得具有域不变性，解码器仍然能够估计原始数据。

通过将（7）与（2）相结合，用于训练 2C-VAE 的优化问题可写为

$$\min_{\boldsymbol{\theta}_{\mathrm{Enc}}, \boldsymbol{\theta}_{\mathrm{Dec}}} \max_{\boldsymbol{\theta}_{\mathrm{Dom}}} \mathbb{E}[\mathcal{L}(\boldsymbol{x}, c, d; \boldsymbol{\theta}_{\mathrm{Enc}}, \boldsymbol{\theta}_{\mathrm{Dec}}, \boldsymbol{\theta}_{\mathrm{Dom}})], \quad (8)$$

$$\mathcal{L}(\boldsymbol{x}, c, d; \boldsymbol{\theta}_{\mathrm{Enc}}, \boldsymbol{\theta}_{\mathrm{Dec}}, \boldsymbol{\theta}_{\mathrm{Dom}}) := -l_{\mathrm{VAE}}(\boldsymbol{x}, c, d; \boldsymbol{\theta}_{\mathrm{Enc}}, \boldsymbol{\theta}_{\mathrm{Dec}}) + \lambda_{dom} l_{\mathrm{dom}}(\boldsymbol{x}, d; \boldsymbol{\theta}_{\mathrm{Enc}}, \boldsymbol{\theta}_{\mathrm{Dom}}), \quad (9)$$

其中，$l_{dom} > 0$ 是一个超参数。注意，由于域标签 d 现在作为解码器的输入，因此我们将其添加到了 l_{VAE} 中。这里，$\boldsymbol{\theta}_{\mathrm{Enc}}$ 和$\boldsymbol{\theta}_{\mathrm{Dec}}$ 通过训练来最小化 $\mathcal{L}$，而 $\boldsymbol{\theta}_{\mathrm{Dom}}$ 则通过训练来最大化 $\mathcal{L}$。因此，$\boldsymbol{\theta}_{\mathrm{Dom}}$ 经过训练后能够根据 $\boldsymbol{z}$正确分类领域，而 $\boldsymbol{\theta}_{\mathrm{Enc}}$ 则经过训练以欺骗域分类器。通过这种方式，模型将获得一个领域不变的潜变量。我们在图1中总结了该模型架构。从这里开始，当损失函数中的参数显而易见时，我们将省略其书写。

使用（9）训练模型时，由于目标域中缺乏已观测到的异常数据，因此会产生一个问题。在4.3节和4.4节中，我们将描述这一问题，并介绍一种基于重要性采样的方法 [16] 来解决该问题。

4.3. 2C-VAE损失的近似计算

l_{VAE} 的期望值可以分解为

$$\begin{aligned}
&\mathbb{E}\left[l_{\mathrm{VAE}}\left(\boldsymbol{x}, c, d; \boldsymbol{\theta}_{\mathrm{Enc}}, \boldsymbol{\theta}_{\mathrm{Dec}}\right)\right] \\
&= \int l_{\mathrm{VAE}}\left(\boldsymbol{x}, c=\mathrm{N}, d=\mathrm{S}\right) p\left(\boldsymbol{x}, c=\mathrm{N}, d=\mathrm{S}\right) d\boldsymbol{x} \\
&+ \int l_{\mathrm{VAE}}\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{S}\right) p\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{S}\right) d\boldsymbol{x} \\
&+ \int l_{\mathrm{VAE}}\left(\boldsymbol{x}, c=\mathrm{N}, d=\mathrm{T}\right) p\left(\boldsymbol{x}, c=\mathrm{N}, d=\mathrm{T}\right) d\boldsymbol{x} \\
&+ \int l_{\mathrm{VAE}}\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{T}\right) p\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{T}\right) d\boldsymbol{x}. \qquad (10)
\end{aligned}$$

由于我们没有来自目标域的异常样本，因此无法获得从 $p(\boldsymbol{x}, c=\mathrm{A}\ d=\mathrm{T})$ 中采样的数据。因此，无法利用现有数据计算第四项。这意味着，如果我们使用现有数据优化 $l_{2CVAE}(\boldsymbol{x}, c, d)$，那么来自目标域的已见异常数据的潜变量将不会受到先验分布 $p(\boldsymbol{z}|c=\mathrm{A})$ 的约束。因此，将无法检测来自目标域的已见异常数据。

为解决这一问题，我们将 $p(\boldsymbol{x}, c=\mathrm{A}\ d=\mathrm{T})$ 改写为

$$\begin{aligned}
p\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{T}\right) &= p\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{S}\right) \frac{p\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{T}\right)}{p\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{S}\right)} \\
&= p\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{S}\right) \frac{p\left(d=\mathrm{T} \mid \boldsymbol{x}\right)}{p\left(d=\mathrm{S} \mid \boldsymbol{x}\right)}. \qquad (11)
\end{aligned}$$

这里，我们假设 $p\left(d \mid \boldsymbol{x}, c=\mathrm{A}\right) = p\left(d \mid \boldsymbol{x}, c=\mathrm{N}\right) = p(d \mid \boldsymbol{x})$。这可以解释为类别偏移下的协变量偏移条件，而非域偏移下的标准协变量偏移条件。当数据领域只能根据数据本身估计、而无法判断其属于正常样本还是异常样本时，这是一个合理的假设。

现在，我们可以利用源域异常数据，并将其权重表示为式（11）的第二项，以计算式（10）的第四项。由于我们在训练 2C-VAE 的同时也在训练域分类器，因此式（11）的第二项可以估计为

$$w_{\mathrm{vae}} := \frac{p\left(d=\mathrm{T} \mid \boldsymbol{x}\right)}{p\left(d=\mathrm{S} \mid \boldsymbol{x}\right)} = \frac{\int q_{\boldsymbol{\theta}_{\mathrm{Enc}}}\left(\boldsymbol{z} \mid \boldsymbol{x}\right) p_{\boldsymbol{\theta}_{\mathrm{Dom}}}\left(d=\mathrm{T} \mid \boldsymbol{z}\right) d\boldsymbol{z}}{\int q_{\boldsymbol{\theta}_{\mathrm{Enc}}}\left(\boldsymbol{z} \mid \boldsymbol{x}\right) p_{\boldsymbol{\theta}_{\mathrm{Dom}}}\left(d=\mathrm{S} \mid \boldsymbol{z}\right) d\boldsymbol{z}}. \qquad (12)$$

这里，我们假设 $p(d \mid \boldsymbol{z}, \boldsymbol{x}) = p(d \mid \boldsymbol{z})$，这是一个合理的假设，因为在实际情况下，训练好的域分类器通常只能根据 $\boldsymbol{z}$ 准确估计域标签。除了上述重构之外，我们还设置 $p\ d= = 0.5\ \ p\ c \mid d= = p\ c \mid d= = p_c\ \ (d=\mathrm{S})$ (T) and (T) (S) :，其中 $p(c \mid d=\mathrm{S})$ 根据源域数据计算得出。最后，(10) 可以重构为

$$\begin{aligned}
&\mathbb{E}\left[l_{\mathrm{VAE}}\left(\boldsymbol{x}, c; \boldsymbol{\theta}_{\mathrm{Enc}}, \boldsymbol{\theta}_{\mathrm{Dec}}\right)\right] \\
&= \frac{p_{\mathrm{N}}}{2} \int l_{\mathrm{VAE}}\left(\boldsymbol{x}, c=\mathrm{N}, d=\mathrm{S}\right) p\left(\boldsymbol{x} \mid c=\mathrm{N}, d=\mathrm{S}\right) d\boldsymbol{x} \\
&+ \frac{p_{\mathrm{A}}}{2} \int \left(l_{\mathrm{VAE}}\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{S}\right)\right. \\
&\qquad \left. + w_{\mathrm{vae}}\, l_{\mathrm{VAE}}\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{T}\right)\right) p\left(\boldsymbol{x} \mid c=\mathrm{A}, d=\mathrm{S}\right) d\boldsymbol{x} \\
&+ \frac{p_{\mathrm{N}}}{2} \int l_{\mathrm{VAE}}\left(\boldsymbol{x}, c=\mathrm{N}, d=\mathrm{T}\right) p\left(\boldsymbol{x} \mid c=\mathrm{N}, d=\mathrm{T}\right) d\boldsymbol{x}. \qquad (13)
\end{aligned}$$

此处，所有项都可以通过现有数据计算。请注意，在第二项中，数据 x 是从源域异常数据中采样的，但在 $l_{\mathrm{VAE}}(\boldsymbol{x}, c=\mathrm{A}\ d=\mathrm{T})$ 中，域标签为 $d=\mathrm{T}$。

表 2：实验条件

	正常样本	已见异常	未见异常
情况 1	1、2、3	4、5、6	7、8、9
情况 2	4, 5, 6	7, 8, 9	1, 2, 3
情况 3	7、8、9	1, 2, 3	4, 5, 6

因此，我们将 $d=\mathrm{T}$ 输入解码器以计算这一项，尽管数据 $\boldsymbol{x}$ 将从源域中采样。结果是，如果数据被认为与目标域数据相似，(13) 会为来自缺失类别的数据赋予较大的权重。这与部分领域自适应问题不同，在后者中，重要性采样被用于为来自缺失类别的数据赋予较小的权重 [18]。

4.4. 域分类损失的近似

若 w_{vae} 准确，则 (13) 将是对 (10) 的准确估计。然而，由于无法获取数据，用于计算 w_{vae} 各组成部分的域分类器也无法得到准确训练。为解决这一问题，我们还将重新表述域分类损失。首先，我们按照 (11) 中相同的步骤，将 $p(\boldsymbol{x}, c=\mathrm{A}\ d=\mathrm{T})$ 重写为

$$\begin{aligned}
p\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{T}\right) &= p\left(\boldsymbol{x}, c=\mathrm{N}, d=\mathrm{T}\right) \frac{p\left(\boldsymbol{x}, c=\mathrm{A}, d=\mathrm{T}\right)}{p\left(\boldsymbol{x}, c=\mathrm{N}, d=\mathrm{T}\right)} \\
&= p\left(\boldsymbol{x}, c=\mathrm{N}, d=\mathrm{T}\right) \frac{p\left(c=\mathrm{A} \mid \boldsymbol{x}\right)}{p\left(c=\mathrm{N} \mid \boldsymbol{x}\right)} \\
&:= p\left(\boldsymbol{x}, c=\mathrm{N}, d=\mathrm{T}\right) w_{\mathrm{dom}}. \qquad (14)
\end{aligned}$$

这里，我们假设协变量偏移条件 $p\ \ = p\ c \mid \boldsymbol{x}\ d= = p\ c \mid \boldsymbol{x}$ $(c \mid \boldsymbol{x}, d=\mathrm{T})$ (, S) ( )。

与 [16]中不同，由于我们没有显式地将给定样本分类为正常样本或异常样本，因此无法直接根据模型输出计算 w_{dom}。然而，假设 $p(\boldsymbol{x} \mid c, \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{z})$，则可将 w_{dom} 的分子和分母计算如下：

$$p\left(c=\mathrm{A} \mid \boldsymbol{x}\right) = \int q_{\boldsymbol{\theta}_{\mathrm{Enc}}}\left(\boldsymbol{z} \mid \boldsymbol{x}\right) \frac{p_{\mathrm{A}} p\left(\boldsymbol{z} \mid c=\mathrm{A}\right)}{p\left(\boldsymbol{z}\right)} d\boldsymbol{z}, \qquad (15)$$

$$p\left(c=\mathrm{N} \mid \boldsymbol{x}\right) = \int q_{\boldsymbol{\theta}_{\mathrm{Enc}}}\left(\boldsymbol{z} \mid \boldsymbol{x}\right) \frac{p_{\mathrm{N}} p\left(\boldsymbol{z} \mid c=\mathrm{N}\right)}{p\left(\boldsymbol{z}\right)} d\boldsymbol{z}, \qquad (16)$$

$$p\left(\boldsymbol{z}\right) = p_{\mathrm{N}}\, p\left(\boldsymbol{z} \mid c=\mathrm{N}\right) + p_{\mathrm{A}}\, p\left(\boldsymbol{z} \mid c=\mathrm{A}\right). \qquad (17)$$

最后，域分类损失可以重新表述为

$$\begin{aligned}
&\mathbb{E}\left[l_{\mathrm{dom}}\left(\boldsymbol{x}, d; \boldsymbol{\theta}_{\mathrm{Enc}}, \boldsymbol{\theta}_{\mathrm{Dom}}\right)\right] \\
&= \frac{p_{\mathrm{N}}}{2} \int l_{\mathrm{dom}}\left(\boldsymbol{x}, d=\mathrm{S}\right) p\left(\boldsymbol{x} \mid c=\mathrm{N}, d=\mathrm{S}\right) d\boldsymbol{x} \\
&+ \frac{p_{\mathrm{A}}}{2} \int l_{\mathrm{dom}}\left(\boldsymbol{x}, d=\mathrm{S}\right) p\left(\boldsymbol{x} \mid c=\mathrm{A}, d=\mathrm{S}\right) d\boldsymbol{x} \\
&+ \frac{p_{\mathrm{N}}}{2} \int \left(1 + w_{\mathrm{dom}}\right) l_{\mathrm{dom}}\left(\boldsymbol{x}, d=\mathrm{T}\right) p\left(\boldsymbol{x} \mid c=\mathrm{N}, d=\mathrm{T}\right) d\boldsymbol{x}. \qquad (18)
\end{aligned}$$

现在，将 (13) 和 (18) 代入 (8)，即可得到领域自适应的整体优化问题。

5. 实验

5.1. 实验条件

为了研究所提方法的有效性，我们在域偏移条件下进行了半监督异常检测（SSAD）实验。我们使用了两个手写数字图像数据集：MNIST [21] 作为源域。

表 3：已见异常数据的曲线下面积（%）

方法	带权命题			不带权命题			2C-vMF-VAE-da			2C-vMF-VAE			VAE
用于异常评分的值	RL	KL	ELBO	RL	KL	ELBO	RL	KL	ELBO	RL	KL	ELBO	RL
情况 1	88.2	92.9	**97.4**	92.5	90.1	96.8	80.7	92.2	94.6	79.5	90.8	95.3	95.9
情况 2	79.3	75.6	**88.6**	84.9	71.7	86.9	73.3	74.9	87.0	77.8	76.5	87.3	85.5
情况 3	85.2	77.7	**92.7**	86.8	70.5	91.8	79.7	67.7	88.4	79.8	69.0	90.5	82.3
Avg.	84.2	82.1	**92.9**	88.1	77.4	91.8	77.9	78.3	90.0	79.0	78.8	91.0	87.9

表 4：未见异常数据的曲线下面积值（%）

方法	带权命题			不带权命题			2C-vMF-VAE-da			2C-vMF-VAE			VAE
用于异常评分的值	RL	KL	ELBO	RL	KL	ELBO	RL	KL	ELBO	RL	KL	ELBO	RL
情况 1	83.7	72.7	89.7	88.5	70.4	90.2	77.4	79.7	86.0	77.5	77.8	86.5	**90.6**
情况 2	80.0	78.2	89.1	87.8	76.3	**90.1**	76.6	84.1	87.6	79.8	85.9	89.0	88.8
情况 3	86.4	60.4	86.8	86.8	60.9	87.1	81.9	58.5	83.9	84.8	60.6	85.5	**90.0**
Avg.	83.4	70.4	88.5	87.7	69.2	89.1	78.6	74.1	85.8	80.7	74.8	87.0	**89.8**

领域和旋转MNIST [22] 作为目标域。旋转MNIST是将MNIST数据集旋转45度得到的。与 [9], 类似，我们将数字分为三类来创建一个SSAD任务：正常样本、已见异常和未见异常。我们研究了表2总结的三种情况。在训练中，我们假设源域中有正常数据和已见异常数据，而目标域中只有正常数据可用。此处，所提方法旨在比无监督异常检测方法更准确地检测目标域中的已见异常数据，同时保持对未见异常数据的检测性能。

对于所提方法的原始模型（带权命题），我们使用了双类别 vMF-VAE [9]。如第 4 节所述，我们在该原始模型中添加了一个域分类器，并将域标签作为解码器的额外输入。为了评估重要性加权的有效性，我们还进行了重要性权重设为 0 的所提方法实验，即 $w_{vae} = w_{dom} = 0$ （不带权命题）。对于传统方法，我们评估了基于 VAE 的 UAD 方法（VAE）和基于原始 2C-vMF-VAE 的 SSAD 方法（2C-vMF-VAE）。我们使用目标域的正常数据训练 VAE，并使用所有可用数据训练 2C-vMF-VAE，即源域的正常数据和已见异常数据，以及目标域的正常数据。我们还评估了一种将标准领域自适应技术 DANN [14] 集成到原始 SSAD 方法中的方法（2C-vMF-VAE-da）。具体而言，我们仅在 2C-vMF-VAE 模型中引入了一个域分类器，并进行了域对抗学习。对于 VAE，我们使用重构损失（RL）作为异常分数。对于其他方法，我们使用 RL、 $p(\boldsymbol{z} \mid c = \mathrm{N})$ 的 KL 散度损失（KL）以及证据下界（ELBO）作为异常分数。对于每个模型的编码器和解码器，我们均使用了一个包含 50 个单元的单层隐藏线性层。对于域分类器，我们同样使用了一个包含 50 个单元的单层隐藏线性层。

我们使用 Adam [23] 训练了所有模型。变分自编码器（VAE）训练了 200 个训练轮次，学习率设为 0.01。所有其他模型均训练了 500 个训练轮次，学习率设为 0.0001。对于需要域分类器的模型，我们设置了 $\lambda_{dom} = 0.01$。对于所提方法，我们对 w_{dom} 和 w_{vae} 的值进行了正则化，使其满足 $\tilde{w} = 1/2 \min\{w, 2\}$，其中 $w \in \{w_{\mathrm{dom}}, w_{\mathrm{vae}}\}$。这是必要的，因为在训练初期，这些值可能不稳定。

5.2. 结果

我们比较了针对目标域数据进行异常检测的受试者工作特征曲线下面积（曲线下面积）分数。这里，分数分别针对已见异常数据和未见异常数据计算。

表 3 展示了在目标域中检测已见异常数据的曲线下面积（AUC）值。对于除变分自编码器（VAE）以外的所有方法，使用证据下界（ELBO）作为异常分数取得的曲线下面积高于 RL 或 KL。可能的原因是，目标领域中的一些已见异常样本与源域中的样本相对相似，而另一些则并非如此。在这种情况下，与源域样本相似的样本可以通过 KL 检测，而其他样本则可以通过 RL 检测。因此，将这两种分数结合为证据下界后，便能检测出这两类异常样本，并取得最佳性能。因此，我们将主要关注证据下界取得的曲线下面积值。首先，2C-vMF-VAE-da 的曲线下面积值低于 2C-vMF-VAE。这一结果表明，当目标领域中没有已见异常数据时，单纯应用领域自适应将彻底失效。相比之下，所提方法（带权命题）的表现优于所有其他方法，包括基于变分自编码器的无监督异常检测方法。这一结果表明，在这种情况下，所提方法成功地将模型自适应到了目标领域。此外，所提方法的曲线下面积值高于不带权方法（不带权命题），这也证明了重要性权重的有效性。

表 4 显示了在目标域中检测未见异常数据的曲线下面积值。在两种情况下，变分自编码器的曲线下面积值最高。不过，变分自编码器与所提方法的曲线下面积值差异相对较小。此外，在一种情况下，所提方法甚至优于变分自编码器。因此，所提方法基本保持了对未见异常数据的检测性能。

总体而言，与无监督方法相比，所提方法在目标域中无法获得已见异常数据时，提高了对已见异常数据的检测性能，同时保持了对未见异常数据的检测性能。因此，当目标域缺少异常数据时，所提方法实现了半监督异常检测模型的领域自适应。

6. 结论

我们提出了一种用于半监督异常检测（SSAD）的领域自适应方法，该方法适用于目标域中没有可用异常数据的情况。首先，我们对基于变分自编码器（VAE）的半监督异常检测模型进行领域对抗训练，以获得领域不变潜变量。由于解码器无法仅凭领域不变潜变量重建原始数据，因此我们使用域标签对解码器进行条件约束。为弥补数据不可用的问题，我们引入了一种基于重要性采样的加权损失函数，以近似理想损失函数。实验结果表明，与无监督异常检测（UAD）相比，所提方法能够更好地检测已见异常数据，同时保持对未见异常数据的检测性能。这一结果表明，当目标域中没有可用异常数据时，所提方法有助于使半监督异常检测模型适应目标域。

7. 参考文献

[1] Y. Koizumi、S. Saito、H. Uematsu、Y. Kawachi 和 N. Harada，《基于深度学习和内曼-皮尔逊引理的无监督异常声音检测》，IEEE 音频、语音与语言处理汇刊，第27卷，第1期，第212–224页，2019年。[2] K. Suefusa、T. Nishida、H. Purohit、R. Tanabe、T. Endo 和 Y. Kawaguchi，《基于插值深度神经网络的异常声音检测》，载于 IEEE 国际声学、语音与信号处理会议论文集（ICASSP），2020年5月，第271–275页。[3] T. Xiao、C. Zhang 和 H. Zha，《学习检测监控视频中的异常》，IEEE 信号处理快报，第22卷，第9期，第1477–1481页，2015年。[4] B. Sch¨olkopf、J. C. Platt、J. Shawe-Taylor、A. J. Smola 和 R. C. Williamson，《高维分布支持集的估计》，Neural Computation，第13卷，第7期，第1443–1471页，2001年。[5] C. Zhou 和 R. C. Paffenroth，《利用鲁棒深度自编码器进行异常检测》，载于国际知识发现与数据挖掘会议（KDD）论文集，2017年8月，第665–674页。[6] J. An 和 S. Cho，《基于变分自编码器并利用重构概率进行异常检测》，Special Lecture on IE，第2卷，第1期，第1–18页，2015年。[7] N. G¨ornitz、M. Kloft、K. Rieck 和 U. Brefeld，《迈向有监督异常检测》，人工智能研究杂志，第46卷，第235–262页，2013年。[8] Y. Kawachi、Y. Koizumi 和 N. Harada，《用于有监督异常检测的互补集变分自编码器》，载于 IEEE 国际声学、语音与信号处理会议论文集（ICASSP），2018年4月，第2366–2370页。[9] Y. Kawachi、Y. Koizumi、S. Murata 和 N. Harada，《用于有监督异常检测的双类别超球面自编码器》，载于 IEEE 国际声学、语音与信号处理会议论文集（ICASSP），2019年5月，第3047–3051页。[10] W. Liu、W. Luo、Z. Li、P. Zhao 和 S. Gao，《通过少量异常样本的间隔学习嵌入式预测进行视频异常检测》，载于国际人工智能联合会议（IJCAI）论文集，2019年8月，第3023–3030页。[11] C. Zhang、G. Li、L. Su、W. Zhang 和 Q. Huang，《利用开放数据过滤器和领域自适应进行视频异常检测》，载于 IEEE 国际视觉通信与图像处理会议（VCIP）论文集，2020年12月，第395–398页。[12] L. Ruff、R. A. Vandermeulen、N. G¨ornitz、A. Binder、E. M¨uller、K. M¨uller 和 M. Kloft，《深度半监督异常检测》，载于国际学习表征会议（ICLR）论文集，2020年4月，第1–23页。[13] Y. Zhou、X. Song、Y. Zhang、F. Liu、C. Zhu 和 L. Liu，《利用自编码器进行特征编码以实现弱监督异常检测》，IEEE 神经网络与学习系统汇刊，第33卷，第6期，第2454–2465页，2022年。[14] Y. Ganin、E. Ustinova、H. Ajakan、P. Germain、H. Larochelle、F. Laviolette、M. Marchand 和 V. Lempitsky，《神经网络的领域对抗训练》，机器学习研究杂志，第17卷，第1期，第2096–2030页，2016年。[15] Z. Yang、I. S. Bozchalooi 和 E. Darve，《利用领域自适应进行异常检测》，arXiv 预印本 arXiv:2006 03689，2020年。。

[16] M. Ishii、T. Takenouchi 和 M. Sugiyama，“针对缺失类别的不完整目标域数据的部分零样本领域自适应”，载于 IEEE/CVF 冬季计算机视觉应用会议论文集，2020年，第3052–3060 页。[17] M. Yamaguchi、Y. Koizumi 和 N. Harada，“AdaFlow：应用于异常检测和非配对跨域转换的领域自适应密度估计器”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，2019年5月，第 3647–3651 页。[18] M. Long、Y. Cao、J. Wang 和 M. Jordan，“利用深度自适应网络学习可迁移特征”，载于国际机器学习会议（ICML）论文集。PMLR，2015年7月，第 97–105 页。[19] J. Zhang、Z. Ding、W. Li 和 P. Ogunbona，“用于部分域自适应的重要性加权对抗网络”，载于 IEEE 计算机视觉与模式识别会议（CVPR）论文集，2018年6月，第 8156–8164 页。[20] Z. Cao、L. Ma、M. Long 和 J. Wang，“部分对抗领域自适应”，载于欧洲计算机视觉会议（ECCV）论文集，2018年9月，第 135–150 页。[21] Li L. Deng，“用于机器学习研究的手写数字图像 MNIST 数据库”，IEEE 信号处理杂志，第 29 卷，第 6 期，第 141–142 页，2012 年。[22] M. Ghifary 、 W. B. Kleijn 、 M. Zhang 和 D. Balduzzi，“利用多任务自编码器实现目标识别的领域泛化”，载于 IEEE 国际计算机视觉会议（ICCV）论文集，2015年12月。[23] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，载于国际学习表征会议（ICLR）论文集，2015年4月。