毕业设计（论文）外文文献翻译

学　　院：	计算机科学与技术学院
专业班级：	软件工程
姓　　名：	李俊杰
学　　号：	2025121015
指导教师：（签 名）	辛明海

翻译文献（见附件）来源：

Xuehan Bai, Yan Li, Yanhua Cheng, Wenjie Yang, Quan Chen, Han Li: Cross-Domain Product Representation Learning for Rich-Content E-Commerce. CoRR abs/2308.05550 (2023)

面向丰富内容电子商务的跨领域产品表示学习

Xuehan Bai Yan Li Yanhua Cheng Wenjie Yang Quan Chen Han Li

Kuaishou Technology

摘 要 短视频和直播平台的激增彻底改变了消费者的网购方式。消费者现在不再浏览产品页面，而是转向内容丰富的电子商务，他们可以通过短视频和直播等动态互动媒体购买产品。这种新兴的在线购物形式带来了技术挑战，因为产品可能在不同的媒体领域以不同的方式呈现。因此，统一的产品表示对于实现跨领域产品识别至关重要，以确保最佳的用户搜索体验和有效的产品推荐。尽管行业迫切需要统一的跨领域产品表示，但之前的研究主要集中在产品页面上，而没有考虑短视频和直播。为了填补内容丰富的电子商务领域的空白，本文引入了一个大规模的跨领域产品识别数据集，称为 ROPE。ROPE 涵盖了广泛的产品类别，包含超过 18 万种产品，对应数百万短视频和直播。这是第一个同时涵盖产品页面、短视频和直播流的数据集，为跨不同媒体领域建立统一的产品表示提供了基础。此外，我们提出了一个跨领域产品表示框架，即 COPE，该框架通过文本和视觉等多模态学习来统一不同领域的产品表示。下游任务的大量实验证明了 COPE 在学习所有产品域的联合特征空间方面的有效性。

中图法分类号 arXiv:2308.05550v1 [cs.CV] DOI 号 https://doi.org/10.48550/arXiv.2308.05550

Cross-Domain Product Representation Learning for Rich-Content E-Commerce

Xuehan Bai Yan Li Yanhua Cheng Wenjie Yang Quan Chen Han Li

Kuaishou Technology

Abstract The proliferation of short video and live-streaming platforms has revolutionized how consumers engage in online shopping. Instead of browsing product pages, consumers are now turning to rich-content e-commerce, where they can purchase products through dynamic and interactive media like short videos and live streams. This emerging form of online shopping has introduced technical challenges, as products may be presented differently across various media domains. Therefore, a unified product representation is essential for achieving cross-domain product recognition to ensure an optimal user search experience and effective product recommendations. Despite the urgent industrial need for a unified cross-domain product representation, previous studies have predominantly focused only on product pages without taking into account short videos and live streams. To fill the gap in the rich-content e-commerce area, in this paper, we introduce a large-scale cRoss-dOmain Product rEcognition dataset, called ROPE. ROPE covers a wide range of product categories and contains over 180,000 products, corresponding to millions of short videos and live streams. It is the first dataset to cover product pages, short videos, and live streams simultaneously, providing the basis for establishing a unified product representation across different media domains. Furthermore, we propose a CrossdOmain Product rEpresentation framework, namely COPE, which unifies product representations in different domains through multimodal learning including text and vision. Extensive experiments on downstream tasks demonstrate the effectiveness of COPE in learning a joint feature space for all product domains.

1 引言

近年来，随着消费者在娱乐媒体上花费时间的变迁，消费者们线上购物的方式已经发生了显著的变化，并且丰富内容电子商务正变得越来越受欢迎。在富内容电商领域，产品不仅以传统的产品页面销售，还以动态和互动的媒体形式销售，如短视频和直播。因此，消费者越来越依赖这些形式来做出明智的购买决定。这种转变促进了更有吸引力的购物体验，弥合了消费者和卖家之间的差距，同时为平台提供了新的机会。

尽管内容丰富的电子商务具有优势，但它也提出了一些技术挑战。最重要的挑战之一是产品在不同媒体领域的表现不一致。例如，相比于传统的商品页面，一个商品在直播中可能会以完全不同的形式出现。在工业场景中，跨不同领域建立统一的产品表示是解决不一致问题的迫切需要。如图 1 所示，当用户搜索特定产品时，统一的产品表示确保了愉快的搜索体验，即返回的产品页面、短视频和实时流准确地描述了同一产品。平台在为用户推荐产品时，统一的产品表示有利于挖掘用户在不同媒体上的消费行为，进行全面的产品推荐。

尽管行业迫切需要统一的跨领域产品表示，但之前的努力只集中在产品页面领域。学习产品表示最常用的方法是用产品图像和标题训练一个产品分类模型[12,14,26,27]。然而，在内容丰富的电子商务中，这种表现方式是远远不能被接受的。具体来说，产品页面上展示的图片一般都是专业人士拍摄的，而在短视频和直播中，产品的姿势和在场景中的位置往往会发生很大的变化。此外，在直播和短视频中，并不能保证产品每时每刻都是可见的。短视频可能与故事情节混在一起，而直播可能包含卖家和观众之间的聊天。这些内容通常与产品无关。为了弥补这一差距并推动相关研究，我们从网上购物平台收集了大量的真实数据，并提出了第一个大规模的跨领域产品识别数据集 ROPE。我们的数据集包含 3,056,624 个产品页面，5,867,526 个短视频，以及 189,958 个不同产品的 3,495,097 个直播流。它涵盖了网上购物场景的所有产品类别。据我们所知，ROPE 是一个内容丰富的电子商务数据集，包括产品页面、短视频和直播。我们希望 ROPE 的出版能够吸引更多的研究人员进入内容商务领域，推动相关技术的发展。

除了 ROPE 数据集之外，我们还提出了一个跨域产品表示基线，COPE，它将产品页面、短视频和实时流映射到相同的特征空间，以构建统一的产品表示。

基于 ROPE 数据集，我们对 COPE 模型在跨域检索和少镜头分类任务上进行了评估。实验结果表明，与现有的技术水平相比，该方法有了显著的改进。

图 1 说明跨领域产品表示对富内容电子商务的重要性。这种新型电商有两个坚实的需求;1)平台需要提供与用户查询相对应的产品页面、短视频、直播等精准的产品结果;2)平台能够根据用户的行为历史向用户推荐感兴趣的同类产品。这两个任务都高度依赖于高性能的跨领域产品表示。例子来自流行的富内容电商平台，包括抖音、葵和淘宝。

综上所述，我们的贡献如下:

1)据我们所知，我们的工作是第一次尝试在产品页面、短视频和直播中建立统一的产品表示，以满足新兴的富内容电子商务的迫切行业需求。

2)收集在线电子商务平台的真实数据，构建大规模跨领域产品识别数据集 ROPE。它包含 3,056,624 个产品页面，5,867,526 个短视频，以及属于 189,958 个不同产品的 3,495,097 个直播。所包含的产品类别涵盖了在线购物场景的全部范围。

3)提出了一种跨领域产品表示模型(COPE)来学习跨领域产品表示。实验结果证明了 COPE 模型相对于现有方法的优越性。

2 相关工作

2.1 电子商务数据集

已经提出了大量的电子商务数据集来推动该领域的技术发展[2,6,8,11,25,32,33]。早期的数据集通常规模有限。Corbiere 等人在 2017 年引入了 Dress Retrieval[6]数据集，该数据集包含 20000 对产品图像和文本对。Rostamzadeh 等人提出了 fashionengen[25]数据集，该数据集包含 293,000 个样本，仅涵盖 48 个产品类别。近年来，随着基于深度学习方法的发展，引入了大规模的产品识别数据集。产品 1M[33]将训练样本的规模提升到百万级，但所有样本均来自 48 个化妆品品牌。产品的覆盖范围很有限。MEP-3M[2]数据集包含 300 多万个样本，每个样本由产品图像、产品标题和分层分类标签组成。然而，所有这些数据集都只关注产品页面域。在实验部分，我们将证明在产品页面领域学习到的表示不足以处理跨领域的产品识别任务。与我们的 ROPE 数据集最相关的数据集是 M5Product[8]和 MovingFashion[11]。M5Product 包含 600 万个样本，对于每个样本，它提供产品图像、产品标题、类别标签、属性表、分配的广告视频和从视频中提取的音频。然而，M5Product 中提供的视频与 ROPE 数据集中引入的实时流有很大不同。M5Product 中的视频全部来自产品页面，通常与广告产品密切相关，产品展示在中间，并贯穿始终。相比之下，在 ROPE 的直播中，卖家和观众之间有很多与产品无关的聊天内容。此外，产品的姿势和位置在直播流中变化很大，这使得 ROPE 数据集对产品识别更具挑战性。MovingFashion[11]也专注于视频和产品页面的对齐。它只包含 15000 个视频，涵盖 13 个产品类别。

2.2 现有的跨域检索方法通常学习视觉域和文本域之间的统一表示。

一些最流行的模型遵循单流架构，如 VL-bert[28]，Imagebert[23]，Videobert[29]，Visualbert[13]和 Uniter[4]。这些模型将视觉和文本特征连接起来，然后使用二元分类器来预测图像-文本对是否匹配。

MovingFahsion 的规模比我们的 ROPE 数据集要小得多，ROPE 数据集涵盖了 1000 多个产品类别，并提供了产品页面、短视频和直播域的百万级样本。

虽然这些方法通常性能较好，但它们的推理效率较低。ViLBERT[18]、LXMERT[30]、CLIP[24]和 CoOp[34]采用了两流架构。该方法使用独立的编码器提取视觉和文本特征，并使用点积运算有效地计算视觉和文本相似度。提出的 COPE 模型使用对比损失学习不同域的表示，以确保有效的跨域检索。

3 ROPE 数据集

3.1 数据收集和清理

我们从在线电子商务平台收集了超过 1300 个产品类别的数据。构造 ROPE 数据集需要三个步骤。

第一，我们从产品页面域、短视频域和直播域收集了大量的无监督多模态样本。对于产品页面域名，我们提供产品图片和标题;对于短视频和直播领域，提供了提取的帧和自动速度识别文本。由此产生的数据集包括超过 2 亿个样本，并被定义为 Draw。

第二，对 Draw 的一小部分(0.1%，200K 数据点)进行采样并定义为 Dsample。对于 Dsample 中的每个样本，我们要求人类注释者从 Draw 中找到具有相同产品的其他样本。为了降低注释成本，所提取的特征具有公共性的中国 CLIP 模型[31]寻找相关样本进行进一步的人检。带注释的样本用于训练基线 COPE 模型。

第三，对于 Draw 中剩余的未注释样本，使用基线 COPE 模型过滤掉相关样本，只保留匹配分数

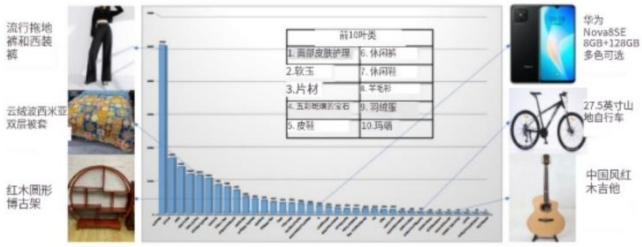

图 2 训练样本在产品类别上的分布。它是有偏见和长尾的。

大于 0.7 的样本。之后，属于同一产品的产品页面、短视频和直播流将被聚合。我们只保留完整的成对样本，包括来自所有三个域的数据。

3.2 数据集的统计数据

最终的 ROPE 数据集包括 3,056,624 个产品页

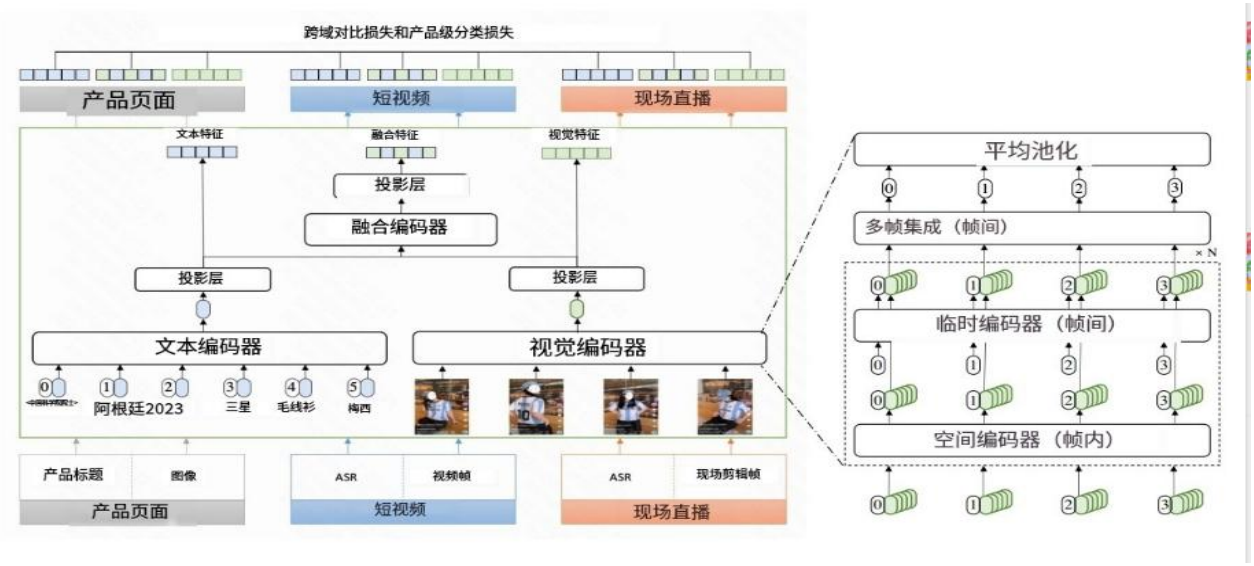

图 3 提出的 COPE 模型的总体框架。利用文本编码器和视觉编码器从单模态中提取特征，并采用融合编码器对两种特征进行聚合。为了对视频和实时流中的时间信息建模，我们将跨帧通信转换器插入到视觉编码器的每个块中。多帧集成转换器放置在视觉编码器的顶部，以总结整个视频的表示。

面，5,867,526 个短视频，以及与 189,958 个产品相关的 3,495,097 个直播流。表 1 比较了 ROPE 和以前的产品数据集。我们将 ROPE 数据集分为训练集和测试集。火车套装有 187431 种产品，3025236 个产品页面，5837716 个短视频，3446116 个直播。平均每个产品有 16 个产品页面、31 个短视频和 18 个直播。跨产品类别的训练样本分布如图 2 所示，显示了反映在线购物兴趣的长尾模式。前五名分别是地毯、书画、被套、祖母绿和床单。

表 1 与其他产品数据集的比较。“-”表示未提及。

资料组	样本	类别	产品	域
FashionGen [25]	293,008	48	78850	产品页面
Dress Retrieval [6]	20200	50	20200	产品页面
Product1M [33]	1182083	458	92200	产品页面
MEP-3M [2]	3012959	599	-	产品页面
M5Product [8]	6313067	6232	-	产品页面
MovingFashion [11]	1500	-	-	产品页面/短视频
ROPE(ours)	12027068	1368	187431	产品页面/短视频/现场直播

4 方法

所建议的 COPE 模型的总体框架如图 3 所示。它包括视觉编码器、文本编码器、融合编码器和域投影层。视觉编码器和文本编码器在三个域中共享，每个域中的域投影层参数不共享。

4.1 建筑设计

如第 3.2 节所述，我们为每个域提供具有多种模态的训练样本。具体来说，我们为产品领域提供产品标题和图像，而对于短视频和直播领域，我们提供提取的帧和 ASR(自动语音识别)文本。COPE 模型设计了一个两流管道来处理视觉和文本模式。这些提取的特征被输入到三个特定于领域的投影层中，以获得特定于领域的表示。此外，我们采用融合编码器模块，然后是投影层，以聚合视觉和文本特征。融合编码器的参数是跨域共享的，而投影层是特定于域的。值得注意的是，我们没有使用 ASR 文本，并且在 COPE 的初始版本中，我们删除了短视频和直播域的 textmodal 相关模块。原始 ASR 文本中过多的噪声信息会对视频和直播流的最终演示产生负面影响。在我们未来的工作中，我们将探索从原始文本中提取与产品相关的关键字来利用 ASR 文本的可能方法。

视觉编码器遵循与[21]相同的架构，由 N 个跨帧通信变压器(CCT)模块和一个多帧集成变压器

(MIT)模块组成。CCT 模块是修改后的 ViT[9]块，通过插入时间编码器来实现时间信息的可交换性。MIT 模块位于 N 堆栈 CCT 模块的顶部，将帧级特征序列整合到统一的视频表示中。

给定一个输入视频 $V\in\mathbb{R}^{T*H*W*3}$(该产品图像可视为只有一帧的视频)，其中 T 表示帧数。H、W 表示视频的空间分辨率;我们将第 t 帧分割成 M 个不重叠的小块，$\mathbf{X}_{vis}\in\mathbb{R}^{M*3}$。在 patch 序列的开头插入可学习类标记，并将空间位置编码添加到 patch 序列中。在形式上,

$$\mathbf{z}_t^{(0)}=[e_{vis}^{cls};\mathbf{X}_{vis}]+e^{spa}$$

然后我们将 z (0) t 输入到 N 个 CCT 模块中以获得帧级表示:

$$\mathbf{z}_t^{(n)}=\mathrm{CCT}^{(n)}(\mathbf{z}_t^{(n-1)}),n=1,\ldots,N$$
$$=[h_{t,cls}^{(n),vis},h_{t,1}^{(n),vis},h_{t,2}^{(n),vis},,\ldots,,h_{t,M}^{(n),vis}]$$

其中，n 为 CCT 模块索引。

我们取第 N 个 CCT 模块的类标记的最终输出，$h_{t,cls}^{(N),vis}$，来表示第 t 帧。然后利用 MIT 模块对帧级特征进行聚合，得到视频的全局表示。在形式上,

$$Z_{vis}=\mathrm{AvgPool}(\mathrm{MIT}([h_{1,cls}^{(N),vis},\ldots,h_{T,cls}^{(N),vis}]+e^{temp}))$$

其中 AvgPool 和e^{temp}表示平均池化操作符和时间位置编码; 分别。$Z_{vis}\in\mathbb{R}^{d}$作为输入产品图像或视频的视觉表示。

文本编码器是一个三层 RoBERTa[7,15]模型。

首先对输入的原始文本进行标记并定义为 $X_{txt}\in\mathbb{R}^{L}$，其中 L 表示标记序列的长度。然后在序列的开头插入类标记，并添加位置嵌入来重新训练位置信息。最终获得的文本序列被输入文本编码器以提取文本表示。在形式上,

$$\mathbf{H}_{txt}=\mathrm{RoBERTa}([e^{cls};\ X_{txt}]+e^{pos})$$
$$=[h_{cls}^{txt},h_1^{txt},h_2^{txt},\ldots,,h_L^{txt}]$$

其中e^{cls}和e^{pos}分别表示输入类标记嵌入和位置嵌入。$h_{cls}^{txt}\in\mathbb{R}^{d}$表示提取的类令牌特征。我们使用$h_{cls}^{txt}$类作为输入原始文本的文本表示。

使用共享的视觉和文本编码器提取三个域的视觉表示Z_{vis}和文本表示h_{cls}^{txt}类，尽管不同域的样本差异很大。该方案有望提高基本特征提取器的泛化能力。通过利用不同的投影层将一般表示转换为特定于领域的表示，保留和放大每个领域的特征。对于每个域，投影层是一个权值为 W，偏置为 b 的线性层，得到域特定表示为:

$$E_{vis}^{P}=\mathbf{W}_{vis}^{P}Z_{vis}^{P}+b_{vis}^{P}$$
$$E_{txt}^{P}=\mathbf{W}_{txt}^{P}h_{txt}^{P}+b_{txt}^{P}$$
$$E_{vis}^{V}=\mathbf{W}_{vis}^{V}Z_{vis}^{V}+b_{vis}^{V}$$
$$E_{vis}^{L}=\mathbf{W}_{vis}^{L}Z_{vis}^{L}+b_{vis}^{L}$$

式中，P、V、L 分别表示产品页面域、短视频域和直播域。值得注意的是，在短视频领域和直播领域，我们不包括文本模态，只使用视觉表示，即E_{vis}^{V}和E_{vis}^{L}。

最后，提出了融合编码器，然后是投影层，用于聚合视觉和文本表示。融合编码器采用自注意层实现，投影层为线性层。

此外，在 COPE 的初始版本中，融合操作仅应用于产品页面域。在形式上,

$$E_{fus}^{P}=\mathrm{SelfAtten}([E_{vis}^{P};E_{txt}^{P}])$$
$$E_{fus}^{P}=\mathbf{W}_{fus}^{P}E_{fus}^{P}+b_{fus}^{P}$$

其中 SelfAttn 表示自注意层。E_{fus}^{P}是获得的产品页面的多模态表示。对于另外两个域 V 和 L，可视化表示E_{vis}^{V}和E_{vis}^{L}是它们最终得到的表示。

4.2 培养目标

为了学习跨不同领域的统一产品表示，我们首先利用对比学习来训练提出的 COPE 模型，该模型遵循之前的自监督学习方法[3,10,22]。对比损失函数的基本表达式[22]定义为:

$$\mathcal{L}_{con}=-\log\frac{\exp(s_{qk_+}/\tau)}{\sum_{i=0}^{i=K}\exp(s_{qk_i}/\tau)}$$

其中s_{qk_i}表示样本 q 和样本k_i之间的余弦相似度。阳性样品k_+表示与 q 有相同产品标签的样品。

相似度s_{qk}可以用不同形式的表示(视觉、文本或融合)来计算，样本 q 和 k 可以来自不同的领域(产品页面、短视频或直播)。在本文中，我们选择了 7 种不同的相似度s_{qk}实现，得到了 7 种不同的对比损失函数。表 2 总结了实现的细节。基于这七个对比损失函数，我们将跨域损失定义为它们的和。在形式上,

$$\mathcal{L}_{cd}=\sum_{n=1}^{n=7}\alpha_n\mathcal{L}_{con}^{n}$$

其中α_n为第 n 个对比学习损失函数的权值。

除了跨域损失，我们还采用了产品分类损失来训练我们的 COPE 模型。具体来说，我们使用具有共享参数的 MLP(多层感知器)来预测具有特定领域表示的每个领域的产品分类分数。对于产品页面

域，采用了多模态表示方法。对于短视频和直播域，分别采用E^{V}_{vis}和E^{L}_{vis}可视化表示。在形式上，

$$s^{\mathrm{P}} = \mathrm{MLP}(E^{\mathrm{P}}_{fus})$$
$$s^{\mathrm{V}} = \mathrm{MLP}(E^{\mathrm{V}}_{vis})$$
$$s^{\mathrm{L}} = \mathrm{MLP}(E^{\mathrm{L}}_{vis})$$

其中 s 表示每个域的分类得分。

然后使用标准的 softmax 损失对模型进行训练。在形式上，

$$\mathcal{L}_{cls} = -\left(\log \frac{e^{s_i^{\mathrm{P}}}}{\sum_j^C e^{s_j^{\mathrm{P}}}} + \log \frac{e^{s_i^{\mathrm{V}}}}{\sum_j^C e^{s_j^{\mathrm{V}}}} + \log \frac{e^{s_i^{\mathrm{L}}}}{\sum_j^C e^{s_j^{\mathrm{L}}}}\right)$$

训练 COPE 模型的总损失是跨域损失和分类损失的结合。

在形式上，

$$\mathcal{L}_f = \mathcal{L}_{cd} + \beta\mathcal{L}_{cls}$$

式中，β 为分类损失权重。

表 2 不同相似函数 sqk 的实现。

相似性 Sqk	领域	形式
$< E^{\mathrm{P}}_{fus}(q), E^{\mathrm{V}}_{vis}(k) >$	产品视频	融合视觉
$< E^{\mathrm{P}}_{fus}(q), E^{\mathrm{V}}_{vis}(k) >$	产品直播	融合视觉
$< E^{\mathrm{P}}_{fus}(q), E^{\mathrm{V}}_{vis}(k) >$	视频直播	融合视觉
$< E^{\mathrm{P}}_{fus}(q), E^{\mathrm{V}}_{vis}(k) >$	产品直播	融合视觉
$< E^{\mathrm{P}}_{fus}(q), E^{\mathrm{V}}_{vis}(k) >$	产品直播	文本视觉
$< E^{\mathrm{P}}_{fus}(q), E^{\mathrm{V}}_{vis}(k) >$	产品直播	文本视觉
$< E^{\mathrm{P}}_{fus}(q), E^{\mathrm{V}}_{vis}(k) >$	产品直播	文本视觉

4.3 实验结果

在本节中，我们将评估我们提出的 COPE 模型，并将其与 ROPE 数据集上的最先进模型进行比较。考虑了跨域产品检索任务和单域分类任务。由于没有现有的方法可以精确地适用于我们的跨域设置，我们将 COPE 模型与多模态视觉语言模型[5,16,19,20,31]进行比较，这些模型没有在我们的数据集上进行微调。通过平均图像和文本特征来获得这些模型的产品页面表示。通过平均所有帧的表示提取短视频和直播表示。

在表 3 的第一个隔间中比较了用一般图像文本对训练的视觉语言模型。

我们可以看到，在两个评估任务的每一个设置上，它们的性能都不如我们的 COPE 模型。
在检索任务和分类任务中，在 P→L、L→P、V→L、L→V 等动态相关设置下的表现明显低于其他设置。

COPE 模型在 P→V 检索任务中获得 82.58% R@1，在 P→V 分类任务中获得 59.84% Acc。相比之下，COPE 在 P→L 设置下的性能分别为 54.06%和 34.95%。实时流中的产品的规模和视图与产品页面不同。

4.4 分类损失有效性

在本节中，我们将研究分类损失对模型的影响。由于我们的数据集中有大量的类别，我们使用了 Partial-FC[1]来提高训练效率。如表 6 所示，包含分类损失可以大大提高模型在所有检索任务中的性能。在 P2V 和 L2P 任务上，有 1cl 的模型在 rank-1 准确率上分别比没有 1cl 的模型高出 30%和 19%。它为分类损失的有效性提供了令人信服的证据。

4.5 采样策略

在表 7 中，我们给出了随机抽样和产品平衡抽样之间的比较。在有 N 个样本的小批量中，随机抽样是指从训练集中随机选择 N 个样本。相比之下，产品平衡抽样选择 P 个产品，然后从每个产品中抽样 K 个实例，得到 N = P x K 样本。

实验结果表明，均衡采样显著提高了模型的性能。

4.6 可视化

在图 4 中，我们展示了产品页面、短视频和直播流嵌入的 t-SNE 可视化。我们随机选择了 30 种产品及其对应的产品页面、短视频和直播来生成这种可视化。可视化清楚地说明了同一产品的嵌入紧密地定位在一起。

表 3 COPE 上的检索和分类结果。P、V、L 分别代表产品页面域、短视频域和直播域。

		跨域检索						few-shot 分类
模型	跨域设置	R@1	R@5	R@10	R @20	R @50	R @mean	Top1 Acc
CLIP4CLIP [19]	P2V	59.06	79.31	86.02	91.01	95.03	82.08	27.94
	V2P	38.48	52.25	59.16	66.54	74.65	58.21	26.55
	P2L	23.68	38.14	45.32	54.27	66.79	45.64	9.97
	L2P	14. ,46	24.52	30.77	38.09	48.91	31.35	10.75
	V2L	18.10	29.83	35.65	42.22	52.01	35.56	9.47
	L2V	20.14	33.51	40.44	48.05	58.68	40.16	7.22
TS2-Net [16]	P2V	57.42	77.88	85.29	90.44	94.92	81.19	26.11
	V2P	36.56	50.93	58.02	65.12	73.89	56.9	24.09
	P2L	22.85	38.49	45.91	54.11	65.89	45.45	9.83
	L2P	14.16	24.52	30.5	37.52	48.37	31.01	10.57
	V2L	17.69	29.63	34.84	41.27	50.95	34.87	9.68
	L2V	20.55	33. .80	40.91	48. ,46	59.16	40.57	7.40
X-CLIP [20]	P2V	56.61	77.46	84.84	90.11	94.51	80.70	26.97
	V2P	35.29	49.41	56.82	64.13	72.54	55.63	23.55
	P2L	22.66	37.47	44.33	52.11	63.38	43.98	9.72
	L2P	13.52	23.08	28.92	35.98	46.14	29.52	8.88
	V2L	17.64	28.71	34.03	40.17	49.67	34.04	9.05
	L2V	19.60	32.73	39.51	47.07	57.25	39.23	7.42
ChineseCLIP [31]	P2V	56.93	79.80	87.43	92.48	96.51	82.65	31.44
	V2P	40.48	57.85	66.74	75.25	84.03	64.87	29.10
	P2L	34.37	50.83	58.66	67.05	78.57	57.89	19.23
	L2P	22.49	37.11	46.78	56.30	68.14	46.16	15.73
	V2L	25.51	38.28	45.02	52.27	62.53	44.72	13.24
	L2V	28.28	45.87	53.67	62.18	72.27	52.45	14.16
FashionClip [5]	P2V	44.31	67.06	75.25	82.57	89.29	71.69	18.59
	V2P	25.51	40.75	48.71	56.63	65.94	47.50	15.88
	P2L	19.54	31.14	36.98	43.91	54.39	37.19	8.70
	L2P	11.22	24.23	31.90	40.05	50.96	31.67	7.57
	V2L	15.55	24. .88	29.51	35.07	42.68	29.53	6.80
	L2V	21.20	35.72	42.55	49.60	58.77	41.56	10.40
COPE (Ours)	P2V	82.58	94.88	97.54	98.89	99.65	94.70	59.84
	V2P	65.20	76.56	82.04	86.86	91.69	80.47	57.12
	P2L	54.06	71.07	77.14	82.86	89.70	74.96	34.95
	L2P	42.33	56.48	63.67	71.11	80.22	62.76	36.51
	V2L	45.95	63.63	70.64	77.50	85.47	68.63	30.43
	L2V	48.28	67.20	74.70	81.52	89.15	72.17	33.30

表 4 Product1M 上的 COPE 检索结果

model	mAP@10	mAP@50	mAP@100	mAR@10	mAR@50	mAR@100	Prec@10	Prec@50	Prec@100
SOTA	79.36	74.79	74.63	34.69	30.04	30.08	73.97	72.12	73.86
COPE (Ours)	86.02	80.51	77.35	53.53	57.03	58.03	80.30	72.39	66.58

表 5 COPE 在 m5 产品上的检索结果

模型	mAP@l	mAP@5	Prec@l	Prec@5
SOTA(I+T)	62.20	66.97	62.20	49.85
SOTA(ALL)	69.25	74.08	69.25	58.76
COPE(Ours)	80.89	83.66	80.89	75.96

图 4 COPE 嵌入的 t-SNE 可视化。同一产品的点颜色相同。

表 6 分类损失显著提高了所有任务的性能。

任务	loss	R@1	R@5	R@10	R@20	R @50
P2V	w/o $\mathcal{L}_{cls}$	51.88	76.45	84.58	90.58	95.50
	w $\mathcal{L}_{cls}$	82.58	94.88	97.54	98.89	99.65
V2P	w/o $\mathcal{L}_{cls}$	44.17	60.01	68.24	75.86	84.29
	w $\mathcal{L}_{cls}$	65.2	76.56	82.04	86.86	91.69
P2L	w/o $\mathcal{L}_{cls}$	26.41	44.76	53.25	62.72	75.26
	w $\mathcal{L}_{cls}$	54.06	71.07	77.14	82.86	89.70
L2P	w/o $\mathcal{L}_{cls}$	23.11	38.28	47.97	57.88	71.04
	w $\mathcal{L}_{cls}$	42.33	56.48	63.67	71.11	80.22
V2L	w/o $\mathcal{L}_{cls}$	29.39	47.54	55.88	64.47	75.81
	w $\mathcal{L}_{cls}$	45.95	63.63	70.64	77.50	85.47
L2V	w/o $\mathcal{L}_{cls}$	29.50	52.07	62.60	72.30	83.12
	w $\mathcal{L}_{cls}$	48.28	67.20	74.70	81.52	89.15

表 7 随机抽样(rs)和产品平衡抽样(pb)两种抽样策略的比较

任务	策略	R@1	R@5	R@10	R@20	R@50
P2V	rs	70.08	88.49	93.18	96.22	98.40
	pb	82.58	94.88	97.54	98.89	99.65
V2P	rs	55.26	68.74	75.44	81.87	88.45
	pb	65.20	76.56	82.04	86.86	91.69
P2L	rS	40.85	60.39	68.51	76.42	85.79
	pb	54.06	71.07	77.14	82.86	89.70
L2P	rs	33.10	48.67	57.39	66.03	76.70
	pb	42.33	56.48	63.67	71.11	80.22
V2L	rs	37.66	56.10	64.28	72.30	82.04
	pb	45.95	63.63	70.64	77.50	85.47
L2V	rs	38.31	60.06	68.99	77.40	86.41
	pb	48.28	67.20	74.70	81.52	89.15

此外，图 5 显示了一些检索结果。值得注意的是，大多数假阳性结果与查询属于同一类别，并且具有相似的视觉特征

图 5 检索结果的可视化，其中红框表示假阳性

5 总结

为了创建统一的跨领域产品表示，我们引入了一个大型电子商务跨领域数据集，该数据集包括三个领域(产品页面、短视频和直播)和两种模式(视觉和语言)。它是第一个包含电子商务场景中各个领域的数据集。我们提出了我们的 COPE 作为基准，并在跨域检索和小样本分类任务上对其进行了评估。最后，我们对结果进行了分析和可视化。该任务适用于大多数电子商务平台，数据集和提出的框架都将激发跨领域产品表示的研究。

参 考 文 献

[1] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In ICCV, pages 1445 – 1449, 2021. 8

[2] Delong Chen, Fan Liu, Xiaoyu Du, Ruizhuo Gao, and Feng Xu. Mep-3m: A large-scale multi-modal e-commerce products dataset. In IJCAI, volume 21, 2021. 2, 3

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597 – 1607. PMLR, 2020. 6

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In ECCV, pages 104 – 120. Springer, 2020. 3

[5] Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhaes, Diogo Goncalves, Ciro ˜ Greco, and Jacopo Tagliabue. Fashionclip: Connecting language and images for product representations. arXiv preprint arXiv:2204.03972, 2022. 6, 7

[6] Charles Corbiere, Hedi Ben-Younes, Alexandre Rame, and ´ Charles Ollion. Leveraging weakly annotated data for fashion image retrieval and label prediction. In ICCV, pages 2268 – 2274, 2017. 2, 3

[7] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3504 – 3514, 2021. 5, 6

[8] Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. M5product: Selfharmonized contrastive learning for e-commercial multimodal pretraining. In CVPR, pages 21252 – 21262, 2022. 2, 3, 7

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 5

[10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In EMNLP, pages 6894 – 6910, 2021. 6

[11] Marco Godi, Christian Joppi, Geri Skenderi, and Marco Cristani. Movingfashion: a benchmark for the video-to-shop challenge. In WACV, pages 1678 – 1686, 2022. 2, 3

[12] Brendan Kolisnik, Isaac Hogan, and Farhana Zulkernine. Condition-cnn: A hierarchical multi-label fashion image classification model. Expert Systems with Applications, 182:115195, 2021. 2

[13] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019. 3

[14] Huidong Liu, Shaoyuan Xu, Jinmiao Fu, Yang Liu, NingXie, Chien-Chih Wang, Bryan Wang, and Yi Sun. Cmaclip: Cross-modality attention clip for image-text classification. arXiv preprint arXiv:2112.03562, 2021. 2

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019. 5, 6

[16] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In ECCV, pages 319 – 335. Springer, 2022. 6, 7

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In ICLR. 6

[18] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In NeurIPS, pages 13 – 23, 2019. 3

[19] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing, 508:293 – 304, 2022. 6, 7

[20] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In ACMMM, pages

638 – 647, 2022. 6, 7

[21] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. 2022. 5, 6

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018. 6

[23] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966, 2020. 3

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748 – 8763. PMLR, 2021. 3

[25] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. 2018. 2, 3

[26] Majuran Shajini and Amirthalingam Ramanan. An improved landmark-driven and spatial – channel attentive convolutional neural network for fashion clothes classification. The Visual Computer, 37(6):1517 – 1526, 2021. 2

[27] Majuran Shajini and Amirthalingam Ramanan. A knowledge-sharing semi-supervised approach for fashion clothes classification and attribute prediction. The Visual Computer, 38(11):3551 – 3561, 2022. 2

[28] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visuallinguistic representations. In ICLR. 3

[29] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In ICCV, pages 7464 – 7473, 2019. 3

[30] Hao Tan and Mohit Bansal. Lxmert: Learning crossmodality encoder representations from transformers. In EMNLP-IJCNLP, pages 5100 – 5111, 2019. 3

[31] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. arXiv preprint arXiv:2211.01335, 2022. 3, 6, 7

[32] Wenjie Yang, Yiyi Chen, Yan Li, Yanhua Cheng, Xudong Liu, Quan Chen, and Han Li. Cross-view semantic alignment for livestreaming product recognition. arXiv preprint arXiv:2308.04912, 2023. 2

[33] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In ICCV, pages 11782 – 11791, 2021. 2, 3, 7

[34] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In CVPR, pages 16816 – 16825, 2022. 3