Kaleido-BERT: 时尚领域视觉-语言预训练模型

诸葛鸣晨 [1,‡] 高德宏 [1,‡] 范登平 [2, ✉]
金林波 [1] 陈犇 [1] 周昊明 [1] 邱明辉 [1] 邵岭 [2]
[1] 阿里巴巴集团 [2] 阿联酋起源人工智能研究院（IIAI）

http://dpfan.net/Kaleido-BERT/

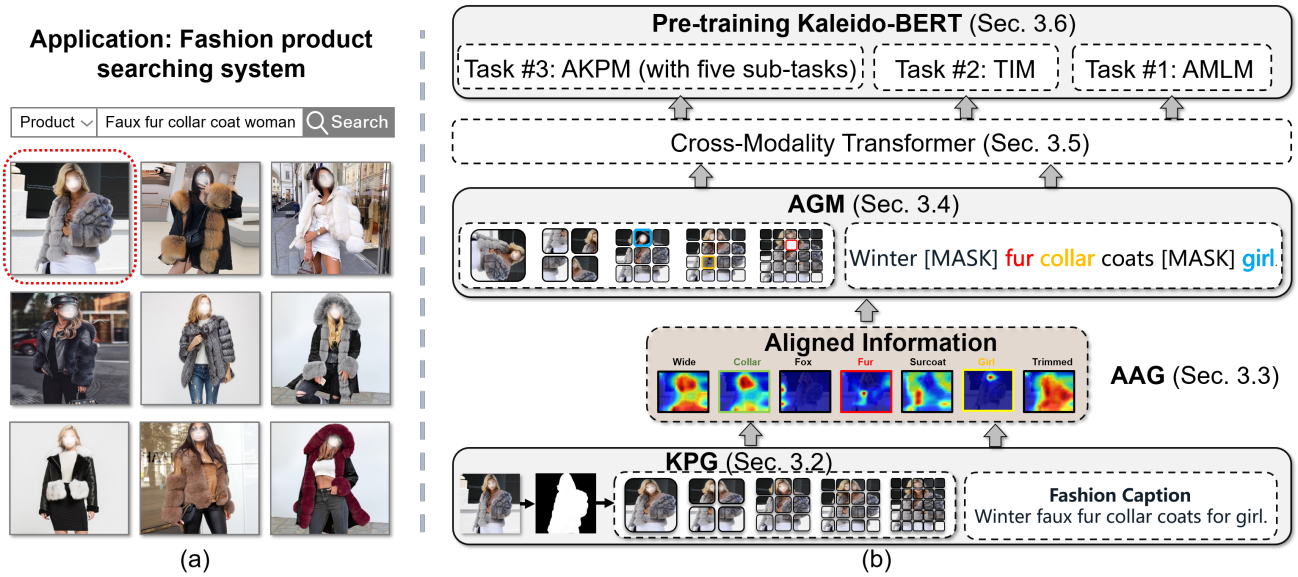

图 1: **时尚领域下视觉-语言 (VL) 预训练模型图示。**本文提出了一个新的多模态预训练模型 (Kaleido-BERT)，它由 Kaleido 图像块生成器 (KPG)，注意力对齐生成器，和预对齐掩码策略组成，用以更好地学习多模态特征。Kaleido-BERT 在 Fashion-Gen 数据集上取得了最好的效果，并且已经部署在真实场景中。

摘要

本文提出了名为 *Kaleido-BERT* 的视觉-语言 *(VL)* 预训练模型，引入百变 *(Kaleido)*策略生成不同尺度的图像特征，旨在帮助 *Transformer* 更好地学习时尚领域的多模态表征。与现有的视觉-语言多模态模型所使用的随机掩码策略不同的是，本文设计了预对齐掩码策略，它能进一步关注图像-文本对的语义关系。在此基础上，本文又针对不同尺度的图像块分别设计了五种自监督任务，分别为旋转 *(Rotation)*、拼图 *(Jigsaw)*、伪装 *(Camouflage)*、着色 *(Grey-to-Color)*、修复 *(Blank-to-Color)*。*Kaleido-BERT* 简而易行，可以轻松地嵌入到 *BERT* 框架中，并在四个下游任务上取得了最佳的效果，如文本检索 *(R@1: 4.03%* 绝对提升*)*, 图像检索 *(R@1: 7.13%* 绝对提升*)*, 类目识别 *(ACC: 3.28%* 绝对提升*)* 以及时尚描述（*Bleu4*：*1.2* 绝对提升)。本文在大规模电商网站检验了其性能，充分挖掘了它在实际场景中的应用前景。

‡: 相同贡献。通讯作者: 范登平 *(dengpfan@gmail.com)*. 本文为 CVPR2021 [85] 论文翻译版.

1. 引言

Transformers [14, 68], 最初被设计并用于自然语言处理 (NLP) 领域，随后又在其它领域大获成功 [5, 11]，包含视觉领域 (如 Selfie [66], DETR [6], ViT [34] 以及 PVT [69])，视觉-语言 (VL) 多模态领域 (ViLBERT [45], VL-BERT [60], OSCAR [42])。当前的多模态预训练模型 (PTM), 如 VL-BERT [60] 和 UNITER [9]，都注重于学习通用的视觉-语言表征 (如，粗粒度匹配)，这对通用领域多模态表征学习有较大帮助。

No.	预训练模型 (PTM)	年份	出版物	结构	训练数据集	核心思想	视觉领域	是否预训练?	Finetune 任务	视觉特征	代码
1	VisualBERT [40]	2019	arXiv	单分支	Coco	首个图像-文本预训练模型 *AFAK*	图像	✓	U	RoI	Torch
2	CBT [63]	2019	arXiv	双分支	Kinetics [30]	引入噪声对比估计损失	视频	✓	U/G/O	Frame	N/A
3	VideoBERT [64]	2019	ICCV	单分支	SC	首个视频-文本预训练模型 *AFAK*	视频	✓	G/O	Frame	N/A
4	B2T2 [2]	2019	EMNLP	单分支	CC	送入 PTM 前, 隐式地关联 RoI 与文本信息	图像	✓	U	RoI	Tensorflow
5	LXMERT [38]	2019	EMNLP	双分支	VG+Coco	三个分别用于 RoIs, 语言, 跨模态特征的解码器	图像	✓	U	RoI	Torch
6	ViLBERT [45]	2019	NeurIPS	双分支	CC	跨模态协同注意力层	图像	✓	U	RoI	Torch
7	ImageBERT [55]	2020	arXiv	单分支	CC+VG+SC	收集大规模的图像-文本对作为预训练数据集	图像	✓	U	RoI	N/A
8	Unicoder-VL [38]	2020	AAAI	单分支	CC+SBU	图像掩码目标分类任务	图像	✓	U	RoI	N/A
9	VLP [82]	2020	AAAI	单分支	CC	统一的视觉语言预训练 (VLP) 模型	图像	✓	U/G	RoI	Torch
10	VL-BERT [61]	2020	ICLR	单分支	CC	视觉特征向量融合词向量	图像	✓	U	RoI	Torch
11	VD-BERT [70]	2020	EMNLP	单分支	VisDial [12]	Video-Dialog 预训练	视频	✓	O	RoI	Torch
12	VLN-BERT [48]	2020	ECCV	双分支	Matterport3D [7]	引入预训练模型到 VL 导航	图像		O	RoI	N/A
13	HERO [39]	2020	EMNLP	多分支	TV+HT100M	视频-子标题匹配 & 帧顺序建模	视频	✓	U	Frame	N/A
14	XGPT [73]	2020	arXiv	单分支	CC+SC	增强 VL 生成能	图像	✓	U/G	RoI	N/A
15	InterBERT [43]	2020	arXiv	单分支	Coco+CC+SBU	组掩码建模	图像	✓	U	RoI	Torch
16	VILLA [20]	2020	NeurIPS	双分支	Coco+CC+SBU	对抗预训练和 Finetune	图像	✓	U/O	RoI	Torch
17	ActBERT [83]	2020	CVPR	单分支	HT100M	全局和局部目标区域 & Tangled Transformer	视频	✓	U/O	Frame & RoI	N/A
18	PREVALENT [24]	2020	CVPR	双分支	Matterport3D [7]	用图像-文本-动作三元组进行预训练	图像	✓	O	Image	Caffe & C++
19	12-IN-1 [46]	2020	CVPR	双分支	ES	多任务学习	图像	✓	U	RoI	Torch
20	Pixel-BERT [27]	2020	arXiv	单分支	Coco+VG	像素级 VL 语义对齐	图像	✓	U	Pixel	N/A
21	FashionBERT [21]	2020	SIGIR	单分支	Fashion-Gen [58]	图像块 & 自适应损失函数	图像	✓	U	Patch	Tensorflow
22	UNITER [9]	2020	ECCV	单分支	Coco+VG+CC+SBU	条件掩码 & 词区域对齐	图像	✓	U	RoI	Torch
23	VisDial-BERT [50]	2020	ECCV	双分支	CC+VQA [4]	采用 ViLBERT 用于 Visual Dialog	图像		O	RoI	Torch
24	OSCAR [42]	2020	ECCV	单分支	ES	目标分类标签作为锚点	图像	✓	U/G	RoI	Torch
25	ERNIEL-VIL [78]	2020	arXiv	双分支	CC+SBU	知识增强的 ERNIE [80] 构建 VL 预训练模型	图像	✓	U	RoI	Paddle
26	RVL-BERT [10]	2020	arXiv	单分支	VDR [44]	用 VL-BERT 作视觉关系检测	图像		U	RoI	Torch
27	UniVL [47]	2020	arXiv	双分支	HT100M	5 种预训练目标和 2 种预训练策略	视频	✓	U/G	Frame	N/A
28	MMFT-BERT [32]	2020	EMNLP	多分支	TV	多模态融合的 PTM	图像	✓	U	RoI	Torch
29	**Kaleido-BERT (OUR)**	2021	CVPR	单分支	Fashion-Gen [58]	Kaleido 图像块 & 预对齐掩码策略	图像	✓	U/G	Patch & Coordinate	Tensorflow

表 1: **总结 28 个代表性多模态方法以及本文的 Kaleido-BERT 模型. 训练数据集:** Coco = *MSCOCO Caption* [8]. VG = *Visual Genome* [35]. CC = *Conceptual Caption* [59]. SBU = *SBU Captions* [53]. TV = *TVQA* [37]. HT100M = *HowTo100M* [49]. SC: *Self Collection*. ES: *12-in-1 and OSCAR* 分别了集成 *12, 5+* 个数据集. **Finetune:** U = 理解任务 (*e.g.* classification). G = 生成任务 (*e.g.* image caption). O = 其它 (*e.g.* 行为预测任务).

不过, 在诸多电商环境下 (如: 配件, 衣服, 玩具), 最主要的目的是学习细粒度表征 (如: 短袖, 棉质及针织物) 而不是通用场景下的粗粒度表征 (是什么, 在哪)。然而, 当前在通用领域下的 VL 模型 [9, 60] 仅可作为时尚领域任务 [1, 26, 67] 的次优解, 这类模型往往从全局出发, 并不利于迁移至属性相关的任务, 比如对特定的时尚领域 [75] 的描述生成和类目预测 [15]。因为它们更需要从图像文本对中提取细粒度特征或者相似性信息 [65]。

在本文的研究中, 提出了一个应用在时尚领域的模型 (具体可见图. 1)。它的核心思想是聚焦细粒度表征学习并且减轻图文间的语义隔阂。为实现这一目的, 本文先引入了高效的“百变 (Kaleido)”策略, 它在图像侧提取了一系列不同尺度的细粒度图像块。因此, 所提出的模型命名为“**Kaleido-BERT**”。该策略将单一尺度扩展为多尺度图像特征, 并用于预训练, 在很大程度上可以减轻通用领域粗粒度表征带来的问题。除此之外, 为了减轻跨模态的语义隔阂, 引入 SAT 网络 [74] 生成 Kaleido 图像块和文本词例的对齐信息。这些预对齐信息可为训练模型的掩码策略提供帮助。因此, Kaleido-BERT 可以隐式学习不同模态间的语义信息。总而言之, 本文的贡献归结为:

- **Kaleido 图像块生成器**: 本文提出了 Kaleido 图像块生成器用来生成多尺度的细粒度图像块。并对不同尺度的图像块设计了不同的预训练自监督任务, 如旋转 (Rotation)、拼图 (Jigsaw)、伪装 (Camouflage)、着色 (Grey-to-Color) 和修复 (Blank-to-Color) 任务。它们帮助 Kaleido-BERT 学习多模态细粒度信息。在时尚领域下, 优于基于固定尺度图像块或者 RoI 的 VL 模型。
- **注意力对齐生成器**: Kaleido-BERT 引入了预对齐策略去产生 Kaleido 图像块与文本词例之间的跨模态对齐信息。这些预对齐的图像文本能够很好的帮助减轻多模态间的隔阂。
- **预对齐掩码策略**: 本文还提出了预对齐指导掩码策略, 让 Kaleido-BERT 隐式地学习视觉与语言的语义关联。实验证明了注意力对齐生成器和预对齐掩码策略的重要性。

2. 相关工作

目前, 已有大量的 VL 工作被提出 [3, 4, 23, 28, 31, 54, 62, 77, 79]。在本节中, 将简要介绍基于 Transformer 的 VL 模型, 更多细节可参见表. 1.

2.1. 视觉-语言预训练

近期基于 Transformer 的预训练模型, 诸如 BERT [14], GPT2 [57], XLNet [76], 以及 GPT3 [5], 影响了众多 NLP 任务。受这些工作启发, 很多解决视觉-语言学习 (*e.g.*, 视频 (图像)/ 文本对) 所设计的多模态模型被提出。针对视频-文本对的模型, CBT [63]

https://github.com/mczhuge/Kaleido-BERT/.

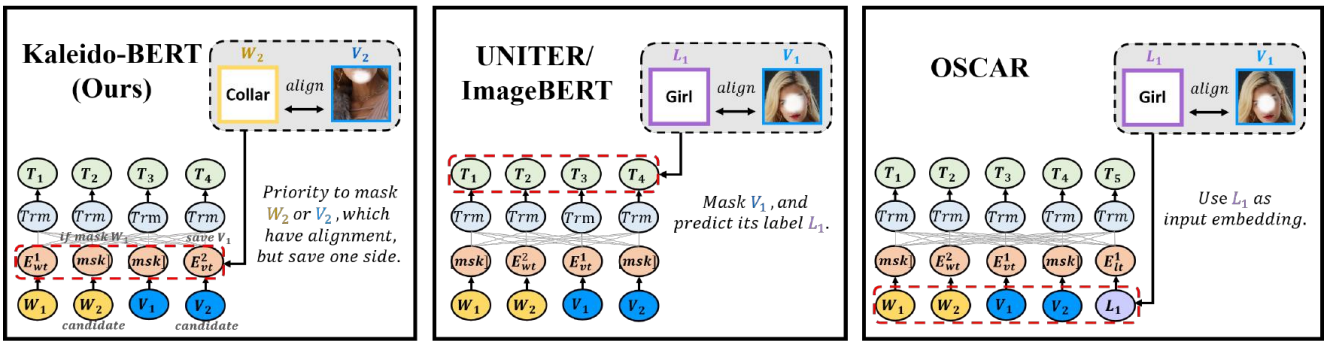

图 2: **现有 VL 预训练模型中，不同的预对齐信息策略对比。** **T** = *Task.* **W** = *Word.* **V** = *Visual Token.* **L** = *Object Detection Label.* **E** = *Embedding.* **Trm** = *Transformer Block.*

和 VideoBERT [64] 是两个探索预训练在多模态领域应用的先驱工作。ActBERT [83] 和 HERO [39] 则更多关注于下游应用，而 UniVL [47] 同时聚焦于视频-语言理解和生成任务。

对于图像-文本模型，根据网络模型处理不同模态输入的方式，可分为单分支结构 [2,9,10,21,27,38,40–43,55,61,70,73,82]，双分支结构 [13,15,24,38,45,48,50,78]，和多分支结构 [32]。对于单分支结构的模型，不同模态的特征会直接送入 Transformer 中。双分支结构的模型则不同，它们首先使用两个子网络处理单模态特征，再将处理好的特征送入 Transformer 中学习。以此类推，多分支结构也遵循这一分类原则。ViLBERT [45] 认为双分支结构优于单分支结构，而 VL-BERT [61] 则认为单分支结构能取得更为理想的结果，因为单分支结构模型有更为充分的信息交互。VisualBERT [40] 和 B2T2 [2] 是为视觉-语言理解任务所设计的单分支结构模型。由于受到通用领域 VL 任务的启发，一些基于 BERT 的模型被提出，如 Unicoder-VL [38], VLP [82], ViLBERT [45], VL-BERT [61]。多模态模型在通用领域的任务 (*e.g.*, VCR [9,43,78], VQA [32,38]) 如雨后春笋般发展，而其他一些任务如视觉关系检测 (RVL-BERT [10]), 视觉导航 (*i.e.*, PERVALENT [24] 和 VLN-BERT [48]), 以及视觉对话 (*e.g.*, VisualD [50], VD-BERT [70]) 还处于萌芽阶段。最近，Lu 等人 [46] 展示了相较于独立任务学习，多任务 VL 学习可以取得显著的提升。在图像描述任务上，XGPT [73] 取得了很好效果；而在图像检索任务，相关算法如 Image-BERT [55] 则表现优异。近期提出的 OSCAR [42] 则在众多相关的 VL 任务上取得国际领先的表现。

VirTex [13] 通过引入密集的语义描述来学习视觉表征，在图像分类、目标检测和实例分割领域取得优异的效果。而另一值得注意的工作中 [41]，作者创新性的证明了多头注意力可以执行实体和句法推理。区别于上述工作（常用区域级图像特征），Pixel-BERT [27] 使用像素级图像特征进行 VL 对齐。

如图. 2所示，之前的模型往往将预对齐信息使用在任务层 (*e.g.*, LXMERT [38] 和 UNITER [9]) 或输入层，如 OSCAR [42]。而本文提出的 Kaleido-BERT 将预对齐信息应用于掩码策略上，从而能够更好的学习时尚领域任务上的细粒度表征。

2.2. 时尚领域任务

如 § 2.1描述，现有的大多数 VL 模型仅关注相对粗糙的通用表征，它们没有将注意力放在时尚领域细粒度的表征学习上。目前有两个与本文相似的研究工作 [15,21]。其中, FashionBERT [21] 是第一个针对时尚领域设计的预训练模型。另一项同期的工作, MAAF [15] 旨在推导一种模态不可知的注意力融合策略，以解决无差别的文本和图像检索任务。与 FashionBERT 利用固定尺寸的图像块不同，MAAF 采用了图像级的注意力机制。本研究认为它们都限制了预训练模型的表征能力，尤其是在细粒度理解的时尚任务。因此，设计多尺度图像块作为细粒度输入的工作是学界/工业界中急需的。

本文提出的 Kaleido-BERT 是首个使用预对齐掩码策略来隐式关联图像-文本语义的模型。

3. 提出的 Kaleido-BERT

在这一章节中，将详细介绍 Kaleido-BERT，相比于学习通用场景中的粗粒度表征, 它注重于学习时尚领域的细粒度视觉-语言特征。本研究采用在 NLP 中典型的 BERT 模型作为框架，使其能在多种基于 Transformer 的多模态学习任务上扩展。

3.1. 模型概览

Kaleido-BERT 的模型结构可见图. 1。它包含 5 个步骤: (1) 在输入阶段, Kaleido-BERT 有两种模态

的特征输入：文本输入 (*e.g.*, 商品图像描述) 以及由 Kaleido 图像块生成器 (**KPG**) 所产生的对应的图像输入。与 LXMERT [38] 相似，每个文本描述被表征为一系列的词例 (token)，而每一张与文本对应的图像被表示为一系列 Kaleido 图像块。(2) 在图文特征向量生成的阶段，本研究使用了注意力对齐生成器 (**AAG**) 去产生词例与 Kaleido 图像块的预对齐信息，以便图像和文本隐式地进行语义对齐。(3) 在交互阶段，与现有的随机掩码策略不同，本文提出采用预对齐掩码策略 (**AGM**) 以缓解跨模态语义交互难度。(4) 词例和 Kaleido 图像块的特征向量在 Kaleido-BERT 得到充分交互后，模型渐进式的学习视觉-语言的语义信息并产生多模态细粒度表征。(5) 除了掩码语言模型 (Masked Language Modeling, MLM) 和图文匹配任务 (Image-Text Matching, ITM) 外，本工作还使用了 5 种新型的预对齐 Kaleido 模型 (Aligned Kaleido Patch Modeling, AKPM)，即: 旋转, 拼图, 伪装, 着色和修复任务。本文提出的模型基于 `EasyTransfer /Huggingface` 库构建。建议读者参考这些标准库以获得实施细节。

3.2. Kaleido 图像块生成器

以一张商品图片作为输入，并将其送入 Kaleido 图像块生成器 (**KPG**)。如图. 3所示，KPG 使用了显著性检测网络(*e.g.*, BAS [56], EGNet [81], ICON [84] 或参考在论文 [17] 所介绍的模型) 去提取前景分割图，并以前景图为依据框定主体目标。受空间包络 (spatial envelop) [52] 以及分块策略 [18, 19] 的启发，本文探索将单张图像切分不同的尺度 (即, 1×1, 2×2, . . . , 5×5)。这些图像块就是“Kaleido(百变)”图像块。除此之外，也可以根据特定任务的难度去考虑更为细致的划分 (如 6×6，或像是 Pixel-BERT [27] 的 N×N 划分)。最终，每一张图像被划分为 55 块 Kaleido 图像块。为了生成这些图像块的特征向量，本文采用 ResNet-50 [25] 作为骨干网络进行模型的特征提取。

3.3. 注意力对齐生成器

注意力对齐生成器 (**AAG**) 目的是产生文本词例 (token) 与 Kaleido 图像块之间的模糊对齐。如图. 4中，直接使用了著名的 SAT 网络 [74]，将其在 FashionGen 数据集上重新训练。之后，它作为文本生成器，自动描述图像的内容。在图像描述阶段，SAT 网络会对每一个词例生成注意力热图，以

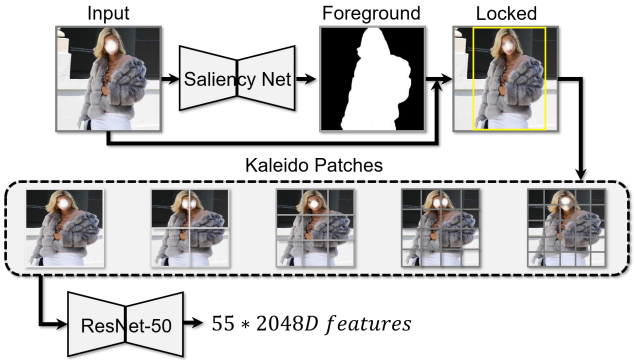

图 3: **Kaleido 图像块生成器 (KPG)**。更多信息参见 § 3.2。

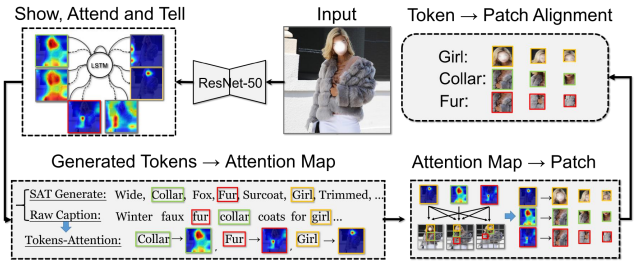

图 4: **注意力对齐信息生成器 (AAG)**。更多细节见 § 3.3。

这些热图为依据可以推断生成的词与图像区域的关系。若生成的描述和原本描述有共现的单词，将依照共现单词的注意力热图来判断该单词倾向于与哪一 Kaleido 图像块关联。从而得到一部分原始描述中的单词与 Kaleido 图像块的对齐信息。

3.4. 预对齐掩码策略

通过注意力对齐生成器，模型获得了关联好的 ⟨token, patch⟩ 对。虽然这些对齐信息并不十分精确，但它提供了不同模态间潜在的语义关联。至此，可依照这些信息修改原始的随机掩码策略。将这些信息利用到预训练阶段，它能更好地帮助 Kaleido-BERT 隐式地探索跨模态语义关系。图. 2 (左) 所示，与随机掩码策略不同，预对齐掩码策略 (**AGM**) 会给予更高优先级去掩码有预对齐信息的词例或图像块。当选中了某一预对齐 ⟨token, patch⟩ 进行掩码时，会随机掩码图像或文本中的其中一侧，这有利于 Kaleido-BERT 通过现有信息 (单模态保留的特征) 去推测另一模态丢失的特征。当所有预对齐图像文本对都被遍历后，仍然出现没有足够的预对齐图像-文本对进行预对齐掩码策略时，则重新采用随机掩码策略补足所需要的掩码个数。通过这样的方式，得到了词例 (token) 与图像块 (patch) 的候选掩码。AGM 策略在 Kaleido 图像块中的 3×3、4×4、

Tensorflow: https://github.com/alibaba/EasyTransfer
Pytorch: https://github.com/huggingface/transformers
为简化, 本文仅使用非常简单 UNet 样式的结构作为前景分割网络。

图 5: **预对齐 Kaleido 图像块模型 (AKPM).** (I) 旋转: Rotation recognition。(II) 拼图: Jigsaw puzzles solving。(III) 伪装: Camouflaged prediction。(IV) 着色: Grey-to-color modeling。(V) 修复: Blank-to-color modeling。可放大以供更好的阅览。更多信息参见 § 3.6。

5×5 层级生效。本文研究工作没有将掩码策略应用于 1×1、2×2 这两种尺度是因为掩码大的图像块会增加模型的预训练难度 (且意义不大)。根据经验，本文分别在 3×3 图像块挑出 1 块，4×4 图像块挑出 2 块，5×5 图像块挑出 3 块进行掩码。

3.5. 多模态 Transformer

本文使用原始的 BERT [14] 构建多模态 Transformer，这使得 Kaleido-BERT 易于开发和迁移。具体而言，在文本侧沿用了 FashionBERT [21] 的做法，即将词例序列 (*i.e.*, 由 WordPieces [72] 产生) 的位置信息编码为 $0, 1, 2, 3, \ldots, N$。在 BERT 中，每一个文本训练语料是由其本身的词嵌入、语义特征、位置编码特征相加而来，再接一个归一化层 (LN Layer) 生成最后的特征向量。而对于图像训练特征，先将每一个图像块的位置信息编码成五维的特征 ($[x_1, x_2, y_1, y_2, w * h]$)。然后将图像块特征与它的位置编码特征分别送入到一个全连接层 (FC)，将它们映射到同一个维度上。最后，采用相加通过全连接层后的特征 (*i.e.*, FC (seg_id), FC (img_feature), FC (pos_emb))的方式，可以得到每一个图像块的视觉特征向量，最后将它们送入 LN 层。

3.6. 预训练

为了缓解视觉与语言的语义隔阂，促进多模态表征学习，本文设计了三种训练任务促进预训练过程，分别是: 预对齐掩码语言模型 (**AMLM**)、图文匹配任务 (**ITM**) 以及提出的预对齐 Kaleido 图像块模型 (**AKPM**) (包含 5 个子任务)。

任务 #1: AMLM 采用预对齐掩码策略，可以得到词例和图像块的掩码候选。当确定了掩码候选集后，将掩码的词例进行以下处理: 以 10% 的概率替换成随机的单词, 10% 保持不变, 以及 80% 用 `[MSK]` 替换。将被掩码处理过后的序列表示为: $T_i = \{t_1, ...[MSK], ..., t_T\}$，即词例 t_i 被掩码。随后将词例序列送入模型，最后取出掩码词例在最后一层的输出 (hidden output)，将它们送入分类器，并基于标准的 BERT 词库预测原始词例。AMLM 的目标即是通过周围的词例特征和图像块特征，来还原被掩码掉的词例。它的目标函数如下:

$$\mathcal{L}_{AMLM} = \sum CE(t_i, \mathcal{F}(T, K, \theta)_{MSK\_hidden}), \quad (1)$$

其中，CE 代表交叉熵 (cross-entropy) 损失函数。$\mathcal{F}$ 代表送入 Kaleido-BERT 后的模型所做的特征操作。$\mathcal{F}(\cdot)_{MSK\_hidden}$ 为被掩码词例经过 Kaleido-BERT 的最后一层的特征、K 代表被掩码的 Kaleido 图像块序列编号。

任务 #2: TIM 即文本-图像匹配任务，是由原始 BERT 中 NSP 任务迁移 (NSP 原本是判断两句子是否为上下句关系)。在这个任务中，`[CLS]` 被用来当做融合后的表征的特征头。后续过程中，将 `[CLS]` 送入模型后得到的输出，再送入一层全连接层，并使用 Sigmoid 函数将预测分值统一到 0 到 1 之间。此处，正样本中图像和文本对取自同一个产品，而负样本的图像或文本二者之一取自不同产品。TIM 的目标函数为:

$$\mathcal{L}_{ITM} = CE(y_m, \mathcal{F}(T, K, \theta)_{CLS\_hidden}), \quad (2)$$

这里 y_m 表示文本和图像匹配的真值标定。

任务 #3: AKPM Kaleido 图像块序列是由一系列不同尺度图像块组成的 $\{K_1, K_2, ..., K_N\}$，在这里 N 表示 Kaleido 的不同层级。如图. 5所示, AKPM 针对每一层级的 Kaleido 图像块都有独立的任务。

子任务 #I: 旋转 (RR). 最近的两个工作 [22, 29] 比较了不同的自监督学习策略，并认为预测图像旋转是最有效的任务。受此启发，本文引入 RR 到预训练预测中。具体而言，层级 1 中的 1×1 的图像块将随机旋转 4 个角度，即 $\theta \in \{0°, 90°, 180°, 270°\}$。在训练过程中，使用图像旋转的角度作为真值标定。$K_1$ 图像块在经过模型后的输出被送入一个 FC 层，紧接 Softmax 激活函数。最后经过 Softmax 激活函数的输出被用来预测角度。RR 任务的损失函数为:

$$\mathcal{L}_{RR} = CE(y_r, \mathcal{F}(T, K, \theta)_{K_1\_hidden}), \quad (3)$$

此处 y_r 为旋转的角度。

与 BERT 中的 'segment embeddings' 类似, 将不同模态特征通过编码 ('T' 代表文本, 'I' 代表图像)，以便模型区分。

子任务 #II: 拼图 (JPS). 拼图 [29, 51] 被认为非常适合自监督表征学习。这样的代理任务 (pretext task) 可以挖掘图像块之间的空间关联。基于这样的视角, 本文借鉴了拼图策略促使 Kaleido-BERT 还原打乱的 2×2 图像块序列以学习到它们潜在的关联。为简化该过程，本文将拼图问题建模为 24 种分类任务，即 2×2 图像块拥有 24 种排列方式 (4! = 24)。

$$\mathcal{L}_{JPS} = CE(y_j, \mathcal{F}(T, K, \theta)_{K_2\_hidden}), \quad (4)$$

y_j 指代拼图的排列方式。

子任务 #III: 伪装 (CP). 为了增强模型的判别能力，本研究在模型中引入另一个任务——伪装。该任务用来判断哪一张图像块被替换过。通过图像和文本的现有信息的帮助，这一任务希望模型在训练过程中能观察 3×3 图像块之间的异同。其本质是用另一张其它商品的图像块伪装在正常图像块中, 因此命名它为伪装 (CP)。预训练时加入该任务，可让模型获得较强识别不同产品的能力。伪装检测依然视为分类问题，监督函数如下:

$$\mathcal{L}_{CP} = CE(y_c, \mathcal{F}(T, K, \theta)_{K_3\_hidden}), \quad (5)$$

在此，y_c 代表着伪装图像块的索引。

子任务 #IV: 着色 (G2CM). 与现有模型的图像掩码策略不同，即它们通常直接替换图像特征为同一维度的 0 填充 (可认为是空白图)，本文先将彩色图块置为灰色, 并引入更平滑的灰图着色任务。后续的训练过程中，使用 KL 散度进行监督，并计算回归出的图像块特征与原始彩色特征的差异，这一过程可被认为是灰图着色。该任务有利于多模态模型图像侧的自监督学习。G2CM 的目标函数为:

$$\mathcal{L}_{G2CM} = \sum KLD(k_{4i}, \mathcal{F}(T, K, \theta)_{K_4\_hidden}), \quad (6)$$

这里 KLD 代表 KL 散度，它衡量重建后的特征分布与真值图像块特征的差异，并在训练过程中使其减小。k_{4i} 为 K_4 图像块序列中被掩码后的图像块。

子任务 #V: 修复 (B2CM). 最后一个子任务为空白图块修复 (B2CM)。它与以往的预训练模型一致，将图像特征用同一维度的 0 填充替代，本文也采用了同样的图像块掩码方式。该任务极端考验模型捕捉上下文信息的能力，其具体目标就是减小训练过程中 B2CM 损失:

$$\mathcal{L}_{B2CM} = \sum KLD(k_{5i}, \mathcal{F}(T, K, \theta)_{K_5\_hidden}), \quad (7)$$

此处 k_{5i} 代表着被掩码的图像块。

伪装图像块与原始图像块有着相同的尺度，从其他的产品里随机选取。

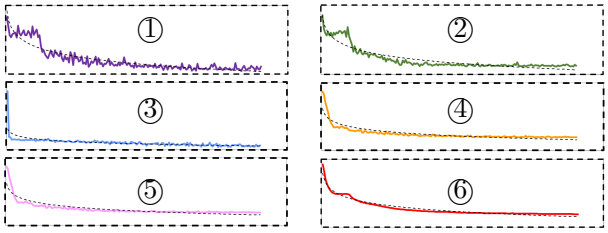

图 6: **损失的变化过程**。图中数值变化均在验证集上测试。①: *Rotation loss*。②: *Jigsaw loss*。③: *Camouflage loss*。④: *Grey-to-Color loss*。⑤: *Blank-to-Color loss*。⑥: *Total Loss = AKPM + ITM + AMLM*。这展示了 Kaleido-BERT 可以通过 Kaleido 策略更好地学习。

总而言之, 本文所引入的预对齐图像块模型可以增强模型对于空间结构上下文的捕捉能力 (*i.e.*, RR and JPS), 分类能力 (*i.e.*, CP) 以及图像生成能力 (*i.e.*, G2CM and B2CM)。最终，模型在训练过程中使得以下损失不断减小:

$$\mathcal{L}_{total} = \mathcal{L}_{AMLM} + \mathcal{L}_{ITM} + \mathcal{L}_{RR} + \mathcal{L}_{JSP} + \mathcal{L}_{CP} + \mathcal{L}_{G2CM} + \mathcal{L}_{B2CM}. \quad (8)$$

在训练过程中, 不同的 Kaleido 任务在验证集上的损失变化曲线如图. 6。由此可见, 损失下降的比较平稳, 证明了预训练过程是正常进行的, 在 Kaleido-BERT 中，所设计的任务能够很好的学习。

4. 实验

本文在预训练好的模型进行端到端微调, 并在四种视觉-语言任务上验证了 Kaleido-BERT 的性能。

4.1. 预训练设置

数据集. 为公平比较，本文遵循了与 Fashion-BERT [21] 相同的设定，并在 Fashion-Gen数据集上预训练 Kaleido-BERT 。

该数据集包含 67,666 个产品。每个产品包含一到六张不同角度的图像及相应的描述。在使用图像文本对时，和 [21] 一样，本文使用 260,480 组图像文本对进行训练，35,528 组图像文本对进行测试。

实施细节. 提出的 Kaleido-BERT 具体参数为: L=12, H=768, A=12。L 是堆叠的 Transformer 块, H 代表隐藏单元数, A 代表注意力头个数。本研究工作在 Tensorflow 平台上进行实验, 共使用 8 张 Tesla V100 显卡进行预训练。学习率为 $2e-5$，Adam 优化器的梯度衰减设为 $1e-4$。并在初始的 5K 步使用 Warming-up 策略。

https://fashion-gen.com/

表 2: **FashionGen 数据集上的检索性能对比。** Sum$\mathcal{R}$=(Rank@1+Rank@5+Rank@10)*100。查看 § 4.3 获得更多细节。

任务		VSE [33]	VSE++ [16]	SCAN [36]	PFAN [71]	ViLBERT [60]	VLBERT [45]	FashionBERT [21]	ImageBERT [55]	OSCAR [42]	Kaleido-BERT 本文模型
1.ITR	Rank@1 ↑	4.010%	4.590%	4.590%	4.290%	20.97%	19.26%	23.96%	22.76%	23.39%	**27.99%**(+4.030%)
	Rank@5 ↑	11.03%	14.99%	16.50%	14.90%	40.49%	39.90%	46.31%	41.89%	44.67%	**60.09%**(+13.78%)
	Rank@10 ↑	22.14%	24.10%	26.60%	24.20%	48.21%	46.05%	52.12%	50.77%	52.55%	**68.37%**(+15.82%)
2.TIR	Rank@1 ↑	4.350%	4.600%	4.300%	6.200%	21.12%	22.63%	26.75%	24.78%	25.10%	**33.88%**(+7.130%)
	Rank@5 ↑	12.76%	16.89%	13.00%	20.79%	37.23%	36.48%	46.48%	45.20%	49.14%	**60.60%**(+11.46%)
	Rank@10 ↑	20.91%	28.99%	22.30%	31.52%	50.11%	48.52%	55.74%	55.90%	56.68%	**68.59%**(+11.91%)
	Sum$\mathcal{R}$ ↑	75.20	94.16	87.29	101.90	218.13	212.84	251.36	241.30	251.53	**319.52**

4.2. 下游任务

本文在四种视觉-语言下游任务上进行验证，包含文本检索、图像检索、类目预测和时尚描述生成。这四种任务都在现实工业场景中有广阔的应用前景。

1. 文本检索 (ITR). 文本检索作为一种下游任务，需要模型判断一个句子是否准确地描述一张图片。本文在 Fashion-Gen [58] 采样了一些商品图像和标题作为图像文本对，并使用原始的产品信息作为正样本。与此同时，打乱数据集并使用不匹配的图像文本对作为负样本。为增加难度，正负样本均采自同样的子类目，因此它们会较难被 PTM 区分。此外，本文使用 Rank@1, Rank@5, Rank@10 评估检索性能。

2. 图像检索 (ITR). 图像检索任务以文本描述为线索，对最相关的商品图像进行排序。与文本检索类似，本文使用真正的商品图像文本对作为正样本，并从同子类目中的商品中随机选取 100 个不相关的描述作为负样本。通过预测样本的匹配分数，本文依旧使用 Rank@1, @5, @10 作为评价指标。

3. 类目/子类目预测 (CR & SUB). 类目是描述商品至关重要的信息，这些信息在现实应用中非常有价值。本文使用分类任务来进行此任务，目的是预测商品的类目和子类目，比如 {`HOODIES`, `SWEATERS`}, {`TROUSERS`, `PANTS`}。在实施过程中，直接在 `[CLS]` 后接一层全连接层来进行该任务。

4. 时尚描述 (FC). 图像描述生成是一项很重要的研究话题，在计算机视觉领域中也有广泛的工作基于此展开。时尚描述的准确率可以衡量多模态模型的生成能力。

4.3. 模型比较

在表. 2和表. 3上, 记录了模型在每个任务上的具体效果。(*i*) 本文的 Kaleido-BERT 几乎在所有指标上都取得较大的性能提升, 证明了它在时尚领域卓越的理解和生成能力。(*ii*) 可以观察到 FashionBERT 方法相较于 ViLBERT 与 VLBERT 有更好的表现。主要的区别在于 FashionBERT 使用图像块作为图像输入特征，而 ViLBERT 和 VLBERT 提取 RoIs 作为特征。这证明了在时尚领域，更好的输入图像特征方式是基于图像块的方法。(*iii*) ImageBERT 和 Oscar 这两大模型，都额外使用了提取 RoIs 时所获得的图像类别标签，相较于 VLBERT 和 ViLBERT 取得了更好的表现。这两种模型都使用了额外的图像侧信息，在一定程度上指明了更多有效的图像语义信息 (*e.g.* 图像特征, 图像监督任务) 可以被用来优化模型的学习。在本文的 Kaleido-BERT 模型中，使用了 Kaleido 策略，它扩展在 FashionBERT [21] 仅利用单一尺度图像块的方式。并且，本文的预对齐掩码策略和 AKPM 任务可促进模型在图像侧进行充分的语义理解。结合以上因素，Kaleido-BERT 在时尚领域的 VL 理解和生成任务上具备优越的性能。

表 3: **类目预测和时尚描述在 FashionGen 数据集上的表现.** 在此, Sum $\mathcal{CLS}$=(ACC+macro-$\mathcal{F}$)*100 and Sum $\mathcal{CAP}$=Bleu-4+METEOR+ROUGE-L+CIDEr. 更多信息请参考 § 4.3。

任务		FashionBERT [21]	ImageBERT [55]	OSCAR [42]	Kaleido-BERT 本文模型
3.CR	ACC ↑	91.25%	90.77%	91.79%	**95.07%** (+3.28%)
	macro-$\mathcal{F}$ ↑	0.705	0.699	0.727	**0.714** (−0.013)
3.SUB	ACC ↑	85.27%	80.11%	84.23%	**88.07%** (+2.80%)
	macro-$\mathcal{F}$ ↑	0.620	0.575	0.591	**0.636** (+0.016)
	Sum $\mathcal{CLS}$ ↑	309.02	298.28	307.82	**318.14**
4.FC	Bleu-4 ↑	3.30	-	4.50	**5.70** (+1.2)
	METEOR ↑	9.80	-	10.9	**12.8** (+1.9)
	ROUGE-L ↑	29.7	-	30.1	**32.9** (+2.8)
	CIDEr ↑	30.1	-	30.7	**32.6** (+1.9)
	Sum $\mathcal{CAP}$ ↑	72.9	-	76.2	**84.0**

4.4. 消融实验

有三个影响 Kaleido-BERT 性能表现的主要因素，它们分别在不同阶段起作用。输入层: Kaleido 图像跨生成器 (KPG)；向量层: 预对齐掩码策略 (AGM)；以及任务层: 对齐 Kaleido 图像块模型。因此本文实施了针对这些因素的消融实验, 去进一步分析这些组件/策略。实验的结果展示在表. 4和图. 7中。

KPG. 本研究尝试了 3 种方法生成 Kaleido 图像块。**方法-1:** 与 FashionBERT [21] 或者 ViT [34]

表 4: **基于 3 种重要因素的消融实验**。更多信息参见 § 4.4。

指标	Kaleido 图像块生成器 (KPG)			预对齐掩码策略 (AGM)		预对齐 Kaleido 图像块模型 (AKPM)						
	Scale-fixed	Kaleido.	Kaleido.+SOD	Random	AGM	B	B+I	B+I∼II	B+I∼III	B+I∼IV	B+I∼V	B+V
1. Rank@1 ↑	24.71	26.73(+8.2%)	27.99(+13.3%)	26.55	27.99(+5.4%)	25.37	25.07(-1.2%)	26.03(+2.6%)	26.88(+6.0%)	26.20(+3.3%)	27.99(+10.3%)	24.62(-2.9%)
1. Rank@5 ↑	50.05	54.55(+9.0%)	60.09(+20.1%)	55.13	60.09(+8.9%)	54.97	55.14(+0.3%)	56.31(+2.4%)	58.34(+6.1%)	59.13(+7.6%)	60.09(+9.3%)	53.78(-2.2%)
1. Rank@10 ↑	58.93	65.44(+11.0%)	68.37(+16.0%)	64.92	68.37(+5.3%)	62.13	62.90(+1.2%)	63.37(+2.0%)	67.79(+9.1%)	67.99(+9.4%)	68.37(+10.0%)	60.88(-2.0)%
2. Rank@1 ↑	30.17	32.19(+6.7%)	33.88(+12.0%)	32.14	33.88(+5.4%)	31.09	30.98(-0.4%)	32.22(+3.6%)	33.17(+6.7%)	33.80(+8.7%)	33.88(+9.0%)	30.77(-1.0%)
2. Rank@5 ↑	52.29	58.40(+11.7%)	60.60(+15.9%)	56.99	60.60(+6.3%)	57.35	57.44(+0.2%)	58.73(+2.4%)	58.55(+2.1%)	60.57(+5.6%)	60.60(+5.7%)	55.95(-2.4%)
2. Rank@10 ↑	60.82	66.49(+9.3%)	68.59(+12.8%)	63.77	68.59(+7.6%)	64.79	65.65(+1.3%)	64.16(-1.0%)	67.92(+4.8%)	68.41(+5.6%)	68.09(+5.1%)	61.70(-4.8%)
Sum $\mathcal{R}$ ↑	**276.97**	**303.80(+9.7%)**	**319.52(+16.2%)**	**299.50**	**319.52(+6.7%)**	**295.70**	**297.18(+0.5%)**	**300.82(+1.7%)**	**312.65(+5.7%)**	**316.10(+6.9%)**	**319.02(+7.9%)**	**287.70(-2.7%)**
3. ACC ↑	93.44%	93.45%(+0.0%)	95.07%(+1.7%)	92.71%	95.07%(+2.5%)	90.94%	90.82%(-0.1%)	91.40%(+0.5%)	93.91%(+3.3%)	94.05%(+3.4%)	95.07%(+4.5%)	88.87(-2.3%)
3. macro-$\mathcal{F}$ ↑	0.701	0.705(+0.6%)	0.714(+1.9%)	0.711	0.714(+0.4%)	0.690	0.692(+0.3%)	0.721(+4.5%)	0.713(+3.3%)	0.710(+2.9%)	0.714(+3.5%)	0.701(+1.4%)
4. ACC ↑	86.89%	87.61%(+0.8%)	88.07%(+1.4%)	87.20%	88.07(+1.0%)	81.66%	81.25%(-0.5%)	84.44%(+3.4%)	86.49%(+5.9%)	88.53%(+8.4%)	88.07%(+7.9%)	81.64(+0.0%)
4. macro-$\mathcal{F}$ ↑	0.630	0.634(+0.6%)	0.636(+1.0%)	0.633	0.636(+0.5%)	0.558	0.575(+3.0%)	0.596(+6.8%)	0.636(+14.0%)	0.633(+13.4%)	0.636(+14.0%)	0.596(+8.4%)
Sum $\mathcal{CLS}$ ↑	**313.43**	**314.96(+0.5%)**	**318.14(+1.5%)**	**314.31**	**318.14(+1.2%)**	**297.40**	**298.77(+0.4%)**	**307.54(+3.4%)**	**315.30(+6.0%)**	**316.88(+6.5%)**	**318.14(+7.0%)**	**300.21(+0.9%)**
5. Bleu-4 ↑	4.9	5.2(+6.1%)	5.7(+16.3%)	5.3	5.7(+7.5%)	4.9	5.2(+6.1%)	5.2(+6.1%)	5.1(+4.1%)	5.6(+14.3%)	5.7(+16.3%)	5.3(+8.2%)
5. METEOR ↑	11.0	11.7(+6.4%)	12.8(+16.4%)	11.3	12.8(+13.3%)	11.6	11.6(+0.0%)	11.8(+1.7%)	12.6(+8.6%)	12.8(+10.3%)	12.8(+10.3%)	11.4(-1.7%)
5. ROUGE-L ↑	29.8	31.5(+5.7%)	32.9(+10.4%)	30.3	32.9(+8.6%)	30.4	30.7(+1.0%)	30.8(+1.3%)	31.9(+4.9%)	32.7(+7.6%)	32.9(+8.2%)	30.6(+0.7%)
5. CIDEr ↑	30.9	31.3(+1.3%)	32.6(+5.5%)	31.7	32.6(+2.8%)	31.0	31.5(+1.6%)	31.4(+1.3%)	32.0(+3.2%)	32.3(+4.2%)	32.6(+5.2%)	31.3(+1.0%)
Sum $\mathcal{CAP}$ ↑	**76.6**	**79.7(+4.0%)**	**84.0(+9.7%)**	**78.6**	**84.0(+6.9%)**	**77.9**	**79.0(+1.4%)**	**79.2(+1.7%)**	**81.6(+4.7%)**	**83.4(+7.1%)**	**84.0(+7.8%)**	**78.6(+0.9%)**

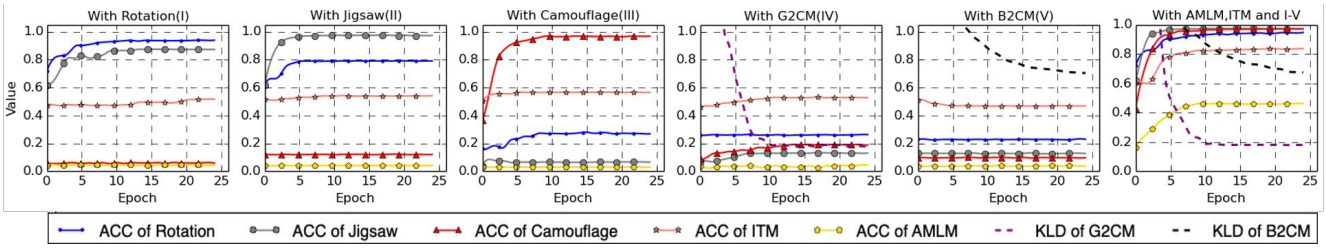

图 7: **单任务分析**。 ACC = Accuracy, KLD = Kullback-Leibler Divergence. 更多信息参见 § 4.4。

类似，第一种方法是直接将时尚图片切分成固定尺寸。用这样的图像块来训练，本文模型取得了 276.97 Sum $\mathcal{R}$，313.43 Sum $\mathcal{CLS}$ 以及 76.6 Sum $\mathcal{CAP}$ 的分数。**方法-2:** 本文模型实施了 Kaleido 图像块生成方式 (尺度变化)，相较于方法-1，在每项指标取得了 +9.7%、+0.5% 和 4.0% 的相对提升。也就是说，与方法-1 相比，该方法可以更好地捕捉细粒度表征。**方法-3:** 进一步引入显著性检测 (SOD) 的方法，避免出现所切分的块出现大量空白 (tabula rasa)。可以观察到，这样的方式相比于方法-1 取得了 16.2%、+1.5% 和 +9.7% 的相对提升。

AGM. 当前最主流的掩码策略，都是以一定选中概率在图像和文本侧，无关联的随机选取掩码候选。这样的掩码策略称为随机掩码。在实验中，本文对比了预对齐掩码策略 (AGM) 和随机掩码 (Random)。AGM 取得了 +6.7%、+1.2% 与 6.9% 的相对提升。这不足为奇，因为相对于随机掩码，AGM 考虑了更多语义关联的掩码，这有利于 Kaleido-BERT 更好地理解多模态信息。

AKPM. 为验证所提出的 AKPM 性能，本文实施了 7 个消融实验 (见图. 7)。基线 (简称 B) 仅仅包含常规的图文匹配任务 (ITM) 和掩码语言模型 (AMLM)。然后逐步增加 5 个 AKPM 子任务到模型的预训练过程中。例如，在图中:“B + I ∼ IV” 等同于 “B + I + II + III + IV”。值得一提的是，现有模型 [21] 通常使用 “ITM + AMLM + B2CM” (B + V) 的组合进行预训练学习。在表. 4中，可以看到这样的学习方式提升很有限，仅在 Sum $\mathcal{CLS}$ 提升了 +0.9%，甚至在 Sum$\mathcal{R}$ 产生副作用 (-2.7%)。有趣的是，简单地将 V (B2CM) 替换成 I (RR)，可以在所有下游任务上取得 +0.5%, +0.4% 与 +1.4% 的提升。逐渐地，当依次加入不同的子任务时，可以观察到性能持续上升。在这个过程中，V 的副作用被削弱，本文认为: 通过加入 I∼IV 学习后，Kaleido-BERT 全面地理解了图像特征，使得 V 可以真正发挥价值。

5. 总结与展望

本文在时尚领域提出了一个新颖的视觉-语言预训练模型，叫做 **Kaleido-BERT**。它包含类 Kaleido 图像块生成器，注意力对齐生成器以及预对齐掩码策略。这些组件紧密关联且易于实施，能学习到模态内与模态间的图-文关联特征。相比于现有模型，本文设计的 Kaleido-BERT 更为高效，并取得了卓越的性能，在下游任务诸如图像-文本检索、类目预测以及时尚描述上取得了长足进步。

参考文献

[1] Ziad Al-Halah and Kristen Grauman. From paris to berlin: Discovering fashion style influences around the world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10136–10145, 2020.

[2] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*. Springer, 2020.

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[9] 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.

[10] Meng-Jiun Chiou, Roger Zimmermann, and Jiashi Feng. Rvl-bert: Visual relationship detection with visual-linguistic knowledge from pre-trained representations. *arXiv preprint arXiv:2009.04965*, 2020.

[11] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Weller Adrian. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[12] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, pages 326–335, 2017.

[13] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*, 2020.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[15] Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020.

[16] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[17] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, pages 186–202, 2018.

[18] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4548–4557, 2017.

[19] Deng-Ping Fan, ShengChuan Zhang, Yu-Huan Wu, Yun Liu, Ming-Ming Cheng, Bo Ren, Paul L Rosin, and Rongrong Ji. Scoot: A perceptual metric for facial sketches. In *ICCV*, pages 5612–5622, 2019.

[20] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NIPS*, 2020.

[21] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *SIGIR*, pages 2251–2260, 2020.

[22] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, pages 6391–6400, 2019.

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017.

[24] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *CVPR*, pages 13137–13146, 2020.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[26] Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, and Kristen Grauman. Fashion++: Minimal edits for outfit improvement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5047–5056, 2019.

[27] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[28] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[29] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *PAMI*, 2020.

[30] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[31] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014.

[32] Aisha Urooj Khan, Amir Mazaheri, Niels da Vitoria Lobo, and Mubarak Shah. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. *arXiv preprint arXiv:2010.14095*, 2020.

[33] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[34] Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16-16 words: Transformers for image recognition at scale. In *arXiv preprint arXiv:2010.11929*, 2020.

[35] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[36] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018.

[37] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[38] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020.

[39] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

[40] 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.

[41] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *ACL*, pages 5265–5275, 2020.

[42] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137. Springer, 2020.

[43] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. Interbert: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint arXiv:2003.13198*, 2020.

[44] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869. Springer, 2016.

[45] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NIPS*, pages 13–23, 2019.

[46] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, pages 10437–10446, 2020.

[47] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. Univilm: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[48] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. *arXiv preprint arXiv:2004.14973*, 2020.

[49] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019.

[50] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *ECCV*, 2019.

[51] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pages 69–84. Springer, 2016.

[52] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[53] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, pages 1143–1151, 2011.

[54] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015.

[55] 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.

[56] Xuebin Qin, Deng-Ping Fan, Chenyang Huang, Cyril Diagne, Zichen Zhang, Adrià Cabeza Sant'Anna, Albert Suàrez, Martin Jagersand, and Ling Shao. Boundary-aware segmentation network for mobile and web applications. *arXiv preprint arXiv:2101.04704*, 2021.

[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[58] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018.

[59] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018.

[60] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020.

[61] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020.

[62] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019.

[63] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[64] 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.

[65] Reuben Tan, Mariya I Vasileva, Kate Saenko, and Bryan A Plummer. Learning similarity conditions without explicit supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10373–10382, 2019.

[66] Trieu H Trinh, Minh-Thang Luong, and Quoc V Le. Selfie: Self-supervised pretraining for image embedding. *arXiv preprint arXiv:1906.02940*, 2019.

[67] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 390–405, 2018.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[69] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[70] Yue Wang, Shafiq Joty, Michael R Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. Vd-bert: A unified vision and dialog transformer with bert. In *EMNLP*, 2020.

[71] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. *arXiv preprint arXiv:1907.09748*, 2019.

[72] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[73] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou. Xgpt: Cross-modal generative pre-training for image captioning. *arXiv preprint arXiv:2003.01473*, 2020.

[74] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[75] Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. *arXiv preprint arXiv:2008.02693*, 2020.

[76] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *NIPS*, pages 5753–5763, 2019.

[77] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016.

[78] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020.

[79] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019.

[80] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*, 2019.

[81] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, pages 8779–8788, 2019.

[82] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, pages 13041–13049, 2020.

[83] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, pages 8746–8755, 2020.

[84] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *arXiv preprint arXiv:2101.07663*, 2021.

[85] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *CVPR*, 2021.