SLAN：自助定位辅助视觉-语言理解网络

翟江天 [1*] Qi Zhang[2*] Tong Wu[2] Xing-Yu Chen[2] 刘姜江 [1†] 程明明 [1†]
[1]VCIP, CS, 南开大学 [2] 腾讯优图实验室
{jtzhai30,j04.liu}@gmail.com, townswu@tencent.com, cmm@nankai.edu.cn

Abstract

学习细粒度的视觉与语言之间的相互作用有助于更准确地理解视觉-语言任务。然而，根据文本提取关键图像区域进行语义对齐仍然具有挑战性。大多数现有的方法要么受限于使用固定的区域生成模块获取与文本无关并且冗余的区域，要么由于在预训练检测器时过于依赖稀缺的（黄金）定位数据而无法进一步扩展。为了解决这些问题，我们提出了一种用于视觉-语言理解任务并无需任何额外黄金数据的自定位辅助网络（*SLAN*）。*SLAN* 包括一个区域过滤器和一个区域适配器，以根据不同的文本定位感兴趣的区域。通过聚合视觉-语言信息，区域过滤器选择关键区域，区域适配器根据文本指导更新其坐标。通过详细地区域-词语对齐，*SLAN* 可以轻松推广到许多下游任务。它在五个视觉-语言理解任务上取得了相当有竞争力的结果（例如，在 *COCO* 图像-文本和文本-图像的检索任务上分别达到了 *85.7%*和 *69.2%*，超越了先前的 *SOTA* 方法）。*SLAN* 还展示了在两个定位任务中的强大的零样本和微调的可迁移性。代码可在以下网址找到：*https://github.com/scok30/SLAN*。

1. 引言

近年来，研究者们对探索视觉和语言模态之间的关系越来越感兴趣。其快速发展推动了许多应用的发展，例如多模态搜索引擎 [3, 7, 12] 和推荐系统 [6, 34, 35]。这激发了研究人员寻找两种模态之间的语义对应关系，并弥合它们之间的视觉-语义差异。一些早期的工作 [14, 16, 24, 31] 侧重于学习两种模态的联合嵌入，而更近期的工作 [17, 25, 46, 47] 已经转向考虑在区域和词语级别上的潜在视觉-语言对齐。

(a) 图像-文本检索 (b) 图像描述

(c) 物体检测 (d) 短语定位

图 1. 四个不同任务的可视化。我们在（a）中可视化了文本到图像检索任务的激活图。对于（b）中的图像描述任务，我们可视化了模型选择的区域。除了视觉-语言理解任务外，SLAN 可以转移到定位任务，如（c）和（d）所示，并列出了每个区域的置信度分数。

为了实现细粒度的视觉-语言对齐，一些研究 [20, 21, 26] 使用目标检测器从图像中提取关键区域。然而，这些检测器仅作为黑匣子处理，仅支持固定词汇的目标检测。同时，由于检测器参数的冻结，提取的区域无法根据不同的文本信息进行适应。为了缓解这个问题，VinVL [46] 使用一个包含超过

*表示贡献相同，这项工作是翟江天和刘姜江在腾讯优图所做的工作。
†刘姜江和程明明是通讯作者

2000 个类别和属性的预训练目标检测器，以丰富局部视觉表示。然而，相比现实场景中的自由文本，扩展的标签集仍然会限制目标检测器在视觉-语言理解中的感知能力。

最近，更多的研究开始尝试使用可学习的区域定位器进行视觉-语言任务，这些定位器根据不同的文本条件提取感兴趣的区域。与先前使用冻结物体检测器的方法不同，MDETR [17] 在带有区域到词语标注的数据集上构建了一个端到端的框架。GLIP [25] 直接提出了基于语言-图像的预训练，用于学习对象级、语言感知和语义丰富的视觉表示。这些方法通过引入可训练的定位器，展示了在视觉-语言推理方面的有效性。然而，为了监督定位器的训练，这些方法需要一定数量的区域到单词的锚定标注（黄金数据），这些数据基于繁重和昂贵的注释工作。这限制了它们在现有大规模视觉-语言数据集上的应用，这些数据集具有丰富但粗粒度的图像和文本对。

为了解决上述问题，我们提出了用于视觉-语言理解的自定位辅助网络（SLAN）。设计的自定位器能够根据不同的文本准确地定位感兴趣的区域。具体来说，自定位器由区域过滤器和区域适配器组成，以选择重要的区域并通过文本指导更新区域的坐标。通过将自定位器融入我们的框架中，SLAN 执行上下文感知的区域提取和视觉-语言特征融合。此外，SLAN 仅在带有成对图像和文本的数据集上进行训练，使其能够扩展到更大的预训练设置，以进一步提高性能。通过细粒度的区域-词汇对齐，SLAN 对于视觉和语言模态之间的交互具有更详细的理解。

综上所述，我们的贡献可以从三个方面总结：

- 我们提出了一个名为 SLAN 的框架，以捕捉视觉和语言模态之间更精细的相互作用。引入了自定位器来进行文本引导的区域适应，为视觉-语言理解任务实现了动态的区域-词汇对齐，如 Fig. 1 所示。

- 我们证明 SLAN 可以轻松应用于视觉-语言数据集的大规模预训练，无需使用黄金数据进行训练。由于其定位图像中关键区域的能力，SLAN 也可以自然地推广到典型的定位任务，如物体检测和短语定位。

- 在五个视觉-语言理解任务和两个定位任务上的实验证明了我们方法的有效性。例如，SLAN 在 COCO 图像-文本检索任务上实现了最先进的性能。

2. 相关工作

2.1. 视觉-语言任务

先前的研究已经探索了视觉和文本模态之间的关系，并将这些知识应用于各种下游多模态任务。诸如 DeViSE [13]、TBNN [36] 和 [48] 等方法提出了损失函数和网络结构来学习语义视觉-语言对齐。其他方法，如 SGG [41] 和 ViSTA [8]，则利用先前的工具或知识进行图像-文本匹配分析。

最近，借助视觉主干网络 [11, 15, 40] 和语言编码器 [18] 在更大规模的数据集上进行视觉-语言预训练变得越来越受欢迎。CLIP [31] 使用来自网络上的 4 亿图像-文本对进行预训练，建立图像和文本之间的全局关系。BLIP [23] 借助丰富的网络数据进行视觉-语言理解和生成任务。Beit-3 [37] 在大规模单模态和多模态数据上采用掩码-然后预测的自监督训练方式学习内部的视觉-语言依赖关系。

然而，这些方法受到细粒度区域-词语数据集成本的限制，在预训练期间难以直接提供局部的区域-词语匹配关系来获得更准确的跨模态知识。这种知识使模型能够根据相应的词汇精确定位对象，为下游任务提供线索。

2.2. 视觉-语言任务的定位

图像区域和句子中的词汇的定位有助于模型学习局部对齐。基于用于视觉-语言任务区域生成模块（region proposal module）是否冻结，存在两种方法类型。

第一种方法使用在 Visual Genomes 上预训练的冻结物体检测器（如 Faster R-CNN）来提取详细的视觉表示。一些后续的工作（如 VinVL [46]、Oscar [26]）增加了检测标签的数量，并引入了属性信息来补充视觉概念。

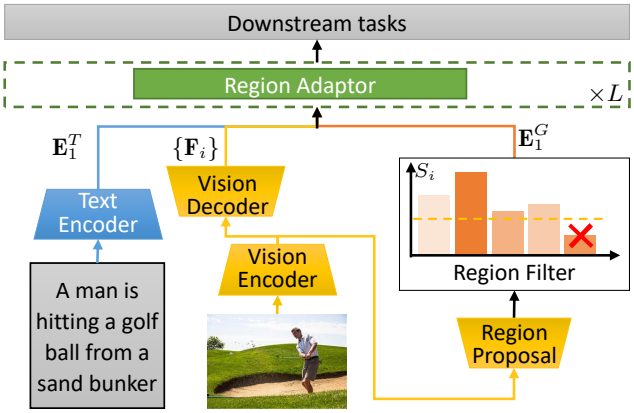
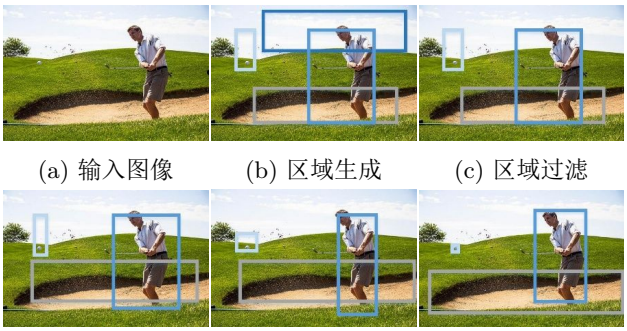


图 2. SLAN 框架。两个单模态编码器分别提取文本和视觉表示。自定位器自动生成、过滤图像区域，并通过迭代适应实现细粒度的区域-词汇对齐。所学习的视觉和语言特征可用于下游任务。



(a) 输入图像 (b) 区域生成 (c) 区域过滤

(d) 第 1、2 和 3 个区域适应器后的结果

图 3. 自定位器的样本中间结果。

另一种方法依赖于视觉-语言数据集的细粒度注释进行预训练。MDETR [17] 引入了一个带有多模态数据集的调制检测器，该数据集在文本中的词语与图像中的对象之间具有精确的对齐关系。GLIP [25] 将基于锚定的预训练应用于学习对象级、语言感知和语义丰富的视觉表示。然而，这些方法需要具有细粒度注释的视觉-语言数据，限制了它们在更大规模的预训练设置上的应用。

3. 自定位辅助网络（SLAN）

SLAN 的框架如 Fig. 2所示。我们首先简要介绍两个单模态编码器，然后再介绍其他组件的详细结构。SLAN 通过文本引导自适应地提出和选择信息丰富的区域，如 Fig. 3所描述。最后，我们列出我们的预训练目标。相关的符号在 Tab. 1中进行了说明。

3.1. 单模态编码

两个单模态编码器以 D 维度学习文本和视觉表示。对于文本特征提取，我们使用 BERT [18] 作为我们的文本编码器，将单词编码为共享的语义空间。编码后的嵌入 $\mathbf{E}^T \in \mathbb{R}^{N^T \times D}$ 概括了整个句子，包括来自 BERT 分类标记的文本标记 $\mathbf{T}_t \in \mathbb{R}^D$ 和 $N^T - 1$ 个词嵌入。

对于图像特征提取，我们使用经典的视觉主干网络（例如 ResNet50 [15]、ViT-Base [11]、ViT-Large 和 ViT-Huge）对图像进行编码，获得具有高级语义的视觉特征图 $\mathbf{V}$。

3.2. 用于视觉-语言理解的自定位器

由于细粒度的区域-词语对齐对于视觉-语言关系的探索非常重要，我们的自定位器遵循区域生成网络 [32] 的方法来输出区域，其中每个区域 i 包含空间坐标 (x, y, w, h) 和相应的区域嵌入 $\mathbf{E}_i^G \in \mathbb{R}^D$。通过使用 RoIAlign 从 $\mathbf{V}$ 中提取与文本相关的局部特征 $\mathbf{E}_i^G$。然后，通过对 $\mathbf{V}$ 进行全局平均池化，获得一个视觉标记 $\mathbf{T}_v$，作为该图像的全局摘要信息。

与大多数使用预定义的标签集的传统目标检测任务不同，视觉-语言任务通常具有更广泛的词汇表和自由形式的文本表达。因此，我们的自定位器引入了区域筛选器用于区域重要性预测，并引入了区域适配器用于渐进式区域回归。通过用区域重要性预测代替固定词汇表的预测，我们的自定位器为每个区域分配一个显著性分数 S_i，以估计该区域在对齐过程中的有效性概率。

对于传统的检测设置，回归目标是区域坐标。由于在我们的设置中没有人工标注的定位信息，我们在多阶段区域适配器中提出了渐进式区域回归，产生每个级别中的中间更新区域。然后，这些更新的区域用于监督内部的区域生成模块。如 Fig. 3所示，SLAN 在 $L = 3$ 个级别动态地调整区域嵌入，从而比全局视觉特征图或来自视觉 Transformer 的 patch 嵌入产生更加灵活和准确的视觉表示。

符号	维度	含义
D	1×1	token 的维度
L	1×1	层数/阶段数
K	1×1	每轴的网格数
N_i^h, N_i^w	1×1	邻域网格尺寸
S_i	1×1	区域的显著性得分
p_{w_i}, p_{h_i}	1×1	缩放参数
$\mathbf{E}^T$	$N^T \times D$	文本嵌入
$\mathbf{E}^G$	$N^G \times D$	区域嵌入
$\mathbf{F}_i$	$H_i \times W_i \times D$	金字塔特征图
$\mathbf{G}_i$	$N^G \times 4$	区域坐标
$\mathbf{T}_v, \mathbf{T}_t$	$1 \times D$	全局视觉/文本 token
$\mathbf{A}_i$	$N^G \times N^T$	交叉注意力图

表 1. 符号的维度和含义。

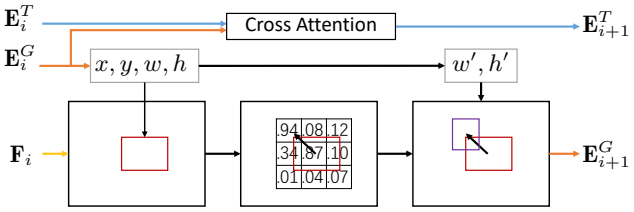


图 4. 区域适配器的第 i 阶段。区域适配器使用文本指导更新每个区域的坐标。我们使用来自视觉解码器的特征图提取区域嵌入，并探索潜在的区域-词语对齐。

3.2.1 视觉解码器：金字塔特征提取

我们提出的自定位器采用了粗到精的方式进行回归设计，需要多尺度的视觉特征。考虑到这些特点，我们在全局视觉特征之后采用了一个视觉解码器，以提取多尺度的特征图 $\mathbf{F}_i$，其中 $i \in 1, 2, ..., L$。$\mathbf{F}_i$ 表示解码器特征的第 i 层，$L = 3$ 是自定位器默认的层数。然后，$\mathbf{F}_i$ 被送入第 i 层的区域适配器。视觉编码器和解码器的结构遵循特征金字塔网络 [27]。

3.2.2 区域过滤器：区域重要性预测

在描述图像时，人们通常会集中关注图像中有限的显著区域 [9,10]。然而，区域生成模块 [32] 通常会为一张图像输出大量的区域（例如 100 个）。直接选择所有区域将导致不必要的计算成本，并可能使模型从一些无意义的区域到词语的配对中进行学习。控制最大选定区域数量的策略有三个步骤。(a) 对所有区域的显著性分数进行归一化。经过此过程，得分表示为 $S = S_1, ..., S_k$，其中 $S_i \in [0, 1]$。(b) 根据它们的显著性分数按降序对这些区域进行排序。(c) 我们选择不超过前 T 个显著性分数高于阈值 h 的区域。最后，我们通过这些分数加权区域嵌入。每个区域的显著性分数会根据下游视觉-语言监督的梯度进行更新，这将在 Sec. 3.3中描述。

3.2.3 区域适配器：渐进式区域回归

区域适配器的目标是调整生成区域的坐标，使其与具有相同语义的词语对齐。它的困难之处在于没有标注的文本参考区域作为引导。我们将这个问题转化为一个 L 层级的级联粗到精的渐进回归过程，默认情况下 $L = 3$。如 Fig. 4所示，区域回归过程的第 i 层级接收三个输入：词嵌入 $\mathbf{E}_i^T \in \mathbb{R}^{N^T \times D}$，区域嵌入 $\mathbf{E}_i^G \in \mathbb{R}^{N^G \times D}$ 以及它们的坐标 $\mathbf{G}_i \in \mathbb{R}^{N^G \times 4}$，以及全局解码器特征图 $\mathbf{F}_i \in \mathbb{R}^{H_i \times W_i \times D}$，其中 N^T 和 N^G 分别表示词语和选择的区域数量。D 表示嵌入的维度。

渐进式区域回归的详细过程在算法 1中描述。视觉-语言的多头注意力层融合了区域和词嵌入，并对它们的交互建模如下：

$$\begin{aligned} \mathbf{A}_i &= \frac{\mathbf{E}_i^G \mathbf{E}_i^{T\top}}{\sqrt{D}}, \\ \mathbf{E}_{i+1}^G &= Softmax(\mathbf{A}_i)\mathbf{E}_i^T, \\ \mathbf{E}_{i+1}^T &= Softmax(\mathbf{A}_i^\top)\mathbf{E}_i^G. \end{aligned} \tag{1}$$

通过视觉-语言的语义对齐，更新后的视觉感知词嵌入 $\mathbf{E}_i^T$ 能够通过在原始区域周围搜索高度相关的区域来指导区域坐标的更新。具体而言，区域 $g = (x, y, w, h)$ 的邻域被定义为以它为中心，大小为 (N_i^h, N_i^w) 的区域，其中 N_i^h 和 N_i^w 是第 i 层级区域回归过程的预定义参数。这个邻域被分割成 $K \times K$ 个区域，以计算区域-词语相似性。如 Fig. 4所示，通过 RoIAlign 提取每个区域嵌入，然后从 F_i 进行平均池化。

由于对不同词语有不同的响应分数，邻域区域将上下文信息聚合到中心区域。中心区域的坐标更新以

Algorithm 1 自定位

输入: 图像 I，区域嵌入 $\mathbf{E}_i^G$，文本嵌入 $\mathbf{E}_i^T$，金字塔特征图 $\mathbf{F}_i$，邻域尺寸 (N_i^h, N_i^w)，总区域回归层数 L。

p_{w_i}, p_{h_i} 是每个层级中每个区域独立的可学习参数。

输出: 更新后的区域坐标 $\mathbf{G}_{out}$，对区域生成模块的区域监督坐标 $\overline{\mathbf{G}}$，视觉 token $\mathbf{T}_v$，文本 token $\mathbf{T}_t$。

$\mathbf{G}_1, \mathbf{E}_1^G \leftarrow$ 区域生成（I）

$\mathbf{G}_1, S, \mathbf{E}_1^G \leftarrow$ 区域重要性预测（$\mathbf{G}_1, \mathbf{E}_1^G$）

for $i \in \{1, 2, ..., L\}$ **do**

  $\mathbf{E}_{i+1}^G, \mathbf{E}_{i+1}^T \leftarrow$ 交叉注意力（$\mathbf{E}_i^G, \mathbf{E}_i^T$）

  $\mathbf{E}_i^N \leftarrow$ 邻域嵌入（$N_i^h, N_i^w, \mathbf{G}_i$）

  $\Delta x_i, \Delta y_i \leftarrow$ 偏移量（相似度（$\mathbf{E}_i^N, \mathbf{E}_{i+1}^T$））

  $\mathbf{G}_{i+1} \leftarrow$ 更新（$\mathbf{G}_i, \Delta x_i, \Delta y_i, p_{w_i}, p_{h_i}$）

  $\mathbf{E}_{i+1}^G \leftarrow$ 嵌入（$\mathbf{G}_{i+1}, \mathbf{F}_i$）

end for

$\mathbf{T}_v, \mathbf{T}_t \leftarrow$ 提取 CLS（$\mathbf{E}_{L+1}^G, \mathbf{E}_{L+1}^T$）

$\mathbf{G}_{out} \leftarrow \mathbf{G}_{L+1}$

$\overline{\mathbf{G}} \leftarrow (\sum_{i=2}^{L+1} \mathbf{G}_i)/L$

邻域中心点的坐标加权求和的形式进行，如 Equ. (2) 所示：

$$\begin{aligned}
\Delta x &= \sum_{j=0}^{K^2-1} M_j N_j^h (\lfloor \frac{j}{K} \rfloor - \lfloor \frac{K}{2} \rfloor), \\
\Delta y &= \sum_{j=0}^{K^2-1} M_j N_j^w (j \bmod K - \lfloor \frac{K}{2} \rfloor), \\
x' &= x + \Delta x, \quad y' = y + \Delta y, \\
w' &= p_w w, \quad h' = p_h h,
\end{aligned} \tag{2}$$

其中 $\lfloor \cdot \rfloor$ 是向下取整操作。区域适配器的所有层级中的每个区域都有其自己的 p_w 和 p_h，它们被设置为可学习的参数。M_j 是第 j 个邻域区域嵌入与所有词嵌入之间的最大余弦相似度。Equ. (2) 中第一、二行中的最后一个项的目的是将 1D 索引映射到 2D 索引（例如，从 $0, 1, .., 8$ 到 $(0,0), (0,1), ..., (2,2)$）。

对于每个原始区域 g，设 g_i 表示它在区域回归的第 i 层后的更新版本。我们取它们的平均作为实际坐标标签，并应用 L_1 和 GIoU 回归损失：

$$\begin{aligned}
\overline{g} &= \frac{\sum_{i=2}^{L+1} g_i}{L}, \\
\mathcal{L}_{reg}(g) &= \mathcal{L}_{L1}(g, \overline{g}) + \mathcal{L}_{GIoU}(g, \overline{g}).
\end{aligned} \tag{3}$$

3.3. 使用 SLAN 进行预训练目标

SLAN 在图像-文本对上进行预训练，通过三种常见损失的监督来学习精细的区域-词语对齐。

图像-文本匹配损失（ITM） 用于预测给定的图像-文本对是否为正样本，这可以看作是一个二元分类问题。将视觉和文本 token $(\mathbf{T}_v, \mathbf{T}_t)$ 连接并传递给线性层 f_c。ITM 损失定义如下：

$$\mathcal{L}_{itm}(\mathbf{I}, \mathbf{T}) = H(f_c(cat(\mathbf{T}_v, \mathbf{T}_t)), y_{v,t}), \tag{4}$$

其中 $y_{v,t}$ 表示匹配关系（1 表示匹配，0 表示不匹配），H 是用于分类的交叉熵损失。我们直接从数据集中选择正样本，并使用批次抽样构建困难负样本，遵循 ALBEF [24]。

图像-文本对比损失（ITC） 确保视觉和文本嵌入共享相同的语义空间，并且正（匹配）的图像-文本对比负（不匹配）的对更近。我们使用两个队列 I_q, T_q 来保存最新访问的图像和文本样本。对于每个图像-文本对 $(\mathbf{I}, \mathbf{T})$，计算 softmax 归一化的视觉-语言相似性如下：

$$\begin{aligned}
p_{i2t}(\mathbf{I}, \mathbf{T}, T_q) &= \frac{exp(sim(\mathbf{T}_v, \mathbf{T}_t)/\tau)}{\sum_{\mathbf{T}' \in T_q} exp(sim(\mathbf{T}_v, \mathbf{T}'_t)/\tau)}, \\
p_{t2i}(\mathbf{T}, \mathbf{I}, I_q) &= \frac{exp(sim(\mathbf{T}_t, \mathbf{T}_v)/\tau)}{\sum_{\mathbf{T}'_v \in I_q} exp(sim(\mathbf{T}_t, \mathbf{T}'_v)/\tau)},
\end{aligned} \tag{5}$$

其中 τ 是温度参数，$sim(\cdot)$ 衡量视觉-语言相似性，通过图像和文本嵌入之间的点积来实现。

依照 ALBEF [24] 的做法，我们计算 ITC 损失如下：

$$\mathcal{L}itc(\mathbf{I}, \mathbf{T}) = -\log(pi2t(\mathbf{I}, \mathbf{T}, T_q)) - \log(p_{t2i}(\mathbf{T}, \mathbf{I}, I_q)). \tag{6}$$

语言建模损失（LM） 鼓励模型使用上下文信息预测被屏蔽的词语。我们随机屏蔽 15% 的文本标记，并应用屏蔽的语言建模损失如下：

$$\mathcal{L}_{lm}(\mathbf{I},\mathbf{T}) = H(p_{mask}(\mathbf{T}_v,\mathbf{T}_t), y_{mask}), \tag{7}$$

其中 y_{mask} 表示被屏蔽的待预测词语，$p_{mask}(\mathbf{I},\mathbf{T})$ 是其预测概率。$\mathcal{L}_{ds}$ 是下游损失，由前述三个损失的总和计算得出：

$$\mathcal{L}_{ds}(\mathbf{I},\mathbf{T}) = \mathcal{L}_{itm}(\mathbf{I},\mathbf{T}) + \mathcal{L}_{itc}(\mathbf{I},\mathbf{T}) + \mathcal{L}_{lm}(\mathbf{I},\mathbf{T}). \tag{8}$$

完整的预训练目标是下游损失和对渐进式区域回归的约束的结合，计算如下：

$$\mathcal{L} = \mathcal{L}_{ds} + \mathcal{L}_{reg}. \tag{9}$$

$\mathcal{L}_{reg}$ 表示所有区域中 Equ. (3) 中回归损失的总和。模型在训练过程中受到 $\mathcal{L}$ 的监督。

4. 实验

首先，SLAN 在一个包含来自五个数据集的 1400 万个图像-文本对的组合数据集上进行预训练：COCO [28]，Visual Genome [19]（不包括 COCO 图像），Conceptual Captions [5]，Conceptual [5] 和 SBU Captions [29]。我们通过将其与其他最先进的跨模态方法在多个下游任务上进行比较来评估 SLAN 的性能。我们还进行了广泛的消融研究，以探究 SLAN 的每个组成部分如何影响性能。

4.1. 实现细节

我们选择 BERT_{base} [18] 作为我们的文本编码器，其初始化来自 HuggingFace [39]。对于视觉编码器，我们尝试了四种设计选择：一个基于 CNN 的模型（即 ResNet50）和三个基于 Transformer 的模型（即 ViT-Base、ViT-Large 和 ViT-Huge），它们都是随机初始化的。至于每个区域适配器的邻域大小，我们使用比率 r_i 来表示：$(N_i^h, N_i^w) = (r_i H_i, r_i W_i)$，其中 $r_1, r_2, r_3 = 1, 0.5, 0.25$。我们对 SLAN 进行了 20 个批次的预训练。对于不同的视觉编码器选择，批次大小分别设置为 1280、960、640、640，用于 ResNet50、ViT-Base、ViT-Large 和 ViT-Huge。

我们采用 AdamW 优化器，初始学习率为 3e-4，学习率线性衰减至 0。我们将输入图像调整大小为 224×224。

4.2. 下游任务比较

我们将 SLAN 与其他最先进的方法在五个具有挑战性的视觉-语言理解任务上进行了比较，包括图像-文本检索、图像描述、视觉问答、自然语言视觉推理、零样本视频-文本检索。我们还将 SLAN 推广到两个定位任务：目标检测和短语定位。默认的视觉编码器是 ViT-Huge，除非另有说明。

4.2.1 图像-文本检索

在给定图像的情况下，检索任务期望通过输入图像从文本库中检索出对应的文本，反之亦然。我们在 Flickr30k [30] 上进行了评估，分别在零样本和微调设置下使用了 Karpathy 分割，并以 Recall@k 作为性能评价指标。比较结果如 Tab. 2所示。

具体来说，在相同的预训练设置下，SLAN 在 COCO 数据集上的平均召回率 @1 上比 BLIP [23] 高出 3.3%。

4.2.2 图像描述

在给定输入图像的情况下，图像描述任务生成一个句子描述，以详细描述图像内容。我们使用 COCO 的 Karpathy 分割进行微调和评估。如 Tab. 3所示，在这个高效的设置下，SLAN 优于大多数现有方法。

4.2.3 视觉问答

视觉问答（Visual Question Answering，VQA）[1] 要求模型从图像-问题对中预测出答案。我们遵循 [23] 的方法，将 VQA 视为开放式的问题生成任务。我们将图像嵌入与问题嵌入融合，然后将它们发送到问题解码器以获取结果。如 Tab. 3所示，SLAN 在 VQAv2 的测试集（test-dev 和 test-std）上的表现优

方法	Backbone	预训练数据	零样本学习						微调					
			图像 → 文本			文本 → 图像			图像 → 文本			文本 → 图像		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ALIGN [16]	EfficientNet	1.8B	88.6	98.7	99.7	75.7	93.8	96.8	95.3	99.8	100.0	84.9	97.4	98.6
FILIP [43]	ViT-Large	300M	89.8	99.2	99.8	75.0	93.4	96.3	96.6	100.0	100.0	87.1	97.7	99.1
BLIP [23]	ViT-Large	14M	94.8	99.7	100.0	84.9	96.7	98.3	96.6	99.8	100.0	87.2	97.5	98.8
Beit-3 [37]	ViT-Giant	21M	94.9	99.9	100.0	81.5	95.6	97.8	98.0	100.0	100.0	**90.3**	98.7	99.5
我们的方法	ViT-Huge	14M	**96.0**	**100.0**	**100.0**	**86.1**	**97.0**	**98.5**	**98.1**	**100.0**	**100.0**	90.2	**99.0**	**99.6**

表 2. 与 Flickr30k 数据集上最先进的图像-文本检索方法的比较。我们在零样本学习和微调设置下使用召回率（Recall@k）作为评估指标。

方法	Backbone	预训练数据	检索 (COCO)		描述 (COCO)				VQA (VQAv2)		NLVR (NLVR2)	
			I2T R@1	T2I R@1	B@4	M	C	S	test-dev	test-std	dev	test-P
Oscar [26]	ResNet101	6.5M	73.5	57.5	37.4	30.7	127.8	23.5	73.6	73.8	79.1	80.3
VinVL [46]	ResNeXt152-C4	8.9M	75.4	58.8	38.5	30.4	130.8	23.4	76.5	76.6	82.6	83.9
SimVLM [38]	ViT-Huge	1.8B	-	-	40.6	33.7	143.3	25.4	80.0	80.3	84.5	85.1
GLIPv2-H [45]	Swin-Huge	16M	-	-	-	-	131.0	-	74.6	74.8	-	-
CoCa [44]	ViT-Giant	4.8B	-	-	40.9	33.9	143.6	24.7	82.3	82.3	86.1	87.0
BLIP [23]	ViT-Large	14M	82.4	65.1	40.4	-	136.7	-	78.2	78.3	82.1	82.2
Beit-3 [37]	ViT-Giant	21M	84.8	67.2	44.1	32.4	147.6	25.4	84.2	84.0	**91.5**	**92.5**
我们的方法	ViT-Huge	14M	**85.7**	**69.2**	**44.2**	**34.3**	**147.8**	**25.8**	**84.5**	**84.7**	91.0	91.7

表 3. 在更多下游任务上的比较。对于 COCO 检索，I2T 和 T2I 分别表示图像到文本和文本到图像检索任务。对于 COCO 图像描述，我们报告了在 Karpathy 测试集上的 BLEU@4 (B@4)，METEOR (M)，CIDEr (C) 和 SPICE (S) 分数。对于 VQA，我们在 VQAv2 的 test-dev 和 test-std 集上评估 vqa-score。对于 NLVR，我们报告了在 NLVR2 开发集（dev）和公共测试集（test-P）上的准确率。

方法	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
ClipBERT [22]	22.0	46.8	59.9	6
VideoCLIP [42]	30.9	55.4	66.8	-
FiT† [2]	43.3	65.6	74.7	2
BLIP† [23]	43.3	65.6	74.7	2
我们的方法 †	**46.8**	**70.5**	**83.6**	**1.5**

表 4. 在 MSRVTT 数据集的 1k 测试集上进行的文本-视频检索任务的比较。† 表示零-shot 设置，其他均为微调设置。

于 Beit-3，并且后者采用了更大的视觉主干网络，并需要更多的预训练数据。

4.2.4 自然语言视觉推理

自然语言视觉推理（Natural Language Visual Reasoning，NLVR2）[33] 用于衡量一个句子是否描述了一对图像。我们从图像-文本输入中提取图像和文本嵌入，然后通过交叉注意力层将它们融合。我们使用一个二元分类模块来预测它们之间的关系。SLAN 在 NLVR2 任务上超过了大部分现有方法，与 Beit-3 的性能相当，显示出学习细粒度的视觉语言对齐的重要性。

4.2.5 零样本视频-文本检索

除了上述提到的图像-文本任务外，SLAN 还可以推广到视频-文本检索任务。我们从视频输入中随

方法	Backbone	预训练数据 (百万)		目标检测 (COCO)		短语定位 (Flickr30k)		
		图像-文本	区域-词语	零样本	微调	R@1	R@5	R@10
DETR [4]$_{\text{ECCV'20}}$	ResNet50	0	0	-	42.0	-	-	-
MDETR [17]$_{\text{ICCV'21}}$	ResNet101	0	0.2	-	-	84.3	93.9	95.8
GLIP [25]$_{\text{CVPR'22}}$	Swin-Large	24	3	49.8	60.8	87.1	96.9	98.1
GLIPv2 [45]$_{\text{NeurIPS'22}}$	Swin-Huge	16	3	-	60.2	87.7	97.3	98.5
Beit-3 [37]$_{\text{CVPR'23}}$	ViT-Giant	21	0	-	**63.7**	-	-	-
我们的方法	ResNet50	14	0	46.9	59.2	86.8	96.6	97.4
	ViT-Base	14	0	47	59.6	87.4	96.9	98.2
	ViT-Large	14	0	48.5	60.5	89.1	98.0	98.9
	ViT-Huge	14	0	**50.1**	63.5	**90.6**	**98.6**	**99.3**

表 5. 两个定位任务的比较：COCO 上的目标检测和 Flickr30k 上的短语定位。预训练数据包括图像-文本对和词汇特定区域注释。我们在目标检测上评估零样本和微调设置。我们使用 Recall@k 分数评估短语定位任务。

可训练区域生成	适配器数量	COCO		Flickr30k	
		TR@1	IR@1	TR@1	IR@1
✘	0	68.5	53.5	85.0	74.1
✔	0	69.1	53.8	86.7	76.2
✔	1	70.0	57.2	88.3	77.4
✔	2	70.8	57.5	88.7	78.1
✔	3	**72.1**	**58.3**	**90.3**	**78.9**

表 6. 在 SLAN 的可训练区域生成模块和区域适配器上的消融实验。第一列中的 ✘ 表示应用冻结的区域生成模块并且没有使用自定位器。TR@1 和 IR@1 分别表示从图像到文本和从文本到图像的召回率 @1。为了评估自定位器对冻结的区域生成模块的影响，我们加载在 COCO 检测任务上预训练的权重，并与我们的方法进行比较（第 1 行对比第 2 行）。其余实验均从头开始训练。ViT-Base 作为视觉编码器。

Top K	Threshold	COCO		Flickr30k	
		TR@1	IR@1	TR@1	IR@1
-	-	69.4	54.1	85.9	74.7
10	-	70.6	56.8	87.5	77.3
10	0.3	71.2	57.6	89.1	78.2
10	0.5	**72.1**	**58.3**	**90.3**	**78.9**

表 7. 不同区域筛选设置的消融实验结果。

机选择 m 帧，并将它们连接起来以获得图像-文本序列，然后直接将其输入到我们的图像-文本检索模型中。如 Tab. 4所示，SLAN 在零样本视频-文本检索任务中实现了与其他方法相当的性能，表明 SLAN 中学习的视觉-语言知识是语义丰富的。

4.2.6 定位任务

我们在两个定位任务上进行了实验：COCO 上的目标检测和 Flickr30k 上的短语定位。在目标检测任务的文本输入中，我们使用一个由 COCO 的标签组成的提示（例如，“detect: 人，自行车，汽车，...，牙刷”）。我们采用区域适应器的最后一层的输出。如 Tab. 5所示，SLAN 在定位任务上表现出色。

例如，在以 ViT-Base 为主干的目标检测任务中，SLAN 的性能与需要更大主干和 300 万金数据的 GLIP 相当。尽管 SLAN 并非专为定位任务设计，但以 ViT-Huge 为主干的 SLAN 的性能超越了几乎所有对比方法。

4.3. 消融研究

4.3.1 自定位器的有效性

可学习区域生成模块的重要性。 如 Tab. 6所示，第 1 行表示用在 COCO 检测任务上预训练的冻结检测器替代自定位器，第 2 行是我们的可学习区域生成模块。我们没有使用预训练权重初始化区域生成模块，而只在下游任务的数据集上对其进行微调。我们的方法在 COCO 和 Flickr30k 的图像-文本和文本-图像检索任务上平均分别提升了约 0.5% 和 2%。

Method	Backbone	Params(M)	FLOPs(G)	COCO	
				TR@1	IR@1
BLIP	ViT-Base	370	558	81.9	64.3
BLIP	ViT-Large	810	1594	82.4	65.1
Coca	ViT-Giant	2100	4103	83.0	65.5
Beit-3	ViT-Giant	1900	-	84.8	67.2
Ours	ResNet50	322	324	**85.1**	**68.9**

表 8. 在视觉-语言检索任务中对参数数量和 FLOPs 的比较。FLOPs 是在输入图像分辨率为 384x384 的情况下计算的。“Backbone” 表示视觉编码器。

用于区域回归的区域适应器数量。 区域适应器对区域生成模块输出的区域进行逐步回归，以为视觉-语言理解任务提供更准确的区域定位。如 Tab. 6所示，当区域适应器的数量从 0 增加到 3 时，检索性能可以显著提高，平均提升超过 3%。

用于显著性预测的区域筛选器。 Tab. 7展示了区域筛选器对 COCO 和 Flickr30k 检索任务性能的影响。我们从零开始训练可学习的区域生成模块，并将区域适应器的数量设置为 3。前两行显示，当按照显著性得分对区域进行排序并仅选择一定数量（前 K 个）时，我们可以在每个数据集上获得约 2% 的性能提升。结合显著性得分阈值使用时，我们的区域筛选器能够移除对视觉-语言适应产生负面影响的冗余区域，并获得更高的性能。

4.3.2 计算成本

Tab. 8展示了 SLAN 和其他最先进方法在计算成本上的比较。可以看出，由于在这些实验中我们的视觉主干网络是相对轻量级的 ResNet50，SLAN 的参数数量和 FLOPs 最少。然而，我们的检索性能明显优于其他方法。我们相信上述现象表明了我们提出的 SLAN 方法的高效性和有效性。

4.4. 可视化分析

4.4.1 文本引导的区域适应

如 Fig. 5所示，我们的区域适应器生成了具有相对高置信度的文本特定结果。当我们更改句子的详细描述，例如将“穿红色外套的男人”改为“穿黑色裤子的男人”，有趣的现象是我们的自定位器的注意力区域也相应地以相对高的置信度进行了移动。

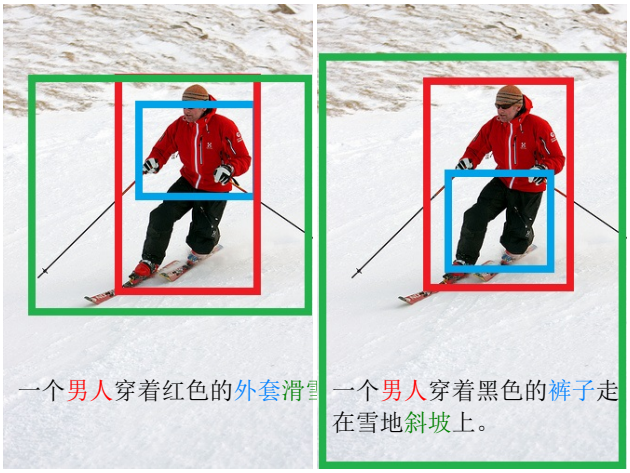


图 5. 文本特定区域适应的示意图。我们对每个句子着色三个词，并使用相应的颜色标记具有最高匹配分数的区域。这突出了 SLAN 的能力，即建议自适应的与文本相关的区域。

图 6. 区域适应的由粗到细的过程的示意图。我们还展示了区域与对应单词之间的匹配分数。请注意，在区域适应器的不同层级中，我们的实现中每个区域都具有独立的缩放和移动行为。

4.4.2 粗到精的区域适应

为了验证区域适应的校准效果，我们在 Fig. 6中展示了一张带有文本描述的图像。在经过三个级别的区域适应器后，模型会定位具有更高相似性分数的更准确的兴趣区域。这表明我们的自定位器能够分层次地细化与提供的单词对应的相关区域。

5. 结论与未来工作

在本文中，我们介绍了自定位辅助网络（SLAN），该网络利用自定位器来适应生成的区域，用于实现视觉-语言对齐，无需额外的区域对应于单词的标签。我们的目标是进一步研究并优化自定位器在各种定位应用中的性能。

致谢. 本研究得到了国家自然科学基金（项目编号：62225604、62176130）和南开大学基本科研业务费专项资金（项目编号：070-63233089）的支持。南开大学超级计算中心提供了计算资源。

参考文献

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Int. Conf. Comput. Vis.*, 2015. 6

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Int. Conf. Comput. Vis.*, 2021. 7

[3] Yue Cao, Mingsheng Long, Jianmin Wang, and Shichen Liu. Collective deep quantization for efficient cross-modal retrieval. In *AAAI*, 2017. 1

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, 2020. 8

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 6

[6] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1

[7] Kan Chen, Trung Bui, Chen Fang, Zhaowen Wang, and Ram Nevatia. Amc: Attention guided multimodal correlation learning for image search. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1

[8] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, et al. Vista: Vision and scene text aggregation for cross-modal retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[9] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(3):569–582, 2015. 4

[10] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995. 4

[11] Alexey Dosovitskiy, Lucas Beyer, 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 16x16 words: Transformers for image recognition at scale, 2020. 2, 3

[12] Benjamin Elizalde, Shuayb Zarar, and Bhiksha Raj. Cross modal audio search and retrieval with joint embeddings based on text and audio. In *ICASSP*, 2019. 1

[13] Andrea Frome, Gregory S Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 2

[14] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *NIPS*, 2020. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2, 3

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Int. Conf. Mach. Learn.*, 2021. 1, 7

[17] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multimodal understanding. In *Int. Conf. Comput. Vis.*, 2021. 1, 2, 3, 8

[18] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 2, 3, 6

[19] 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. *Int. J. Comput. Vis.*, 2017. 6

[20] Chia-Wen Kuo and Zsolt Kira. Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1

[21] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Eur. Conf. Comput. Vis.*, 2018. 1

[22] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 7

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Int. Conf. Mach. Learn.*, 2022. 2, 6, 7

[24] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Gotmare, Shafiq R Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NIPS*, 2021. 1, 5

[25] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 1, 2, 3, 8

[26] 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 *Eur. Conf. Comput. Vis.*, 2020. 1, 2, 7

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 4

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, 2014. 6

[29] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *NIPS*, 2011. 6

[30] 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 *Int. Conf. Comput. Vis.*, 2015. 6

[31] 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 *Int. Conf. Mach. Learn.*, 2021. 1, 2

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015. 3, 4

[33] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *ACL*, 2019. 7

[34] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. Multi-modal knowledge graphs for recommender systems. In *CIKM*, 2020. 1

[35] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1

[36] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018. 2

[37] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks.

In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2, 7, 8

[38] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *Int. Conf. Learn. Represent.*, 2021. 7

[39] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *EMNLP*, 2020. 6

[40] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. P2t: Pyramid pooling transformer for scene understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 2

[41] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[42] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *EMNLP*, 2021. 7

[43] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *Int. Conf. Learn. Represent.*, 2021. 7

[44] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 7

[45] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *NIPS*, 2022. 7, 8

[46] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2, 7

[47] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1

[48] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *TOMM*, 2020. 2