Alpha-CLIP: A CLIP Model Focusing on Wherever You Want（CVPR 2024）

Abstract：

这篇文章主要提出了一种加强对局部注意力的CLIP方法，成功做到了即插即用。

Conclusion：

SOTA。

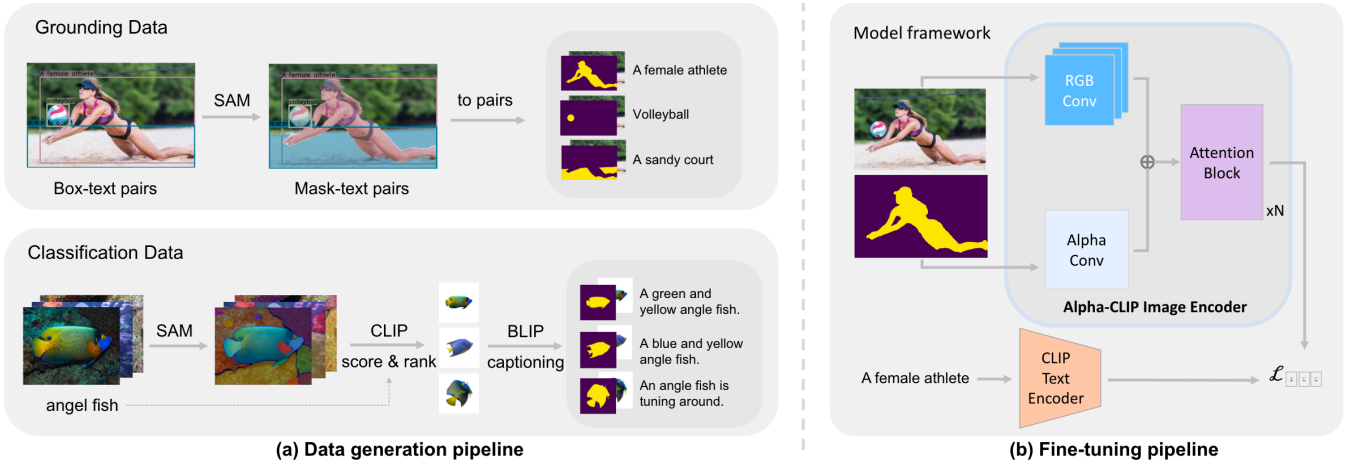
Introduction&Related Work：

为了做到以上目的，之前的工作尝试了以下方法

- 直接将无关部份mask掉或者crop掉。然而这样会省略很多语义信息。
- 使用attention或者辅助网络等机制帮助local。然而这样的结果并不好。
- 给感兴趣的物体画个红圈圈。然而这样会破坏原图的语义信息。

本文的Alpha-CLIP使用了另一种方法：引入额外的Alpha通道构建RGBA图像，这样可以在关注感兴趣物体的同时不丢失语义信息。简单来说就是用SAM找物体进行分割，然后用BLIP-2这种生成新标签，构成region-text pairs，最后与原图的image-text pairs合并。

Method：



(a) Data generation pipeline　　(b) Fine-tuning pipeline

Figure 3. **The pipeline of our data generation method and model architecture**. **(a)** Our method generates millions of RGBA-region text pairs. **(b)** Alpha-CLIP modifies the CLIP image encoder to take an additional alpha channel along with RGB.

(1)构建GRBA数据集

- Grounding data pipeline：该分支专用于生成区域文本对，原始图片来自于GRIT数据集，用GLIP和CLIP给box region打标签，作者再采用SAM给每个box region弄出来伪掩码
- Classification data pipeline：图片来自于imagenet，首先，采用SAM提取出前景，将背景丢掉，再将前景对齐中央，放大。然后使用 CLIP 计算每个mask所属图像的相应类标签的分数。接下来，根据分数按类别对面具进行排序，并

选择分数最高的排名靠前的mask。关于text部份，为了确保每个掩码的标题不仅仅是 ImageNet类标签，需要将前景对象放置在纯白色背景上。然后我们使用 BLIP-2 来注释这些掩码。最后，将ImageNet 类标签与 BLIP-2生成的图像特定标题合并，从而产生数百万个 RGBA 区域文本对。

（2）框架

作者引入了一个与 RGB Conv 层平行的附加 Alpha Conv 层，这使得 CLIP 图像编码器能够接受额外的 Alpha 通道作为输入。 Alpha 通道输入设置为 [0, 1] 范围，其中 1 表示前景，0 表示背景。将 Alpha Conv 内核权重初始化为零，确保初始 Alpha-CLIP 忽略 alpha 通道作为输入。

在训练过程中，保持 CLIP 文本编码器固定，完全训练 Alpha-CLIP 图像编码器。与处理 alpha 通道输入的第一个卷积层相比，作者对后续的 Transformer 块应用较低的学习率。为了保持 CLIP 对完整图像的全局识别能力，作者在训练过程中采用了特定的数据采样策略，设置样本比率，表示为 rs = 0.1，偶尔用原始图像-文本对替换生成的 RGBA-文本对，并将 alpha 通道设置为全 1（？？？）。经过训练后，Alpha-CLIP具备了聚焦指定区域并进行受控编辑的能力。 Alpha-CLIP 可以以即插即用的方式增强 CLIP 在各种基线上的性能，涵盖各种下游任务，如识别、MLLM 和 2D/3D 生成。

局限性

虽然 Alpha-CLIP 在需要区域聚焦的各种场景中展示了有效的性能，但其当前的结构和训练过程限制了其关注多个对象、或不同对象之间的模型关系的能力。此外，当前的训练方法限制 alpha 通道泛化到除了 0 和 1 的二进制值之外的中间值。（？？？）因此，用户无法指定注意力的幅度。Alpha-CLIP 和原始 CLIP 的另一个限制是分辨率低，这阻碍了 Alpha-CLIP 识别小物体的方式。可以在未来的工作中解决这些限制并扩展 CLIP 输入分辨率。