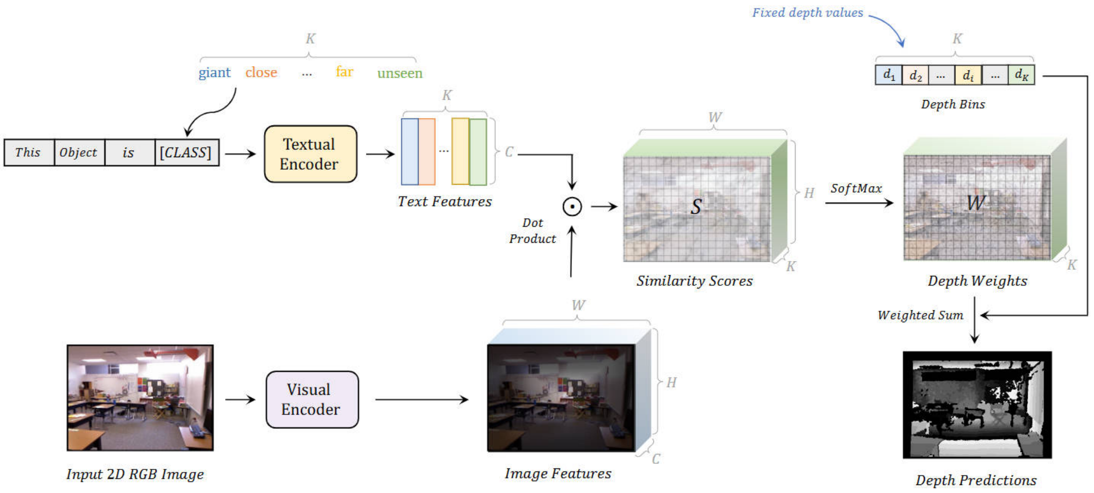
Can Language Understand Depth?

动机：

将CLIP迁移到monocular depth estimation（单目深度评估）领域中

但CLIP对概念和背景的不敏感（在本文limitation中提到的），而对前景物体十分敏感

Architecture：



技巧：convert the depth value representation to a distance classification task

手动设计了Depth bins对应着[CLASS]，在本是：['giant', 'extremely close', 'close', 'not in distance', 'a little remote', 'far', 'unseen']对应着[1.00, 1.50, 2.00, 2.25, 2.50, 2.75, 3.00]