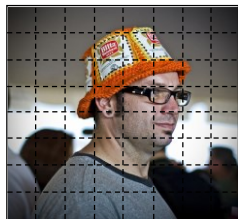


(a)提取名词短语



$n \times n$

文本语言
编码器

图片编码器

特征融合及
目标识别

$n \times n$

$\{v_2$ ,

v_5 : 

(b)提取视觉目标