

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(10) 国际公布号

WO 2023/201990 A1

(43) 国际公布日

2023 年 10 月 26 日 (26.10.2023)

WIPO | PCT

(51) 国际专利分类号:

G06T 5/00 (2006.01)

G06T 7/70 (2017.01)

G06T 9/00 (2006.01)

区吴中经济开发区郭巷街道官浦路 1 号 9 幢, Jiangsu 215000 (CN)。

(21) 国际申请号:

PCT/CN2022/122335

(22) 国际申请日:

2022 年 9 月 28 日 (28.09.2022)

(25) 申请语言:

中文

(26) 公布语言:

中文

(30) 优先权:

202210407177.8

2022 年 4 月 19 日 (19.04.2022) CN

(71) 申请人: 苏州浪潮智能科技有限公司

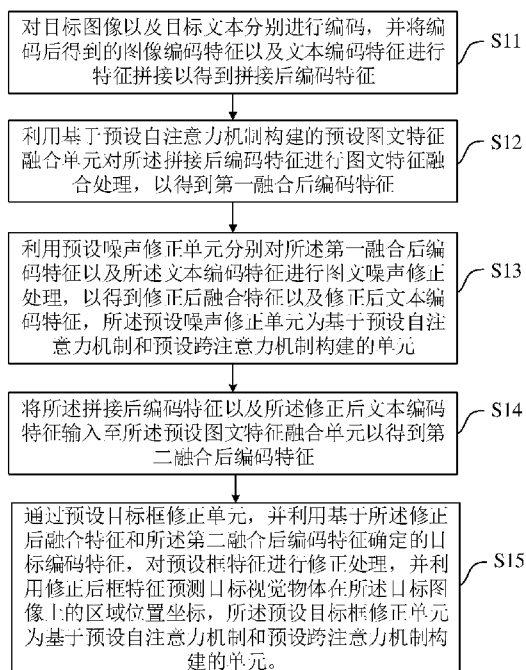
(INSPUR SUZHOU INTELLIGENT TECHNOLOGY

CO., LTD.) [CN/CN]; 中国江苏省苏州市吴中

(72) 发明人: 李晓川(LI, Xiaochuan); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路 1 号 9 幢, Jiangsu 215000 (CN)。李仁刚(LI, Rengang); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路 1 号 9 幢, Jiangsu 215000 (CN)。赵雅倩(ZHAO, Yaqian); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路 1 号 9 幢, Jiangsu 215000 (CN)。郭振华(GUO, Zhenhua); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路 1 号 9 幢, Jiangsu 215000 (CN)。范宝余(FAN, Baoyu); 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路 1 号 9 幢, Jiangsu 215000 (CN)。

(54) Title: VISUAL POSITIONING METHOD AND APPARATUS, DEVICE, AND MEDIUM

(54) 发明名称: 一种视觉定位方法、装置、设备及介质



S11 Respectively code a target image and a target text, and perform feature splicing on an image coded feature and a text coded feature obtained after coding to obtain spliced coded features

S12 Perform image-text feature fusion processing on the spliced coded features by using a preset image-text feature fusion unit constructed on the basis of a preset attention mechanism to obtain a first fused coded feature

S13 Respectively perform image-text noise correction processing on the first fused coded feature and the text coded feature by using a preset noise correction unit to obtain a corrected fused feature and a corrected text coded feature, the preset noise correction unit being a unit constructed on the basis of a preset self-attention mechanism and a preset cross-attention mechanism

S14 Input the spliced coded features and the corrected text coded feature into the preset image-text feature fusion unit to obtain a second fused coded feature

S15 By means of a preset target frame correction unit, correct a preset frame feature by using a target coded feature determined on the basis of the corrected fused feature and the second fused coded feature, and predict a regional position coordinate of a target visual object on the target image by using the corrected frame feature, the preset target frame correction unit being a unit constructed on the basis of the preset self-attention mechanism and the preset cross-attention mechanism

图 1

(57) Abstract: The present application relates to the technical field of artificial intelligence, and discloses a visual positioning method and apparatus, a device, and a medium. The method comprises: performing feature splicing on an image coded feature and a text coded feature; performing feature fusion on the spliced coded features to obtain a first fused coded feature; respectively performing noise correction on the first fused coded feature and the text coded feature on the basis of a preset cross-attention mechanism to obtain a corrected fused feature and a corrected text coded feature, and performing feature fusion on the spliced coded feature and the corrected

(74) 代理人: 北京市万慧达律师事务所 (BEIJING WANHUIDA LAW FIRM); 中国北京市海淀区中关村南大街 1 号友谊宾馆颐园写字楼 2 楼, Beijing 100873 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

text coded feature to obtain a second fused coded feature; and correcting a preset frame feature by using a target coded feature determined on the basis of the corrected fused feature and the second fused coded feature to predict a regional position coordinate of the target visual object. Hence, according to the present application, image-text noise is corrected on the basis of the preset cross-attention mechanism, the impact of noise is weakened by reducing the attention on the noise part in the text, and anti-noise visual positioning is achieved.

(57) 摘要: 本申请公开了一种视觉定位方法、装置、设备及介质, 涉及人工智能技术领域, 该方法包括: 对图像编码特征与文本编码特征进行特征拼接; 对拼接后编码特征进行特征融合, 得到第一融合后编码特征; 基于预设跨注意力机制分别对第一融合后编码特征与文本编码特征进行噪声修正, 得到修正后融合特征与修正后文本编码特征, 对拼接后编码特征与修正后文本编码特征进行特征融合得到第二融合后编码特征; 利用基于修正后融合特征和第二融合后编码特征确定的目标编码特征对预设框特征进行修正, 以预测目标视觉物体的区域位置坐标, 可见, 本申请基于预设跨注意力机制对图文噪声进行修正, 通过降低对文本中噪声部分的关注度削弱了噪声的影响, 实现抗噪视觉定位。

一种视觉定位方法、装置、设备及介质

相关申请的交叉引用

本申请要求于 2022 年 04 月 19 日提交中国专利局，申请号为 202210407177.8，申请名称为“一种视觉定位方法、装置、设备及介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及人工智能技术领域，特别涉及一种视觉定位方法、装置、设备及介质。

背景技术

近年来，多模态（Multi Modal，MM）成为人工智能领域非常重要的研究方向。由于其重视视觉和文本、语音等信息的融合，多模态相关的各种算法也层出不穷：基于卷积神经网络（Convolutional Neural Networks，CNN）、注意力机制（attention）的各种方法均各有广泛的适用面，均在诸如视觉常识推理（Visual Commonsense Reasoning，VCR）、视觉问答（Visual Question Answering，VQA）、视觉定位（Visual Grounding，VG）等领域成为主流方法。

视觉定位任务是多模态人工智能领域重要的研究方向之一，该任务旨在根据描述定位出图片中的相关物体并给出该物体的坐标位置。然而，现有视觉定位任务在落地过程中仍然存在一些问题，而这些问题在“实验室”研究中是极易被忽略的，比如“文本”中出现的错误，也称为噪声。文本错误是由“人”的原因导致的文本失真。通常情况下，人的口误、人在描述物体时产生的主观性偏差、描述语句的歧义等原因均会导致文本的错误，这些错误在日常生活中非常普遍，但在 AI 算法设计过程中很容易被忽略，这就成为现有方法和落地之间的障碍。简言之，当输入的文本存在某些错误的情况下，现有方法很难找到这句话本身想描述的物体并定位。

为此，避免因人类语言文本错误而产生的噪声对视觉定位的影响，实现抗噪视觉定位是本领域亟待解决的问题。

发明内容

本申请提供一种视觉定位方法，包括：

对目标图像以及目标文本分别进行编码，并将编码后得到的图像编码特征以及文本编码特征进行特征拼接以得到拼接后编码特征；

利用基于预设自注意力机制构建的预设图文特征融合单元对拼接后编码特征进行图文特征融合处理，以得到第一融合后编码特征；

利用预设噪声修正单元分别对第一融合后编码特征以及文本编码特征进行图文噪声修正处理，以得到修正后融合特征以及修正后文本编码特征，预设噪声修正单元为基于预设自注意力机制和预设跨注意力机制构建的单元；

将拼接后编码特征以及修正后文本编码特征输入至预设图文特征融合单元以得到第二融合后编码特征；和

通过预设目标框修正单元，并利用基于修正后融合特征和第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理，并利用修正后框特征预测目标视觉物体在目标图像上的区域位置坐标，预设目标框修正单元为基于预设自注意力机制和预设跨注意力机制构建的单元。

在一个或多个实施例中，利用基于预设自注意力机制构建的预设图文特征融合单元对拼接后编码特征进行图文特征融合处理之前，还包括：

利用基于预设自注意力机制构建的自注意力运算单元、层归一化单元、特征删除单元以及特征相加单元构建图文特征融合子单元；

通过将第一预设数量的图文特征融合子单元进行依次串接，以构建得到预设图文特征融合单元；

相应的，利用基于预设自注意力机制构建的预设图文特征融合单元对拼接后编码特征进行图文特征融

合处理，以得到第一融合后编码特征，包括：

将预设图文特征融合单元中的第一个图文特征融合子单元作为当前图文特征融合子单元，并将拼接后编码特征作为待处理特征；

将待处理特征输入至当前图文特征融合子单元中；

利用当前图文特征融合子单元对待处理特征依次进行自注意力运算、层归一化运算、特征删除运算以及特征相加运算，以得到相应的当前运算处理结果；和

根据当前运算处理结果获得第一融合后编码特征。

在一个或多个实施例中，根据当前运算处理结果获得第一融合后编码特征，包括：

判断当前图文特征融合子单元是否为最后一个；

响应于当前图文特征融合子单元不为最后一个，将当前图文特征融合子单元更新为下一个图文特征融合子单元，将待处理特征更新为当前运算处理结果，并返回执行将待处理特征输入至当前图文特征融合子单元中的步骤；和

响应于当前图文特征融合子单元为最后一个，将当前运算处理结果作为第一融合后编码特征。

在一个或多个实施例中，利用当前图文特征融合子单元对待处理特征依次进行自注意力运算、层归一化运算、特征删除运算以及特征相加运算，以得到相应的当前运算处理结果，包括：

利用当前图文特征融合子单元中的自注意力运算单元对待处理特征进行自注意力运算，得到第一运算特征；

利用当前图文特征融合子单元中的层归一化单元对第一运算特征进行层归一化处理，得到第二运算特征；

利用当前图文特征融合子单元中的特征删除单元并按照预设比例对第二运算特征进行特征删除运算，以得到第三运算特征；和

利用当前图文特征融合子单元中的特征相加单元对第三运算特征与待处理特征进行特征相加运算，以得到当前图文特征融合子单元中的运算处理结果。

在一个或多个实施例中，利用预设噪声修正单元分别对第一融合后编码特征以及文本编码特征进行图文噪声修正处理之前，还包括：

利用基于预设自注意力机制构建的自注意力运算单元、特征删除单元、层归一化单元、特征相加单元构建第一噪声修正子单元；

利用基于预设跨注意力机制构建的跨注意力运算单元、特征删除单元、层归一化单元、特征相加单元构建第二噪声修正子单元；和

通过将第一噪声修正子单元以及第二预设数量的第二噪声修正子单元进行依次串接，以构建得到预设噪声修正单元。

在一个或多个实施例中，利用预设噪声修正单元分别对第一融合后编码特征以及文本编码特征进行图文噪声修正处理，以得到修正后融合特征以及修正后文本编码特征，包括：

将第一融合后编码特征以及文本编码特征输入至预设噪声修正单元中的第一噪声修正子单元，以便对第一融合后编码特征以及文本编码特征均分别进行自注意力运算、特征删除运算、层归一化运算以及特征相加运算，以得到第一融合后编码特征以及文本编码特征各自对应的第一运算处理结果；

将预设噪声修正单元中的第一个第二噪声修正子单元作为当前第二噪声修正子单元，并将第一融合后编码特征以及文本编码特征各自对应的第一运算处理结果均作为当前的待处理特征；

将待处理特征输入至当前第二噪声修正子单元中；

利用当前第二噪声修正子单元对待处理特征依次进行跨注意力运算、特征删除运算、层归一化运算以及特征相加运算，以得到第一融合后编码特征以及文本编码特征各自对应的当前第二运算处理结果；和

根据当前第二运算处理结果获得修正后文本编码特征。

在一个或多个实施例中，根据当前第二运算处理结果获得修正后文本编码特征，包括：

判断当前第二噪声修正子单元是否为最后一个；

响应于当前第二噪声修正子单元不为最后一个，将当前第二噪声修正子单元更新为下一个第二噪声修正子单元，将待处理特征更新为当前第二运算处理结果，并返回执行将待处理特征输入至当前第二噪声修正子单元；

正子单元中的步骤；和

响应于当前第二噪声修正子单元为最后一个，将第一融合后编码特征以及文本编码特征各自对应的当前第二运算处理结果分别作为修正后融合特征以及修正后文本编码特征。

在一个或多个实施例中，通过预设目标框修正单元，并利用基于修正后融合特征和第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理之前，还包括：

利用基于预设自注意力机制构建的自注意力运算单元、特征删除单元、层归一化单元、特征相加单元构建第一目标框修正子单元；

利用基于预设跨注意力机制构建的跨注意力运算单元、特征删除单元、层归一化单元、特征相加单元构建第二目标框修正子单元；和

通过将第一目标框修正子单元以及第三预设数量的第二目标框修正子单元进行依次串接，以构建得到预设目标框修正单元。

在一个或多个实施例中，通过预设目标框修正单元，并利用基于修正后融合特征和第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理，包括：

将目标编码特征以及预设框特征输入至预设目标框修正单元中的第一目标框修正子单元，以便对目标编码特征以及预设框特征均分别进行自注意力运算、特征删除运算、层归一化运算以及特征相加运算，以得到目标编码特征以及预设框特征各自对应的第三运算处理结果；

将预设目标框修正单元中的第一个第二目标框修正子单元作为当前第二目标框修正子单元，并将目标编码特征以及预设框特征各自对应的第三运算处理结果均作为当前的待处理特征；

将待处理特征输入至当前第二目标框修正子单元中；

利用当前第二目标框修正子单元对待处理特征依次进行跨注意力运算、特征删除运算、层归一化运算以及特征相加运算，以得到目标编码特征以及预设框特征各自对应的当前第四运算处理结果；和

根据当前第四运算处理结果获得修正后框特征。

在一个或多个实施例中，根据当前第四运算处理结果获得修正后框特征，包括：

判断当前第二目标框修正子单元是否为最后一个；

响应于当前第二目标框修正子单元不为最后一个，将当前第二目标框修正子单元更新为下一个第二目标框修正子单元，将待处理特征更新为当前第四运算处理结果，并返回执行将待处理特征输入至当前第二目标框修正子单元中的步骤；和

响应于当前第二目标框修正子单元为最后一个，将当前第四运算处理结果作为修正后框特征。

在一个或多个实施例中，利用基于修正后融合特征和第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理之前，还包括：

将修正后融合特征或第二融合后编码特征或修正后融合特征与第二融合后编码特征进行预设运算处理后得到的特征确定为目标编码特征；其中，预设运算处理包括修正后融合特征与第二融合后编码特征进行特征相加或特征拼接。

在一个或多个实施例中，特征相加的公式为：

$$f_{cat} = f_{modify} + f_{denoise} ;$$

其中， f_{modify} 为修正融合特征， $f_{denoise}$ 为降噪融合特征， f_{cat} 为特征相加后的输出。

在一个或多个实施例中，特征拼接的公式为：

$$f_{cat} = [f_{modify}; f_{denoise}] ;$$

其中， f_{modify} 为修正融合特征， $f_{denoise}$ 为降噪融合特征， f_{cat} 为特征拼接后的输出。

在一个或多个实施例中，利用修正后框特征预测目标视觉物体在目标图像上的区域位置坐标，包括：

将修正后框特征输入至基于第一全连接层和第二全连接层构建的坐标预测器；第一全连接层为用于预测初始目标框的置信度的全连接层，第二全连接层为用于对初始目标框进行坐标回归处理的全连接层；

利用坐标预测器以及修正后框特征，确定每个初始目标框的置信度；和

根据每个初始目标框的置信度以及每个初始目标框的坐标确定目标视觉物体在目标图像上的区域位置坐标。

在一个或多个实施例中，根据每个初始目标框的置信度以及每个初始目标框的坐标确定目标视觉物体在目标图像上的区域位置坐标，包括：

对置信度进行降序排序，然后从降序排序结果中筛选置信度最高的初始目标框，并将筛选出的初始目标框的坐标确定为目标视觉物体在目标图像上的区域位置坐标。

在一个或多个实施例中，一种视觉定位方法还包括：

基于第一预设公式创建预设自注意力机制；

其中，第一预设公式为：

$$attn_{self}(f) = \text{soft max}\left(\frac{(W_q \cdot f)^T \times (W_k \cdot f)}{\sqrt{\text{size}(f)}}\right) \times (W_v \cdot f) ;$$

其中， f 为每个预设自注意力机制的输入， W_q 、 W_k 以及 W_v 表示映射矩阵， $\text{size}(f)$ 表示维度， $attn_{self}(f)$ 表示预设自注意力机制的输出，softmax 表示激活函数。

在一个或多个实施例中，一种视觉定位方法还包括：

基于第二预设公式创建预设跨注意力机制；

其中，第二预设公式为：

$$attn_{cross}(f, g) = \text{soft max}\left(\frac{(W_q \cdot f)^T \times (W_k \cdot g)}{\sqrt{\text{size}(g)}}\right) \times (W_v \cdot g) ;$$

其中， f ， g 分别表示预设跨注意力机制中跨注意力层中每次参与跨注意力运算的两个输入特征， $\text{size}(g)$ 表示维度， $attn_{cross}(f, g)$ 表示预设跨注意力机制的输出，softmax 表示激活函数。

本申请还提供一种视觉定位装置，包括：

特征拼接模块，用于对目标图像以及目标文本分别进行编码，并将编码后得到的图像编码特征以及文本编码特征进行特征拼接以得到拼接后编码特征；

第一特征融合模块，用于利用基于预设自注意力机制构建的预设图文特征融合单元对拼接后编码特征进行图文特征融合处理，以得到第一融合后编码特征；

噪声修正模块，用于利用预设噪声修正单元分别对第一融合后编码特征以及文本编码特征进行图文噪声修正处理，以得到修正后融合特征以及修正后文本编码特征，预设噪声修正单元基于预设自注意力机制和预设跨注意力机制构建而成；

第二特征融合模块，用于将拼接后编码特征以及修正后文本编码特征输入至预设图文特征融合单元以得到第二融合后编码特征；和

目标框修正单元，用于通过预设目标框修正单元，并利用基于修正后融合特征和第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理，并利用修正后框特征预测目标视觉物体在目标图像上的区域位置坐标，预设目标框修正单元为基于预设自注意力机制和预设跨注意力机制构建的单元。

本申请还提供一种电子设备，包括存储器及一个或多个处理器，存储器中储存有计算机可读指令，计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行上述任一项一种视觉定位方法的步

骤。

本申请还提供一个或多个存储有计算机可读指令的非易失性计算机可读存储介质，计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行上述任一项一种视觉定位方法的步骤。

附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据提供的附图获得其他的附图。

图1为本申请一个或多个实施例公开的一种视觉定位方法流程图；

图2为本申请一个或多个实施例公开的一种具体的视觉定位方法子流程图；

图3为本申请一个或多个实施例公开的一种具体的视觉定位方法子流程图；

图4为一种传统的视觉定位任务结构示意图；

图5为本申请一个或多个实施例公开的一种抗噪视觉定位任务结构示意图；

图6为一种传统的视觉定位方法流程图；

图7为本申请一个或多个实施例公开的一种具体的视觉定位方法流程图；

图8为本申请一个或多个实施例公开的一种预设图文特征融合单元结构示意图；

图9为本申请一个或多个实施例公开的一种预设噪声修正单元结构示意图；

图10为本申请一个或多个实施例公开的一种预设目标框修正单元结构示意图；

图11为本申请一个或多个实施例公开的一种坐标预测器结构示意图；

图12为本申请一个或多个实施例公开的一种视觉定位装置结构示意图；

图13为本申请一个或多个实施例公开的一种电子设备结构图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

在视觉定位任务中，当输入的文本存在某些错误的情况下，现有方法很难找到这句话本身想描述的物体并定位。

为此，本申请实施例提出一种视觉定位方案，能够避免因人类语言文本错误而产生的噪声对视觉定位的影响，实现抗噪视觉定位。

本申请实施例公开了一种视觉定位方法，参见图1所示，该方法包括：

步骤S11：对目标图像以及目标文本分别进行编码，并将编码后得到的图像编码特征以及文本编码特征进行特征拼接以得到拼接后编码特征。

本实施例中，对目标图像以及目标文本进行编码的编码器可以采用经典模型，例如：对目标图像进行编码的图像编码器可以采用卷积神经网络ResNet (Residual Neural Network)、ResNext等，对目标文本进行编码的图像编码器可以采用 Roberta、Bert (Bidirectional Encoder Representation from Transformers) 等。

本实施例中，将编码后得到的图像编码特征以及文本编码特征进行特征拼接以得到拼接后编码特征，如此一来，可以将所述图像编码特征以及所述文本编码特征作为一个整体输入至下一处理单元。

步骤S12：利用基于预设自注意力机制构建的预设图文特征融合单元对所述拼接后编码特征进行图文特征融合处理，以得到第一融合后编码特征。

本实施例中，在利用基于预设自注意力机制构建的预设图文特征融合单元对所述拼接后编码特征进行图文特征融合处理之前，需要利用基于预设自注意力机制构建的自注意力运算单元、层归一化单元、特征删除单元以及特征相加单元构建图文特征融合子单元；通过将第一预设数量的所述图文特征融合子单元进行依次串接，以构建得到所述预设图文特征融合单元，所述特征删除单元用来按一定比例对特征进行随机删除，如此一来，可以防止系统过拟合。

本实施例中，在构建得到所述预设图文特征融合单元之后，将所述预设图文特征融合单元中的第一个图文特征融合子单元作为当前图文特征融合子单元，并将所述拼接后编码特征作为待处理特征；将所述待处理特征输入至当前图文特征融合子单元中；利用当前图文特征融合子单元对所述待处理特征依次进行自注意力运算、层归一化运算、特征删除运算以及特征相加运算，以得到相应的当前运算处理结果；判断当前图文特征融合子单元是否为最后一个；若否，则将当前图文特征融合子单元更新为下一个图文特征融合子单元，将所述待处理特征更新为当前所述运算处理结果，并返回执行所述将所述待处理特征输入至当前图文特征融合子单元中的步骤；若是，则将当前所述运算处理结果作为所述第一融合后编码特征。

需要指出的是，利用当前图文特征融合子单元对所述待处理特征依次进行自注意力运算、层归一化运算、特征删除运算以及特征相加运算，以得到相应的运算处理结果的过程具体包括：利用当前图文特征融合子单元中的所述自注意力运算单元对所述待处理特征进行自注意力运算，得到第一运算特征；利用当前图文特征融合子单元中的所述层归一化单元对所述第一运算特征进行层归一化处理，得到第二运算特征；利用当前图文特征融合子单元中的所述特征删除单元并按照预设比例对所述第二运算特征进行特征删除运算，以得到第三运算特征；利用当前图文特征融合子单元中的所述特征相加单元对所述第三运算特征与所述待处理特征进行特征相加运算，以得到当前图文特征融合子单元中的所述运算处理结果。

需要指出的是，本实施例在利用基于预设自注意力机制构建的预设图文特征融合单元对所述拼接后编码特征进行第一次图文特征融合处理时，侧重于描述图文匹配关系，也即，基于所述图文匹配关系推测目标图像与目标文本之间不匹配的部分，也即噪声部分。

步骤S13：利用预设噪声修正单元分别对所述第一融合后编码特征以及所述文本编码特征进行图文噪声修正处理，以得到修正后融合特征以及修正后文本编码特征，所述预设噪声修正单元为基于预设自注意力机制和预设跨注意力机制构建的单元。

需要指出的是，相较于传统的视觉定位方法，本实施例新增了基于预设自注意力机制和预设跨注意力机制构建的预设噪声修正单元，由此可以基于预设跨注意力机制对所述第一融合后编码特征以及所述文本编码特征进行图文噪声修正处理，以降低对文本中噪声部分的关注度，如此一来，削弱了噪声的影响，实现了抗噪视觉定位。本实施例中，利用预设噪声修正单元分别对所述第一融合后编码特征以及所述文本编码特征进行图文噪声修正处理之前，还包括：利用基于预设自注意力机制构建的自注意力运算单元、特征删除单元、层归一化单元、特征相加单元构建第一噪声修正子单元；利用基于预设跨注意力机制构建的跨注意力运算单元、特征删除单元、层归一化单元、特征相加单元构建第二噪声修正子单元；通过将所述第一噪声修正子单元以及第二预设数量的所述第二噪声修正子单元进行依次串接，以构建得到所述预设噪声修正单元。

步骤S14：将所述拼接后编码特征以及所述修正后文本编码特征输入至所述预设图文特征融合单元以得到第二融合后编码特征。

本实施例中，在利用基于预设自注意力机制构建的预设图文特征融合单元对所述拼接后编码特征以及所述修正后文本编码特征进行第二次图文特征融合处理时，侧重于融合图文特征，由前述公开的内容可知，在第一次图文特征融合时，确定了目标图像与目标文本之间不匹配的部分，也即噪声部分，因此本实施例在已知噪声部分的前提下侧重于融合图文特征。

步骤S15：通过预设目标框修正单元，并利用基于所述修正后融合特征和所述第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理，并利用修正后框特征预测目标视觉物体在所述目标图像上的区域位置坐标，所述预设目标框修正单元为基于预设自注意力机制和预设跨注意力机制构建的单元。

本实施例中，在对预设框特征进行修正处理之前，需要将所述修正后融合特征或所述第二融合后编码特征或所述修正后融合特征与所述第二融合后编码特征进行预设运算处理后得到的特征确定为所述目标编码特征；其中，所述预设运算处理包括所述修正后融合特征与所述第二融合后编码特征进行特征相加或特征拼接，如此一来，可以利用所述目标编码特征对预设框特征进行修正处理。

本实施例中，利用修正后框特征预测目标视觉物体在所述目标图像上的区域位置坐标具体包括：将所述修正后框特征输入至基于第一全连接层和第二全连接层构建的坐标预测器；所述第一全连接层为用于预测初始目标框的置信度的全连接层，第二全连接层为用于对所述初始目标框进行坐标回归处理的全连接层；利用所述坐标预测器以及所述修正后框特征，确定每个所述初始目标框的置信度；对所述置信度进行

降序排序，然后从降序排序结果中筛选置信度最高的所述初始目标框，并将筛选出的所述初始目标框的坐标确定为目标视觉物体在所述目标图像上的区域位置坐标。

本实施例中，在通过预设目标框修正单元，并利用基于所述修正后融合特征和所述第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理之前，还包括：利用基于预设自注意力机制构建的自注意力运算单元、特征删除单元、层归一化单元、特征相加单元构建第一目标框修正子单元；利用基于预设跨注意力机制构建的跨注意力运算单元、特征删除单元、层归一化单元、特征相加单元构建第二目标框修正子单元；通过将所述第一目标框修正子单元以及第三预设数量的所述第二目标框修正子单元进行依次串接，以构建得到所述预设目标框修正单元。

可见，本申请公开了一种视觉定位方法，包括：对目标图像以及目标文本分别进行编码，并将编码后得到的图像编码特征以及文本编码特征进行特征拼接以得到拼接后编码特征；利用基于预设自注意力机制构建的预设图文特征融合单元对所述拼接后编码特征进行图文特征融合处理，以得到第一融合后编码特征；利用预设噪声修正单元分别对所述第一融合后编码特征以及所述文本编码特征进行图文噪声修正处理，以得到修正后融合特征以及修正后文本编码特征，所述预设噪声修正单元为基于预设自注意力机制和预设跨注意力机制构建的单元；将所述拼接后编码特征以及所述修正后文本编码特征输入至所述预设图文特征融合单元以得到第二融合后编码特征；通过预设目标框修正单元，并利用基于所述修正后融合特征和所述第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理，并利用修正后框特征预测目标视觉物体在所述目标图像上的区域位置坐标，所述预设目标框修正单元为基于预设自注意力机制和预设跨注意力机制构建的单元，可见，本申请是通过基于预设自注意力机制和预设跨注意力机制构建的噪声修正单元来进行图文噪声修正处理的，由于在基于跨注意力机制进行处理的过程中，文本相对于图像的差异在图像中无法找到匹配关系，从而降低了对图文噪声成分的关注度，由此削弱了图文噪声对视觉定位准确度的影响，也即实现了抗噪视觉定位。

进一步的，本实施例针对前述实施例步骤 S13 中的利用预设噪声修正单元分别对第一融合后编码特征以及文本编码特征进行图文噪声修正处理的过程，进行详细的介绍和说明。具体的，参见图 2 所示，上述步骤 S13 包括：

步骤 S131：将所述第一融合后编码特征以及所述文本编码特征输入至所述预设噪声修正单元中的所述第一噪声修正子单元，以便对所述第一融合后编码特征以及所述文本编码特征均分别进行自注意力运算、特征删除运算、层归一化运算以及特征相加运算，以得到所述第一融合后编码特征以及所述文本编码特征各自对应的第一运算处理结果。

步骤 S132：将所述预设噪声修正单元中的第一个第二噪声修正子单元作为当前第二噪声修正子单元，并将所述第一融合后编码特征以及所述文本编码特征各自对应的第一运算处理结果均作为当前的待处理特征。

步骤 S133：将所述待处理特征输入至当前第二噪声修正子单元中。

步骤 S134：利用当前第二噪声修正子单元对所述待处理特征依次进行跨注意力运算、特征删除运算、层归一化运算以及特征相加运算，以得到所述第一融合后编码特征以及所述文本编码特征各自对应的当前第二运算处理结果。

步骤 S135：判断当前第二噪声修正子单元是否为最后一个。

步骤 S136：若否，则将当前第二噪声修正子单元更新为下一个第二噪声修正子单元，将所述待处理特征更新为当前所述第二运算处理结果，并返回执行所述将所述待处理特征输入至当前第二噪声修正子单元中的步骤。

步骤 S137：若是，则将所述第一融合后编码特征以及所述文本编码特征各自对应的当前所述第二运算处理结果分别作为所述修正后融合特征以及所述修正后文本编码特征。

也即，本实施例基于预设自注意力机制以及预设跨注意力机制对所述第一融合后编码特征以及所述文本编码特征进行图文噪声修正处理，以降低对文本中噪声部分的关注度，如此一来，削弱了噪声的影响，实现了抗噪视觉定位。

可见，本申请是通过基于预设自注意力机制和预设跨注意力机制构建的噪声修正单元来进行图文噪声修正处理的，由于在基于跨注意力机制进行处理的过程中，文本相对于图像的差异在图像中无法找到匹配

关系，从而降低了对图文噪声成分的关注度，由此削弱了图文噪声对视觉定位准确度的影响，也即实现了抗噪视觉定位。

进一步的，本实施例针对前述实施例步骤 S15 中的对预设框特征进行修正处理的过程进行详细的介绍和说明。参见图 3 所示，上述步骤 S15 中的对预设框特征进行修正处理的过程，具体包括：

步骤 S151：将所述目标编码特征以及所述预设框特征输入至所述预设目标框修正单元中的所述第一目标框修正子单元，以便对所述目标编码特征以及所述预设框特征均分别进行自注意力运算、特征删除运算、层归一化运算以及特征相加运算，以得到所述目标编码特征以及所述预设框特征各自对应的第三运算处理结果。

步骤 S152：将所述预设目标框修正单元中的第一个第二目标框修正子单元作为当前第二目标框修正子单元，并将所述目标编码特征以及所述预设框特征各自对应的第三运算处理结果均作为当前的待处理特征。

步骤 S153：将所述待处理特征输入至当前第二目标框修正子单元中。

步骤 S154：利用当前第二目标框修正子单元对所述待处理特征依次进行跨注意力运算、特征删除运算、层归一化运算以及特征相加运算，以得到所述目标编码特征以及所述预设框特征各自对应的当前第四运算处理结果。

步骤 S155：判断当前第二目标框修正子单元是否为最后一个。

步骤 S156：若否，则将当前第二目标框修正子单元更新为下一个第二目标框修正子单元，将所述待处理特征更新为当前所述第四运算处理结果，并返回执行所述将所述待处理特征输入至当前第二目标框修正子单元中的步骤。

步骤 S157：若是，则将当前所述第四运算处理结果作为所述修正后框特征。

也即，本实施例中，基于预设自注意力机制与预设跨注意力机制构建得到所述预设目标框修正单元，如此一来，对预设框特征进行修正，修正后的特征用于预测目标视觉物体在所述目标图像上的区域位置坐标。

图4为当人类语言正确时的传统的视觉定位任务结构示意图，图5为本申请公开的当人类语言存在错误时的抗噪视觉定位任务结构示意图。

图5中的输入文本存在噪声，也即，图中女孩衣服的颜色被说成了红色。抗噪视觉定位任务需要在这种情况下，根据文本中“女孩”、“裙子”等正确的信息推测出“红色”为噪声，进而准确理解文本原本想要描述的目标，并定位其位置坐标。

图6为一种传统的视觉定位方法流程图，首先设置图像编码器和文本编码器对输入的图像以及文本进行编码，并将编码后得到的图像特征以及文本特征进行拼接，然后将拼接后得到的特征进行融合编码得到融合编码特征；设置一组预设框用来模拟框信息，通常，设置一个大小为 $[n, d]$ 的零矩阵， n 表示预设的数量， d 表示特征的维度；对融合编码特征和预设框特征进行交叉编码，旨在对预设框特征进行修正；将修正后的框信息进行置信度预测和坐标位置回归，并将置信度最高的框坐标确定为视觉定位系统的最终输出。需要指出的是，由于传统的视觉定位方法在处理过程中无法解决文本中的噪声问题，因此当文本存在噪声时，利用传统的视觉定位方法进行视觉定位存在误差。

图7为本申请公开的一种具体的抗噪视觉定位方法流程图，较传统方法，本申请加入噪声修正单元来实现抗噪的目的，进而实现抗噪视觉定位的功能。本申请的视觉定位方法流程具体包括图像和文本的编码器、预设图文特征融合单元、预设噪声修正单元、交叉编码单元，坐标预测器。

(1)、预设图文特征融合单元

预设图文特征融合单元，也称融合编码器，用来融合图像和文本特征进而学习图文之间的匹配关系。该模块有两个作用，一方面可以用来编码图文之间的关系，因此对于文本带噪的情况来说，该编码器可实现图文之间的差异编码，最终用来生成图文匹配关系。另一方面可以用来进行图文特征融合，这对于视觉定位任务本身有意义，因此可放于最终的定位模块之前做编码器；该模块在本申请所涉及的系统中出现两次，第一次侧重描述图文匹配关系，第二次用来融合图文特征，由于两个功能并不冲突，为了节约算力，两个模块权重共享。

本申请所设计的预设图文特征融合单元结构示意图参见图8所示。首先，对图像编码特征 f_{raw} 和文本编码特征 g_{raw} 进行拼接；之后，将其输入到第一预设数量的图文特征融合子单元中进行编码，并得到最终的融合后编码特征。每个图文特征融合子单元包括一个自注意力层、一个层归一化、一个随机删除层、一个相加模块。在自注意力层中，基于第一预设公式创建预设自注意力机制。

所述第一预设公式为：

$$attn_{self}(f) = \text{soft max}\left(\frac{(W_q \cdot f)^T \times (W_k \cdot f)}{\sqrt{\text{size}(f)}}\right) \times (W_v \cdot f) ;$$

其中， f 为每个所述自注意力运算单元的输入， W_q 、 W_k 以及 W_v 表示映射矩阵，

$$\text{soft max}(x) = \frac{e^x}{\sum_{j=1}^n e^x}, \quad \text{size}(f) \text{ 表示维度。}$$

随机删除层用来按一定比例对特征进行随机删除，该层的作用是防止系统过拟合。

(2)、预设噪声修正单元

预设噪声修正单元，也称修正模块，主要功能是修复噪声，是降噪过程中的关键步骤，其输入为预设图文特征融合单元的输出“融合后编码特征”和文本编码器的输出“文本编码特征”，输出为修正后融合特征以及修正后文本编码特征，如图9所示。预设噪声修正单元包括一个第一噪声修正子单元以及第二预设数量的所述第二噪声修正子单元。第一噪声修正子单元包括一个自注意力层、一个层归一化、一个随机删除层以及一个相加模块，其中，在自注意力层中，基于第一预设公式创建预设自注意力机制；第二噪声修正子单元包括一个跨注意力层、一个层归一化、一个随机删除层以及一个相加模块，其中，在跨注意力层中，基于第二预设公式创建预设跨注意力机制。

所述第二预设公式为：

$$attn_{cross}(f, g) = \text{soft max}\left(\frac{(W_q \cdot f)^T \times (W_k \cdot g)}{\sqrt{\text{size}(g)}}\right) \times (W_v \cdot g) ;$$

其中， f ， g 分别表示跨注意力层中每次参与跨注意力运算的两个输入特征， $\text{size}(g)$ 表示维度。

预设噪声修正单元通过设计跨注意力机制对图文之间的噪声进行削弱。例如，对于图5来说，文本中的“红色”由于无法在图像中找到对应的匹配关系，其注意力会偏低，经过若干层的叠加，模型对“红色”的关注度会越来越低，最终“红色”的噪声不足以影响模型性能。因此，本模块设计了跨注意力层并通过反复叠加实现了降噪的功能。

(3)、预设目标框修正单元

预设目标框修正单元，也称交叉编码模块，用于对预设框特征进行修正，修正后的特征用于预测目标视觉物体在所述目标图像上的区域位置坐标。参见图10所示，预设目标框修正单元包括一个第一预设目标框修正子单元以及第三预设数量的所述第二预设目标框修正子单元。第一预设目标框修正子单元包括一个自注意力层、一个层归一化、一个随机删除层以及一个相加模块，其中，在自注意力层中，基于第一预设公式创建预设自注意力机制；第二预设目标框修正子单元包括一个跨注意力层、一个层归一化、一个随机删除层以及一个相加模块，其中，在跨注意力层中，基于第二预设公式创建预设跨注意力机制。

此外，本申请提出了编码特征的融合，因为对交叉编码模块的输入而言，有两个融合特征均具有视觉

定位的潜力：一是修正模块修正后的“修正融合特征” f_{modify} ；二是将修正模块的输出“修正文本特征”与图像编码特征再次使用融合编码器进行编码，得到的“降噪融合特征” f_{denoise} 。因此，本申请在此处，提供三种交叉编码的输入设置。1、使用 f_{modify} ；2、使用 f_{denoise} ；3、将二者进行特征拼接或特征相加。

所述特征拼接与特征相加的公式分别为：

$$f_{\text{cat}} = [f_{\text{modify}}; f_{\text{denoise}}];$$

$$f_{\text{cat}} = f_{\text{modify}} + f_{\text{denoise}}.$$

(4) 坐标预测器

坐标预测器基于第一全连接层和第二全连接层构建而成，参见图 11 所示，所述第一全连接层为用于预测初始目标框的置信度的全连接层，第二全连接层为用于对所述初始目标框进行坐标回归处理的全连接层。利用所述坐标预测器以及所述修正后框特征，确定每个所述初始目标框的置信度；对所述置信度进行降序排序，然后从降序排序结果中筛选置信度最高的所述初始目标框，并将筛选出的所述初始目标框的坐标确定为目标视觉物体在所述目标图像上的区域位置坐标。具体的，通过对置信度排序，可以得到修正特征（大小为 $[n, d]$ ）中置信度最高的框 k ，并将第 k 个坐标输出作为视觉定位系统的最终输出。

相应的，本申请实施例还公开了一种视觉定位装置，参见图12所示，该装置包括：

特征拼接模块 11，用于对目标图像以及目标文本分别进行编码，并将编码后得到的图像编码特征以及文本编码特征进行特征拼接以得到拼接后编码特征；

第一特征融合模块 12，用于利用基于预设自注意力机制构建的预设图文特征融合单元对所述拼接后编码特征进行图文特征融合处理，以得到第一融合后编码特征；

噪声修正模块 13，用于利用预设噪声修正单元分别对所述第一融合后编码特征以及所述文本编码特征进行图文噪声修正处理，以得到修正后融合特征以及修正后文本编码特征，所述预设噪声修正单元基于预设自注意力机制和预设跨注意力机制构建而成；

第二特征融合模块 14，用于将所述拼接后编码特征以及所述修正后文本编码特征输入至所述预设图文特征融合单元以得到第二融合后编码特征；

目标框修正单元 15，用于通过预设目标框修正单元，并利用基于所述修正后融合特征和所述第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理，并利用修正后框特征预测目标视觉物体在所述目标图像上的区域位置坐标，所述预设目标框修正单元为基于预设自注意力机制和预设跨注意力机制构建的单元。

其中，关于上述各个模块更加具体的工作过程可以参考前述实施例中公开的相应内容，在此不再进行赘述。

可见，本申请公开了一种视觉定位方法，包括：对目标图像以及目标文本分别进行编码，并将编码后得到的图像编码特征以及文本编码特征进行特征拼接以得到拼接后编码特征；利用基于预设自注意力机制构建的预设图文特征融合单元对所述拼接后编码特征进行图文特征融合处理，以得到第一融合后编码特征；利用预设噪声修正单元分别对所述第一融合后编码特征以及所述文本编码特征进行图文噪声修正处理，以得到修正后融合特征以及修正后文本编码特征，所述预设噪声修正单元为基于预设自注意力机制和预设跨注意力机制构建的单元；将所述拼接后编码特征以及所述修正后文本编码特征输入至所述预设图文特征融合单元以得到第二融合后编码特征；通过预设目标框修正单元，并利用基于所述修正后融合特征和所述第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理，并利用修正后框特征预测目标视觉物体在所述目标图像上的区域位置坐标，所述预设目标框修正单元为基于预设自注意力机制和预设

跨注意力机制构建的单元，可见，本申请是通过基于预设自注意力机制和预设跨注意力机制构建的噪声修正单元来进行图文噪声修正处理的，由于在基于跨注意力机制进行处理的过程中，文本相对于图像的差异在图像中无法找到匹配关系，从而降低了对图文噪声成分的关注度，由此削弱了图文噪声对视觉定位准确度的影响，也即实现了抗噪视觉定位。

进一步的，本申请实施例还提供了一种电子设备。图13是根据一示例性实施例示出的电子设备20结构图，图中的内容不能认为是对本申请的使用范围的任何限制。

图13为本申请实施例提供的一种电子设备20的结构示意图。该电子设备20，具体可以包括：至少一个处理器21、至少一个存储器22、显示屏23、输入输出接口24、通信接口25、电源26、和通信总线27。其中，所述存储器22用于存储计算机可读指令，所述计算机可读指令由所述处理器21加载并执行，以实现前述任一实施例公开的视觉定位方法中的相关步骤。另外，本实施例中的电子设备20具体可以为电子计算机。

本实施例中，电源26用于为电子设备20上的各硬件设备提供工作电压；通信接口25能够为电子设备20创建与外界设备之间的数据传输通道，其所遵循的通信协议是能够适用于本申请技术方案的任意通信协议，在此不对其进行具体限定；输入输出接口24，用于获取外界输入数据或向外界输出数据，其具体的接口类型可以根据具体应用需要进行选取，在此不进行具体限定。

另外，存储器22作为资源存储的载体，可以是只读存储器、随机存储器、磁盘或者光盘等，其上所存储的资源可以包括计算机可读指令221，存储方式可以是短暂存储或者永久存储。其中，计算机可读指令221除了包括能够用于完成前述任一实施例公开的由电子设备20执行的视觉定位方法的计算机可读指令之外，还可以进一步包括能够用于完成其他特定工作的计算机可读指令。

进一步的，本申请实施例还公开了一个或多个存储有计算机可读指令的非易失性计算机可读存储介质，用于存储计算机可读指令；其中，所述计算机可读指令被一个或多个处理器执行时实现前述公开的视觉定位方法。

关于该方法的具体步骤可以参考前述实施例中公开的相应内容，在此不再进行赘述。

本申请书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其它实施例的不同之处，各个实施例之间相同或相似部分互相参见即可对于实施例公开的装置而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。

专业人员还可以进一步意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

结合本文中所公开的实施例描述的方法或算法的步骤可以直接用硬件、处理器执行的软件模块，或者二者的结合来实施。软件模块可以置于随机存储器（RAM）、内存、只读存储器（ROM）、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

以上对本申请所提供的一种视觉定位方法、装置、设备、存储介质进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

权 利 要 求 书

1、一种视觉定位方法，其特征在于，包括：

对目标图像以及目标文本分别进行编码，并将编码后得到的图像编码特征以及文本编码特征进行特征拼接以得到拼接后编码特征；

利用基于预设自注意力机制构建的预设图文特征融合单元对所述拼接后编码特征进行图文特征融合处理，以得到第一融合后编码特征；

利用预设噪声修正单元分别对所述第一融合后编码特征以及所述文本编码特征进行图文噪声修正处理，以得到修正后融合特征以及修正后文本编码特征，所述预设噪声修正单元为基于预设自注意力机制和预设跨注意力机制构建的单元；

将所述拼接后编码特征以及所述修正后文本编码特征输入至所述预设图文特征融合单元以得到第二融合后编码特征；和

通过预设目标框修正单元，并利用基于所述修正后融合特征和所述第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理，并利用修正后框特征预测目标视觉物体在所述目标图像上的区域位置坐标，所述预设目标框修正单元为基于预设自注意力机制和预设跨注意力机制构建的单元。

2、根据权利要求1所述的视觉定位方法，其特征在于，所述利用基于预设自注意力机制构建的预设图文特征融合单元对所述拼接后编码特征进行图文特征融合处理之前，还包括：

利用基于预设自注意力机制构建的自注意力运算单元、层归一化单元、特征删除单元以及特征相加单元构建图文特征融合子单元；

通过将第一预设数量的所述图文特征融合子单元进行依次串接，以构建得到所述预设图文特征融合单元；

相应的，所述利用基于预设自注意力机制构建的预设图文特征融合单元对所述拼接后编码特征进行图文特征融合处理，以得到第一融合后编码特征，包括：

将所述预设图文特征融合单元中的第一个图文特征融合子单元作为当前图文特征融合子单元，并将所述拼接后编码特征作为待处理特征；

将所述待处理特征输入至当前图文特征融合子单元中；

利用当前图文特征融合子单元对所述待处理特征依次进行自注意力运算、层归一化运算、特征删除运算以及特征相加运算，以得到相应的当前运算处理结果；和

根据所述当前运算处理结果获得所述第一融合后编码特征。

3、根据权利要求2所述的视觉定位方法，其特征在于，所述根据所述当前运算处理结果获得所述第一融合后编码特征，包括：

判断当前图文特征融合子单元是否为最后一个；

响应于当前图文特征融合子单元不为最后一个，将当前图文特征融合子单元更新为下一个图文特征融合子单元，将所述待处理特征更新为当前所述运算处理结果，并返回执行所述将所述待处理特征输入至当前图文特征融合子单元中的步骤；和

响应于当前图文特征融合子单元为最后一个，将当前所述运算处理结果作为所述第一融合后编码特征。

4、根据权利要求2所述的视觉定位方法，其特征在于，所述利用当前图文特征融合子单元对所述待处理特征依次进行自注意力运算、层归一化运算、特征删除运算以及特征相加运算，以得到相应的当前运算处理结果，包括：

利用当前图文特征融合子单元中的所述自注意力运算单元对所述待处理特征进行自注意力运算，得到第一运算特征；

利用当前图文特征融合子单元中的所述层归一化单元对所述第一运算特征进行层归一化处理，得到第二运算特征；

利用当前图文特征融合子单元中的所述特征删除单元并按照预设比例对所述第二运算特征进行特征删除运算，以得到第三运算特征；和

利用当前图文特征融合子单元中的所述特征相加单元对所述第三运算特征与所述待处理特征进行特征相加运算，以得到当前图文特征融合子单元中的所述运算处理结果。

5、根据权利要求1所述的视觉定位方法，其特征在于，所述利用预设噪声修正单元分别对所述第一融合后编码特征以及所述文本编码特征进行图文噪声修正处理之前，还包括：

利用基于预设自注意力机制构建的自注意力运算单元、特征删除单元、层归一化单元、特征相加单元构建第一噪声修正子单元；

利用基于预设跨注意力机制构建的跨注意力运算单元、特征删除单元、层归一化单元、特征相加单元构建第二噪声修正子单元；和

通过将所述第一噪声修正子单元以及第二预设数量的所述第二噪声修正子单元进行依次串接，以构建得到所述预设噪声修正单元。

6、根据权利要求5所述的视觉定位方法，其特征在于，所述利用预设噪声修正单元分别对所述第一融合后编码特征以及所述文本编码特征进行图文噪声修正处理，以得到修正后融合特征以及修正后文本编码特征，包括：

将所述第一融合后编码特征以及所述文本编码特征输入至所述预设噪声修正单元中的所述第一噪声修正子单元，以便对所述第一融合后编码特征以及所述文本编码特征均分别进行自注意力运算、特征删除运算、层归一化运算以及特征相加运算，以得到所述第一融合后编码特征以及所述文本编码特征各自对应的第一运算处理结果；

将所述预设噪声修正单元中的第一个第二噪声修正子单元作为当前第二噪声修正子单元，并将所述第一融合后编码特征以及所述文本编码特征各自对应的第一运算处理结果均作为当前的待处理特征；

将所述待处理特征输入至当前第二噪声修正子单元中；

利用当前第二噪声修正子单元对所述待处理特征依次进行跨注意力运算、特征删除运算、层归一化运算以及特征相加运算，以得到所述第一融合后编码特征以及所述文本编码特征各自对应的当前第二运算处理结果；和

根据所述当前第二运算处理结果获得所述修正后文本编码特征。

7、根据权利要求6所述的视觉定位方法，其特征在于，所述根据所述当前第二运算处理结果获得所述修正后文本编码特征，包括：

判断当前第二噪声修正子单元是否为最后一个；

响应于当前第二噪声修正子单元不为最后一个，将当前第二噪声修正子单元更新为下一个第二噪声修正子单元，将所述待处理特征更新为当前所述第二运算处理结果，并返回执行所述将所述待处理特征输入至当前第二噪声修正子单元中的步骤；和

响应于当前第二噪声修正子单元为最后一个，将所述第一融合后编码特征以及所述文本编码特征各自对应的当前所述第二运算处理结果分别作为所述修正后融合特征以及所述修正后文本编码特征。

8、根据权利要求1所述的视觉定位方法，其特征在于，所述通过预设目标框修正单元，并利用基于所述修正后融合特征和所述第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理之前，还包括：

利用基于预设自注意力机制构建的自注意力运算单元、特征删除单元、层归一化单元、特征相加单元构建第一目标框修正子单元；

利用基于预设跨注意力机制构建的跨注意力运算单元、特征删除单元、层归一化单元、特征相加单元构建第二目标框修正子单元；和

通过将所述第一目标框修正子单元以及第三预设数量的所述第二目标框修正子单元进行依次串接，以构建得到所述预设目标框修正单元。

9、根据权利要求8所述的视觉定位方法，其特征在于，所述通过预设目标框修正单元，并利用基于所述修正后融合特征和所述第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理，包括：

将所述目标编码特征以及所述预设框特征输入至所述预设目标框修正单元中的所述第一目标框修正子单元，以便对所述目标编码特征以及所述预设框特征均分别进行自注意力运算、特征删除运算、层归一化运算以及特征相加运算，以得到所述目标编码特征以及所述预设框特征各自对应的第三运算处理结果；

将所述预设目标框修正单元中的第一个第二目标框修正子单元作为当前第二目标框修正子单元，并将所述目标编码特征以及所述预设框特征各自对应的第三运算处理结果均作为当前的待处理特征；

将所述待处理特征输入至当前第二目标框修正子单元中；

利用当前第二目标框修正子单元对所述待处理特征依次进行跨注意力运算、特征删除运算、层归一化

运算以及特征相加运算，以得到所述目标编码特征以及所述预设框特征各自对应的当前第四运算处理结果；和

根据所述当前第四运算处理结果获得所述修正后框特征。

10、根据权利要求 9 所述的视觉定位方法，其特征在于，所述根据所述当前第四运算处理结果获得所述修正后框特征，包括：

判断当前第二目标框修正子单元是否为最后一个；

响应于当前第二目标框修正子单元不为最后一个，将当前第二目标框修正子单元更新为下一个第二目标框修正子单元，将所述待处理特征更新为当前所述第四运算处理结果，并返回执行所述将所述待处理特征输入至当前第二目标框修正子单元中的步骤；和

响应于当前第二目标框修正子单元为最后一个，将当前所述第四运算处理结果作为所述修正后框特征。

11、根据权利要求 1 所述的视觉定位方法，其特征在于，所述利用基于所述修正后融合特征和所述第二融合后编码特征确定的目标编码特征，对预设框特征进行修正处理之前，还包括：

将所述修正后融合特征或所述第二融合后编码特征或所述修正后融合特征与所述第二融合后编码特征进行预设运算处理后得到的特征确定为所述目标编码特征；其中，所述预设运算处理包括所述修正后融合特征与所述第二融合后编码特征进行特征相加或特征拼接。

12、根据权利要求 11 所述的视觉定位方法，其特征在于，所述特征相加的公式为：

$$f_{cat} = f_{modify} + f_{denoise} ;$$

其中， f_{modify} 为修正融合特征， $f_{denoise}$ 为降噪融合特征， f_{cat} 为特征相加后的输出。

13、根据权利要求 11 所述的视觉定位方法，其特征在于，所述特征拼接的公式为：

$$f_{cat} = [f_{modify}; f_{denoise}] ;$$

其中， f_{modify} 为修正融合特征， $f_{denoise}$ 为降噪融合特征， f_{cat} 为特征拼接后的输出。

14、根据权利要求 1 至 13 任一项所述的视觉定位方法，其特征在于，所述利用修正后框特征预测目标视觉物体在所述目标图像上的区域位置坐标，包括：

将所述修正后框特征输入至基于第一全连接层和第二全连接层构建的坐标预测器；所述第一全连接层为用于预测初始目标框的置信度的全连接层，第二全连接层为用于对所述初始目标框进行坐标回归处理的全连接层；

利用所述坐标预测器以及所述修正后框特征，确定每个所述初始目标框的置信度；和

根据所述每个所述初始目标框的置信度以及每个所述初始目标框的坐标确定目标视觉物体在所述目标图像上的区域位置坐标。

15、根据权利要求 14 任一项所述的视觉定位方法，其特征在于，所述根据所述每个所述初始目标框的置信度以及每个所述初始目标框的坐标确定目标视觉物体在所述目标图像上的区域位置坐标，包括：

对所述置信度进行降序排序，然后从降序排序结果中筛选置信度最高的所述初始目标框，并将筛选出的所述初始目标框的坐标确定为目标视觉物体在所述目标图像上的区域位置坐标。

16、根据权利要求 1 任一项所述的视觉定位方法，其特征在于，所述方法还包括：

基于第一预设公式创建所述预设自注意力机制；

其中，所述第一预设公式为：

$$attn_{self}(f) = soft \max(\frac{(W_q \cdot f)^T \times (W_k \cdot f)}{\sqrt{size(f)}}) \times (W_v \cdot f) ;$$

其中， f 为每个所述预设自注意力机制的输入， W_q 、 W_k 以及 W_v 表示映射矩阵， $size(f)$ 表示

维度, $attn_{self}(f)$ 表示所述预设自注意力机制的输出, softmax 表示激活函数。

17、根据权利要求 1 任一项所述的视觉定位方法, 其特征在于, 所述方法还包括:

基于第二预设公式创建所述预设跨注意力机制;

其中, 所述第二预设公式为:

$$attn_{cross}(f, g) = \text{softmax}\left(\frac{(W_q \cdot f)^T \times (W_k \cdot g)}{\sqrt{\text{size}(g)}}\right) \times (W_v \cdot g);$$

其中, f , g 分别表示预设跨注意力机制中跨注意力层中每次参与跨注意力运算的两个输入特征, $\text{size}(g)$ 表示维度, $attn_{cross}(f, g)$ 表示所述预设跨注意力机制的输出, softmax 表示激活函数。

18、一种视觉定位装置, 其特征在于, 包括:

特征拼接模块, 用于对目标图像以及目标文本分别进行编码, 并将编码后得到的图像编码特征以及文本编码特征进行特征拼接以得到拼接后编码特征;

第一特征融合模块, 用于利用基于预设自注意力机制构建的预设图文特征融合单元对所述拼接后编码特征进行图文特征融合处理, 以得到第一融合后编码特征;

噪声修正模块, 用于利用预设噪声修正单元分别对所述第一融合后编码特征以及所述文本编码特征进行图文噪声修正处理, 以得到修正后融合特征以及修正后文本编码特征, 所述预设噪声修正单元基于预设自注意力机制和预设跨注意力机制构建而成;

第二特征融合模块, 用于将所述拼接后编码特征以及所述修正后文本编码特征输入至所述预设图文特征融合单元以得到第二融合后编码特征; 和

目标框修正单元, 用于通过预设目标框修正单元, 并利用基于所述修正后融合特征和所述第二融合后编码特征确定的目标编码特征, 对预设框特征进行修正处理, 并利用修正后框特征预测目标视觉物体在所述目标图像上的区域位置坐标, 所述预设目标框修正单元为基于预设自注意力机制和预设跨注意力机制构建的单元。

19、一种电子设备, 其特征在于, 包括存储器及一个或多个处理器, 所述存储器中储存有计算机可读指令, 所述计算机可读指令被所述一个或多个处理器执行时, 使得所述一个或多个处理器执行如权利要求 1-17 任意一项所述方法的步骤。

20、一个或多个存储有计算机可读指令的非易失性计算机可读存储介质, 其特征在于, 所述计算机可读指令被一个或多个处理器执行时, 使得所述一个或多个处理器执行如权利要求 1-17 任意一项所述方法的步骤。

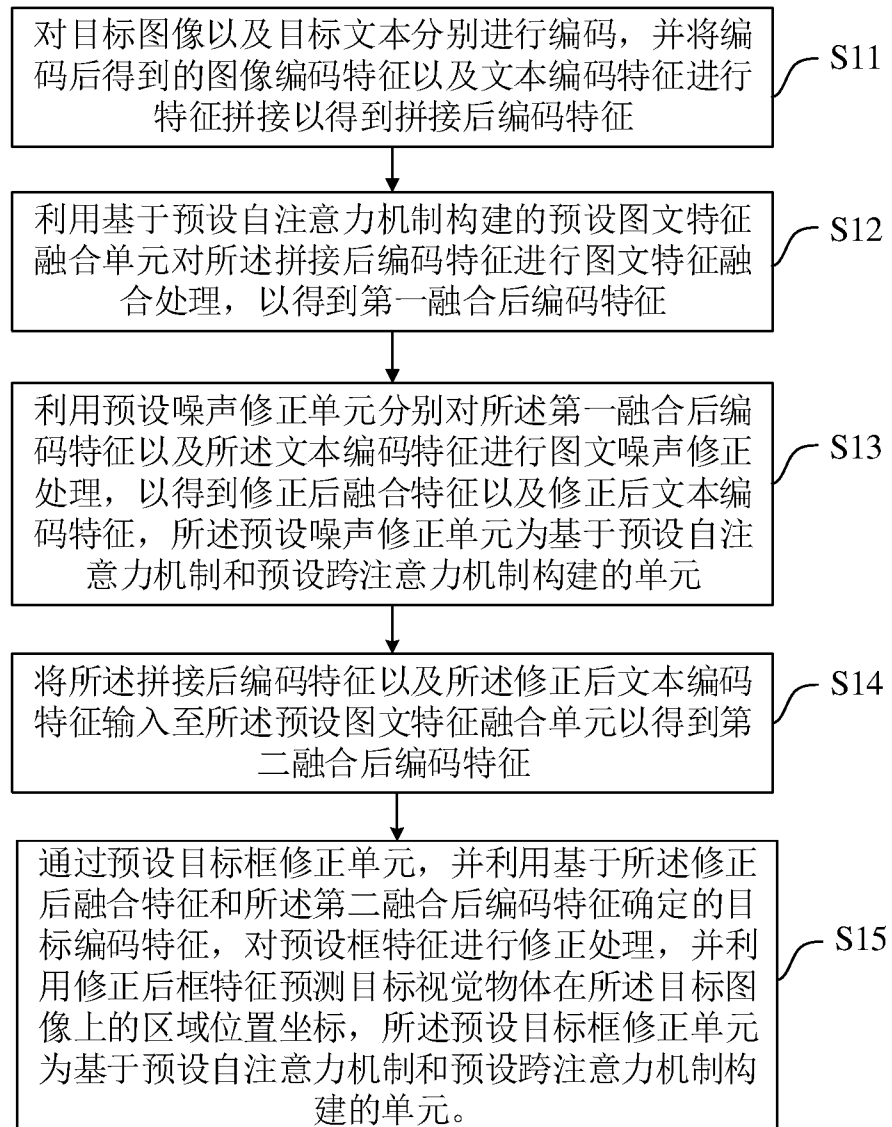


图 1

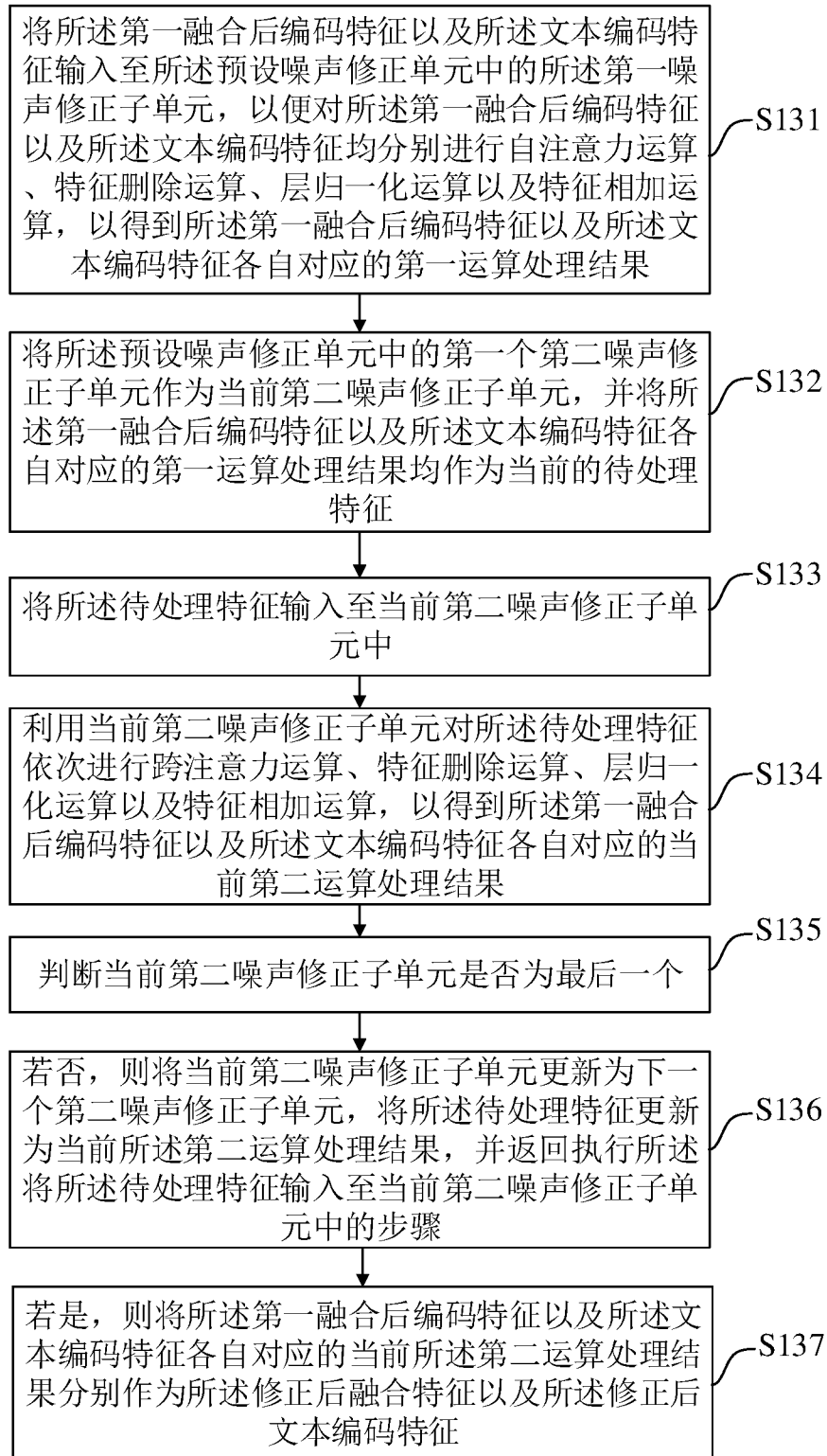


图 2

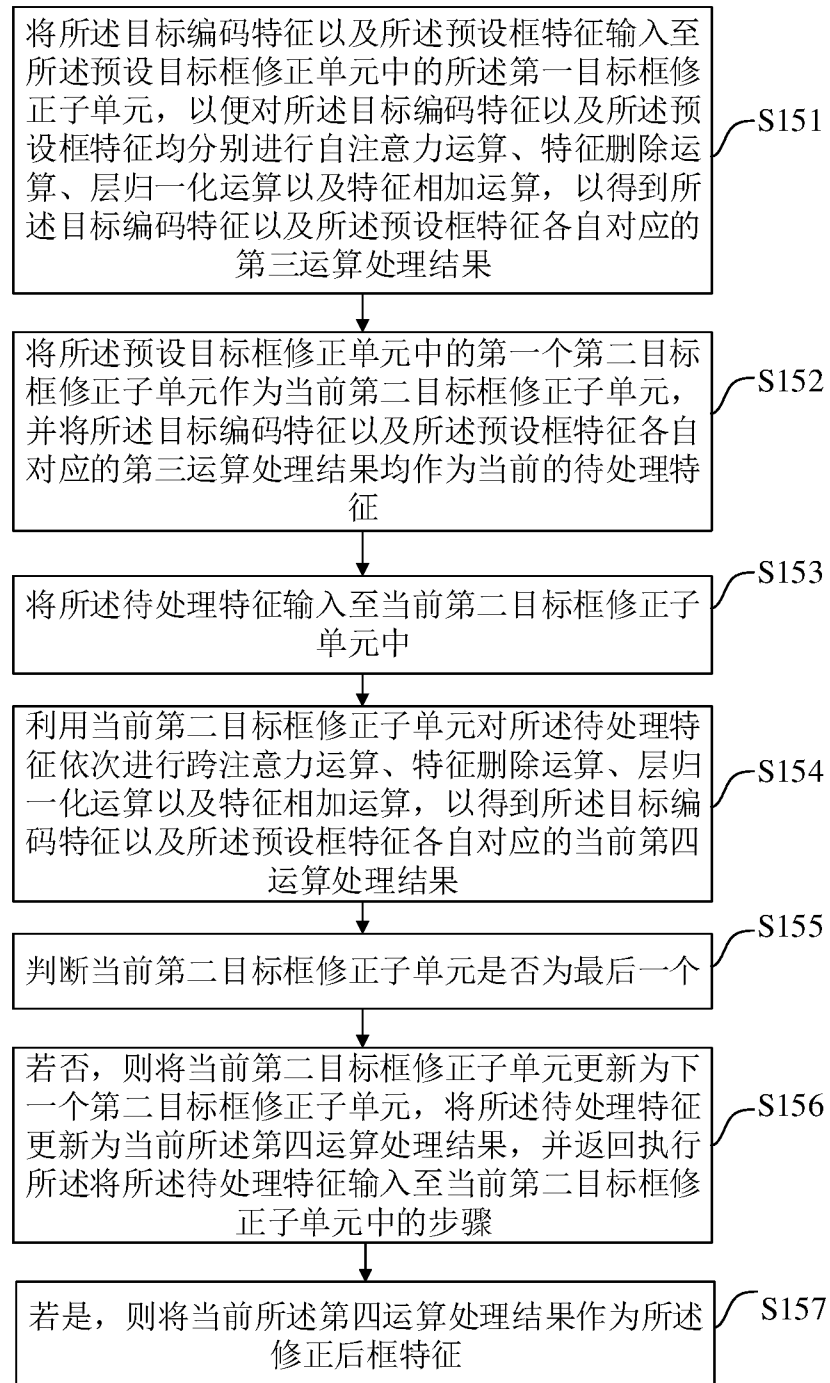


图 3

女孩穿的是白色的裙子

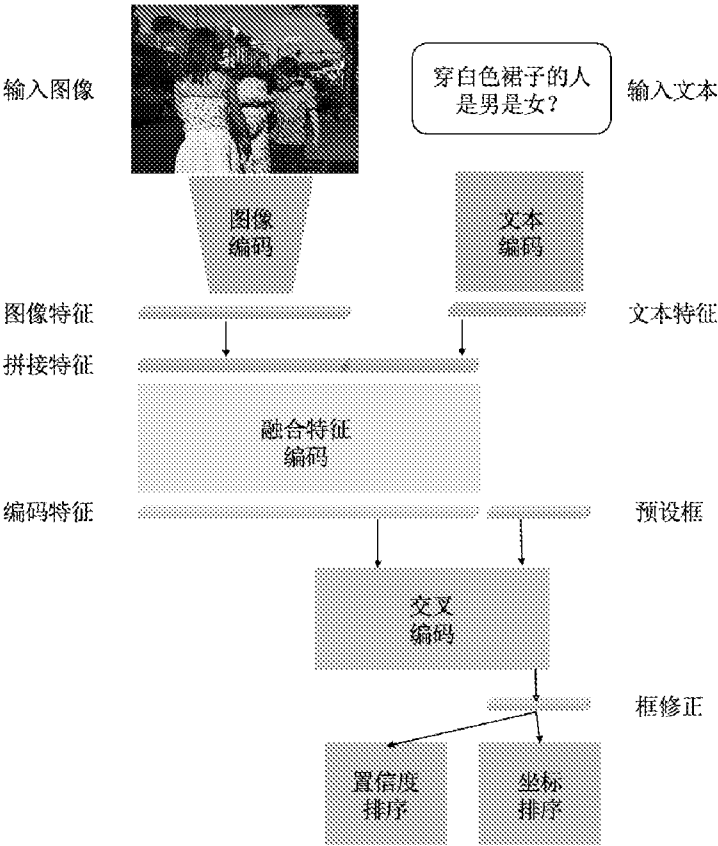


图 4

女孩穿的是红色的裙子



图 5



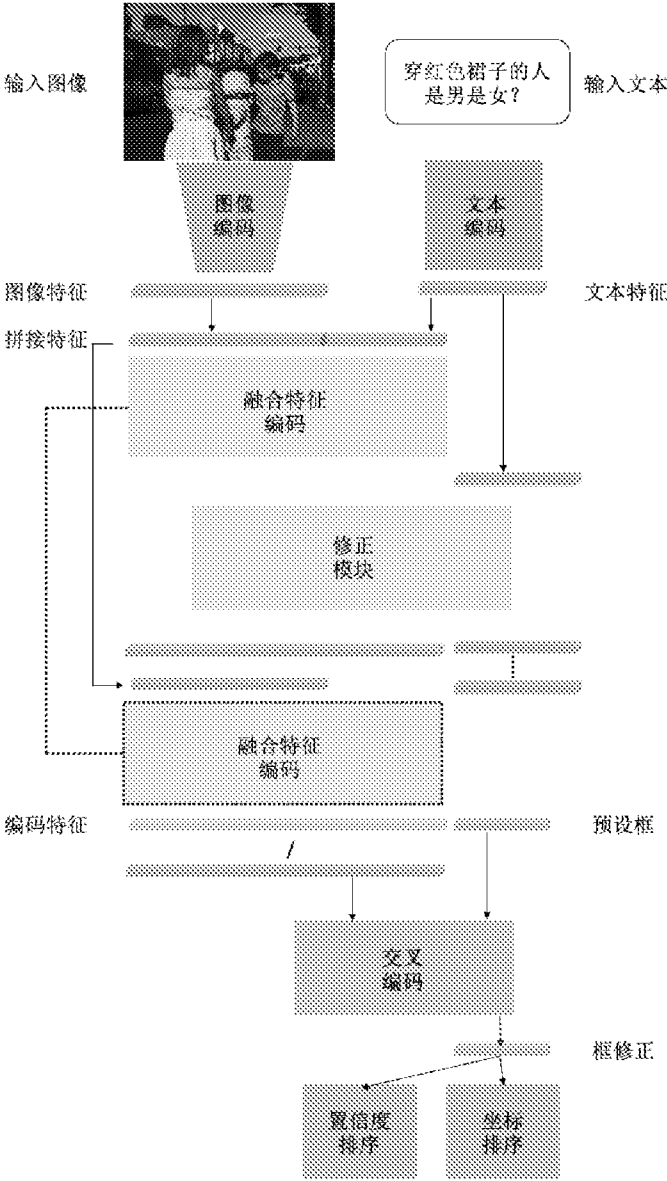


图 7

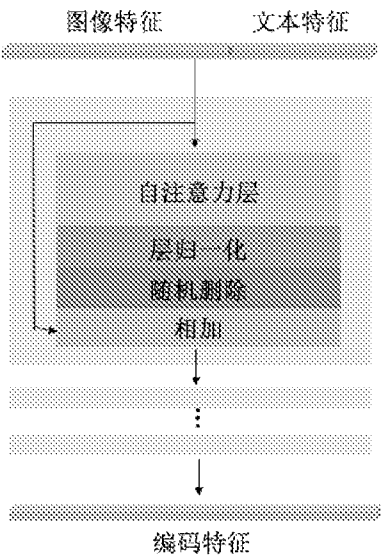


图 8

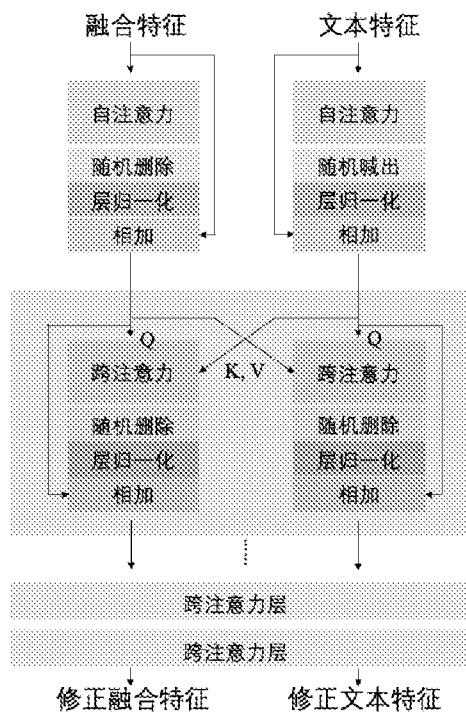


图 9

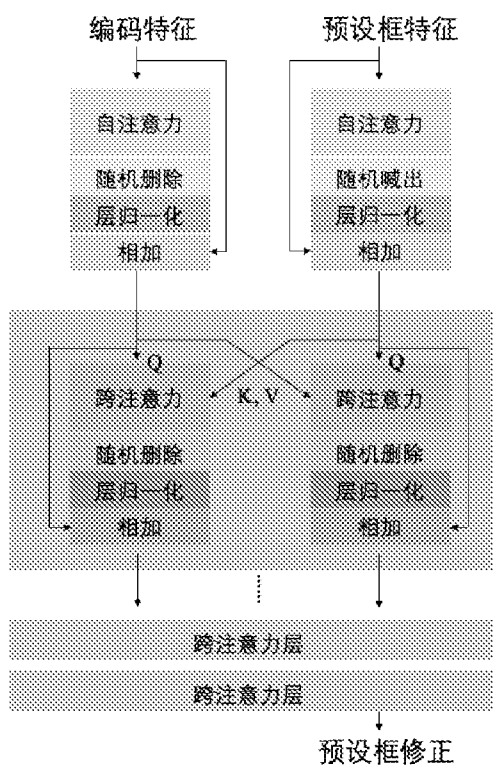


图 10

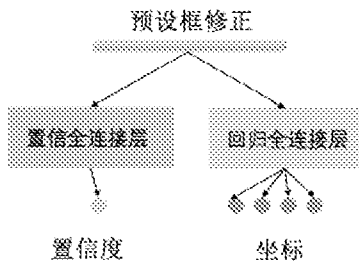


图 11



图 12

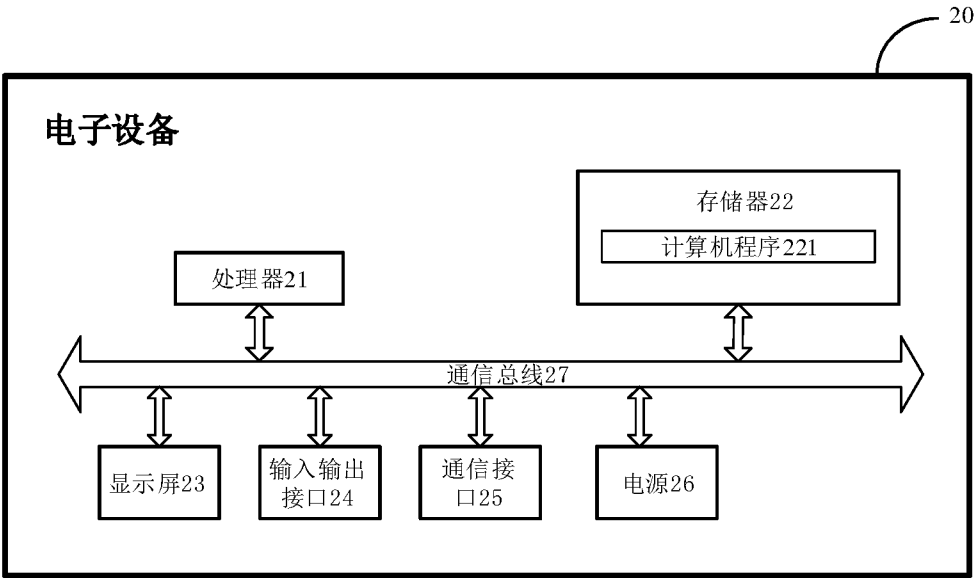


图 13

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/122335

A. CLASSIFICATION OF SUBJECT MATTER

G06T 5/00(2006.01)i; G06T 9/00(2006.01)i; G06T 7/70(2017.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06T

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; CNKI; WPABS; DWPI; USTXT; WOTXT; EPTXT; IEEE: 苏州浪潮智能科技有限公司, 李晓川, 视觉, 定位, 图像, 文本, 编码, 特征, 拼接, 融合, 注意力, 修正, 噪声, visual, position, image, text, code, feature, splicing, fusion, attention, correction, noise

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 114511472 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 17 May 2022 (2022-05-17) claims 1-10, and description, paragraphs [0005]-[0077]	1-20
A	CN 113850201 A (GUANGZHOU HUADUO NETWORK TECHNOLOGY CO., LTD.) 28 December 2021 (2021-12-28) description, paragraphs [0005]-[0192]	1-20
A	CN 113837102 A (GUANGZHOU HUADUO NETWORK TECHNOLOGY CO., LTD.) 24 December 2021 (2021-12-24) entire document	1-20
A	CN 113095435 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 09 July 2021 (2021-07-09) entire document	1-20
A	CN 112800782 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES) 14 May 2021 (2021-05-14) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

15 November 2022

Date of mailing of the international search report

05 January 2023

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/122335

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KR 102279797 B1 (INDUSTRY FOUNDATION OF CHONNAM NATIONAL UNIVERSITY) 21 July 2021 (2021-07-21) entire document	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/122335

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	114511472	A	17 May 2022	CN 114511472 B	08 July 2022
CN	113850201	A	28 December 2021	None	
CN	113837102	A	24 December 2021	None	
CN	113095435	A	09 July 2021	None	
CN	112800782	A	14 May 2021	None	
KR	102279797	B1	21 July 2021	None	

国际检索报告

国际申请号

PCT/CN2022/122335

A. 主题的分类

G06T 5/00(2006.01)i; G06T 9/00(2006.01)i; G06T 7/70(2017.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06T

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称,和使用的检索词(如使用))

CNABS;CNTXT;CNKI;WPABS;DWPI;USTXT;WOTXT;EPTXT;IEEE: 苏州浪潮智能科技有限公司, 李晓川, 视觉, 定位, 图像, 文本, 编码, 特征, 拼接, 融合, 注意力, 修正, 噪声, visual, position, image, text, code, feature, splicing, fusion, attention, correction, noise

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 114511472 A (苏州浪潮智能科技有限公司) 2022年5月17日 (2022 - 05 - 17) 权利要求第1-10项, 说明书第[0005]-[0077]段	1-20
A	CN 113850201 A (广州华多网络科技有限公司) 2021年12月28日 (2021 - 12 - 28) 说明书第[0005]-[0192]段	1-20
A	CN 113837102 A (广州华多网络科技有限公司) 2021年12月24日 (2021 - 12 - 24) 全文	1-20
A	CN 113095435 A (平安科技深圳有限公司) 2021年7月9日 (2021 - 07 - 09) 全文	1-20
A	CN 112800782 A (中国科学院自动化研究所) 2021年5月14日 (2021 - 05 - 14) 全文	1-20
A	KR 102279797 B1 (UNIV NAT CHONNAM IND FOUND) 2021年7月21日 (2021 - 07 - 21) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件
---	---

国际检索实际完成的日期

2022年11月15日

国际检索报告邮寄日期

2023年1月5日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)
中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

顾明海

电话号码 (86-512) 88995878

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/122335

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	114511472	A	2022年5月17日	CN 114511472 B	2022年7月8日
CN	113850201	A	2021年12月28日	无	
CN	113837102	A	2021年12月24日	无	
CN	113095435	A	2021年7月9日	无	
CN	112800782	A	2021年5月14日	无	
KR	102279797	B1	2021年7月21日	无	