基于包围盒回归的图像构图推荐

杨国焯, 周文洋, 刘兰, 张松海*

(清华大学计算机科学与技术系 北京 100084)
(shz@tsinghua.edu.cn)

摘要: 图像构图推荐旨在找到一幅图像中最具构图美学价值的裁剪, 可以辅助拍照者拍摄出构图优美、雅致、协调的照片。由于该任务较难精确、完整地标注出所有优秀构图包围盒, 故此前基于神经网络的方法多数并不直接回归构图包围盒, 而是通过枚举预制的构图包围盒回归分数, 取分数最大者为构图结果。而这一做法会对回归结果精度和算法效率造成负面影响。通过提出一个由特征提取模块、包围盒回归模块和分数回归模块组成的可回归构图包围盒的端到端神经网络模型, 并设计相应的数据构造方法、训练方法和损失函数, 克服上述难点。选择 FCDB 和 FLMS 这 2 个公开数据集作为测试集, 与现有方法相比, 该方法在 IoU 和 Disp 测试指标上均达到最优。

关键词: 构图推荐; 神经网络; 包围盒回归

中图法分类号: TP391.41 DOI: 10.3724/SP.J.1089.2021.18560

Bounding Box Regression Based Image Composition Recommendation

Yang Guoyao, Zhou Wenyang, Liu Lan, and Zhang Songhai*

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

Abstract: Image composition recommendation aims to find the most aesthetically valuable crop in an image, which can assist the photographer to take beautiful, elegant, and coordinated photos. However, most of the previous methods based on neural network do not generate the composition boxes directly because it is difficult to accurately and completely mark all excellent composition bounding boxes. Instead, these methods first enumerate some pre-made boxes, regress their score and return the box with highest score, which will have a negative impact on the accuracy of regression results and algorithm efficiency. By proposing an end-to-end neural network model consisting of a feature extraction module, a bounding box regression module, and a score regression module, which can regress the composition bounding box, and designing the corresponding data construction method, training method and loss function, the above difficulties have been successfully overcome. In this paper, two public datasets, FCDB and FLMS, are selected as the test set. Compared with the existing methods, this method achieves the best IoU and Disp.

Key words: composition recommendation; neural network; bounding box regression

随着手机、相机的大规模普及, 拍照留念已经成为人们生活中不可或缺的一部分, 随手拿出手机或相机拍照记录生活已经成为了人们的一种习

惯。然而, 对于缺乏摄影经验和技巧的普通用户, 拍摄出构图优秀的照片具有一定的困难和挑战性。在这一背景下, 图像构图推荐作为一个重要的研

收稿日期: 2020-08-06; 修回日期: 2020-10-20. 基金项目: 国家重点研发计划(2017YFB1002604); 国家自然科学基金(61772298, 61832016). 杨国焯(1997—), 男, 博士研究生, 主要研究方向为计算机图形学、图像分析、计算机视觉; 周文洋(1998—), 男, 博士研究生, 主要研究方向为计算机图形学、图像分析、计算机视觉; 刘兰(1996—), 女, 硕士, 主要研究方向为计算机视觉应用; 张松海(1978—), 男, 博士, 副教授, 博士生导师, CCF 会员, 论文通讯作者, 主要研究方向为图像视频处理、几何计算。

究方向,吸引着研究者和工业界的眼光. 图像构图推荐任务的目标是从一幅图像中通过简单的裁剪操作得到一幅或多幅构图较优、美学价值较高的子图.

2015 年以前,大多数工作^[1-4]是通过人工构建特征评价一幅图像的美学价值. 近年来,卷积神经网络在计算机视觉中图像分类和目标检测等诸多任务上表现出了优异的性能. 在图像构图推荐任务中,很多方法^[5-9]同样使用了卷积神经网络并得到较好的结果.

本文注意到,目标检测任务与图像构图推荐任务的回归目标比较类似,都是从图像中回归出若干个包围盒. 但是此前鲜有工作借鉴目标检测任务的思路,通过直接回归构图包围盒的方式完成图像构图推荐任务,这是因为图像构图推荐任务比目标检测任务有以下特殊性: (1) 一幅图像中可能有多个较好的构图包围盒,如果直接使用目标检测任务中的方法进行学习,则需要保证数据集中尽量不能遗落好的构图包围盒,否则会对网络回归的目标造成混淆; (2) 目标检测任务中包围盒的定义为物体的 4 个边界,比较明确,而图像构图推荐任务中构图包围盒边界的定义相比之下较不明确. 鉴于 2 点特殊性,多数已有的工作^[5-9]会标注若干包围盒对的构图优劣关系,并通过这种偏序关系训练一个单幅图像的构图评分网络,在推断时对枚举或预制的包围盒回归得到若干分数,并将得分最高的包围盒作为最终的结果. 而这一做法会使算法的性能在很大程度受到包围盒采样密度的影响.

本文提出能够直接回归得到构图包围盒的卷积神经网络模型及其训练方法,相对于以往的方法在 FLMS(Fang, Lin, Mech, and Shen)数据集^[10]和 FCDB(Flickr cropping dataset)数据集^[11]中取得了最优的效果.

1 相关工作

1.1 图像构图推荐

图像构图推荐任务的目标是用户输入一幅图像,算法自动从图像中裁剪出符合摄影构图评价、具有美学价值的子图. 图像构图推荐的方法可以分为传统方法和神经网络方法.

1.1.1 传统方法

文献[12]基于常用构图规则提出了一个启发式的优化算法,以确定前景图像的最佳位置. 文献[13]基于优秀构图模板库完成图像构图推荐的任

务,同时设计了一种精确衡量照片构图距离的方法实现模板匹配. 文献[14]认为,视觉信息相似的图像通常具有相似的显著性,通过在已标注数据集中检索相似视觉信息图像,构建显著性分类器,随后提取图像中显著性高的区域作为缩略图. 文献[15]设计了一种视觉注意力的评价指标,以保证充分反映图像的信息,并基于视觉注意力完成图像自动裁剪的任务. 文献[1]设计了一套构建高层次特征的方法,用于对图像质量进行评估,可将高质量和低质量的图像进行分类. 文献[3]提出了一套基于内容感知的局部和全局的评估照片质量的方法. 文献[4]从构图规则、内容和天空照明 3 个属性构建评估图像质量的特征. 解决这一问题的传统方法多是基于规则或模板匹配进行构图推荐,对于更复杂多样的真实数据泛化能力较差. 于是,近年来,更多的研究者开始使用神经网络方法解决这一问题.

1.1.2 神经网络方法

文献[11]首次提出在图像裁剪采用排序方法,其强调在美学价值上,图像没有绝对的评价指标,只有相对的优劣. 文献[5]基于成对训练的模式构建出视图查找网络(view finding network, VFN),通过滑动窗口的方式找到图像的最优子图. 文献[6]将强化学习技术引入图像裁剪任务,根据历史状态和当前状态决策出下一步裁剪包围盒的移动,多次迭代输出最优结果. 文献[7]提出了基于比较的图像构图(comparative photo composition, CPC)数据集,并构建构图评分网络(view evaluation networks, VEN)和构图提议网络(view proposal networks, VPN)完成构图推荐的任务, VEN 负责学习构图好的图像所具备的特征, VPN 在 VEN 的监督下对 895 个预制包围盒进行评分,并反馈给用户最好的构图包围盒. 文献[8-9]在此基础上对网络结构、损失函数和数据集进行了改进. 此前多数基于深度学习解决构图推荐问题的方法会先训练一个构图评价网络,推断时通过枚举包围盒或使用预制构图包围盒的方式回归得到每个包围盒的分数,并取分数最高的包围盒作为结果. 而本文提出的网络则可以直接回归得到构图包围盒,因此可以取得更好的结果.

1.2 目标检测

目标检测任务的目标是从一幅图像中检测到不同物体并回归得到其位置.

在深度学习大规模应用之前,传统方法^[16]目标检测的过程一般是区域选择、特征提取和分类.

区域选择一般采用滑动窗口, 通过设置不同的大小和长宽比遍历整幅图像, 但是这样做非常耗时, 且选择的区域有大量不合理的情况. 特征提取阶段常用的有方向梯度直方图(histogram of oriented gradient, HOG)^[17]、尺度不变特征变换(scale-invariant feature transform, SIFT)^[18]等. 分类阶段通常使用支持向量机(support vector machine, SVM)分类器^[19].

文献[20]采用选择性搜索的方法在输入图像中生成若干个区域提议, 并将提议中每幅子图像缩放为固定大小后送入 SVM 模型进行分类预测. 文献[21]使用神经网络代替 SVM 完成分类任务, 同时引入多任务, 建立了端到端的目标检测框架. 文献[22]提出区域提议网络和 RoI Pooling(region of interest pooling)模块, 建立了完整的端到端的目标检测架构. 文献[23]提出了一套实例分割架构, 在文献[22]的基础上添加了分割任务, 同时在 RoI Pooling 的基础上进行改进并提出了 RoI Align(region of interest align)模块, 在目标检测任务上取得了更优的性能.

本文借鉴了部分目标检测方法的思路, 并克服了将其转移至图像构图推荐任务上的难点, 提出了本文模型, 取得了更好的结果.

2 基于包围盒回归的卷积神经网络模型

本文提出了一种基于包围盒回归的卷积神经网络模型, 其使用基于 CPC^[7]构造的数据集进行训练. 本节首先介绍该数据集的构造方法, 随后介绍本文模型以及损失设计.

2.1 数据集构造方法

CPC^[7]是一个基于比较的图像构图数据集, 该

数据集中包含 10 800 条数据, 每条数据包含一幅图像、24 个包围盒以及约 100 对包围盒间的构图优劣偏序关系. 本文在此基础上对每条数据构造了一个优秀包围盒集合, 并随机选取了 300 条数据作为交叉验证集, 其余 10 500 条数据作为训练集.

对于 CPC 数据集中的每幅图像 I , 通过算法生成了包围盒集合 $B=[B_1, \dots, B_n]$; 其中, $n=24$, 每个包围盒表示一个对于 I 的构图裁剪. 随后, 该数据集分为 2 个阶段标注了若干包围盒对的构图优劣偏序关系 $P=[P_1, \dots, P_m]$; 其中, $m \approx 100$, $P_i=(p_{i0}, p_{i1})$ 表示由包围盒 $B_{p_{i0}}$ 裁剪 I 所构成子图的构图, 其优于由包围盒 $B_{p_{i1}}$ 裁剪 I 所构成子图的构图.

本文在以上数据的基础上对每条数据构造了一个优秀包围盒集合. CPC 数据集中每幅图像由 6 位标注人员标注, 在标注的第 2 阶段, 每个标注人员会在 24 个包围盒中选择 3 个包围盒作为构图最优的包围盒. 为了训练本文提出的网络, 本文从 24 个包围盒中选择了一部分构成了优秀包围盒集合 $G=[G_1, \dots, G_k]$, $G_i \in G$ 表示 B_{G_i} 是被标注为构图最优包围盒次数最多的包围盒之一.

本文使用偏序关系集合 P 和优秀包围盒集合 G 对提出的网络进行训练.

2.2 网络模型的推断与训练

本文网络模型的推断流程如图 1 所示, 训练流程如图 2 所示. 该模型由特征提取模块、包围盒回归模块和分数回归模块 3 部分组成. 在特征提取模块中, 首先将输入图像的长宽缩放为 256×256 , 然后通过由 VGG16(visual geometry group networks 16)^[24]构成的特征提取网络得到 $16 \times 16 \times 512$ 的特征图, 以供另 2 个模块使用.

图 1 本文网络模型的推断流程图

图 2 本文网络模型的训练流程图

2.2.1 包围盒回归模块

一幅图像中可能有多种较好的构图裁剪, 直接使用目标检测的方法解决图像构图推荐问题可能因为遗漏较优构图包围盒而导致网络回归目标混淆. 本文通过观察图像构图推荐任务的特性, 发现在裁剪范围被限制在一定范围内的情况下, 最优的构图有较大概率是唯一的.

基于以上发现, 本文将图像的所有裁剪方案划分为不同的裁剪集合, 称每个裁剪集合为一个局部裁剪集, 并对每个局部裁剪集回归出一个最优的裁剪方案. 下面介绍局部裁剪集的划分方式.

图像的一个裁剪可以由这个裁剪左上角及右下角 2 个点的坐标表示. 如图 3 所示, 本文将一幅图像划分为 8×8 的网格, 并把每个由若干网格组成的子矩形 R_i 作为一个局部裁剪集 C_i , 每个局部裁剪集内包含所有左上角位于 R_i 左上 $1/4$ 部分且右下角位于 R_i 右下 $1/4$ 部分的裁剪. 已有工作^[7-8]证明, 一幅图像中较好的构图包围盒不会过小, 于是本文去掉了对应子矩形短边小于 4 个网格的局部裁剪集, 共得到局部裁剪集 225 个, 通过这种方式划分的裁剪集合不会遗漏除短边长度较小的任何裁剪.

图 3 局部裁剪集划分示例

对于每个局部裁剪集 C_i , 本文使用 RoI Align^[23]从特征图中提取出子矩形 R_i 区域的特征图, 并回归得到 C_i 中的最优构图包围盒. 最优构图包围盒可以由左上角及右下角 2 点坐标表示, 本文将回归目标转化为其对应子矩形 R_i 的左上角和右下角 $1/4$ 小矩形区域中心点的相对坐标, 并进行了归一化, 设回归目标为 $(d_{xi1}, d_{yi1}, d_{xi2}, d_{yi2})$. 设 R_i 左上角和右下角坐标分别为 (x_{i1}, y_{i1}) 和 (x_{i2}, y_{i2}) , 宽为 $w_i = x_{i2} - x_{i1}$, 高为 $h_i = y_{i2} - y_{i1}$; 则 R_i 左上角 $1/4$ 小矩形中心点坐标为 $(x_{i1} + \frac{w_i}{8}, y_{i1} + \frac{h_i}{8})$, R_i 右下角 $1/4$ 小矩形中心点坐标为 $(x_{i2} - \frac{w_i}{8}, y_{i2} - \frac{h_i}{8})$, 回归得到最优构图包围盒左上角 (p_{xi1}, p_{yi1}) 和右下角坐标 (p_{xi2}, p_{yi2}) 的计算式分别为

$$\begin{aligned} p_{xi1} &= x_{i1} + \frac{w_i}{8} + (d_{xi1} - 0.5) \times \frac{w_i}{4}, \\ p_{yi1} &= y_{i1} + \frac{h_i}{8} + (d_{yi1} - 0.5) \times \frac{h_i}{4}, \\ p_{xi2} &= x_{i2} + \frac{w_i}{8} + (d_{xi2} - 0.5) \times \frac{w_i}{4}, \\ p_{yi2} &= y_{i2} + \frac{h_i}{8} + (d_{yi2} - 0.5) \times \frac{h_i}{4}. \end{aligned}$$

为了将包围盒的角点限制在 $1/4$ 小矩形区域内, 本文对网络的输出进行了截断, 即 $d_{xi1} = \max(0, \min(1, d_{xi1}))$, 对于 $d_{yi1}, d_{xi2}, d_{yi2}$ 同理. 为保证截断后网络不会出现梯度消失的情况, 本文设置了截断损失(continuous)为

$$L_C = \frac{1}{4N} \sum_{i=1}^N \sum_{s \in \{d_{xi1}, d_{yi1}, d_{xi2}, d_{yi2}\}} \max(|s - 0.5| - 0.5, 0).$$

其中, $N = 225$ 表示局部裁剪集的个数.

本文使用优秀包围盒集合 G 监督包围盒的回归, 设优秀包围盒 G_j 的左上角和右下角坐标分别为 (g_{xj1}, g_{yj1}) , (g_{xj2}, g_{yj2}) , 如果一个优秀包围盒 G_j 在局部裁剪集 C_i 的回归范围内, 则把 C_i 的回归结果与 G_j 的 L_1 距离计入包围盒回归损失(box regress), 即

$$L_{BR} = \frac{1}{M} \sum_{i=1}^N \sum_{j|G_j \in C_i} \frac{1}{4} (|p_{xi1} - g_{xj1}| + |p_{yi1} - g_{yj1}| + |p_{xi2} - g_{xj2}| + |p_{yi2} - g_{yj2}|).$$

其中, $G_j \in C_i$ 表示优秀包围盒 G_j 在局部裁剪集 C_i 的回归范围内; M 表示所有优秀包围盒在局部裁剪集内的对数. 值得注意的是, 因为本文合理地划分了局部裁剪集, 所以绝大多数情况下一个局部裁剪集的回归范围内只会至多有一个优秀包围盒.

2.2.2 分数回归模块

为了比较不同局部裁剪集回归得到结果的优劣, 本文设计了分数回归模块, 用于预测不同裁剪结果的构图分数. 在推断时, 该模块接受包围盒回归模块得到的 225 组裁剪结果, 并回归得到不同裁剪结果的得分.

本文参考了 VEN^[7]的训练方法训练所提网络的分数回归功能. CPC 数据集每幅图像 I 对应的标注数据包含一个包围盒集合 $B = [B_1, \dots, B_n]$ 和一个构图优劣偏序关系集合 $P = [P_1, \dots, P_m]$; 其中, $P_i = (p_{i0}, p_{i1})$ 表示由包围盒 $B_{p_{i0}}$ 裁剪 I 所构成子图的构图. 其优于由包围盒 $B_{p_{i1}}$ 裁剪 I 所构成子图的构图.

训练时, 本文使用 RoI Align^[23]从特征图中提取出 B 内所有包围盒对应区域的特征图, 然后对于每个包围盒 B_i 回归得到一个分数 f_i , 并使用分数距离损失(score distance)对网络的分数回归功能进行训练, 即

$$L_{SD} = \frac{1}{m} \sum_{i=1}^m \max(1 + f_{p_{i1}} - f_{p_{i0}}, 0).$$

2.2.3 损失函数设计

综上所述, 本文的总损失函数为

$$L = \lambda_C L_C + \lambda_{BR} L_{BR} + \lambda_{SD} L_{SD}.$$

其中, $\lambda_C, \lambda_{BR}, \lambda_{SD}$ 分别为各项损失的权重. 根据实验, 本文设置 $\lambda_C = 10.0, \lambda_{BR} = 10.0, \lambda_{SD} = 1.0$.

2.3 实现细节

本文提出的框架会将输入图像尺寸缩放至

256×256, 并将 VGG16 网络去掉最后的 3 个全连接层及最后一个最大池化层的剩余网络作为特征提取模块的主干网络. 包围盒回归模块包含 3 个全连接层: 第 1 个全连接层输入维度为 25 088, 输出维度为 4 096; 第 2 个全连接层输入维度为 4 096, 输出维度为 4 096; 第 3 个全连接层输入维度为 4 096, 输出维度为 4, 表示 2 点坐标偏移量. 分数回归模块与之类似, 包含 3 个全连接层: 第 1 个全连接层输入维度为 25 088, 输出维度为 4 096; 第 2 个全连接层输入维度为 4 096, 输出维度为 4 096; 第 3 个全连接层输入维度为 4 096, 输出维度为 1, 表示包围盒构图得分.

本文中的网络使用在 MS COCO(Microsoft common objects in context)数据集^[25]上训练的 Faster R-CNN^[22]进行初始化, 并使用随机梯度下降优化器训练本文中的网络, 设置参数批大小(batch size)为 1, 随机梯度下降冲量(stochastic gradient descent momentum)为 0.9, 权重衰减(weight decay)为 0.000 5, 初始学习率为 0.000 1, 每 100 000 次迭代将学习率乘 0.1. 本文使用一块 NVIDIA GeForce 1080 Ti 显卡进行训练和测试, 共训练 150 000 轮次, 平均每次迭代需要 0.18 s. 本文模型使用清华大学 Jittor^[26]深度学习框架实现.

3 实验结果与讨论

3.1 测试数据集

本文在 FCDB^[11]和 FLMS^[10]2 个公开的图像构图推荐数据集上评测本文的效果.

参考已有的工作^[7,11], 使用交并比(intersection over union, IoU), 即

$$\text{IoU} = \frac{1}{N} \sum_{i=1}^N \frac{A(W_i^g \cap W_i^c)}{A(W_i^g \cup W_i^c)},$$

和边界位移误差(boundary displacement error, Disp)

$$\text{Disp} = \frac{1}{N} \sum_{i=1}^N \sum_{j \in \{l, r, b, u\}} \frac{\|B_{i,j}^g - B_{i,j}^c\|}{4}.$$

作为性能评估指标. 其中, W_i^g 表示第 i 幅图像所标注的最优构图包围盒; W_i^c 表示第 i 幅图像通过算法生成的最优构图包围盒; $A(W_i^g \cap W_i^c)$ 代表 2 个包围盒相交部分的面积; $A(W_i^g \cup W_i^c)$ 代表 2 个包围盒相并部分的面积; N 为数据集中图像的数量; $B_{i,j}^g$ 表示第 i 幅图像所标注最优构图包围盒的边界 j ; $B_{i,j}^c$ 表示第 i 幅图像通过算法生成最优构图包围盒的边界 j ; (l, r, b, u) 分别代表包围盒的

左、右、下、上边界。

在 FCDB 数据集上, 本文使用其提供的脚本获取了 321 幅图像, 每幅图像都被标注了一个包围盒作为该图像的最优构图包围盒。

FLMS 数据集共包含 500 幅图像, 每幅图像都被 10 个不同的标注人员分别标注了一个最优构图包围盒。与已有的工作^[6]一样, 本文在测试时对于每个方法输出的最优包围盒, 都会从该图像的 10 个标注包围盒中选一个与最优包围盒 IoU 最大的包围盒, 将算法输出的包围盒与该包围盒计算 IoU 与 Disp。

3.2 实验结果对比

3.2.1 效果对比

在 FCDB 数据集上分别对 VFN^[5], VEN^[7], A2-RL (aesthetics aware reinforcement learning)^[6], VPN^[7], GAIC(grid anchor based image cropping)^[8]和 LVRN(listwise view ranking networks)^[9]这 6 种方法进行了评测。在 FLMS 数据集上分别对 4 种方法 VEN^[7], A2-RL^[6], VPN^[7]和 GAIC^[8]进行了评测, 其中, VFN^[5]和 LVRN^[9]方法的论文中没有提供在 FLMS 数据集上的实验结果或可执行的源码, 故无法在该数据集上对其进行比较。在图像构图推荐任务中, 预制包围盒的个数很大程度决定了网络的性能, 如 LVRN 通过增加预制包围盒个数得到了较大的性能提升。为了更好地反映方法本身的性能优劣, 本文比较了各方法预制包围盒个数不超过 1000 的版本。值得强调的是, 本文方法相当于只使用了 225 个预制包围盒。此外, 部分方法的预制包围盒中包含了数据集中标注的最优构图包围盒, 还有部分方法对回归结果进行了针对数据集的后处理。为了公平起见, 本文将比较预制包围盒中不包含标注包围盒且未经后处理的各个方法结果。表 1 和表 2 分别展示了不同方法在 FCDB 数据集和 FLMS 数据集上的性能, 受益于包围盒回归的使用, 本文方法在 2 个数据集上的性能都超过了对比方法, 取得了最优的指标。

表 1 不同方法在 FCDB 数据集上实验结果

方法	IoU ($\uparrow$)	Disp ($\downarrow$)
VFN ^[5]	0.6378	0.0962
A2-RL ^[6]	0.6561	0.0914
VEN ^[7]	0.6407	0.0953
VPN ^[7]	0.6660	0.0849
LVRN ^[9]	0.6841	
GAIC ^[8]	0.6802	0.0804
本文	0.6878	0.0802

注: 加粗表示最优指标。

表 2 不同方法在 FLMS 数据集上实验结果

方法	IoU ($\uparrow$)	Disp ($\downarrow$)
A2-RL ^[6]	0.8136	0.0472
VEN ^[7]	0.8126	0.0438
VPN ^[7]	0.8233	0.0399
GAIC ^[8]	0.8025	0.0470
本文	0.8534	0.0345

注: 加粗表示最优指标。

图 4 展示了各方法在 FLMS 数据集上的结果示例, 每 1 列展示了不同方法的输出, 绿色包围盒展示了方法的输出, 红色包围盒展示了 FLMS 数据集中与输出包围盒最接近的标注结果。图 5 展示了各方法在 FCDB 数据集上的结果示例, 第 1 列中红色包围盒为标注结果, 第 2~7 列的绿色包围盒展示了不同方法的输出。

图 4 5 种方法在 FLMS 数据集的结果对比

3.2.2 效率对比

本文比较了 VFN^[5], VEN^[7], A2-RL^[6], VPN^[7], GAIC^[8], LVRN^[9]这 6 种方法的效率, 如表 3 所示。本文方法虽然在效率上与部分算法相比没有优势, 但是能够取得更好的效果, 并且已经达到了实时的速度。本文将把效率优化作为未来的工作之一。

3.3 消融实验

为了验证本文所提模型中各个模块的有效性, 本文在所提模型的基础上进行了消融实验。

图 5 6 种方法在 FCDB 数据集的结果对比

表 3 不同方法效率对比

方法	每秒平均帧数($\uparrow$)
GAIC ^[8]	125.0
LVRN ^[9]	125.0
VPN ^[7]	75.0
VEN ^[7]	0.2
A2-RL ^[6]	4.1
VFN ^[5]	0.5
本文	24.0

为了验证分数回归模块的有效性, 本文设置 $\lambda_{SD} = 0$ 并训练网络, 得到 Ours_w/o_SD, 结果如表 4 的第 1 行所示. 该结果表明本文的分数回归模块能够较好地回归得到不同构图裁剪的得分, 帮助网络找到最优的构图裁剪.

为了验证本文提出的包围盒回归功能的有效性, 本文实验了将包围盒回归模块的输出结果固定时网

络的效果. 本文分别将包围盒回归模块的输出固定为局部裁剪集回归范围内的最大裁剪(本文_big)、中等裁剪(本文_middle)和最小裁剪(本文_small), 并保持特征提取模块和分数回归模块的参数不变. 表 4 的后 4 行展示了实验的结果, 结果表明本文所提出包围盒回归的概念确实改进了网络的性能, 模型能够回归得到更多样且质量更高的构图包围盒, 算法性能不再受到预制包围盒的约束.

表 4 本文方法在 FCDB 数据集上消融实验结果

方法	IoU ($\uparrow$)	Disp ($\downarrow$)
本文_w/o_SD	0.3471	0.1890
本文_small	0.5544	0.1754
本文_middle	0.6116	0.1316
本文_big	0.6770	0.0925
本文	0.6878	0.0802

注: 加粗表示最优指标.

图6展示了本文方法的部分输出结果, 对于相邻的一对图像, 左图为输入图像, 其中的红框为本

文方法回归得到的构图包围盒; 右图为将输入图像按照该构图包围盒裁剪得到的结果。

图6 本文方法结果示例

4 结 语

本文弥补了已有工作中网络性能受制于枚举或预制包围盒的缺陷, 通过消融实验证明了本文所提包围盒回归概念的重要性, 并且通过与已有工作在多个公开数据集上的对比实验, 证明了本文方法在图像构图推荐任务中达到了目前最好的效果。

在实验中注意到现有的工作对于多个主体图像效果较差, 较难对于每个主体都得到较好的结果。此外, 现有数据集的评价指标对于多主体图像回归结果的评价也不够客观, 可能出现算法结果构图优秀但是得分较低的情况。基于此, 未来将考虑在图像构图推荐任务中增加对主体因素的考虑, 并构建更加客观的数据集及评价指标。

参考文献(References):

- [1] Ke Y, Tang X O, Jing F. The design of high-level features for photo quality assessment[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2006: 419-426
- [2] Datta R, Joshi D, Li J, *et al.* Studying aesthetics in photographic images using a computational approach[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2006: 288-301
- [3] Luo W, Wang X G, Tang X O. Content-based photo quality assessment[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2011: 2206-2213
- [4] Dhar S, Ordonez V, Berg T L. High level describable attributes for predicting aesthetics and interestingness[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2011: 1657-1664
- [5] Chen Y L, Klopp J, Sun M, *et al.* Learning to compose with professional photographs on the web[C] //Proceedings of the 25th ACM International Conference on Multimedia. New York: The Association for Computing Machinery, 2017: 37-45
- [6] Li D B, Wu H K, Zhang J G, *et al.* A2-RL: aesthetics aware reinforcement learning for image cropping[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 8193-8201
- [7] Wei Z J, Zhang J M, Shen X H, *et al.* Good view hunting: learning photo composition from dense view pairs[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 5437-5446
- [8] Zeng H, Li L D, Cao Z S, *et al.* Reliable and efficient image cropping: a grid anchor based approach[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 5949-5957
- [9] Lu W R, Xing X F, Cai B L, *et al.* Listwise view ranking for image cropping[J]. IEEE Access, 2019, 7: 91904-91911
- [10] Fang C, Lin Z, Mech R, *et al.* Automatic image cropping using visual composition, boundary simplicity and content preservation models[C] //Proceedings of the 22nd ACM International Conference on Multimedia. New York: The Association for Computing Machinery, 2014: 1105-1108
- [11] Chen Y L, Huang T W, Chang K H, *et al.* Quantitative analysis of automatic image cropping algorithms: a dataset and comparative study[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 226-234
- [12] Zhang F L, Wang M, Hu S M. Aesthetic image enhancement by dependence-aware object recomposition[J]. IEEE Transactions on Multimedia, 2013, 15(7): 1480-1490
- [13] Liang Y, Wang X T, Zhang S H, *et al.* PhotoRecomposer: interactive photo recomposition by cropping[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 24(10): 2728-2742
- [14] Marchesotti L, Cifarelli C, Csurka G. A framework for visual saliency detection with applications to image thumbnailing[C] //Proceedings of the 12th IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2009: 2232-2239
- [15] Stentiford F. Attention based auto image cropping[C] //Proceedings of the 5th International Conference on Computer Vision Systems. Los Alamitos: IEEE Computer Society Press,

- 2007
- [16] Felzenszwalb P F, Girshick R B, McAllester D, *et al.* Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645
- [17] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2005: 886-893
- [18] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110
- [19] Platt J C. Sequential minimal optimization: a fast algorithm for training support vector machines[R]. Microsoft Research, Technical Report MSR-TR-98-14, 1998
- [20] Girshick R, Donahue J, Darrell T, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 580-587
- [21] He K M, Zhang X Y, Ren S Q, *et al.* Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916
- [22] Ren S Q, He K M, Girshick R, *et al.* Faster R-CNN: towards real-time object detection with region proposal networks[C] //Proceedings of Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2015: 91-99
- [23] He K M, Gkioxari G, Dollár P, *et al.* Mask R-CNN[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2961-2969
- [24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2020-08-06]. <https://arxiv.org/pdf/1409.1556.pdf>
- [25] Lin T Y, Maire M, Belongie S, *et al.* Microsoft COCO: common objects in context[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2014: 740-755
- [26] Hu S M, Liang D, Yang G Y, *et al.* Jittor: a novel deep learning framework with meta-operators and unified graph execution[J]. Science China Information Sciences, 2020, 63(12): Article No. 222103