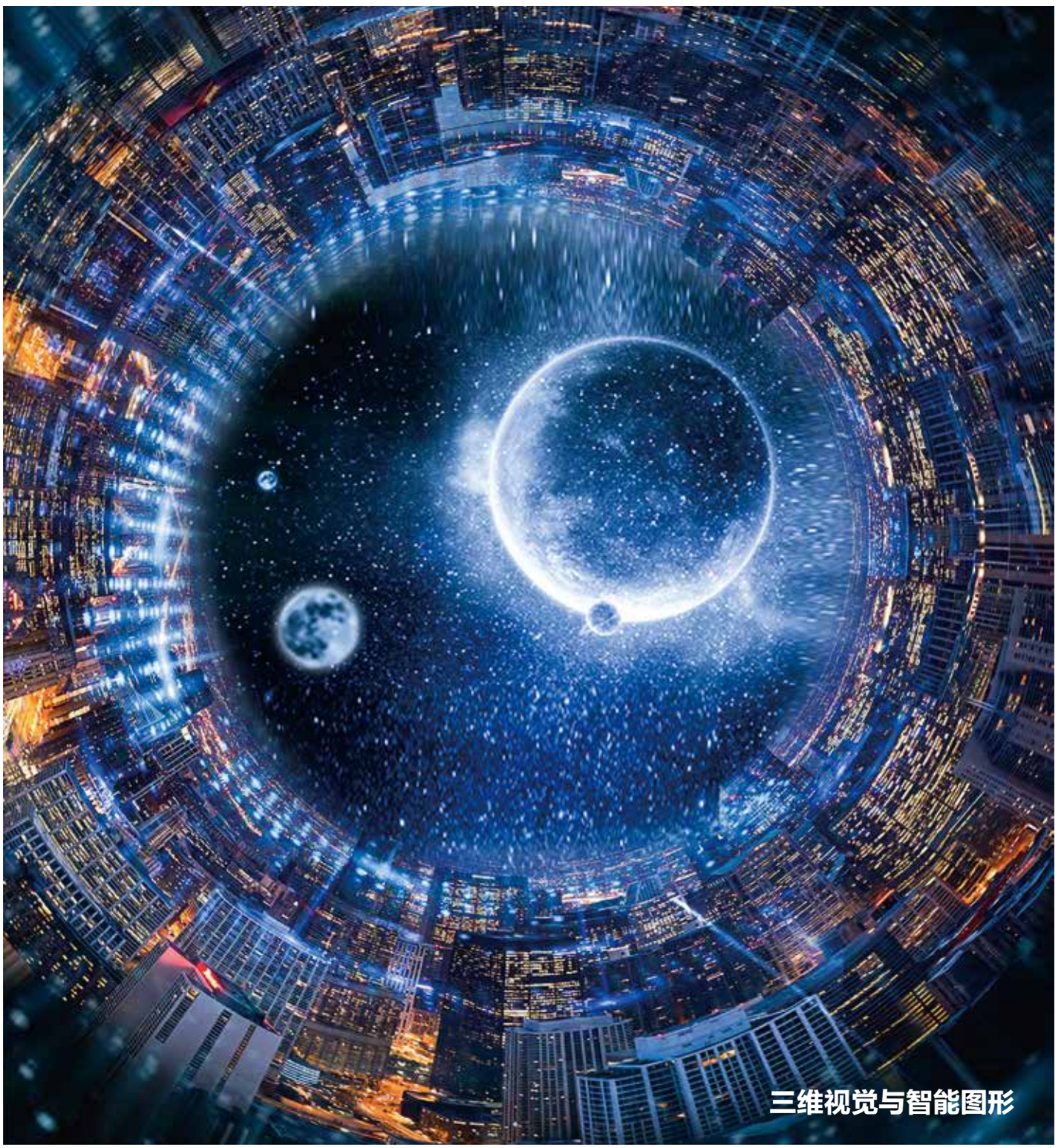
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
02
VOL.27

ISSN1006-8961
CN11-3758/TB



三维视觉与智能图形

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第2期（总第310期）
2022年2月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2022年2月 第27卷 第2期（总第310期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

面向本征图像分解的高质量渲染数据集与非局部卷积网络(第0404页)

高效鲁棒三维结构化重建(第0421页)

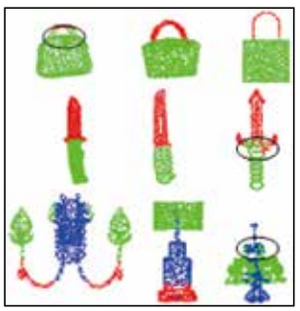
利用隐式解码器的三维模型簇协同分割(第0550页)

Volume 27, Number 2
Published February 16, 2022

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

High quality rendered dataset and non-local graph convolutional network for intrinsic image decomposition(P0404)

Efficient and robust 3D structure-aware reconstruction(P0421)



Co-segmentation of 3D shape clusters based on implicit decoder(P0550)

E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2022)02-0586-16

论文引用格式: Fang Y M, Zhong Y, Yan J B and Liu L X. 2022. Deep attention guided image cropping with fine-grained feature aggregation. Journal of Image and Graphics,27(02):0586-0601(方玉明,钟裕,鄢杰斌,刘丽霞. 2022. 聚合细粒度特征的深度注意力自动裁图. 中国图象图形学报,27(02):0586-0601)[DOI:10.11834/jig.210544]

聚合细粒度特征的深度注意力自动裁图

方玉明*,钟裕,鄢杰斌,刘丽霞

江西财经大学信息管理学院, 南昌　330032

摘　要: **目的**　从图像中裁剪出构图更佳的区域是提升图像美感的有效手段之一,也是计算机视觉领域极具挑战性的问题。为提升自动裁图的视觉效果,本文提出了聚合细粒度特征的深度注意力自动裁图方法(deep attention guided image cropping network with fine-grained feature aggregation,DAIC-Net)。**方法**　整体模型结构由通道校准的语义特征提取(semantic feature extraction with channel calibration,ECC)、细粒度特征聚合(fine-grained feature aggregation,FFA)和上下文注意力融合(contextual attention fusion,CAF)3个模块构成,采用端到端的训练方式,核心思想是多尺度逐级增强不同细粒度区域特征,融合全局和局部注意力特征,强化上下文语义信息表征。ECC模块在通用语义特征的通道维度上进行自适应校准,融合了通道注意力;FFA模块将多尺度区域特征级联互补,产生富含图像构成和空间位置信息的特征表示;CAF模块模拟人眼观看图像的规律,从不同方向、不同尺度显式编码图像空间不同像素块之间的记忆上下文关系;此外,定义了多项损失函数以指导模型训练,进行多任务监督学习。**结果**　在3个数据集上与最新的6种方法进行对比实验,本文方法优于现有的自动裁图方法,在最新裁图数据集GAICD(grid anchor based image cropping database)上,斯皮尔曼相关性和皮尔森相关性指标分别提升了2.0%和1.9%,其他最佳回报率指标最高提升了4.1%。在ICDB(image cropping database)和FCDB(flickr cropping database)上的跨数据集测试结果进一步表明了本文提出的DAIC-Net的泛化性能。此外,消融实验验证了各模块的有效性,用户主观实验及定性分析也表明DAIC-Net能裁剪出视觉效果更佳的裁图结果。**结论**　本文提出的DAIC-Net在GAICD数据集上多种评价指标均取得最优的预测结果,在ICDB和FCDB测试集上展现出较强的泛化能力,能有效提升裁图效果。

关键词:自动裁图;图像美学评价(IAA);感兴趣区域(RoI);空间金字塔池化(SPP);注意力机制;多任务学习

Deep attention guided image cropping with fine-grained feature aggregation

Fang Yuming*, Zhong Yu, Yan Jiebin, Liu Lixia

School of Information Management, Jiangxi University of Finance and Economics, Nanchang　330032, China

Abstract: **Objective** Image cropping is a remarkable factor in composing photography's aesthetics, aiming at cropping the region of interest (RoI) with a better aesthetic composition. Image cropping has been widely used in photography, printing,

收稿日期:2021-07-05;**修回日期**:2021-09-23;**预印本日期**:2021-09-30

*通信作者:方玉明　leo.fangyuming@foxmail.com

基金项目:国家重点研发计划项目(2020AAA0109301);国家自然科学基金项目(61822109);霍英东教育基金会高等院校青年教师基金项目(161061)

Supported by: National Key R&D Program of China (2020AAA0109301); National Natural Science Foundation of China (61822109)

thumbnail generating, and other related fields, especially in image processing/computer vision tasks that need to process a large number of images simultaneously. However, modeling the aesthetic properties of image composition in image cropping is highly challenging due to the subjectivity of image aesthetic assessment (IAA). In the past few years, many researchers tried to maximize the visual important information to crop a target region by feat of salient object detection or eye fixation. The results are often not in line with human preferences due to the lack of consideration of the integrity of image composition. Recently, owing to the powerful representative ability of deep learning (mainly refers to convolutional neural network (CNN)), many data-driven image cropping methods have been proposed and achieved great success. The cropped RoI images have a substantial similarity, making distinguishing the aesthetics between them, which is different from natural IAA, more difficult. Most of existing CNN-based methods only focus on feature corresponding to each cropped RoI and use rough location information, which is not robust enough for complex scenes, spatial deformation, and translation. Few methods consider the fine-grained features and local and global context dependence, which is remarkably beneficial to image composition understanding. Motivated by this, a novel deep attention guided image cropping network with fine-grained feature aggregation, namely, DAIC-Net, is proposed. **Method** In an end-to-end learning manner, the overall model structure of DAIC-Net consists of three modules: semantic feature extraction with channel calibration (ECC), fine-grained feature aggregation (FFA), and global-to-local contextual attention fusion (CAF). Our main idea is to combine the multiscale features and incorporate global and local contexts, which contribute to enhancing informative contextual representation from coarse to fine. First, a backbone is used to extract high-level semantic feature maps of the input in ECC. Three popular architectures, namely, Visual Geometry Group 16-layer network (VGG16), MobileNetV2, and ShuffleNetV2, are tested, and all of the variants achieve competitive performance. The output of the backbone is followed by a squeeze and excitation module, which exploits the attention between channels to calibrate channel features adaptively. Then, an FFA module connects multiscale regional information to generate various fine-grained features. The operation is designed for capturing higher semantic representations and complex composition rules in image composition. Almost no additional running time is observed due to the low-dimensional semantic feature sharing of the FFA module. Moreover, to mimic the human visual attention mechanism, the CAF module is proposed to recalibrate high fine-grained features, generating contextual knowledge for each pixel by selectively scanning from different directions and scales. The input features of the CAF module are re-encoded explicitly by fusing global and local attention features, and it generates top-to-down and left-to-right contextual regional attention for each pixel, obtaining richer context features and facilitating the final decision. Finally, considering the particularity of image cropping scoring regression, a multi-task loss function is defined by incorporating score regression, pairwise comparison, and correlation ranking to train the proposed DAIC-Net. The proposed multi-task loss functions can explicitly rank the aesthetics to model the relations between every two different regions. An NVIDIA GeForce GTX 1060 device is used to train and test the proposed DAIC-Net. **Result** The performance of our method is compared with six state-of-the-art methods on three public datasets, namely, grid anchor based image cropping database (GAICD), image cropping database (ICDB), and flickr cropping database (FCDB). The quantitative evaluation metrics of GAICD contain average Pearson correlation coefficient ($\overline{PCC}$), average Spearman's rank-order correlation coefficient ($\overline{SRCC}$), best return metrics ($Acc_{K/N}$), and rank-weighted best return metrics ($wAcc_{K/N}$) (i.e., higher is better over these metrics). Intersection over union and boundary displacement error are adopted as evaluation metrics in the two other datasets. The GAICD dataset is split into 2 636 training images, 200 validating images, and 500 test images. ICDB and FCDB contain 950 and 348 test images respectively, which are not used for training by all compared methods. Experimental results demonstrate the effectiveness of DAIC-Net compared with other state-of-the-art methods. Specifically, $\overline{SRCC}$ and $\overline{PCC}$ increase by 2.0% and 1.9%, and other best return metrics increase by 4.1% at most on the GAICD. The proposed DAIC-Net outperforms most of the other methods despite very minimal room for improvement on ICDB and FCDB. Qualitative analysis and user study of each method are also provided for comparison. The results demonstrate that the proposed DAIC-Net generates better composition views than the other compared methods. **Conclusion** In this paper, a new automatic image cropping method with fine-grained feature aggregation and contextual attention is presented. The ablation study demonstrates the effectiveness of each module in DAIC-Net, and further experiments show that DAIC-Net can obtain better results than other methods on the GAICD dataset. Comparison experiments on the ICDB and FCDB datasets verify the generalization of DAIC-Net.

Key words: automatic image cropping; image aesthetics assessment (IAA); region of interest (RoI); spatial pyramid pooling (SPP); attention mechanism; multi-task learning

0 引 言

自动裁图是计算机视觉中具有实用性和挑战性的问题,旨在裁剪出图像中构图更佳的区域,提升图像美感。自动裁图广泛应用于摄影、印刷和缩略图技术等相关领域。随着深度学习的发展,利用数据驱动解决自动裁图问题的方法越来越多。

早期研究从人眼关注的图像显著区域(Jian 等,2015)出发,以凸显视觉重要信息为目的进行裁图。Suh 等人(2003)结合显著性检测和人脸高级语义特征指导缩略图的裁剪。Stentiford(2007)将像素平均显著分数作为衡量裁图好坏的标准。Chen 等人(2016a)探索了基于图像显著性的自动裁图搜索算法的复杂度。上述方法的效果均取决于像素级别的分类精度,容易产生误判和漏检。同时,由于缺乏对图像构成的考量,结果往往不符合审美要求。Zhang 等人(2013)提出基于邻接图的概率模型,较早地关注了图像构图完整性。Yan 等人(2013)基于先验规则提取 38 维手工特征,探究了图像在专业裁剪前后的差异。Fang 等人(2014)总结 3 种普遍的图像裁剪规则,提出一个评估图像构成和内容的模型。这类方法对裁图有一定提升,但多数采用滑窗生成候选裁图框的方式,增加了算法时间复杂度。此外,由于图像构图的规则和技巧相对复杂,实际应用过程中存在诸多问题,如局部特征丢失、空间形变失真和平移不敏感等。

卷积神经网络(convolutional neural network, CNN)因性能优异广泛应用于自动裁图领域。Chen 等人(2017a,b)发布了一个用于成对比较(pairwise ranking)的裁图数据集,并基于排序思想提出一个孪生网络(siamese network)裁图模型。Wei 等人(2018)进一步提高裁图数据集的规模,提出首个密集标注的裁图数据集,并采用迁移学习(transfer learning)方法,将复杂模型所学知识迁移到基于目标检测(object detection)的高性能裁图模型,提高了自动裁图的效率。Zeng 等人(2020)基于网格的图像表示和语义约束生成更加合理的候选裁图框,提高了裁图数据集标注的密度,促进模型学习更有区分力的特征表示。

尽管深度学习方法较传统方法取得了更好的效果,但仍很少考虑裁图区域的局部和全局上下文依赖关系、裁图特征的细粒度等与图像构成相关的重要信息,多数深度学习方法对复杂场景、显著目标位置、空间形变以及裁图平移等问题仍然不够鲁棒。为此,本文提出一种聚合细粒度特征的深度注意力自动裁图方法(deep attention guided image cropping network with fine-grained feature aggregation, DAIC-Net)。

主要贡献如下:1)结合区域特征提取和空间金字塔池化技术,提高模型对图像构成的解析,可以有效捕获多尺度结构信息。由于图像低维特征的共享,可以在几乎不增加额外运行时间的条件下,同时应用多种不同细粒度区域特征聚合操作,提高模型的检测精度和效果。2)融合全局和局部多尺度注意力特征,由粗到细地对输入特征进行重编码,引入更多上下文信息,进一步增强对裁图特征的学习,采用通道注意力指导语义特征的校准,可以获得更为丰富的全局关系依赖特征,促进最终决策。3)提出一个涵盖分数回归、成对比较和相关性排名的多任务损失函数,在 3 种不同的特征提取主干网络上性能均有提升,并在多个裁图数据集中取得了最优性能,使模型具有实用价值。

1 相关工作

1.1 美学质量评价

随着深度学习的发展,CNN 对图像信息的表征能力不断增强,基于 CNN 的图像美学质量评价(image aesthetic assessment, IAA)也得以快速发展。IAA 与图像质量评价(image quality assessment, IQA)不同,IAA 关注图像内容和构图对美感的影响程度,强调主观的图像感受;而 IQA 则通过对图像信号进行相关特性分析,评价图像的视觉失真程度(方玉明等,2021),强调客观影响因素。因此 IQA 难以预测图像内容和构图。常见构图方法如图 1 所示。

带有主观美学标注的图像数据集主要有 CUHKPQ(Chinese University of Hong Kong photo quality dataset)(Tang 等,2013)、AVA(aesthetic visual analy-

(a) 三分构图　(b) 中心构图　(c) 对称构图

图 1　不同构图方法示例

Fig. 1　Example images of different composition rules((a) rule of thirds; (b) rule of center; (c) rule of symmetric)

sis database)(Murray 等,2012)和 AADB(aesthetics and attributes database)(Kong 等,2016)等。基于这些美学数据集,Lu 等人(2014)提出一个基于全局、局部和风格属性的 CNN 分类模型,用于判断图像美丑,在当时达到了较高精度。Mai 等人(2016)依据空间金字塔思想,提出自适应卷积神经网络以解决输入图像形变问题,该网络可接收任意尺寸的输入,但无法进行批量训练。类似地,Chen 等人(2020)直接从卷积核结构上解决了图像形变和图像长宽比构图信息的共存问题,设计了一种自适应的空洞卷积(Yu 和 Koltun,2015)结构,在批量训练的同时考虑了不同纵横比图像的缩放问题。Li 等人(2020)提出一种融合大众审美和人格特征的 IAA 模型,探索了个人偏好对个性化 IAA 的影响。Zhu 等人(2020)基于元学习(meta-learning)方法解决个性化 IAA 中的小样本学习问题,有效捕捉了人们在审美判断时共同的先验知识。Sheng 等人(2020)针对一系列图像处理操作(如加噪、模糊和压缩等),发现这些操作的强弱程度与图像美学质量的线性关系,从自监督学习的角度重新审视 IAA 问题。目前,基于美学评价的研究方法已成为自动裁图的主流方向,对同一幅图像中的不同裁图进行美学评价,相比常规的图像美学质量评价对模型的细粒度识别要求更高。

1.2　图像区域特征提取

随着深度学习在目标检测任务中的发展,产生了一系列图像区域特征提取技术。空间金字塔池化(spatial pyramid pooling,SPP)(He 等,2015)和感兴趣区域(region of interest pooling, RoI)(Girshick, 2015)是两个经典的区域特征提取技术。SPP 采用不同粒度的金字塔将区域特征划分为不同网格,然后将所有网格进行池化合并;RoI 池化则仅采用一种网格粒度,输出特征大小可由用户指定,训练更为灵活。二者都能将多尺度特征提取为固定大小的特征输出,从而不影响顶端模块(如全连接层)的执行。目前,感兴趣区域特征提取已衍生出多种方法。RoI Align(He 等,2020)和 RoI Warp(Dai 等,2016)方法不同程度地提高了区域特征提取的精度,Lu 等人(2019)进一步分析了 RoI 池化操作的利弊,提出 RoI Refine 方法降低精度损失。感兴趣区域特征提取模块衔接不同尺度的区域特征和最终的分类/回归层,实现了图像特征的重用,在保持较高检测精度的同时,能显著提高模型训练和测试速度。

Zeng 等人(2020)认为自动裁图不仅需要考虑感兴趣区域,还应考虑裁掉区域(region of discard, RoD)。本文模型采用了相似做法。RoD 池化操作是通过去除感兴趣区域的特征,通过双线性插值算法将余下特征图映射到指定大小,一般是与 RoI 特征相同的空间尺寸,便于后续操作。

1.3　深度视觉注意力模型

视觉注意力机制(visual attention mechanism)是一种用于提升基于循环神经网络(recurrent neural network,RNN)的编码器—解码器结构模型性能的机制,广泛应用于机器翻译、语音识别和语义分割(Yan 等,2021)等领域。近年来,注意力模型也应用于诸多视觉任务。Xu 等人(2015)将注意力机制引入图像描述中,提出一个深度递归注意力模型。Chen 等人(2016b)将注意力机制用于像素级别增强的语义分割任务,并自适应地选择多尺度特征。Liu 等人(2018)提出 PiCANet 用于显著性检测任务,通过对每个像素(区域)产生丰富的上下文注意力信息,以获得更高级的语义特征,从而提高最终的检测效果。由于注意力机制赋予模型更强的分辨能力,能够模拟视觉场景中信息聚焦区域和上下文信息。

本文模型针对复杂场景、多尺度特征的裁图进行设计,结合了全局和局部的区域注意力特征,裁图模型的鲁棒性有了进一步提升。

作为 RNN 的变体,长短期记忆网络(long short term memory,LSTM)改善了 RNN 梯度消失等问题,通过门控机制使循环网络不仅能记忆过去的信息,其隐含层还能选择性遗忘不重要的信息,对长期语境和远距离上下文依赖进行建模。LSTM 常用于时间序列信息的编码,ReNet(Visin 等,2015)的提出使更多研究开始关注空间上下文的依赖关系,Varior 等人(2016)利用空间 LSTM 捕捉人体各部分之间的空间依赖关系,提高行人重识别的准确度;Byeon 等人(2015)和 Liang 等人(2016)则将类似的方法分别应用于语义分割和语义对象解析。本文模型采用空间 LSTM 进行全局上下文特征编码,更好地关联来自卷积层和全连接层的特征。

2 模型设计

本文模型(DAIC-Net)的整体框架如图 2 所示,由通道校准的语义特征提取(semantic feature extraction with channel calibration,ECC)、细粒度特征聚合(fine-grained feature aggregation,FFA)和上下文注意力融合(contextual attention fusion,CAF)3 个模块构成,采用端到端学习策略,对给定图像 $\boldsymbol{I}$ 及对应的候选裁图集合 $\boldsymbol{C}_I$,首先利用 ECC 模块提取图像深度特征 $\boldsymbol{F}_I$,然后通过 FFA 模块将 $\boldsymbol{F}_I$ 映射到所有 $\boldsymbol{C}_I$ 区域,获得所有候选裁图的多尺度区域特征 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]$,$N$ 表示候选裁图集合的大小,$\boldsymbol{x}_i$ 表示第 i 个裁图区域的特征。不同细粒度区域特征经过 CAF 模块,在特征图上任意位置 (w,h) 产生全局注意力权重 $\boldsymbol{A}_{\mathrm{g}}^{w,h}$ 和局部注意力权重 $\boldsymbol{A}_{\mathrm{l}}^{w,h}$,用于上下文特征转化。最后将每个区域的上下文注意力特征输入到两个 1×1 卷积组成的回归层,得到所有区域的美学质量分数。

2.1 通道校准的语义特征提取模块

通道校准的语义特征提取模块旨在获取图像的高级语义特征,服务于后续操作。多项研究证明,某一领域发展较快的网络模型可以很好地迁移至其他领域当中,如图像分类(Simonyan 和 Zisserman,2014;He 等,2016)任务中模型所学知识可以有效地迁移至目标检测(Girshick,2015;Ren 等,2017)和图像质量评价(Zhang 等,2020)等任务。通用特征提取模块耦合度低,因此本文方法能自适应主流的网络框架。对比实验表明,对于不同网络的语义特征,本文方法均能有效提升模型整体的预测结果。

网络层数的加深有助于学习到更为复杂的非线性映射关系和高级语义信息,但也会造成梯度弥散和梯度爆炸,尤其对于具有循环神经网络架构的模型。He 等人(2016)提出了跳跃连接,通过将浅层特

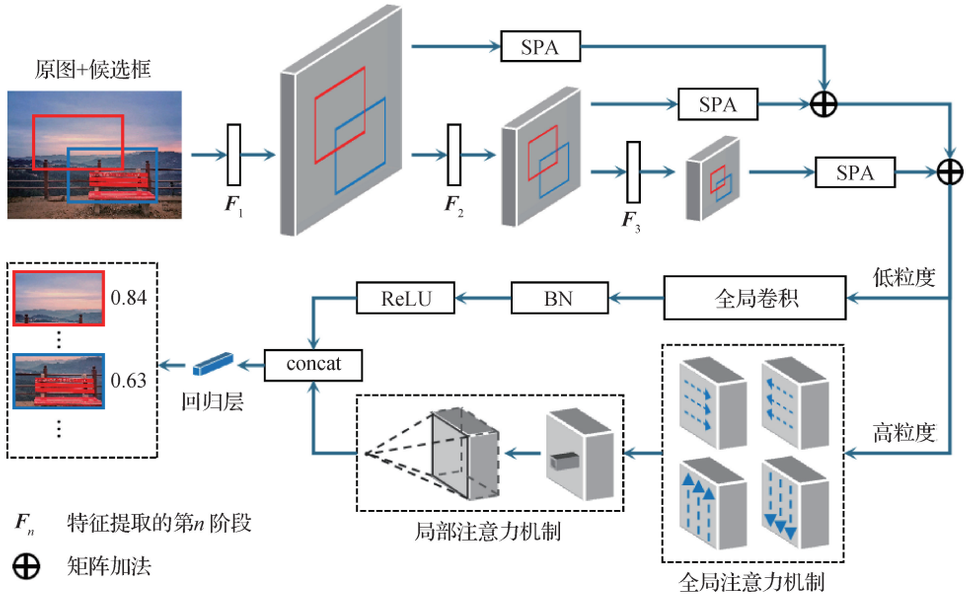


图 2 本文模型 DAIC-Net 的整体框架

Fig. 2 Overall framework of the proposed DAIC-Net

征作为深层网络输入的一部分，提高深层网络的浅层表征。本文使用 3 种不同尺度的语义特征 $\boldsymbol{F}_l^l$、$\boldsymbol{F}_l^m$ 和 $\boldsymbol{F}_l^s$，其中 l 、m 、s 表示 3 种不同的感受野大小，分别对应不同深度的网络层输出。浅层网络用于提取低层次的局部纹理特征，而深层网络用于捕获更复杂的全局场景语义信息（Zeiler 和 Fergus，2014）。

语义特征提取后，为减少过拟合，Zeng 等人（2020）采用 1 ×1 卷积将 3 种尺度特征通道数降至 8 维，该操作能有效降低模型复杂度。然而，特征降维和轻量级结构容易影响模型泛化能力。受 SENet（squeeze-and-excitation networks）（Hu 等，2018）的启发，本文提出一种通道注意力校准的区域金字塔特征提取方法（squeeze and excitation regional pyramid align，SPA），对每个阶段输入的语义特征作预处理，获得通道注意力加权的语义特征 $\omega^l\boldsymbol{F}_l^l$ 、$\omega^m\boldsymbol{F}_l^m$ 和 $\omega^s\boldsymbol{F}_l^s$ 。具体操作如图 3 所示，通过全局自适应平均池化获取输入特征的全局空间信息，然后依次经过通道压缩全连接层、ReLU 激活函数、通道扩张全连接层和 sigmoid 激活函数，得到一个与输入特征通道数相同的通道注意力权重矩阵，将二者逐通道相乘，即完成对输入特征的重新校准。实验中的通道压缩比和扩张比均设置为 16。最后，仍采用 1 ×1 卷积将注意力加权的高维语义特征降至 8 维，在该特征图上提取每个候选框不同细粒度的 RoI 和 RoD 特征。

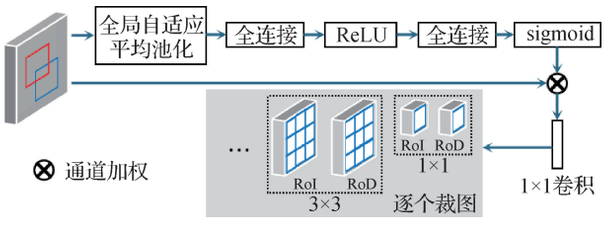


图 3　通道注意力校准的区域金字塔特征提取

Fig. 3　Regional pyramid feature extraction with attention calibration between channels

2.2　细粒度特征聚合模块

构图规则具有复杂多样性，图 1 显示了三分构图、中心构图和对称构图 3 种最常见的构图方法。不同的构图方法需要关注不同位置上的重要信息，例如，三分构图要求人们将重要信息放在 4 个交点处（在摄影中也称黄金分割点），而中心构图将重要信息居中显示，凸显目标的全貌。穷举所有的构图规则是不切实际的，即便是同一种构图，物体本身也会随场景变化造成尺度不一致（Chen 等，2020）。对此，本文基于常见的构图先验知识，将 ECC 提取的低维多尺度语义特征进行分流组合，并行获取所有裁图的 3 种尺度区域特征。同时，为更好地表征学习图像构成，本文采用一种构图敏感的细粒度特征聚合（FFA）模块，即使用 RoI 和 RoD 输出多级区域特征，如图 4 所示。

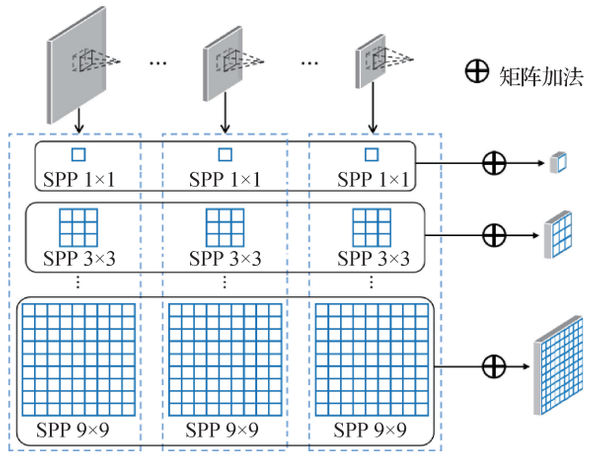


图 4　细粒度特征聚合（FFA）模块

Fig. 4　The fine-grained feature aggregation （FFA） module

全局池化保留了全局显著特征，使模型消除了空间变化带来的不一致性，但也导致局部特征的丢失。出于对图像构成多样性的考量及受空间金字塔池化（He 等，2015）的启发，本文提出多重细粒度特征聚合方法，不同的细粒度反映并作用于不同的构图，具体为

$$\boldsymbol{X} = \sum_{t \in \{l,m,s\}} concat(RoI(\omega^t\boldsymbol{F}_l^t), RoD(\omega^t\boldsymbol{F}_l^t)) \quad (1)$$

式中，$\boldsymbol{X}$ 表示同一种粒度的裁图区域特征，$concat$ 表示特征融合操作，ω 表示语义特征的通道注意力权重。对于相同粒度的 RoI 和 RoD 特征，先按通道维度合并，得到 16 维融合特征，再对 3 种不同尺度的融合特征进行逐元素相加，作为上下文注意力融合模块的输入特征。本文实验使用了 {1，3，5，7，9} 共 5 种粒度的组合。

与生成单一细粒度的池化方法相比，多重细粒度特征聚合方法结合了局部和全局信息，有利于图像构成的解析，提供更丰富的鲁棒性特征，产生更好的效果。由于每种粒度均采用相同处理方式，可用并行化计算，整个过程无需增加额外训练参数。

2.3　上下文注意力融合

上下文注意力由全局注意力（global attention，

GA)和局部注意力(local attention,LA)组成,由粗到细地对输入特征进行重编码,引入更多上下文依赖信息,进一步增强对裁图特征的学习。GA 为输入的每个像素块产生自顶向下、自左向右的上下文区域注意力,可以获得更为丰富的上下文特征,促进最终决策;LA 提取局部细粒度特征,将上下文特征转化到分数回归层的输入空间。

2.3.1 全局注意力机制

GA 的初始输入为同一种粒度的裁图区域特征 $\boldsymbol{X} \in \mathbf{R}^{N\times 16\times k^2}$,$N$ 表示候选裁图集合的大小,$K \in \{1,3,5,7,9\}$,需要产生像素级别全局特征,即提高每个区块像素的感受野。本文采用双向长短期记忆网络(bi-directional long short term memory, BiLSTM)(Graves 等,2013)代替传统 RNN,分别处理上下左右 4 个方向的序列数据,通过横向和纵向交替扫描,混合 4 个方向的上下文信息,每个像素的信息传播至其他像素,同时存储来自任意方向的上下文信息。对于细粒度为 k 的裁图区域特征图上任一位置 (w,h),特征编码产生注意力权重向量 $\boldsymbol{\alpha}^{w,h} \in \mathbf{R}^{N\times k^2}$,将所有 $k\times k$ 个像素产生的注意力权重与初始输入特征内积,转化为全局注意力特征。计算为

$$\boldsymbol{X}_{\mathrm{GA}}^{w,h} = \sum_{i=1}^{k\times k} \boldsymbol{x}_i \odot \boldsymbol{\alpha}_i^{w,h} \tag{2}$$

式中,$\boldsymbol{x}_i \in \mathbf{R}^{N\times 16}$ 表示 N 个裁图在第 i 个位置上的细粒度特征,$\odot$ 表示哈达玛乘积。输出的全局注意力特征与原始输入特征大小相同。

2.3.2 局部注意力机制

以 (w,h) 为中心,LA 只关注其局部区域内的相邻特征,即局部特征块 $\boldsymbol{X}^{w,h} \in \mathbf{R}^{W\times H\times 16}$,$W\times H$ 为每个位置的感受野大小。本文使用全卷积层进行裁图分数回归,为便于处理,采用全卷积操作进行局部注意力特征提取。首先,将多层 $W\times H$ 大小的卷积核应用于输入特征,以细粒度 $k=9$ 为例,采用 4 个 3×3 的卷积核,输出通道数分别是 16、32、64 和 768,随着输出特征尺度降低至 1,最后采用 3 层权值共享的 1×1 卷积降维,输出预测分数。

消融实验表明,并非所有细粒度特征都需要计算 GA 和 LA。细粒度较低的特征本身包含了全局空间信息,同时也会限制 LA 的局部特征块数量,因此,为降低计算复杂度,实验中仅对细粒度最高的特征进行上下文特征转化,如图 2 所示。其他细粒度特征则采用全局卷积、批归一化层和 ReLU 激活函数操作处理,得到一个与 LA 输出相同维度的特征图,所有细粒度特征融合后输入到由两个全卷积层组成的分数回归层,最终预测每个候选框的美学分数。

2.4 损失函数

考虑到分数回归的特殊性,本文提出基于多任务训练的裁图评分模型,总体损失函数由 3 种损失函数加权求和得到,具体为

$$L_{\mathrm{total}} = \omega_1 L_{\mathrm{R}} + \omega_2 L_{\mathrm{P}} + \omega_3 L_{\mathrm{C}} \tag{3}$$

式中,L_{R} 为分数回归损失函数,L_{P} 为成对比较损失函数,L_{C} 为相关性损失函数,ω_1、ω_2 和 ω_3 为各项损失的均衡化权重。

2.4.1 分数回归损失

光滑 L1 损失因对异常值的鲁棒性而广泛应用于目标检测的回归问题(Ren 等,2017),结合平方误差损失和线性误差损失的优点,使训练过程更加稳定。本文使用光滑 L1 损失用于裁图分数回归,具体为

$$L_{\mathrm{R}}(y, f(\boldsymbol{x})) = \begin{cases} \dfrac{1}{2}(y - f(\boldsymbol{x}))^2 & |y - f(\boldsymbol{x})| \leqslant \delta \\ \delta\left(|y - f(\boldsymbol{x})| - \dfrac{1}{2}\delta\right) & \text{其他} \end{cases} \tag{4}$$

式中,δ 是超参数,本文在所有实验中设置 $\delta = 1$,y 是真实分数,$f(\boldsymbol{x})$ 是模型的预测分数。

2.4.2 成对比较损失

成对比较(pairwise comparison)思想广泛应用于排序问题,通过打擂方式,设计一个针对列表中的元素进行两两比较的模型,最终实现排序。本文采用成对比较思想,引导模型学习更具有区分度的特征,使得裁图之间的分数差距更符合人的主观评价。对于同一幅图像中的任意一对候选裁图 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$,假设预期前者的真实美学分数大于后者,则可定义损失函数为

$$L_{\mathrm{P}}(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)) = \max\{0, f(\boldsymbol{x}_2) - f(\boldsymbol{x}_1) + g\} \tag{5}$$

式中,$f(\boldsymbol{x})$ 表示模型预测分数,g 是用于正则化最小间距的超参数,本文在所有实验中设置 $g = 0.5$。

考虑到穷举所有的裁图组合会大幅降低模型的效率,实验过程中,过滤裁图分数间距较小的裁图组合。由于裁图组合数量减少,模型训练过程中仅使用成对比较损失不能得到很好的效果,因此本文将该损失函数作为其他损失函数的辅助项,设置更低

的权重。

2.4.3　相关性损失

尽管分数回归损失和成对比较损失对裁图分数进行了隐式排序，但裁图特征之间本身具有很强的相似性，因此本文仍采用相关性损失监督模型的训练，显式地对不同的裁图进行美学质量排名，以捕获不同裁图间的细节特征。为便于计算，采用皮尔森线性相关系数，计算为

$$L_{\mathrm{C}}(y,f(\boldsymbol{x})) = 1-\frac{\sum_{i=1}^{N}((y_i-\bar{y})(f(\boldsymbol{x}_i)-\overline{f(\boldsymbol{x})}))}{\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{N}(f(\boldsymbol{x}_i)-\overline{f(\boldsymbol{x})})^2}} \tag{6}$$

式中，y_i 是第 i 个裁图的真实分数，$\bar{y}$ 为 y_i 的均值，$f(\boldsymbol{x}_i)$ 是模型对第 i 个裁图的预测分数，$\overline{f(\boldsymbol{x})}$ 为 $f(\boldsymbol{x}_i)$ 的均值，N 为候选裁图框个数。

考虑到裁图的最终目标是找出构图更佳区域，因此，真实主观分数更高的裁图更受关注，模型对其也有更强的区分能力。为此，本文针对不同分数的裁图设置了不同的损失权重，最终的相关性损失为

$$L_{\mathrm{C}}(y,f(\boldsymbol{x})) = 1-\frac{\sum_{i=1}^{N}\mathrm{e}^{(y_i-\bar{y})(f(\boldsymbol{x}_i)-\overline{f(\boldsymbol{x})})}}{\sqrt{\sum_{i=1}^{N}\mathrm{e}^{(y_i-\bar{y})^2}}\sqrt{\sum_{i=1}^{N}\mathrm{e}^{(f(\boldsymbol{x}_i)-\overline{f(\boldsymbol{x})})^2}}} \tag{7}$$

3　实验与分析

实验在 Intel Core i5-3470 CPU、Nvidia GeForce GTX 1060 显卡、16 GB 内存的台式电脑上进行，操作系统采用 Ubuntu 20.04.1 LTS，深度学习框架为 PyTorch 1.6.0，CUDA 和 cuDNN 版本为 CUDA 10.2 和 cuDNN v7.6.5，编程语言为 Python 3.6.12。

3.1　数据集和评估指标

首先在最新公开的裁图数据集 GAICD（grid anchor based image cropping database）（Zeng 等，2020）上进行实验。该数据集共 3 336 幅图像，划分为训练集 2 636 幅、验证集 200 幅和测试集 500 幅。每幅图像包含[52，90]个有效裁图，共 288 069 个。每个裁图都有坐标位置和平均意见得分（mean opinion score，MOS）。评估指标同样采用 GAICD 数据集使用的排名相关性、最佳回报率和排名加权最佳回报率 3 类指标。针对每幅图像的各个裁图分数及对应的模型预测分数，排名相关性指标包括斯皮尔曼相关性（Spearman’s rank-order correlation coefficient，SRCC）和皮尔森相关性（Pearson correlation coefficient，PCC），用于评估模型预测分数与实际分数之间的一致性和相关性；最佳回报率指标 $Acc_{K/N}$，一般取 $K \in \{1,2,3,4\}$，$N \in \{5,10\}$，共组成 8 种指标，衡量模型产生美学质量裁图的概率；排名加权最佳回报率指标 $wAcc_{K/N}$ 在最佳回报率指标的基础上增加了排名权重，该类指标的评估更加准确和严格（Zeng 等，2020）。特别地，本文还使用了平均最佳回报率及其加权指标，计算为

$$\overline{Acc_N} = \frac{1}{4}\sum_{K=1}^{4}Acc_{K/N} \tag{8}$$

$$\overline{wAcc_N} = \frac{1}{4}\sum_{K=1}^{4}wAcc_{K/N} \tag{9}$$

GAICD 数据集的提出打破了经典的裁图评估方法，以往的裁图数据集只有稀疏的裁图标注，使用交并比（intersection over union，IoU）和四边偏移误差（boundary displacement error，BDE）等客观评价指标，如 ICDB（image cropping database）（Yan 等，2013）、FCDB（flickr cropping database）（Chen 等，2017a）和 HCDB（human cropping database）（Fang 等，2014）等。对于每幅图像，IoU 计算模型预测的裁图框与人为标注的最优裁图框之间的面积交并比，而 BDE 计算两者对应四条边的平均偏移量。虽然这两类评价指标具有一定的局限性（Zeng 等，2020），本文仍然在 ICDB 和 FCDB 两种数据集上测试，并与其他裁图算法进行了对比实验。ICDB 包含 950 幅图像，由 3 位专家标注，因此每幅图像包含 3 个最好的裁图，分别反映了不同标注者的主观性；FCDB 包含 348 幅裁图测试图像，每幅图像由多人投票选出一个最好的裁图标注。

3.2　实施细节

3.2.1　基本设置

训练前对数据进行归一化处理有助于加快网络收敛并防止过拟合。首先使用 ImageNet 数据集（Deng 等，2009）的均值和方差对输入图像的像素值进行归一化，然后将裁图的主观分数映射到零均值正态分布。模型参数均采用随机初始化，初始学习率为 0.000 1，训练周期为 40 epochs，学习率每 15 个训练周期衰减为当前的 10%，最初的一个周期内采

用 warmup 策略预热学习率(He 等,2016),训练中使用自适应矩估计优化算法(Kingma 和 Ba,2015)对网络参数进行更新,动量为 0.9,权值衰减系数为 0.000 5。

3.2.2 数据增强

数据增强是一种提高模型鲁棒性、防止过拟合的有效方式。在图像分类任务中,常采用随机裁剪、随机旋转和随机翻转等数据增强操作,这些操作的出发点是使模型具有旋转不变性,有助于图像分类,缺点是会破坏图像的原始构图,不利于美学评价等主观任务。本文采用能保留构图一致性的数据增强方法(Chen 等,2020),通过对输入图像进行缩放,有助于减少训练资源的开销。在训练阶段,不同于常规缩放处理(缩放到同一固定尺寸,如 256×256 像素),本文采用批量的随机尺寸缩放。具体地,将每个样本等比缩放,且最短边缩放到[224,416]范围内、以 32 为步长作为输入。实验还采用了其他对图像原始构图影响小的数据增强方法,如随机调整亮度、对比度、饱和度和色调等,以提高模型的鲁棒性。在测试阶段,通过大量实验发现,将测试图像最短边缩放至 384 时取得的裁图效果最佳,这与大多数图像分类任务将图像缩放至 224 或 256 不同,间接验证了图像缩放对自动裁图和图像美学评价的影响很大(Chen 等,2020)。

3.2.3 训练细节

为了验证本文提出框架的通用性,测试了 3 种自动裁图任务中常用的主干网络(backbone)作为本文的图像语义特征提取模块,包括 VGG16(Visual Geometry Group 16-layer network)(Simonyan 和 Zisserman,2014)、MobileNetV2(Sandler 等,2018)和 ShuffleNetV2(Ma 等,2018)。这 3 种模型涵盖了大多数现有的深度学习技术,其中 VGG16 采用小卷积核结构和深层堆叠方式,有效提高了模型性能,但参数量较大,对硬件要求较高;MobileNetV2 和 ShuffleNetV2 则追求轻量化和高效率,采用了残差模块、批归一化处理和分组卷积等先进的网络优化思想。训练过程中,根据本文提出的数据增强方法,对于每种输入尺寸,通过批量训练方式以提高训练速度,通过大量实验发现,每个小批量为 16 幅图像时训练效果最好,当同时输入的图像很少时,预训练模型中的批归一化层参数将不再改变。本文模型只在 GAICD 训练集上进行训练,从每幅图像预定义的候选裁图中随机挑选最多 64 个预测美学分数,相比通过滑窗预设候选裁图的方法,模型的训练速度大幅提高。为了公平对比所有方法,主干网络均使用 ImageNet 预训练参数进行初始化,所有模型都在同一台服务器上运行,采用相同的复杂度计算方法。

3.3 消融实验与分析

为了验证细粒度特征聚合(FFA)模块的有效性,本文在 GAICD 数据集上进行消融实验,结果如图 5 所示。可以看出,随着细粒度的增加,各项指标呈上升趋势,相比于单一细粒度,多重细粒度的结果更好,其中使用 5 种不同细粒度特征聚合的效果最佳,客观上说明细粒度特征聚合模块使模型学习到了更多构图细节特征。

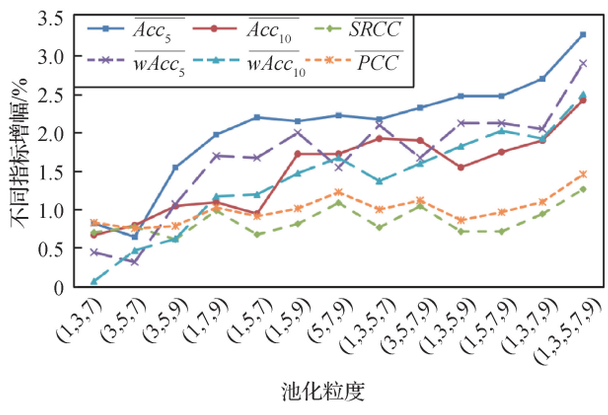


图 5 不同细粒度特征组合的消融实验结果

Fig. 5 Ablation experimental results of different fine-grained combinations

同时,选取不同细粒度的特征图进行上下文注意力融合(CAF)模块的消融实验,以探究特征重编码的作用,验证上下文注意力融合模块对最终模型性能的影响,结果如表 1 所示。

表 1 细粒度特征重编码的消融实验结果

Table 1 Ablation experimental results of fine-grained feature recoding

重编码特征	$\overline{SRCC}$	$\overline{PCC}$	$\overline{Acc_5}$/%	$\overline{Acc_{10}}$/%	$\overline{wAcc_5}$/%	$\overline{wAcc_{10}}$/%
无	0.870	0.895	61.1	80.9	48.2	62.8
3×3	0.872	0.897	62.5	81.7	48.9	63.9
5×5	0.874	0.896	62.8	81.5	49.0	64.3
7×7	0.873	0.897	63.0	82.3	**49.6**	64.5
9×9	**0.876**	**0.901**	**63.3**	**82.4**	**49.6**	**64.7**

注:加粗字体表示各列最优结果。

实验中，模型在无任何上下文注意力融合模块的情况下，均采用全卷积回归层直接产生区域特征的分数映射。从表 1 可以看出，随着重编码特征的细粒度增加，模型在相关性指标上保持相当的水平，而在排名准确性指标上有明显上升趋势，这是由于特征重编码提高了模型对区域特征的判别敏感度。整体判别能力的提升对相关性指标的影响较小，较难突出其作用。但其他指标更关注排名靠前的部分候选裁图，更容易受到特征重编码的影响。

此外，损失函数在模型训练过程中也起着至关重要的作用，不同损失的叠加对模型产生不同的效果。本文研究了多任务损失函数中各分量的影响，消融实验结果如表 2 所示，所有客观指标结果均为 ShuffleNetV2 在 GAICD 验证集上的结果。采用网格超参数优化方法并结合实验验证，最终设置 $\omega_1 = 1$、$\omega_2 = 0.5$ 和 $\omega_3 = 1$。从表 2 可以看出，对单一损失函数，L_C 对排名相关性指标的提升较大，这是由于 L_C 更加关注元素之间的关系，即便两个元素间距很小也可能产生很大的区分度。L_R 控制所有元素的绝对误差，使得总体误差缩小，相比 L_C 对模型的要求更加严格，更有利于提升最佳回报率和排名加权最佳回报率指标。当 L_P 作为辅助损失函数时，$L_R + L_P$ 和 $L_C + L_P$ 对模型性能均有所提升，这是因为引入 L_P 之后，训练的模型不仅要关注所有裁图的预测分数与真实分数的差异，还要考虑裁图与裁图之间的细微变化，显式地学习更多具有区分度的特征，进一步提高预测准确性。当同时使用 3 种损失函数时，模型在所有客观指标上均值最高，客观说明了模型在多项损失函数的引导下，得到的裁图效果最好。

表 2　不同损失函数的消融实验结果

Table 2　Ablation experimental results of different loss function

损失函数	$\overline{SRCC}$	$\overline{PCC}$	$Acc_{1/5}$ /%	$Acc_{4/5}$ /%	$Acc_{1/10}$ /%	$Acc_{4/10}$ /%	$wAcc_{1/5}$ /%	$wAcc_{4/5}$ /%	$wAcc_{1/10}$ /%	$wAcc_{4/10}$ /%
L_C	0.869	0.889	67.5	55.9	85.5	77.9	49.7	45.5	64.4	61.9
L_R	0.866	0.886	67.9	56.6	85.6	78.3	49.7	45.8	64.7	62.4
$L_R + L_P$	0.866	0.888	68.2	56.8	85.8	78.3	50.5	46.1	64.8	62.5
$L_C + L_P$	0.870	0.892	67.8	56.2	85.8	78.0	49.9	45.7	64.4	62.2
$L_R + L_C$	0.873	0.899	68.6	57.2	86.3	78.4	51.7	46.8	66.3	63.0
$L_R + L_P + L_C$	**0.876**	**0.901**	**69.0**	**58.1**	**86.5**	**78.7**	**51.8**	**47.1**	**66.6**	**63.3**

注：加粗字体表示各列最优结果。

3.4　对比实验与方法评估

为进一步验证本文方法的性能，选择 A2-RL（aesthetics aware reinforcement learning）（Li 等，2018）、VPN（view proposal net）（Wei 等，2018）、VFN（view finding network）（Chen 等，2017b）、VEN（view evaluation net）（Wei 等，2018）、LVRN（listwise view ranking network）（Lu 等，2019）和 GAIC（grid anchor based image cropping）（Zeng 等，2020）等 6 种具有代表性且可复现的裁图方法与本文方法进行对比。不同裁图方法的预测结果不同，VPN 和 A2-RL 只能预测一个最优裁图，无法对所有裁图进行美学评分，因此只计算 $Acc_{1/5}$ 和 $Acc_{1/10}$ 两种评估指标；其他方法均可对所有裁图评分，可对比所有评估指标。

3.4.1　模型复杂度分析

本文比较了不同模型的时间复杂度和参数量，如表 3 所示。

表 3　模型复杂度分析

Table 3　Complexity analysis of each model

方法	主干网络	速度/（帧/s）	参数/兆
A2-RL	AlexNet	1.54	24.1
VPN	VGG16	15.7	65.3
VFN	AlexNet	0.32	11.6
VEN	VGG16	0.62	40.9
LVRN	VGG16	44.4	37.9
GAIC	VGG16	48.8	14.7
GAIC	MobileNetV2	77.5	1.81
GAIC	ShuffleNetV2	70.4	0.78
本文	VGG16	46.2	15.1
本文	MobileNetV2	74.1	2.79
本文	ShuffleNetV2	63.7	1.80

由于 GAICD 数据集中每幅图像对应的候选裁

图框均不超过 90 个,因而所有方法的运行效率都有提高。然而 VFN 和 VEN 需要对每个裁图端到端地运行整个网络,才能得到所有裁图的预测分数,因此复杂度会随着候选裁图框数量线性增长;VPN 基于一个高效的目标检测网络,只对预设的候选框进行分类,其预测速度远高于 VEN。A2-RL 通过调整裁图大小进行有限次迭代优化,基于预测的美学分数计算每步的累积奖励,经少量的迭代步数即可找到模型认为的最优裁图。LVRN、GAIC 和本文方法均采用区域特征提取操作,只需要对原图做一次特征提取,极大减少了冗余的前向次数。

3.4.2 主观定性分析

图像自动裁图本质上是为了改善图像的美感,最终获得美学构图最佳的裁图结果。为了比较不同方法产生的裁剪结果的主观感受,本文在几种典型场景上,对不同方法裁图的结果进行了定性比较,如图 6 所示。分析可知,VPN 和 A2-RL 整体裁图效果偏差,这两种方法无法对候选裁图框进行美学评分,易造成过度裁剪。VPN 容易裁剪掉图像的主要目标物体,破坏重要内容;而 A2-RL 对同一幅图像可能得到不同的裁图结果,有时甚至无法对图像进行任何裁剪(将原图预测为最终裁图结果)。VFN 生成的裁图基本保持内容完整性,但仍无法有效地移除非重要区域,构图效果一般。VEN 和 LVRN 具备一定的美学评价能力,裁图效果相对较好,但缺乏构图细节。实验中所有模型均使用 GAICD 数据集中预设的候选裁图框。由于候选裁图框的选取规则,GAIC 和本文方法在部分图像上的裁图效果相近,但从一些图像的裁图结果可知,本文模型结果能够更好地保留构图特征,更符合常见的构图方法(如第 1—3 行)。综合以上分析,本文方法可以更有效地移除原图中的非重要区域,并产生构图更佳的视觉效果。

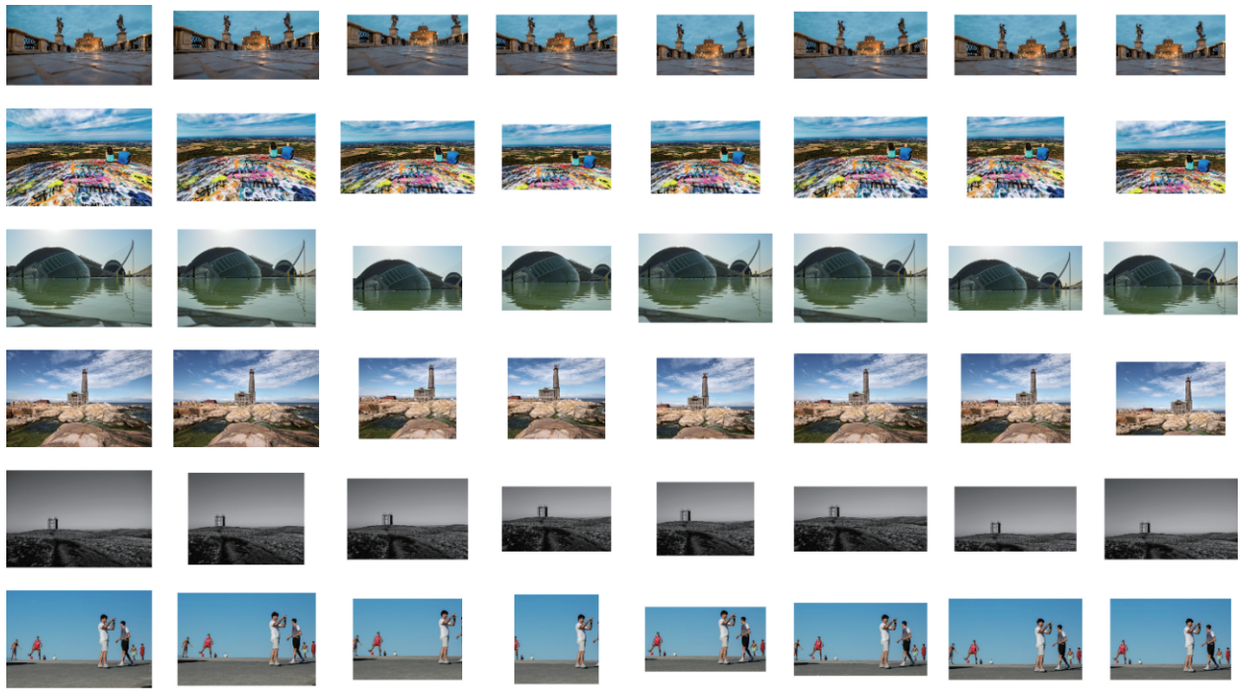


图 6 不同方法预测的最佳裁图效果对比

Fig. 6 Comparison of the best crops generated by the compared methods
((a) original images; (b) A2-RL; (c) VPN; (d) VFN; (e) VEN; (f) LVRN; (g) GAIC; (h) ours)

3.4.3 客观定量分析

为了有效对比数据驱动的主流裁图方法,将本文提出的方法与其他方法进行定量对比分析,通过大量实验证明模型预测结果的有效性。

表 4 和表 5 分别是所有对比方法在 GAICD 数据集的验证集和测试集上的实验结果。可以看出,本文提出的 3 种模型相比现有最先进的模型具有更良好的预测结果。对于排名相关性指标,平均 SRCC 和 PCC 较现有方法的最好结果分别提高了 2.0% 和 1.9%,说明本文提出的模型能更好地区分

裁图好坏并排序;对于最佳回报率指标, $Acc_{K/5}$ 和 $Acc_{K/10}$ 均有较大提升,平均提高 2.5%,最高提升 4.1%,说明模型能更好地预测美学构图最佳的裁图结果。与 GAIC 相比,本文方法的性能在各项指标上均有提高,且训练测试速度和运行效率与 GAIC 相当,这也间接证明了本文模型结构设计的有效性。

表 4　GAICD 验证集结果

Table 4　The results on the validation set of GAICD

方法	主干网络	$\overline{SRCC}$	$\overline{PCC}$	$Acc_{1/5}$ /%	$Acc_{4/5}$ /%	$Acc_{1/10}$ /%	$Acc_{4/10}$ /%	$wAcc_{1/5}$ /%	$wAcc_{4/5}$ /%	$wAcc_{1/10}$ /%	$wAcc_{4/10}$ /%
GAIC	VGG16	0.854	0.879	65.5	55.0	83.5	76.9	48.1	43.5	63.9	60.9
本文	VGG16	**0.874**	**0.898**	**68.5**	**57.5**	**85.6**	**78.4**	**49.6**	**46.2**	**65.5**	**62.0**
GAIC	MobileNetV2	0.861	0.884	67.5	56.4	85.0	78.1	48.6	44.9	64.1	62.0
本文	MobileNetV2	**0.876**	**0.902**	**70.0**	**58.9**	**87.0**	**78.9**	**51.9**	**46.7**	**66.7**	**63.9**
GAIC	ShuffleNetV2	0.863	0.886	67.0	55.8	85.5	77.9	48.7	45.2	64.4	61.9
本文	ShuffleNetV2	**0.876**	**0.901**	**69.0**	**58.1**	**86.5**	**78.7**	**52.8**	**47.1**	**66.6**	**63.3**

注:加粗字体表示主干网络相同时两种方法中的最优结果。

表 5　GAICD 测试集结果

Table 5　The results on the test set of GAICD

方法	主干网络	$\overline{SRCC}$	$\overline{PCC}$	$Acc_{1/5}$ /%	$Acc_{4/5}$ /%	$Acc_{1/10}$ /%	$Acc_{4/10}$ /%	$wAcc_{1/5}$ /%	$wAcc_{4/5}$ /%	$wAcc_{1/10}$ /%	$wAcc_{4/10}$ /%
A2-RL	AlexNet	–	–	23.2	–	39.5	–	15.1	–	25.6	–
VPN	VGG16	–	–	36.0	–	48.5	–	19.1	–	29.4	–
VFN	AlexNet	0.485	0.503	26.6	25.7	40.6	39.3	18.0	11.3	27.9	20.6
VEN	VGG16	0.616	0.662	37.5	34.2	50.5	46.4	20.2	13.4	30.1	23.3
LVRN	VGG16	0.652	0.720	40.2	38.9	54.9	51.6	26.6	21.7	33.4	28.7
GAIC	VGG16	0.842	0.866	67.2	55.9	84.0	76.8	48.2	43.2	63.7	60.8
GAIC	MobileNetV2	0.849	0.874	68.2	58.5	84.4	78.7	48.8	44.1	64.2	61.8
GAIC	ShuffleNetV2	0.850	0.872	68.0	56.6	85.8	77.8	49.2	43.2	65.1	61.4
本文	VGG16	0.852	0.876	67.5	56.4	84.8	76.9	48.9	43.5	64.7	60.7
本文	MobileNetV2	**0.862**	**0.884**	**68.8**	**58.6**	85.2	**79.4**	48.9	**44.7**	65.2	61.9
本文	ShuffleNetV2	0.860	0.881	68.6	57.2	**86.3**	78.3	**49.7**	44.0	**65.3**	**62.2**

注:加粗字体表示各列最优结果,“–”表示未列出结果。

表 6 是本文方法与其他方法在经典裁图数据集 ICDB 和 FCDB 上的 IoU 和 BDE 两项性能指标对比。按 IoU 和 BDE 的定义,IoU 值越高,效果越好;BDE 值越低,效果越好。所有方法均未使用这两个数据集的数据训练。实验结果表明,尽管本文方法在这两个数据集上的各项指标提升有限,但与现有最先进的方法相比,本文模型对大多数图像的预测结果与人为标注的最佳裁图更接近,IoU 值高,BDE 值小,在一定程度上反映出本文模型的裁图效果更符合人的主观审美判断。由于本文模型只在 GAICD 数据集上训练,跨数据集的测试结果表明本文模型具有较高的泛化性能。

3.4.4　用户主观实验

美学裁图具有很强的主观特性,为进一步验证本文所提方法在真实场景下的性能表现,进行用户主观实验。从 GAICD 测试集中随机挑选 45 幅原始图像,采用表 6 中的主干网络得到 7 种方法预测的最佳裁图结果,一轮主观实验中,邀请的 15 名受试者每次对 7 个裁图结果进行排序评分(由低到高 1～5 分),共评45 次。每名受试者进行2 轮主观实验,故一种方法的所有裁图结果共得到 1 350 次评分,最后统计各评分分值与对应次数占比的乘积得

表 6　不同方法在 ICDB 和 FCDB 数据集上的 IoU 和 BDE 指标比较

Table 6　Comparison of IoU and BDE on the ICDB and FCDB datasets among different methods

方法	主干网络	ICDB set 1		ICDB set 2		ICDB set 3		FCDB	
		IoU ↑	BDE ↓	IoU ↑	BDE ↓	IoU ↑	BDE ↓	IoU ↑	BDE ↓
A2-RL	AlexNet	0.802	**0.052**	0.796	0.054	0.790	**0.054**	0.663	0.089
VPN	VGG16	0.802	–	0.791	–	0.778	–	0.711	0.073
VFN	AlexNet	0.785	0.058	0.776	0.061	0.760	0.065	0.684	0.084
VEN	VGG16	0.781	–	0.770	–	0.753	–	**0.735**	0.072
LVRN	VGG16	0.799	0.060	0.805	0.058	0.795	0.060	0.734	0.068
GAIC	VGG16	0.809	0.059	0.807	**0.051**	0.799	0.057	0.675	0.066
本文	VGG16	**0.821**	**0.052**	**0.819**	0.052	**0.807**	0.056	0.723	**0.064**

注：加粗字体表示各列最优结果，“–”表示未列出结果，“↑”表示指标值越高越好，“↓”表示指标值越低越好。

表 7　不同方法的用户主观实验

Table 7　User study to validate each method

方法	1 分	2 分	3 分	4 分	5 分	加权平均分
A2-RL	143	273	405	383	146	3.086
VPN	316	351	300	283	100	2.630
VFN	387	350	294	237	82	2.464
VEN	47	214	370	464	255	3.493
LVRN	25	174	453	500	198	3.498
GAIC	16	152	376	523	283	3.670
本文	2	84	288	595	381	**3.940**

注：加粗字体表示最优结果，第 2—6 列表示各评分分值对应的次数。

到加权平均分，作为 7 种方法的主观评分结果，实验结果如表 7 所示。可以看出，在所有对比的裁图方法中，本文方法能产生最符合人们主观审美的裁剪效果。由于人们在日常生活中对美的评判标准存在较大差异（祝汉城 等，2021），本文裁图方法在一定程度上能够满足大众化图像美学评价标准，但对精确到个体的审美偏好（个性化美学评价）仍然存在局限性。

4　结　论

本文提出一种聚合细粒度特征的深度注意力自动裁图方法 DAIC-Net，基于空间金字塔思想和深度视觉注意力机制，逐级增强多尺度区域特征，融合全局和局部注意力特征，由粗到细地增强上下文语义信息表征。DAIC-Net 主要由通道校准的语义特征提取、细粒度特征聚合和上下文注意力融合模块组成。通道校准的语义特征提取模块用于获取图像的深度语义特征，细粒度特征聚合模块增强互补语义信息并产生富含图像构成和空间位置信息的特征表示，上下文注意力融合模块在此基础上引入更多关系依赖信息，进一步增强对裁图特征的学习。此外，本文定义了多项损失函数以指导模型多任务监督学习。大量实验结果验证了 DAIC-Net 的有效性，跨数据集测试结果进一步表明了 DAIC-Net 良好的泛化能力，模型预测结果与专家标注结果更为接近。

但是，基于美学的自动裁图任务的准确率还有进一步的提升空间。解决裁图数据不足和稀疏标注问题、提高客观评价指标合理性、减少图像形变带来的精度损失，以及结合多任务训练策略进一步提高模型性能、基于个性化美学评价方法的候选裁图框生成策略等，都是未来的研究方向。

参考文献（References）

Byeon W, Breuel T M, Raue F and Liwicki M. 2015. Scene labeling with LSTM recurrent neural networks//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3547-3555 [DOI: 10.1109/CVPR.2015.7298977]

Chen J S, Bai G C, Liang S H and Li Z Q. 2016a. Automatic image cropping: a computational complexity study//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 507-515 [DOI: 10.1109/CVPR.2016.61]

Chen L C, Yang Y, Wang J, Xu W and Yuille A L. 2016b. Attention to

scale: scale-aware semantic image segmentation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3640-3649 [DOI: 10.1109/CVPR.2016.396]

Chen Q Y, Zhang W, Zhou N, Lei P, Xu Y, Zheng Y and Fan J P. 2020. Adaptive fractional dilated convolution network for image aesthetics assessment//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14102-14111 [DOI: 10.1109/CVPR42600.2020.01412]

Chen Y L, Huang T W, Chang K H, Tsai Y C, Chen H T and Chen B Y. 2017a. Quantitative analysis of automatic image cropping algorithms: a dataset and comparative study//Proceedings of 2017 IEEE Winter Conference on Applications of Computer Vision. Santa Rosa, USA: IEEE: 226-234 [DOI: 10.1109/WACV.2017.32]

Chen Y L, Klopp J, Sun M, Chien S Y and Ma K L. 2017b. Learning to compose with professional photographs on the web//Proceedings of the 25th ACM international conference on Multimedia. Mountain View, USA: ACM: 37-45 [DOI: 10.1145/3123266.3123274]

Dai J F, He K M and Sun J. 2016. Instance-aware semantic segmentation via multi-task network cascades//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3150-3158 [DOI: 10.1109/cvpr.2016.343]

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Fang C, Lin Z, Mech R and Shen X H. 2014. Automatic image cropping using visual composition, boundary simplicity and content preservation models//Proceedings of the 22nd ACM International Conference on Multimedia. Orlando, USA: ACM: 1105-1108 [DOI: 10.1145/2647868.2654979]

Fang Y M, Sui X J, Yan J B, Liu X L and Huang L P. 2021. Progress in no-reference image quality assessment. Journal of Image and Graphics, 26(2): 265-286 (方玉明, 眭相杰, 鄢杰斌, 刘学林, 黄丽萍. 2021. 无参考图像质量评价研究进展. 中国图象图形学报, 26(2): 265-286) [DOI: 10.11834/jig.200274]

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

Graves A, Jaitly N and Mohamed A R. 2013. Hybrid speech recognition with deep bidirectional LSTM//Proceedings of 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Olomouc, Czech Republic: IEEE: 273-278 [DOI: 10.1109/ASRU.2013.6707742]

He K M, Gkioxari G, Dollár P and Girshick R. 2020. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 386-397 [DOI: 10.1109/TPAMI.2018.2844175]

He K M, Zhang X Y, Ren S Q and Sun J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1904-1916 [DOI: 10.1109/TPAMI.2015.2389824]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/cvpr.2018.00745]

Jian M W, Lam K M, Dong J Y and Shen L L. 2015. Visual-patch-attention-aware saliency detection. IEEE Transactions on Cybernetics, 45(8): 1575-1586 [DOI: 10.1109/TCYB.2014.2356200]

Kingma D P and Ba J. 2015. Adam: a method for stochastic optimization [EB/OL]. [2021-06-05]. https:// https://arxiv.org/pdf/1412.6980v8.pdf

Kong S, Shen X H, Lin Z, Mech R and Fowlkes C. 2016. Photo aesthetics ranking network with attributes and content adaptation//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 662-679 [DOI: 10.1007/978-3-319-46448-0_40]

Li D B, Wu H K, Zhang J G and Huang K Q. 2018. A2-RL: aesthetics aware reinforcement learning for image cropping//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8193-8201 [DOI: 10.1109/CVPR.2018.00855]

Li L D, Zhu H C, Zhao S C, Ding G G and Lin W S. 2020. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. IEEE Transactions on Image Processing, 29: 3898-3910 [DOI: 10.1109/TIP.2020.2968285]

Liang X D, Shen X H, Xiang D L, Feng J S, Lin L and Yan S C. 2016. Semantic object parsing with local-global long short-term memory//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3185-3193 [DOI: 10.1109/CVPR.2016.347]

Liu N, Han J W and Yang M H. 2018. PiCANet: learning pixel-wise contextual attention for saliency detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: 3089-3098 [DOI: 10.1109/CVPR.2018.00326]

Lu W R, Xing X F, Cai B L and Xu X M. 2019. Listwise view ranking for image cropping. IEEE Access, 7: 91904-91911 [DOI: 10.1109/ACCESS.2019.2925430]

Lu X, Lin Z, Jin H L, Yang J C and Wang J Z. 2014. RAPID: rating pictorial aesthetics using deep learning//Proceedings of the 22nd ACM International Conference on Multimedia. Orlando, USA:

ACM: 457-466 [DOI: 10.1145/2647868.2654927]

Ma N N, Zhang X Y, Zheng H T and Sun J. 2018. ShuffleNet V2: practical guidelines for efficient CNN architecture design//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 116-131 [DOI: 10.1007/978-3-030-01264-9_8]

Mai L, Jin H L and Liu F. 2016. Composition-preserving deep photo aesthetics assessment//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 497-506 [DOI: 10.1109/CVPR.2016.60]

Murray N, Marchesotti L and Perronnin F. 2012. AVA: a large-scale database for aesthetic visual analysis//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 2408-2415 [DOI: 10.1109/CVPR.2012.6247954]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Sandler M, Howard A, Zhu M L, Zhmoginov A and Chen L C. 2018. MobileNetV2: inverted residuals and linear bottlenecks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4510-4520 [DOI: 10.1109/CVPR.2018.00474]

Sheng K K, Dong W M, Chai M L, Wang G H, Zhou P, Huang F Y, Hu B G, Ji R R and Ma C Y. 2020. Revisiting image aesthetic assessment via self-supervised feature learning//Proceedings of the AAAI Conference on Artificial Intelligence, 34(4): 5709-5716 [DOI: 10.1609/aaai.v34i04.6026]

Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2021-06-05]. https://arxiv.org/pdf/1409.1556v1.pdf

Stentiford F. 2007. Attention based auto image cropping//Proceedings of the 5th International Conference on Computer Vision Systems. Bielefeld, Germany: [s. n.] [DOI: 10.2390/biecoll-icvs2007-148]

Suh B, Ling H B, Bederson B B and Jacobs D W. 2003. Automatic thumbnail cropping and its effectiveness//The 16th Annual ACM Symposium on User Interface Software and Technology. Vancouver, Canada: ACM: 95-104 [DOI: 10.1145/964696.964707]

Tang X O, Luo W and Wang X G. 2013. Content-based photo quality assessment. IEEE Transactions on Multimedia, 15(8): 1930-1943 [DOI: 10.1109/TMM.2013.2269899]

Varior R R, Shuai B, Lu J W, Xu D and Wang G. 2016. A siamese long short-term memory architecture for human re-identification//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 135-153 [DOI: 10.1007/978-3-319-46478-7_9]

Visin F, Kastner K, Cho K, Matteucci M, Courville A and Bengio Y. 2015. ReNet: a recurrent neural network based alternative to convolutional networks [EB/OL]. [2021-06-05]. https://arxiv.org/pdf/1505.00393.pdf

Wei Z J, Zhang J M, Shen X H, Lin Z, Mech R, Hoai M and Samaras D. 2018. Good view hunting: learning photo composition from dense view pairs//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5437-5446 [DOI: 10.1109/CVPR.2018.00570]

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R S and Bengio Y. 2015. Show, attend and tell: neural image caption generation with visual attention//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org: 2048-2057

Yan J B, Zhong Y, Fang Y M, Wang Z Y and Ma K D. 2021. Exposing semantic segmentation failures via maximum discrepancy competition. International Journal of Computer Vision, 129(2): 1768-1786 [DOI: 10.1007/s11263-021-01450-2]

Yan J Z, Lin S, Kang S B and Tang X O. 2013. Learning the change for automatic image cropping//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 971-978 [DOI: 10.1109/CVPR.2013.130]

Yu F and Koltun V. 2015. Multi-scale context aggregation by dilated convolutions//Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico: [s. n.]

Zeiler M D and Fergus R. 2014. Visualizing and understanding convolutional networks//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 818-833 [DOI: 10.1007/978-3-319-10590-1_53]

Zeng H, Li L D, Cao Z S and Zhang L. 2020. Grid anchor based image cropping: a new benchmark and an efficient model. IEEE Transactions on Pattern Analysis and Machine Intelligence: #3024207 [DOI: 10.1109/TPAMI.2020.3024207]

Zhang L M, Song M L, Zhao Q, Liu X, Bu J J and Chen C. 2013. Probabilistic graphlet transfer for photo cropping. IEEE Transactions on Image Processing, 22(2): 802-815 [DOI: 10.1109/TIP.2012.2223226]

Zhang W X, Ma K D, Yan J, Deng D X and Wang Z. 2020. Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology, 30(1): 36-47 [DOI: 10.1109/TCSVT.2018.2886771]

Zhu H C, Li L D, Wu J J, Zhao S C, Ding G G and Shi G M. 2020. Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. IEEE Transactions on Cybernetics: 1-14 [DOI: 10.1109/TCYB.2020.2984670]

Zhu H C, Zhou Y, Li L D, Zhao J Q and Du W L. 2021. Recent progress and tend of personalized image aesthetics assessment [J/OL]. Journal of Image and Graphics[DOI:10.11834/jig.210211] (祝汉

城，周勇，李雷达，赵佳琦，杜文亮. 2021. 个性化图像美学评价的研究进展与趋势[J/OL]. 中国图象图形学报[DOI:10.11834/jig.210211])

作者简介

方玉明，1984 年生，男，教授，主要研究方向为计算机视觉、多媒体信号处理和视觉质量评估。
E-mail：leo.fangyuming@foxmail.com

钟裕，男，硕士研究生，主要研究方向为图像美学评价和自动裁图。E-mail：zhystu@qq.com
鄢杰斌，男，博士研究生，主要研究方向为计算机视觉。
E-mail：jiebinyan@foxmail.com
刘丽霞，女，硕士研究生，主要研究方向为图像处理。
E-mail：shrimp.liu@qq.com