Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
E-mail: jzus@zju.edu.cn

智能多媒体内容设计在阿里巴巴的应用

刘奎龙，李为，杨昌源†，杨光

阿里巴巴集团，中国杭州，311121

†E-mail: kuilong.lkl@alibaba-inc.com; pangeng.lw@alibaba-inc.com; hangyuan.yangcy@alibaba-inc.com; qingyun@taobao.com

投稿日期：2019-10-23；录用日期：2019-12-15；Crosschecked：2019-12-15

摘要：多媒体内容是阿里巴巴业务生态中必不可少的组成部分，且需求量巨大。多媒体内容生产通常具有较高技术及资金要求。随着人工智能技术近年飞速发展，众多辅助多媒体内容生产的工具应运而生，人工智能技术与多媒体内容设计的结合在阿里巴巴业务生态中的应用愈加广泛，涉及领域包括辅助设计、平面设计、视频生成和页面制造。本文首先介绍了在阿里巴巴业务生态中人工智能辅助设计工具的通用处理流程，然后在上述 4 个应用领域分别选择一个代表性工具着重介绍。通过这些工具的使用，多媒体内容设计结合人工智能带来的价值在业务中得到很好验证，体现了人工智能技术在促进多媒体内容生产中起到的巨大作用，也预示了其广泛应用前景。

关键词：多媒体内容；阿里巴巴；人工智能；设计；业务应用

本文译自 Liu KL, Li W, Yang CY, et al., 2019. Intelligent design of multimedia content in Alibaba. *Front Inform Technol Electron Eng*, 20(12):1657-1664.
https://doi.org/10.1631/FITEE.1900580
中图分类号：TP391

1 介绍

在阿里巴巴业务生态中，多媒体内容形式多种多样，通常以图像、声音、文本等为基本要素，表现为广告图片、图文海报、商品视频，以及包含以上多种载体的详情页、活动页等网络页面。众多商品以及大规模促销活动对多媒体内容有着强烈的需求。2018 年双十一期间，平面设计需求量达到上亿张。淘宝网生成视频数据上亿条，其中包括热点视频数千万条，日播放量数十亿。

然而，多媒体内容的生产和传播存在诸多问题。以电商短视频为例，视频制作需要经过拍摄（棚拍、街拍）、素材制作、视频剪辑、字幕音效录制、内容审核及产品发布等多个步骤。其程序复杂，场景分散，人员分工细致，设备种类繁多，全部流程需要大量人力、物力以及时间投入，绝大部分商家无法承担如此高昂的成本。

随着人工智能技术的飞速发展，其部分分支领域日趋成熟。人工智能技术在多媒体内容理解技术方面的应用在学术界和工业界均受到广泛关注（Peng et al., 2019）。众多多媒体内容分析生产辅助工具应运而生，多媒体内容的生产成本逐步降低。在图像领域，深度学习技术的分类（Simonyan and Zisserman, 2015; He et al., 2016; Chollet, 2017）、定位（Zhou et al., 2016）、检测（Lin et al., 2017; Ren et al., 2017）,以及分割（He et al., 2017；Chen et al., 2018）技术的发展使得机器可以深入理解素材乃至平面内容设计结构。对抗生成网络（GAN）（Goodfellow et al., 2014）的快速发展，使得字体生成（Azadi et al., 2018）、素材风格迁移（Zhu et al., 2017）、人体姿态迁移（Song et al., 2019）等众多生成技术成为可能。文本是多媒体内容中传递重要信息的媒介，文本内容理解

‡ 通讯作者
ORCID: 刘奎龙, http://orcid.org/0000-0001-9726-8369；杨昌源, http://orcid.org/0000-0003-0065-6272

和摘要提取等自然语言处理技术的应用，能够降低视频匹配字幕的成本。对音乐风格迁移、音乐生成（Bretan et al., 2016）以及音效联动技术的探索，可以使音乐的选择、音效的调整更加灵活方便。除了对单一媒体的研究以外，人工智能技术在多模态信息处理中的应用同样受到广泛关注（Peng et al., 2017, 2018）。基于多模态信息的特征学习（Ngiam et al., 2011）使得视频内容的理解更加准确。跨媒体关系的建模（Huang and Peng, 2019）使得多媒体信息的使用更加便利和广泛。随着人工智能技术日趋成熟，其与传统设计的结合将成为必然趋势。

2 阿里巴巴多媒体内容生成

在阿里巴巴的多项业务应用中，人工智能技术辅助多媒体内容设计及生产的工具已大量落地并成功应用。将人工智能技术与多媒体内容设计相结合，其生产流程大体可分为 5 个环节，如图 1 所示。

1. 分析

该环节主要根据参数、脚本等配置信息对原始素材进行解析。作为基础环节，它在整个流程中起到了至关重要的作用，也是人工智能技术各原子能力应用较多的一个环节，如图像的分类、检测、分割，文本的识别与定位等。在实际应用中，这些算法通常需要根据实际应用场景进行订制化改进与实现。

2. 处理

在理解输入素材的基础上，该环节基于素材的情况及业务需求对其进行必要的处理加工，如图像的美化、修复、抠图，视频的剪切等。从而将原始素材处理成可进一步使用的原子素材。

3. 生成

基于结构化的素材信息，结合实际需求，时下研究热点——GAN 技术在本环节得到大量应用。包括基于图像或视频的文本生成、基于视频流的音乐生产以及字体生成等技术均得到不同程度的应用。

4. 渲染

本环节根据预设的脚本和参数，有效融合所有可使用的素材信息，使文字、声音、图像、视频流等元素的展示协调一致，生成更富有表现力的广告、视频、页面、动效等产品，达到助力商业的目的。

5. 评估

该环节用于对多媒体内容生产的最终结果进行评价。评价指标既包括图像美学、图像质量等算法指标，也包括商业应用指标，生成结果的用户使用率是比较常用的商业指标之一。评估结果对全流程各环节的优化起到积极作用，而对多媒体内容的生产通常没有直接影响。基于对生成的多媒体内容和商业结果数据的分析，开发人员可以对相应模块进行迭代优化以达到更好的商业结果。为简单起见，下文中的所有图片均不再显示评估模块。

基于以上流程的 AI 辅助设计生产工具可大体分为两类：AI 辅助设计生产资料生产工具和 AI 辅助设计生产力工具。前者的产品为设计素材，作为整个生产过程的一部分，通常应用于全流程的某环节。后者的产品通常为直接面向用户的广告、视频、页面等，所以通常包含上述所有环节。

2.1 AI 辅助设计生产资料生产工具

随着人工智能技术的发展，以各种人工智能技术为基础的辅助设计工具得到广泛应用，如辅

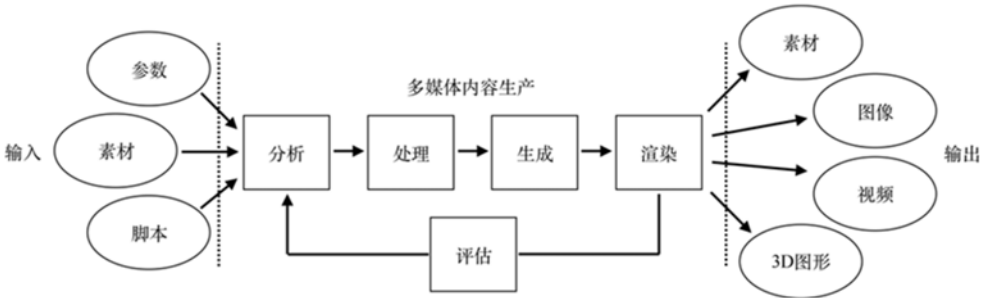

图 1 智能多媒体内容生产流程

助平面设计的海报自动排版工具、自动字体生成工具等。作为平面设计流程中较为基础的一步，图片抠图是将商家上传的数以亿万计的广告图片转化为创意素材的必经之路。然而其工作量大、技术含量低，不仅加重设计师的负担，且令企业耗时耗资不菲。

人工智能技术在图像分割中的成功应用使图片抠图自动化成为可能。顽兔抠图产品抓住这一机遇，只需要用户输入一张分辨率为800×800的RGB图片，就可在2秒钟内将抠图结果返回用户。若一次抠图请求中包含20张图片，并行处理可使单张图片的平均计算时间降至 100 毫秒。作为图片抠图的一站式解决方案，当自动抠图算法无法自动给出可用抠图结果时，顽兔抠图产品同时提供一个交互式抠图工具，用户只需使用前景画笔和背景画笔在图像上进行简单标记，人工智能即可以根据画笔位置在一秒钟之内计算出目标主体边界，用户还可根据实际情况补充画笔以纠正错误的边界位置。除了上述智能画笔以外，顽兔产品还提供了非智能的修补笔和橡皮，可以简单快速地修正自动抠图结果中可能存在的微小错误。

面对电商场景下数以万计的商品品类和多变形态，基于深度学习的显著性分割技术可以识别出图片中主体的像素位置，如图 2 所示。针对深度学习算法分割结果在边缘上过于平滑的问题，结合像素颜色、边缘梯度等信息的传统图像处理技术可以使边缘像素的定位更加准确，如改进的SLIC算法（Kim et al., 2013），grabcut算法（Rother et al., 2004），watershed算法（Bradski and Kaehler, 2008），alpha matting算法（Levin et al., 2007）等。经过边缘优化，抠图结果的边缘过渡更加自然，可用率更高。基于上述算法，顽兔提供了一个可以对图片进行并行处理的 API 调用方式，用户只需提供图片网址列表即可以在短时间内快速完成大量图片的抠图任务。

由于待抠图片的多样性和随机性，没有任何一种抠图算法可以解决所有抠图问题。当在促销活动中数以百万计的图片经过自动抠图技术获得抠图结果后，通常需要对其进行逐一审核以滤除无法使用的抠图结果，在短时间内完成这样的审核任务通常需要大量人力资源的投入。不同于互联网上的其它抠图产品（如 http://www.remove.bg 和 https://www.gaoding.com），顽兔抠图产品考虑了阿里巴巴平台生态中的多种应用场景，为每一个自动抠图结果提供一个置信度分数。该分数基于深度学习模型的结果和边缘优化的结果计算得到。更高的置信度分数通常意味着抠图结果可使用的概率更大。用户可以简单设置一个合适的置信度阈值以滤除大部分不可使用的抠图结果。如需对抠图结果进行人工审核，将待审图片按置信度分数降序逐一审核，可以在审核初期获得较高的可使用抠图结果检出率，这意味着可以用更少的时间获得足够多的可使用抠图结果数量。

自从顽兔产品上线服务以来，先后渗透到集团服务市场、鹿班、飞猪、国际业务、淘宝等多个业务。服务调用量稳步增长，半年累积超过500万次。从某种程度上说，顽兔产品为智能设计全链路自动化奠定了基础。

2.2 AI辅助设计生产力工具

2.1.1 平面智能设计

在平面设计领域，鹿班无疑是阿里巴巴生态中的佼佼者，在2018年双十一期间，其日请求峰值超过5000万，热促期间累积平面设计生成量达到近5亿张。

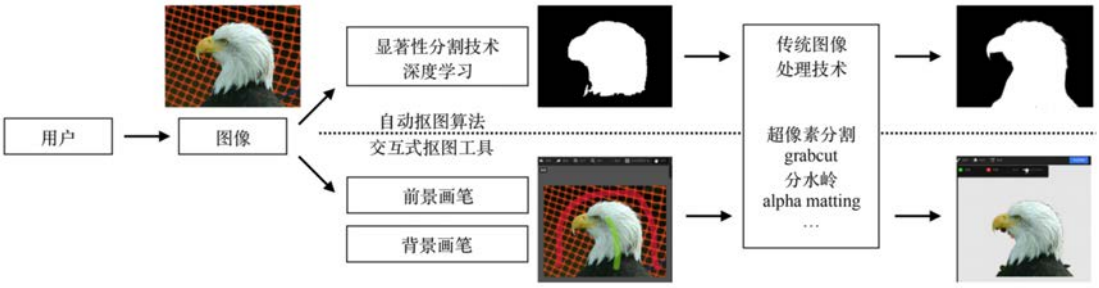

图 2 顽兔的处理流程

鹿班在平面设计上实现了全流程自动化，与https://www.canva.cn和https://www.gaoding.com不同，鹿班不仅是一个平面设计的生产工具，还是一个平面设计的投放工具。鹿班与阿里巴巴商业生态紧密结合，基于对淘宝大数据以及用户行为的深入分析，能够基于不同用户的喜好，自动生成包含不同商品的个性化广告并投放给用户，实现“千人千面”。

在鹿班网站，用户只需要提供商品的图片并简单设置需求分辨率和所属行业，即可生成几种不同风格的商品广告。用户可选择下载最喜欢的生成结果，也可直接将其发布到线上使用。

鹿班创新性地推出了基于场景图的全画幅设计方案。全画幅设计的核心思想是将平面设计分解为场景图设计和文案组团设计，如图 3 所示。场景图设计针对用户输入的场景图片，使用图像分类、检测、生成等多项图像处理技术生成符合生成要求的场景图设计背景。在此基础上，结合设计师预设的文案组团设计，及用户输入的文案信息，依据设计协议进行搭配，生成最终的个性化平面设计。

场景图通常置于一定真实情境中，信息完整且表现力更为饱满。在整个双十一期间，与生成自传统白底图模板的平面设计相比，全画幅场景设计方案的点击率超出 30%以上，结合文案的智能生成技术，总体提升率超过 90%。

2.2.2 视频智能生成

随着视频内容理解技术的发展，智能设计的边界得到拓展，面向视觉的内容生成能力进一步提升。计算机视觉中图像分类、商品检测、人脸检测（Zhang et al., 2017）、肢体检测（Xia et al., 2017; Papandreou et al., 2018）、运动检测（Cao et al., 2017）、摄像机追踪（Ristani and Tomasi, 2018）等技术的发展，使计算机能够更加准确地理解“镜头语言”，捕捉镜头拍摄手法（推、拉、平移），识别拍摄环境（棚拍、街拍，近景、远景），判断场景切换位置等。众多技术构筑了视频生成的实现基础。

业界已经产生了大量视频智能生成工具，如https://gliacloud.com, https://zenvideo.cn 等。这些工具大多专注于基于文案的视频生成。通过分析用户提供的文案，在产品数据库或互联网上搜索匹配的图片或视频，并基于预设模板进行视频合成。而在阿里巴巴生态中，视频合成的素材类型往往更加多样，包括图片、视频片段、音乐甚至关于商品细节的描述字句。为了解决这种基于复杂素材类型的视频自动合成问题，Alibaba Wood 应运而生，它是阿里巴巴推出重要视频生成工具。

在 Alibaba Wood 网站，用户可以提供一个商品详情页，或者一组视频片段，在简单设置视频节奏类型、时长和分辨率信息后，即可得到一段自动合成的短视频。Alibaba Wood 还提供了一个线上交互式视频编辑工具，用户可根据自己的意愿修改自动合成的视频结果。在大规模应用中，Alibaba Wood 提供了 API 的调用方式，可以对用户提供的详情页列表进行并行的批量处理。

淘宝网商品详情页上的所有素材都可以作为 Alibaba Wood 的原始输入素材，如图 4 所示。由于素材形式的复杂性和分析需求的多样性，素材分析是一个相对耗时的过程。在复杂度适中的情况下，该过程通常需要使用 5－10 秒钟时间。基

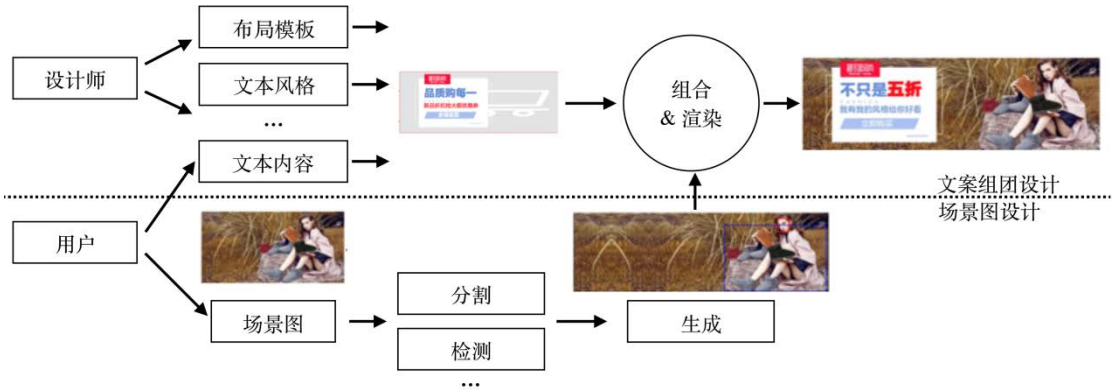

图 3 鹿班的全画幅设计方案

于对输入素材的分析结果，Alibaba Wood 可以理解输入图像或视频中的内容，如主体属性、运动规律、镜头语言等，从而剪切视频，选取图片，并根据用户的使用场景、视频叙事逻辑以及预设脚本协议（如由远景外观推进镜头至细节亮点的聚焦），将得到的关键帧或关键视频片段通过各种动效模板有序连接起来。在渲染过程中，Alibaba Wood 使用了浏览器渲染技术和分段并行渲染技术，可以有效降低视频合成时间。以一个帧分辨率为 700×700，时长为 15 秒的短视频合成为例，简单动效模板的合成时间通常低于 5 秒。如果动效模板较为复杂，通过使用 OpenGL 技术可以将合成时间控制在 10 秒之内。

除了基于图像的人工智能技术，Alibaba Wood 也使用了文本和声音相关的人工智能技术。在视频合成开始之前，Alibaba Wood 能够基于关键帧或其它图像素材自动生成相关描述文案，以供视频合成时适时插入。文案生成过程由两部分组成，首先由商品卖点标签生成商品卖点描述，再由商品卖点描述生成商品推荐理由。商品卖点标签可以通过分析输入图片素材得到，而文案的生成可以通过使用 transformer 模型实现（Vaswani et al., 2017）。

Alibaba Wood 对背景音乐的选择及影响也进行了研究与探索。对比实验表明，合适的背景音乐可以有效提升消费者的购买欲望。当从原始素材中识别出目标商品时，Alibaba Wood 可以从声音数据库中选出一段最为匹配目标商品风格的旋律作为背景音乐，并调整动效模板节奏，使其与旋律节奏一致。Alibaba Wood 也可以自动生成背景音乐。通过使用基于 LSTM（Hochreiter and Schmidhuber, 1997）的 GAN 技术，Alibaba Wood 可以由一段白噪声生成一段饱含指定情感类型的音乐。

自从 Alibaba Wood 上线服务以来，累计为商家产出 2 千多万条短视频，并使得合作商家森马服饰的视频制作成本降低 90%，制作效率提升 95%，两周内效益增加超过 70 万元。

2.2.3 页面内容制造

商品展示网络页面设计，是包含平面图像、3D 图形、视频、音频、文案在内的数字内容的综合智能制造。在电商场景下，活动页面、详情页面均可包含多种形式的信息。面对数以亿万计的商品细节展示需求，页面智能化的设计和生产将成为必然趋势。

AI-detail 作为一种详情页智能生成工具，可以融合平面智能设计和视频智能设计的多种元素，将包括文案、视频、图像等多种要素融入到自动生成页面的流程之中，其处理流程如图 5 所示。不同于阿里巴巴集团外的其它详情页生成工具（如 http://www.deepdraw.cn），AI-detail 深植于阿里巴巴的平台生态中，它的输入数据源和发布路径均与应用场景紧密连接。基于线上使用的商品详情页，AI-detail 可以并行生成大量的新详情页。个人用户也可以编写脚本，使用 AI-detail 提供的 API 接口调用产品功能。

如图 5 所示，AI-de-tail 可以通过使用人工智能技术对现有详情页进行结构化解析，并使用新语言或新风格模板自动批量重构新详情页。如果详情页合成的素材来源于对现有详情页的分析和提取，那么分析环节的时间消耗与前述 Alibaba Wood 基本一致，且占据了详情页自动生成的绝大部分时间。

当详情页作为输入数据时，AI-detail 首先对其进行分割以获得所有组成元素，包括视频、图像位置、文本和表格等。基于分割出的各种素材，AI-detail 可以识别品牌和场景，多角度分析模特姿态，捕捉商品细节。通过 OCR 技术， AI-detail 不

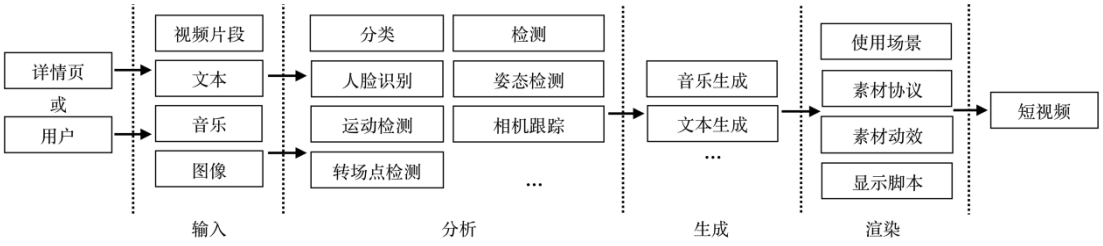

图 4 Alibaba Wood 处理流程

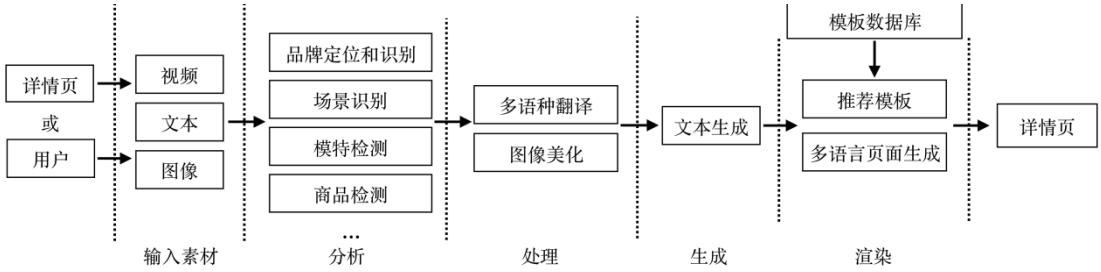

图 5 AI-detail 处理流程

仅可以识别详情页上的文字，也可以识别图片和视频帧中的文字。对详情页信息的结构化存储是 AI-detail 的核心功能。对商品详情页大数据的深入分析可以在电商平台、商品乃至相关行业的深入分析和相关商业决策中发挥重要作用。

在详情页重构应用中，基于由现有详情页获得的素材信息，AI-detail 可以采取美化素材图像、自动生成文案、多语言翻译等多种处理手段，最后结合模板数据库中推荐的模板和投放语言环境，生成新的商品展示详情页。在跨境重构的业务场景中，AI-detail 通过对中文页面进行智能结构化解析，并基于为投放场景订制的模板重构数据库页面、翻译文字至当地语言。上线半年以来，已有效支撑 100 多万商品的境外投放。

3 未来

在阿里巴巴，未来的多媒体内容生成有两个潜在的发展方向，即个性化内容设计和情感计算。

在当前阶段，多媒体内容设计的决策仍大多依赖于设计师的专业能力。机器的优势在于能够快速批量实现设计迁移，可大幅降低规模化设计成本。不同于 2.2.1 节中鹿班的“千人千面”，个性化设计不仅局限于展示的内容，其它的设计元素（如显示风格、模板布局等）均应按照具体用户的偏好实现不同的生产和制造。多媒体内容制作平台（如鹿班、Alibaba Wood）在线上运营过程中均应能够依据商品类目、用户所属人群、地域等信息以用户喜欢的设计风格推荐和展示商品。

直播场景下，情绪（语义、语气、表情）表现更加积极的主播能够带来更多观看量及商品购买量，影视剧热点片段通常具有更强烈的情感表达。情感因素与商业的潜在联系促进了情感计算相关技术的快速发展，基于文本的情感倾向性分析、基于语音的语气情感分析、基于图像的面部表情分析等均引起了工业界的广泛关注。在多媒体内容制作过程中，加入场景化、情感化的分析及表达将成为一个非常值得探索的方向。结合深度图像分析技术和场景 3D 建模技术，在不同场景（办公、家庭、睡眠、运动）的视频片段中可以植入不同的产品。

遵守伦理准则声明

作者声明发表这篇论文没有利益冲突。

参考文献

Azadi S, Fisher M, Kim VG, et al., 2018. Multi-content GAN for few-shot font style transfer. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.7564-7573.

Bradski G, Kaehler A, 2008. Learning OpenCV: Computer Vision with the OpenCV Library. O’Reilly Media, Inc.

Bretan M, Weinberg G, Heck L, 2016. A unit selection methodology for music generation using deep neural networks. https://arxiv.org/abs/1612.03789

Cao Z, Simon T, Wei SE, et al., 2017. Realtime multi-person 2D pose estimation using part affinity fields. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.7291-7299. https://doi.org/10.1109/CVPR.2017.143

Chen LC, Zhu YK, Papandreou G, et al., 2018. Encoderdecoder with atrous separable convolution for semantic image segmentation. https://arxiv.org/abs/1802.02611

Chollet F, 2017. Xception: deep learning with depthwise separable convolutions. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1251-1258. https://doi.org/10.1109/CVPR.2017.195

Goodfellow IJ, Pouget-Abadie J, Mirza M, et al., 2014. Generative adversarial nets. Proc 27th Int Conf on Neural Information Processing Systems, p.2672-2680.

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

He KM, Gkioxari G, Dollàr P, et al., 2017. Mask R-CNN. Proc IEEE Int Conf on Computer Vision, p.2961-2969. https://doi.org/10.1109/ICCV.2017.322

其余文献从略