Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Report:

Intelligent design of multimedia content in Alibaba

Kui-long LIU, Wei LI, Chang-yuan YANG‡, Guang YANG

Alibaba Group, Hangzhou 311121, China

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

Received Oct. 23, 2019; Revision accepted Dec. 15, 2019; Crosschecked Dec. 15, 2019

Abstract: Multimedia content is an integral part of Alibaba's business ecosystem and is in great demand. The production of multimedia content usually requires high technology and much money. With the rapid development of artificial intelligence (AI) technology in recent years, to meet the design requirements of multimedia content, many AI auxiliary tools for the production of multimedia content have emerged and become more and more widely used in Alibaba's business ecology. Related applications include mainly auxiliary design, graphic design, video generation, and page production. In this report, a general pipeline of the AI auxiliary tools is introduced. Four representative tools applied in the Alibaba Group are presented for the applications mentioned above. The value brought by multimedia content design combined with AI technology has been well verified in business through these tools. This reflects the great role played by AI technology in promoting the production of multimedia content. The application prospects of the combination of multimedia content design and AI are also indicated.

Key words: Multimedia content; Alibaba; Artificial intelligence; Design; Business application
https://doi.org/10.1631/FITEE.1900580 **CLC number:** TP391

1 Introduction

In Alibaba's business ecosystem, multimedia content is in various forms, including advertising images, graphic posters, product videos, and detailed pages. These forms usually consist of images, sounds, and texts. Numerous commodities and large-scale promotions have a strong demand for multimedia content. During the "Double 11" (Nov. 11) promotion in 2018, the demand for graphic design was over 100 million. Taobao generated hundreds of millions of videos, including tens of thousands of hotspot videos. Billions of videos were broadcast every day during that time.

However, there are many problems in the production and dissemination of multimedia content. Taking the production of an e-commerce short video as an example, the production process includes multiple steps, such as live-action shooting, material production, video clipping, subtitle production, sound recording, content auditing, and product release. It requires initial investment on different types of equipment as well as on human resources. The shooting process may need to be conducted at different venues. Most businesses cannot afford its lead time and high cost.

With the rapid development of AI technology, some branches of AI have become mature. The application of AI in the understanding of multimedia content has been of considerable interest to academia and industry (Peng et al., 2019). Many auxiliary tools for the analysis and production of multimedia content have emerged. This has brought a reduction in the cost of multimedia content production. The deep learning technique in image classification (Simonyan and Zisserman, 2015; He et al., 2016; Chollet, 2017), localization (Zhou et al., 2016), detection (Lin et al., 2017; Ren et al., 2017), and

‡ Corresponding author
ORCID: Kui-long LIU, http://orcid.org/0000-0001-9726-8369; Chang-yuan YANG, http://orcid.org/0000-0003-0065-6272

segmentation (He et al., 2017; Chen et al., 2018) allows machines to understand material and even the design structure of planar content. The generative adversarial net (GAN) (Goodfellow et al., 2014) makes font generation (Azadi et al., 2018), material style transfer (Zhu et al., 2017), human posture transfer (Song et al., 2019), and many other generation techniques possible. Text is an important medium for transmitting information in multimedia content. The applications of natural language processing techniques, such as text content understanding and abstract extraction, reduce the cost of video subtitling. Music style transfer, music generation (Bretan et al., 2016), and the exploration of sound effect linkage technique make music selection and sound effect adjustment more flexible and convenient. In addition to studies based on a single medium, AI-assisted information process technology between multiple kinds of media is also a hot research field (Peng et al., 2017, 2018). Multi-modal feature learning (Ngiam et al., 2011) can make the understanding of video content more accurate. The relationship modeling of cross-media (Huang and Peng, 2019) makes the use of multimedia information more convenient and extensive. With the maturity of AI technology, the combination with traditional design will become an inevitable trend.

2 Alibaba multimedia content generation

In many of Alibaba's business applications, the tools of multimedia content design and production assisted by AI technology have been successfully applied. Combining AI technology with multimedia content design, the general production process is shown in Fig. 1, which can be roughly divided into five modules:

1. Analysis

This module analyzes mainly the original material according to the configuration information included in parameters, scripts, or presets. As the most basic module, it plays an important role in the whole process. Many capabilities of AI technology are widely used in this module, such as image classification, detection, and segmentation, or optical character recognition (OCR). In practical applications, these algorithms usually need customized improvement for usage scenarios.

2. Processing

Based on the understanding of input materials, necessary processing is carried out in this module according to the material situation and business requirements, such as image beautification, image inpainting, alpha matting, and video cutting. Through this step, the original material is processed into available material elements for future usage.

3. Generation

Based on the structured material information, according to practical requirements, GAN, as one of the research hotspots nowadays, is widely used in this module. Many generation techniques have been applied to different degrees, such as text generation from image or video, music generation from video stream, and font generation.

4. Rendering

According to the preset scripts and parameters, this module effectively integrates all available material elements, such as images, texts, sound, and dynamic effect, to make them be displayed consistently. More expressive multimedia content can promote business more effectively.

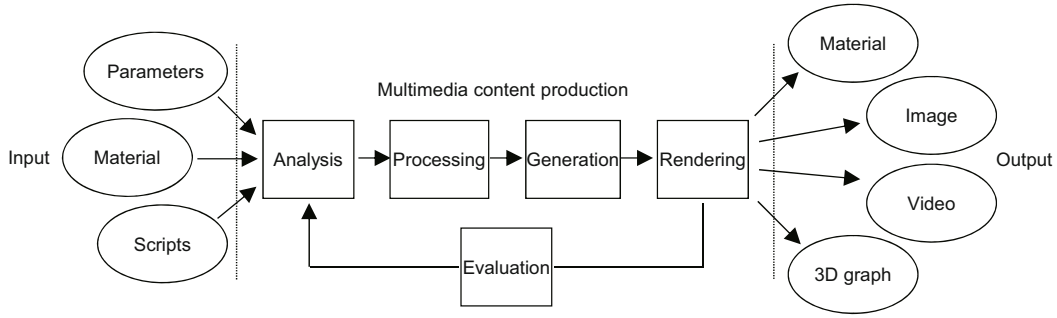

Fig. 1 Overview of the general processing flow

5. Evaluation

The final multimedia content products are evaluated in this module. Not only an evaluation index for the algorithm, such as image aesthetics and image quality, but also business metrics, are applied. For business metrics, the usage rate by users of the generated results is usually one of the typical reference metrics. The evaluation results play a positive role in the optimization of all modules in the whole process. However, this module generally does not work in the process of producing multimedia content. It begins after multimedia content is generated, even after it is put into use. Based on the evaluation of multimedia content itself and the analysis of business data, developers carry out iterative optimization for each process module, to achieve better business results. For brevity, this module is omitted in all figures and descriptions of all specific production tools below.

AI-assisted design and production tools based on the above process can be generally divided into two categories: AI-assisted production tools for design material and AI-assisted production tools for design productivity. The former is applied mainly in a certain module of the above process, because its final product is the design material, considering its role in the entire production process. The latter usually contains all modules of the above process, since its final output is complete multimedia content product, which is usually shown to users, such as banners, videos, and pages.

2.1 AI-assisted production tools for design material

With the development of AI technology, auxiliary design tools based on various AI techniques are widely applied, such as automatic poster layout tools for graphic design and automatic font generation tools. As a relatively basic step in the graphic design process, matting is usually necessary in converting hundreds of millions of advertising images uploaded by businesses into available material elements. However, although it is relatively low-tech, the heavy workload not only increases the burden on designers, but also costs much time and money.

The successful application of AI technology in image segmentation makes the automation of matting possible. The Wantoo product captures this opportunity. In the Wantoo website, as long as users input an RGB image with a resolution of 800×800, the matting result can be returned within 2 s. If a request includes 20 images, it takes an average of 100 ms each because of parallel processing. As a one-stop solution for image matting, the Wantoo website also provides an interactive matting tool to deal with complex situations where automatic matting cannot work perfectly. Users can easily get subjects in images from scratch by simply painting foreground and background brushes in some correct places. Using AI technology we can find a reasonable boundary for the foreground subject automatically within 1 s. The strokes painted by the correct brush in error matting regions can effectively drive the matting boundary to a more proper place. Besides these AI brushes, an unintelligent pair of tools, including a supplement pen and an eraser, is provided to fix minor errors in automatic matting results.

In e-commerce scenarios, the object of matting includes tens of thousands of commodity categories with various forms. To solve this complicated matting problem, a saliency segmentation algorithm based on deep learning is used to identify the pixel position of subjects in images (Fig. 2). Then, accurate localization technology is applied

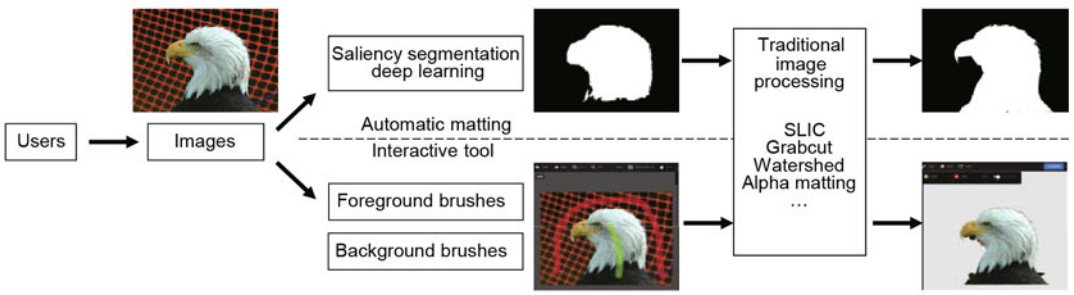

Fig. 2 Processing flow of Wantoo

to recover a detailed local structure, as compensation for smooth responses caused by deep models, such as improved simple linear iterative clustering (SLIC) (Kim et al., 2013), grabcut (Rother et al., 2004), watershed (Bradski and Kaehler, 2008), and alpha matting (Levin et al., 2007). As a result, the subjects matted out have a more accurate segmentation boundary, and more natural transition to their backgrounds. Based on the techniques mentioned above, Wantoo also offers an application programming interface (API) call, by which users can perform parallel processing on a large number of images in a short time with only a list of image uniform resource locators (URLs).

Because of the diversity and randomness of the images to be matted, no matting algorithm can solve all matting problems. When millions of product images are automatically matted for activities or promotions, it will take much time and manpower to review the matting results one by one to remove bad cases. It is even more impractical to complete this review task in a short time. Different from other matting products on the Internet, such as removebg (https://www.remove.bg/) and Gaoding Sheji (https://www.gaoding.com/), combined with the usage scenario in Alibaba, Wantoo introduces a confidence score for automatic matting results. The confidence score is obtained by analyzing the difference between the output of the deep learning model and the final refined result through image features, such as colors and contours. A higher confidence score means a greater likelihood of correct matting results. A suitable threshold for the confidence score can be set to filter out most bad cases. When a manual review is required, by reviewing all matting results in descending order of their confidence scores, a higher detection rate of good cases is reached in the early stage. This means a sufficient number of matting results for activities or promotions can be obtained in less time.

Since Wantoo begins to serve online, it has penetrated into multiple businesses, such as the internal service market, Luban, Feizhu, international business, and Taobao. The volume of service invocations has increased steadily, and accumulated more than five million times in half a year. To some extent, Wantoo lays the foundation for the full-link automation of intelligent design.

2.2 AI-assisted production tools for design productivity

2.2.1 Intelligent graphic design

In the field of graphic design, Luban is undoubtedly the leader in Alibaba. During the "Double 11" period in 2018, its daily request peak exceeded 50 million, and the cumulated graphic designs generated during the promotion reached nearly 500 million.

Luban realizes the automation of the whole process in the graphic design. It is not only a production tool for graphic design, just as Canva (https://www.canva.cn/) and Gaoding Sheji (https://www.gaoding.com/), but also an intelligent banner release tool. Based on big data from Taobao and the analysis of users' behavior, relying on the close connection with Alibaba ecology, Luban could automatically synthesize personalized banners with different commodities for different users and release them without users' knowledge, that is, different recommendations for different people.

In the Luban website, commodity images are the only material that individual users need to prepare. After simple selections of banner scale wanted and their fields, several different styles of banners are produced in seconds. Users can choose favorite ones to download or directly publish them online.

Given a scene graph, a full-frame design scheme is innovatively introduced by Luban for the graphic design generation technique. The design flow is decomposed into scene graph design and text collection design in the full-frame design scheme (Fig. 3). Using image processing techniques including detection and generation, scene graph design generates scene graph backgrounds from the scene images provided by users. On the other hand, in text collection design, multiple templates containing information such as layout, text content, text style, and text font are provided by designers. Combining the scene graph backgrounds generated and the text collection preset, the final personalized graphic design is generated.

The scene graph is usually photographed in a real situation with fuller expressiveness and more abundant information. During the entire "Double 11" period, compared with the graphic design generated from traditional white background templates, the click rate of full-frame scene design exceeds 30%.

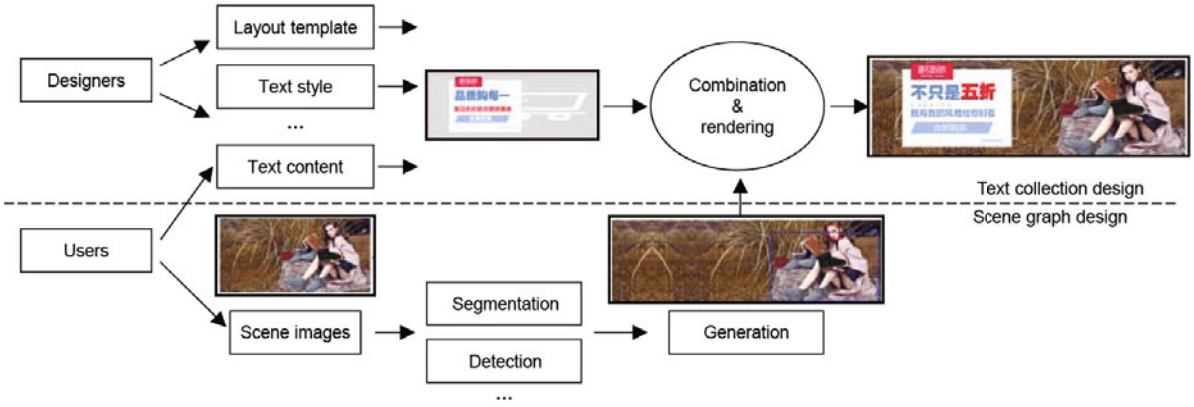

Fig. 3 Full-frame design scheme of Luban

The overall improvement rate is over 90% if intelligent text generation techniques are applied.

2.2.2 Intelligent video generation

With the development of video content understanding techniques, the boundaries of intelligent design have been expanded, and the visual-based content generation capability has been further enhanced, benefiting from the development of computer vision technology in classification, detection, face recognition (Zhang et al., 2017), pose estimation (Xia et al., 2017; Papandreou et al., 2018), motion detection (Cao et al., 2017), and camera tracking (Ristani and Tomasi, 2018). The computer can understand the "lens language" more accurately. The lens shooting methods (push, pull, and pan) can be captured. The shooting environment (shed shot, street shot, close shot, or vision shot) can be identified. The shot boundary can be detected. Numerous techniques construct the basis of video generation.

On the Internet, several intelligent video generation tools have been developed, such as GliaStudio (https://gliacloud.com/) and ZEN VIDEO (https://zenvideo.cn/). These tools focus on story-based video generation. By analyzing a story offered from users, relevant images or clips are searched from the Internet or their databases, and a video is automatically generated according to the preset templates. However, the original material of the video required on Alibaba's e-commerce platform usually includes images, clips, sounds, and sentences from a description of a specific product and its details. To solve this kind of video generation problem based on complex materials, Alibaba Wood comes into being. It is an important tool for video generation in Alibaba.

In the Alibaba Wood website, to generate a short video, the URL of a commodity detail page in the Taobao website is the only input for individual users. Three simple video attributes can be arbitrarily configured, including rhythm, length-width ratio, and duration. In addition, an automatic video clipping function is offered with the same configurations, but users need to upload several video clips in a limited range. When a video is generated, the website provides an online interactive editing tool to improve the video effect according to users' will. In large-scale applications, Alibaba Wood offers an API call, which can process a list of detail page URLs in parallel.

As shown in Fig. 4, all kinds of elements in detail pages can be used as the original material in Alibaba Wood. Due to the complexity of material forms and the variety of analysis requirements, the analysis process of input data may be time-consuming. For moderate complexity, the time taken is 5–10 s. Based on the information obtained from the input data, Alibaba Wood can understand the content of the input images or videos, such as subject properties, motion patterns, and lens language. Key video clips are cut out and informative images are reserved. According to the users' usage scenario, the narrative logic of videos, and the preset script protocol (such as from vision to detail focus), the selected

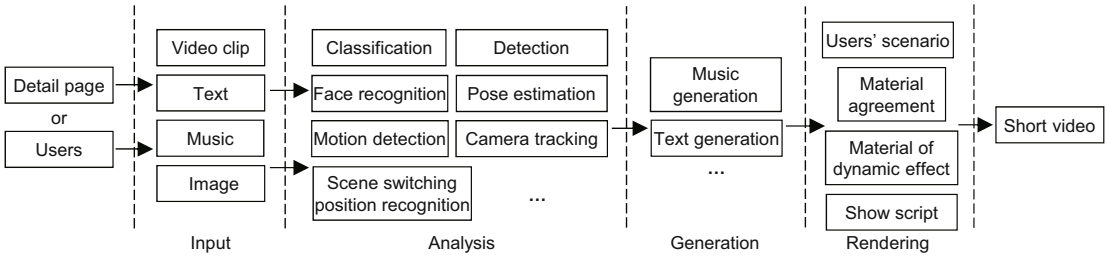

Fig. 4 Processing flow of Alibaba Wood

images and (or) video clips are sequentially connected through various dynamic templates. Browser rendering and segmented synchronous rendering techniques are applied to save composition time. Take a 500 000-pixel-per-frame, 15-s short video as an example. The composition time for simple dynamic effects is within 5 s. If a complicated dynamic template is selected, OpenGL is used to control the composition time within 10 s.

Besides image-based AI technology, text- and sound-relevant AI techniques are applied in Alibaba Wood. Before the video begins to be composed, text descriptions are generated according to some key frames or other image materials for timely insertions, if ready-made texts are not provided from users. The text generation process includes two steps, a generation of selling point descriptions from selling point labels and a generation of recommendation reasons from selling point descriptions. The selling point labels are extracted from images through an image-based classification deep model, and text generation from labels is realized through the transformer model (Vaswani et al., 2017).

The choice and influence of background music are also considered in Alibaba Wood. Comparative experiments show that appropriate background music can effectively enhance consumers' desire to buy. Alibaba Wood maintains a background music library. When the target product is recognized from original materials, a proper music that matches the product style is selected, and the preset dynamic effect template is also adjusted slightly to match its rhythm. Alibaba Wood can also generate its own background music. By training a long short-term memory (LSTM) based (Hochreiter and Schmidhuber, 1997) GAN network, a 10-s piece of music can be generated from a period of white noise with a specified kind of emotion.

Since Alibaba Wood's online service was provided, more than 20 million short videos have been produced. The video production cost of the cooperative merchant Senma apparel was reduced by 90%, as well as 95% improvement in production efficiency. The profit was increased by more than 700 000 CNY in two weeks.

2.2.3 Product detail page production

The design of product detail pages is a comprehensive intelligent manufacturing of digital content, including graphic images, 3D graphics, videos, audios, and texts. In e-commerce scenarios, promotion pages and detail pages contain various forms of information. Faced with the display requirements of hundreds of millions of commodities, making the web page design intelligent will become an inevitable trend.

AI-detail, as a kind of intelligent generation tool for product detail pages, integrates multiple elements of graphic intelligent design and video intelligent design into the process of automatic detail page generation. Different from other independently developed page generation tools outside the group, such as deepdraw (http://www.deepdraw.cn/), AI-detail is deeply rooted in the Alibaba ecosystem. Its data source and distribution path are connected with the application scenarios. Based on analyzing the detail pages in use, parallel generation of large numbers of detail pages is realized. Individual users can also try to use the AI-detail by writing a script to call the API.

As shown in Fig. 5, AI-detail can perform structured analysis on existing detail pages using AI technology, and automatically restructure detail pages in batches with other languages or new style templates using the materials obtained. If the material comes from the analysis of detail pages, the time consumption of the automatic analysis phase, which accounts

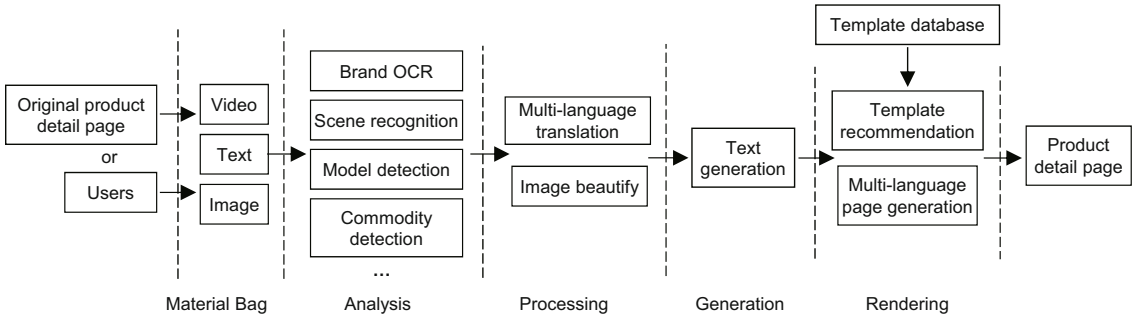

Fig. 5 Processing flow of AI-detail

for the vast majority of the whole process, is the same as in Alibaba Wood. If the material is directly offered from users, the structured analysis process is skipped.

When detail pages are input as materials, page segmentation is first carried out to obtain all elements of the page, including videos, images with exact locations, texts, and tables. Then, in-depth material analysis is conducted based on the elements obtained. The brand and scene can be identified. The model's postures can be estimated in multiple frames or images. The details of the product can be captured from multiple angles. Not only the texts directly printed in detail pages, but also the texts printed on images or frames of videos, are recognized through OCR technology. For AI-detail, the storage of structured information of pages is its core function. The database of detail pages can play an important role in the analysis and promotion of the e-commerce platform, commodities, and even all relevant industries.

In the detail page reconstruction application, based on the material information obtained from detail pages, AI-detail can perform image beautification, automatic generation of texts, and multilingual translation. According to a template recommended from the template database, a new product detail page can be generated. In the business scenario of cross-border reconstruction, AI-detail analyzes Chinese pages and reproduces new detail pages automatically with the local language. Since it has begun to operate online, it has effectively supported the overseas delivery of more than one million commodities in half a year.

3 Future work

Personalized content design and emotional computing may be two potential development directions for future multimedia content generation in Alibaba.

At the current stage, most of the decisions in multimedia content design still depend on the professional competence of designers. For the design tasks that can be implemented by AI technology, machines can quickly accomplish them in parallel on a large scale. The cost can be greatly reduced. Just as the "different recommendations for different people" in Luban mentioned in Section 2.2.1, not only the multimedia content shown, but also the design elements, such as display style and template layout, can be different among users according to their preferences. In online operations, the multimedia content production platforms (such as Luban and Alibaba Wood) can recommend and display products in users' favorite design style, according to product categories, users' population, and location.

In live broadcast scenarios, anchors with more positive emotion (tone, expression) bring more views and purchases. Similarly, the hotspots of films usually have more intense emotional expressions. The potential connection between emotional factors and business promotes the development of emotion computing related techniques. Text-based sentiment orientation analysis, speech-based mood analysis, and image-based facial expression analysis have attracted wide attention from the industry. In the process of multimedia content production, adding scenes and emotion related expression will become a direction which is worth exploring. Combined with depth map analysis techniques and 3D scene modeling techniques, different products can be implanted into video clips of different scenes, such

as office, living room, bedroom, and playground.

Compliance with ethics guidelines

Kui-long LIU, Wei LI, Chang-yuan YANG, and Guang YANG declare that they have no conflict of interest.

References

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. Encoder-decoder 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

Hochreiter S, Schmidhuber J, 1997. Long short-term memory. *Neur Comput*, 9(8):1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang X, Peng YX, 2019. TPCKT: two-level progressive cross-media knowledge transfer. *IEEE Trans Multim*, 21(11):2850-2862. https://doi.org/10.1109/TMM.2019.2911456

Kim KS, Zhang DN, Kang MC, et al., 2013. Improved simple linear iterative clustering superpixels. IEEE Int Symp on Consumer Electronics, p.259-260. https://doi.org/10.1109/ISCE.2013.6570216

Levin A, Lischinski D, Weiss Y, 2007. A closed-form solution to natural image matting. *IEEE Trans Patt Anal Mach Intell*, 30(2):228-242. https://doi.org/10.1109/TPAMI.2007.1177

Lin TY, Dollàr P, Girshick R, et al., 2017. Feature pyramid networks for object detection. Proc Conf on Computer Vision and Pattern Recognition, p.2117-2125. https://doi.org/10.1109/CVPR.2017.106

Ngiam J, Khosla A, Kim M, et al., 2011. Multimodal deep learning. Proc 28th Int Conf on Machine Learning, p.689-696.

Papandreou G, Zhu T, Chen LC, et al., 2018. PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. https://arxiv.org/abs/1803.08225

Peng YX, Zhu WW, Zhao Y, et al., 2017. Cross-media analysis and reasoning: advances and directions. *Front Inform Technol Electron Eng*, 18(1):44-57. https://doi.org/10.1631/FITEE.1601787

Peng YX, Huang X, Zhao YZ, 2018. An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges. *IEEE Trans Circ Syst Video Technol*, 28(9):2372-2385. https://doi.org/10.1109/TCSVT.2017.2705068

Peng YX, Qi JW, Huang X, 2019. Current research status and prospects on multimedia content understanding. *J Comput Res Devel*, 56(1):187-212 (in Chinese). https://doi.org/10.7544/issn1000-1239.2019.20180770

Ren SQ, He KM, Girshick R, et al., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Patt Anal Mach Intell*, 39(6):1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

Ristani E, Tomasi C, 2018. Features for multi-target multi-camera tracking and re-identification. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.6036-6046. https://doi.org/10.1109/CVPR.2018.00632

Rother C, Kolmogorov V, Blake A, 2004. "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Trans Graph*, 23(3):309-314. https://doi.org/10.1145/1015706.1015720

Simonyan K, Zisserman A, 2015. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556

Song SJ, Zhang W, Liu JY, et al., 2019. Unsupervised person image generation with semantic parsing transformation. https://arxiv.org/abs/1904.03379

Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Advances in Neural Information Processing Systems, p.5998-6008.

Xia FT, Wang P, Chen XJ, et al., 2017. Joint multi-person pose estimation and semantic part segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.6769-6778. https://doi.org/10.1109/CVPR.2017.644

Zhang SF, Zhu XY, Lei Z, et al., 2017. S^{3}FD: single shot scale-invariant face detector. Proc IEEE Int Conf on Computer Vision, p.192-201. https://doi.org/10.1109/ICCV.2017.30

Zhou BL, Khosla A, Lapedriza A, et al., 2016. Learning deep features for discriminative localization. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2921-2929. https://doi.org/10.1109/CVPR.2016.319

Zhu JY, Park T, Isola P, et al., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proc IEEE Int Conf on Computer Vision, p.2223-2232. https://doi.org/10.1109/ICCV.2017.244