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(2024)02-0478-13

论文引用格式:Peng H, Zhang J B, Jia D, An T, Cai P and Zhao J Y. 2024. Real-time high-resolution video portrait matting network combined with background image. Journal of Image and Graphics, 29(02):0478-0490(彭泓, 张家宝, 贾迪, 安彤, 蔡鹏, 赵金源. 2024. 结合背景图的高分辨率视频人像实时抠图网络. 中国图象图形学报, 29(02):0478-0490)[DOI:10.11834/jig.230174]

结合背景图的高分辨率视频人像实时抠图网络

彭泓[1],张家宝[1*],贾迪[1,2],安彤[1],蔡鹏[1],赵金源[1]

1. 辽宁工程技术大学电子与信息工程学院,葫芦岛　125105; 2. 辽宁工程技术大学电气与控制工程学院,葫芦岛　125105

摘　要:**目的**　近年来,采用神经网络完成人像实时抠图已成为计算机视觉领域的研究热点,现有相关网络在处理高分辨率视频时还无法满足实时性要求,为此本文提出一种结合背景图的高分辨率视频人像实时抠图网络。**方法**　给出一种由基准网络和精细化网络构成的双层网络,在基准网络中,视频帧通过编码器模块提取图像的多尺度特征,采用金字塔池化模块融合这些特征作为循环解码器网络的输入;在循环解码器中,通过残差门控循环单元聚合连续视频帧间的时间信息,以此生成蒙版图、前景残差图和隐藏特征图,采用残差结构降低模型参数量并提高网络的实时性。为提高高分辨率图像实时抠图性能,在精细化网络中,设计高分辨率信息指导模块,通过高分辨率图像信息指导低分辨率图像的方式生成高质量人像抠图结果。**结果**　与近年来的相关网络模型进行实验对比,实验结果表明,本文方法在高分辨率数据集Human2K上优于现有相关方法,在评价指标(绝对误差、均方误差、梯度、连通性)上分别提升了18.8%、39.2%、40.7%、20.9%。在NVIDIA GTX 1080Ti GPU上处理4 K分辨率影像运行速率可达26帧/s,处理HD(high definition)分辨率影像运行速率可达43帧/s。**结论**　本文模型能够更好地完成高分辨率人像实时抠图任务,可以为影视、短视频社交以及网络会议等高级应用提供更好的支持。

关键词:人像实时抠图;神经网络;多尺度特征;时间信息;高分辨率

Real-time high-resolution video portrait matting network combined with background image

Peng Hong[1], Zhang Jiabao[1*], Jia Di[1,2], An Tong[1], Cai Peng[1], Zhao Jinyuan[1]

1. School of Electronic and Information Engineering, Liaoning Technical University, Huludao　125105, China;

2. Faculty of Electrical and Control Engineering, Liaoning Technical University, Huludao　125105, China

Abstract: **Objective** Video matting is one of the most commonly used operations in visual image processing. It aims to separate a certain part of an image from the original image into a separate layer and further apply it to specific scenes for later video synthesis. In recent years, real-time portrait matting that uses neural networks has become a research hotspot in the field of computer vision. Existing related networks cannot meet real-time requirements when processing high-resolution video. Moreover, the matting results at the edges of high-resolution image targets still have blurry issues. To solve these problems, several recently proposed methods that use various auxiliary information to guide high-resolution image for mask estimation have demonstrated good performance. However, many methods cannot perfectly learn information about the edges and details of portraits. Therefore, this study proposes a high-resolution video real-time portrait matting network com-

收稿日期:2023-04-18;**修回日期**:2023-07-24;**预印本日期**:2023-08-31

*通信作者:张家宝　zhang_jiabao316@163.com

基金项目:国家自然科学基金项目(61601213);辽宁省教育厅项目(LJ2020FWL004)

Supported by: National Natural Science Foundation of China (61601213); Research Foundation of Education Bureau of Liaoning Province (LJ2020FWL004)

bined with background images. **Method** A double-layer network composed of a base network and a refinement network is presented. To achieve a lightweight network, high-resolution feature maps are first downsampled at sampling rate D. In the base network, the multi-scale features of video frames are extracted by the encoder module, and these features are fused by the pyramid pooling module, because the input of the cyclic decoder network is beneficial for the cyclic decoder to learn the multi-scale features of video frames. In the cyclic decoder, a residual gated recurrent unit (GRU) is used to aggregate the time information between consecutive video frames. The masked map, foreground residual map, and hidden feature map are generated. A residual structure is used to reduce model parameters and improve the real-time performance of the network. In the residual GRU, the time information of the video is fully utilized to promote the construction of the masked map of the video frame sequence based on time information. To improve the real-time matting performance of high-resolution images, the high-resolution information guidance module designed in the refinement network, and the initial high-resolution video frames and low-resolution predicted features (masked map, foreground residual map, and hidden feature map) are used as input to pass the high-resolution information guidance module, generating high-quality portrait matting results by guiding low-resolution images with high-resolution image information. In the high-resolution information guidance module, the combination of covariance means filtering, variance means filtering, and pointwise convolution processing can effectively extract the matting quality of the detailed areas of character contours in a high-resolution video frame. Under the synergistic effects of the benchmark and refinement networks, the designed network cannot only fully extract multi-scale information from low-resolution video frames, but can also more fully learn the edge information of portraits in high-resolution video frames. This condition is conducive to more accurate prediction of masked maps and foreground images in the network structure and can also improve the generalization ability of the matting network at multiple resolutions. In addition, the high-resolution image downsampling scheme, lightweight pyramid pooling module, and residual link structure designed in the network further reduce the number of network parameters, improving the real-time performance of the network. **Result** We use PyTorch to implement our network on NVIDIA GTX 1080Ti GPU with 11 GB RAM. Batch size is 1, and the optimizer used is Adam. This study trains the benchmark network on three datasets in sequence: the Video240K SD dataset, with an input frame sequence of 15. After 8 epochs of training, the fine network is trained on the Video240K HD dataset for 1 epoch. To improve the robustness of the model in processing high-resolution videos, the refinement network was further trained on the Human2K dataset, with a downsampling rate D of 0.25 and an input frame sequence of 2 for 50 epochs of training. Compared with related network models in recent years, the experimental results show that the proposed method is superior to other methods on the Video240K SD dataset and the Human2K dataset. On the Video240K SD dataset, 26.1%, 50.6%, 56.9%, and 39.5% of the evaluation indicators (sum of absolute difference (SAD), mean squared error(MSE), gradient error(Grad), and connectivity error(Coon)) were optimized, respectively. In particular, on the high-resolution Human2K dataset, the proposed method is significantly superior to other state-of-the-art methods, optimizing the evaluation indicators (SAD, MSE, Grad, and Coon) by 18.8%, 39.2%, 40.7%, and 20.9%, respectively. Simultaneously achieving the lowest network complexity at 4 K resolution (28.78 GMac). The running speed of processing low-resolution video (512 × 288 pixels) can reach 49 frame/s, and the running speed of processing medium-resolution video (1 024 × 576 pixels) can reach 42.4 frame/s. In particular, the running speed of processing 4 K resolution video can reach 26 frame/s, while the running speed of processing HD-resolution video can reach 43 frame/s on NVIDIA GTX 1080Ti GPU. This value is significantly improved compared with other state-of-the-art methods. **Conclusion** The network model proposed in this study can better complete the real-time matting task of high-resolution portraits. The pyramid pooling module in the benchmark network effectively extracts and integrates multi-scale information of video frames, while the residual GRU module significantly aggregates continuous inter-frame time information. The high-resolution information guidance module captures high-resolution information in images and guides low-resolution images to learn high-resolution information. The improved network effectively enhances the matting information of high-resolution human-oriented edges. The experiments on the high-resolution dataset Human2K show that the proposed network is more effective in predicting high-resolution montage maps. It has high real-time processing speed and can provide better support for advanced applications, such as film and television, short video social networking, and online conference.

Key words: real-time human figure matting; neural network; multiscale features; time information; high resolution

0 引 言

人像实时抠图的应用领域广泛，涉及到影视、短视频社交、网络会议等。对于给定视频帧$\boldsymbol{I}\in\mathbf{R}^{H\times W\times C}$，可将其看做是由前景$\boldsymbol{F}\in\mathbf{R}^{H\times W\times C}$与背景$\boldsymbol{B}\in\mathbf{R}^{H\times W\times C}$按一定透明度$\alpha\in[0,\ 1]$线性叠加而成，即

$$\boldsymbol{I}=\alpha\boldsymbol{F}+(1-\alpha)\boldsymbol{B} \tag{1}$$

式中，H、W和C分别为视频帧的高、宽和通道数量。抠图的主要目的是分离前景与背景，并计算每个像素对应的α，进而生成蒙版图。

视频抠图可分为传统和基于深度学习两类方法，传统抠图方法常将三分图作为辅助信息输入，主要根据图像中的前景与背景色来估算差异。其中，三分图是将图像分为前景$\boldsymbol{F}$、背景$\boldsymbol{B}$和前景—背景相接的未知区域。在应用三分图辅助抠图任务中，只需专注未知区域中前景的透明度，而无需关注前景和背景的具体像素值，减少了问题的解空间，可以辅助预测更为精细的蒙版图。根据样本间的相似度，可将传统方法分为基于采样和基于传播的抠图方法。基于采样的抠图方法（Hong等，2018；Shahrian等，2013；Chuang等，2001；He等，2011；Wang和Cohen，2007）根据样本间的连续性和相似性估计前景色与背景色所占比例，进而求解“未知区域”中的蒙版图。基于传播的抠图方法（Aksoy等，2017；刘天艺 等，2022；He等，2010；Lee和Wu，2011；吴玉娥等，2010；Levin等，2008；Chen等，2013b）根据像素间的相似度将已知区域中的透明度值传播至“未知区域”，求解完整的蒙版图。尽管传统视频抠图方法取得了阶段性成果，但在复杂场景中仍存在处理速度慢、目标图像边缘模糊、细节处理不够理想（如发丝、镜片等）的问题。

随着深度学习的崛起，基于学习的抠图方法在具有挑战性的场景中取得了突破性的进展。Chen等人（2013a）将深度学习与KNN（K-nearest neighbor）相结合，在高维特征空间中应用KNN计算蒙版图，优化了抠图效果，然而该方法在定义高维特征空间时存在难度，很难提高人物边缘的抠图效果。因此一些方法利用三分图中的辅助信息弥补人物边缘抠图信息的不足，为网络提供更多的边缘信息。Xu等人（2017）将三分图作为辅助信息，同时在编码器—解码器网络之后增加一个卷积网络，对预测的蒙版图进行精细化处理，从而解决抠图边缘模糊的问题。Sun等人（2021a）提出基于语义三分图的可学习抠图模式，估计未知区域相应的像素分类置信度图，将传统三分图转化为语义三分图，辅助原始图像估计蒙版图，提高了抠图效果。Park等人（2022）和Liu等人（2023）将三分图作为全局先验知识，并提出基于Transformer的抠图模型，在Transformer块中充分利用三分图信息进一步提高了蒙版图的精度。Liu等人（2021）基于三分图提出一种三方信息挖掘和融合网络，采用三方信息集成模块完成多分支信息之间的交互，实现了全局信息和局部信息之间的协调性，并取得了高质量的效果。弥补了人物边缘模糊的问题。然而制作精准的三分图需要耗费大量的时间，并且三分图的质量也直接影响着抠图的效果。对此，Yu等人（2021）提出一种以粗糙掩码作为指导信息的抠图框架，通过构建自引导模型逐步完善“未知区域”进行回归抠图，降低了人工制作精细三分图的要求。

此外，还有一些学者将工作重点放在了无需三分图的输入上。Sengupta等人（2020）采用拍摄照片前捕获无主题照片作为辅助信息，通过膨胀腐蚀操作生成人像粗分割图像，以此估计蒙版图和前景图，免去了制作三分图的环节并获得了较好的结果。通过形状模板图和X-Y坐标信息作为辅助信息，也可以在无三分图的前提下更好地估计蒙版图（许征波和杨煜俊，2020）。Chen等人（2022）提出高分辨率细节分支和语义上下文分支进行交互，进一步解决了无辅助信息指导的需求。虽然上述方法在抠图细节上已经具备了较好的表现力，但在视频任务中仍不具备连续处理视频帧的能力，更无法达到实时性能。为了能够连续处理视频帧，Sun等人（2021b）基于深度学习提出将初始帧三分图作为参考，通过三分图传播网络引导后续目标帧，并采用时空特征聚合模块获取时域信息，估计对应帧的蒙版图。此外，Seong等人（2022）提出级联三分图模块和蒙版图计算模块，采用传播三分图信息和α值对视频帧进行回归计算蒙版图，降低了视频帧三分图的制作要求。然而该方法仍然依赖人工制作三分图。为了解决人工制作三分图依赖的问题，一些学者在网络中融入了三分图生成网络自动生成三分图辅助视频抠图，Zhang等人（2021）将用户标注的视频关键帧输入到

视频对象分割网络中，用于生成视频三分图，辅助计算蒙版图，消除了手动制作三分图的需求。然而对三分图生成模块的设计降低了网络处理的实时性。为解决三分图生成网络的冗余，Jin 等人(2022)采用在绿幕背景下拍摄视频，有效提高了网络的实时性。Song(2022)和 Lin 等人(2022)采用无辅助信息输入的方式进行视频抠图，进一步加快了网络的实时性。然而，对于高分辨率视频，网络的实时性和抠图精度仍具有挑战性，为此，Lin 等人(2021)提出基于背景的方法，采用两层神经网络，基础网络计算低分辨率图像误差，优化网络在误差图上选择误差较大的图像区域进行优化处理，获得视频帧蒙版图，提高了人像视频抠图质量，但是由于优化网络未考虑全局信息，导致在部分细节区域的效果不佳。

为提高高分辨率视频实时抠图质量，本文在考虑视频帧间序列相关性的基础上，采用背景作为辅助信息，提出一种结合背景图的高分辨率视频人像实时抠图网络，主要贡献点如下：1)给出一种高分辨率视频实时抠图网络结构，能够有效处理人像边缘和细节处的像素信息，提高蒙版图构建的准确率。2)提出一种循环解码器网络，融合连续视频帧间信息，充分利用上下文时序信息提高相邻帧特征的捕获能力，同时引入类残差结构降低了模型的参数量，提高了模型的处理速度。3)对高分辨率信息指导模块进行设计，结合协方差均值滤波、逐点卷积等处理，给出一种精细化网络结构，能够有效提取高分辨率影像中人物轮廓细节区域的抠图质量。

1 方　法

为了实现网络的轻量化，首先将高分辨率特征图按照采样率 D 进行下采样，可以在基准网络中减少 $(D^2-1)/D^2$ 倍计算量，其次在基准网络的特征融合部分采用具有平均池化操作的轻量级金字塔池化模块，进一步减少网络的参数量。解码器部分，在保持高分辨率抠图精度的前提下，采用具有低参数量的线性插值的方式恢复高分辨率图像，从而提高网络的实时性能。

网络结构主要由基准网络和精细化网络两个部分组成，基准网络在低分辨率图像上进行处理，精细化网络根据基准网络的预测结果以原始高分辨率图像作为指导信息生成目标结果，如图 1 所示。融合给定视频帧 $\boldsymbol{I}$ 与背景 $\boldsymbol{B}$，通过降采样获得低分辨率视频帧 $\boldsymbol{I}_l$ 和背景 $\boldsymbol{B}_l$，并将其输入到基准网络中获得低分辨率蒙版图 $\boldsymbol{\alpha}_l$、隐藏特征图 $\boldsymbol{H}_c$ 以及前景残差图 $\boldsymbol{F}_c^R=\boldsymbol{F}-\boldsymbol{I}$。精细化网络采用 $\boldsymbol{\alpha}_l$、$\boldsymbol{F}_c^R$ 和 $\boldsymbol{H}_c$ 及原始视频帧 $\boldsymbol{I}$ 指导低分辨率视频帧 $\boldsymbol{I}_l$ 生成高质量蒙版图与前景图 $\boldsymbol{F}$。本文网络采用多帧图像作为输入，更有利于模型对时间信息的整合。

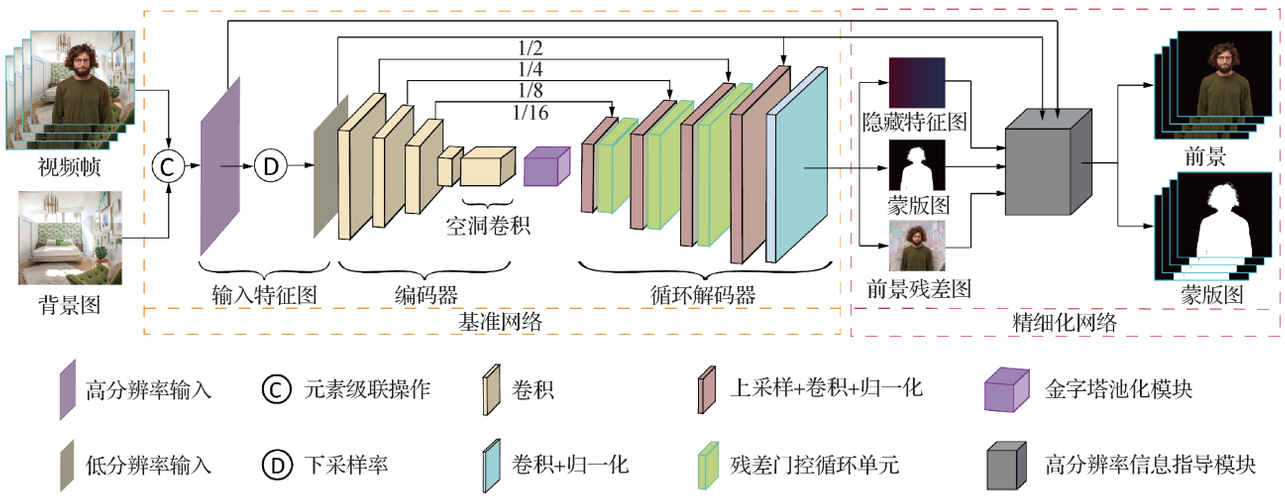

图 1　总体网络架构

Fig. 1　Overall network architecture

1.1　基准网络

基准网络采用全卷积编码—解码网络，主要由编码器模块、金字塔池化模块(pyramid pooling module，PPM)(Zhao 等，2017)及循环解码器模块组成。

1.1.1　编码器模块

为增大感受野保留更多视频帧信息，本文在编码器中采用空洞卷积来提取特征。图像细节特征提取依赖语义分割质量，采用DeepLabV3+架构的主干ResNet50（residual network 50）神经网络作为编码器主体结构，将其第1层卷积设为六通道输入，为保持编码器1/16下采样输出，对编码器的最后一层下采样接入空洞卷积操作，提高编码器的语义分割能力，令$g_{\text{en}}(\cdot)$为编码器模块，以低分辨率图像$\boldsymbol{I}_l$作为输入获得特征图$\boldsymbol{F}_l$，即

$$\boldsymbol{F}_l = g_{\text{en}}(\boldsymbol{I}_l) \tag{2}$$

编码器不仅输出特征图$\boldsymbol{F}_l$，还提取1/2、1/4、1/8和1/16分辨率下的多尺度中间特征图，以此捕获精细结构，为后期循环解码器特征融合提供支持。

1.1.2　金字塔池化模块

多尺度特征提取有利于收集更多的细节和语义信息，进一步提升网络的鲁棒性。如图2所示，金字塔池化模块由多个大小为1×1，2×2，3×3，6×6自适应平均池化操作组成，对其输出进行1×1卷积减少通道数，分别进行双线性上采样操作，并将输出特征图与模块输入图通过Concat()函数进行元素级联，通过1×1卷积输出多尺度特征融合图。

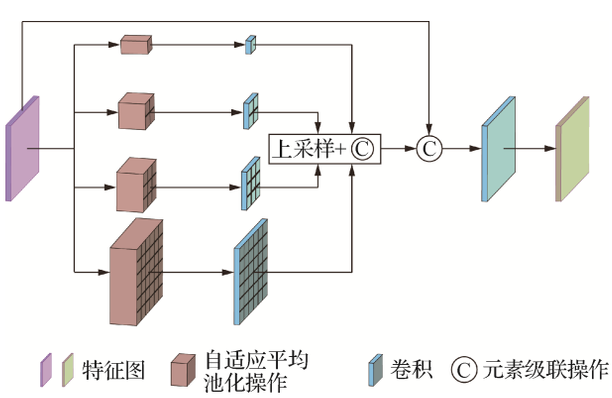

图2　金字塔池化模块

Fig. 2　Pyramid pooling module

1.1.3　循环解码器模块

结合长短期记忆网络进行解码器设计，采用循环架构提高视频流中的重要信息捕获能力，通过自适应学习的方式连续处理视频帧间信息。循环解码器由连续上采样层和残差门控循环单元模块（residual ConvGRU，R-ConvGRU）交替组成。上采样层将上一模块生成的特征图通过双线性上采样生成对应特征图$\boldsymbol{F}_l$，分别与来自编码器提取的多尺度中间特征图$\boldsymbol{F}_{1/16}$、$\boldsymbol{F}_{1/8}$、$\boldsymbol{F}_{1/4}$和$\boldsymbol{F}_{1/2}$合并连接，通过无偏置卷积、批量归一化和ReLU激活函数进行特征融合与通道合并，采用残差门控循环单元模块对特征图的信息流进行迭代更新，通过输出层分离目标预测图。为了更好地聚合信息流中的时间信息，采用如图3所示的残差门控循环单元模块对视频流进行更新，利用卷积门控循环单元（ConvGRU）将分离后的特征图与视频前一帧特征图融合，并对信息流的时空序列信息进行更新。同时，在残差门控循环单元模块中引入类残差连接结构，将分离后的特征图与经ConvGRU融合后的特征信息进行残差合并。上述设计不仅在一定程度上减少了参数量，而且还能使网络更加专注时空序列信息。

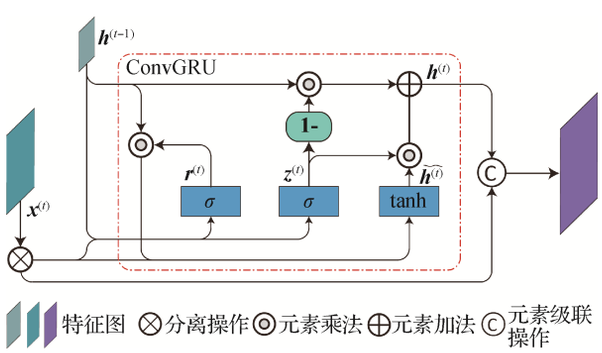

图3　残差门控循环单元（R-ConvGRU）

Fig. 3　Residual gated recurrent unit（R-ConvGRU）

在残差门控循环单元模块中，采用具有整合时空序列信息的卷积门控循环单元融合视频信息流，可表示为

$$\boldsymbol{z}^{(t)} = \sigma(\boldsymbol{w}^{(zx)} \cdot \boldsymbol{x}^{(t)} + \boldsymbol{w}^{(zh)} \cdot \boldsymbol{h}^{(t-1)} + \boldsymbol{b}^{(z)}) \tag{3}$$

$$\boldsymbol{r}^{(t)} = \sigma(\boldsymbol{w}^{(rx)} \cdot \boldsymbol{x}^{(t)} + \boldsymbol{w}^{(rh)} \cdot \boldsymbol{h}^{(t-1)} + \boldsymbol{b}^{(r)}) \tag{4}$$

$$\widetilde{\boldsymbol{h}^{(t)}} = \tanh(\boldsymbol{w}^{(\tilde{h}x)} \cdot \boldsymbol{x}^{(t)} + \boldsymbol{w}^{(\tilde{h}h)} \cdot (\boldsymbol{r}^{(t)} \circ \boldsymbol{h}^{(t-1)}) + \boldsymbol{b}^{(\tilde{h})}) \tag{5}$$

$$\boldsymbol{h}^{(t)} = (1 - \boldsymbol{z}^{(t)}) \circ \boldsymbol{h}^{(t-1)} + \boldsymbol{z}^{(t)} \circ \widetilde{\boldsymbol{h}^{(t)}} \tag{6}$$

式中，运算符·和∘分别为卷积操作与乘积，$\sigma(\cdot)$和$\tanh(\cdot)$分别为sigmoid激活函数和双曲正切激活函数，$\boldsymbol{w}$和$\boldsymbol{b}$分别为3×3卷积核与偏置项，$\boldsymbol{h}^{(t-1)}$为循环架构中上一个循环生成的隐藏状态图，将其作为当前隐藏状态图$\widetilde{\boldsymbol{h}^{(t)}}$的输入，将初始隐藏状态图$\boldsymbol{h}^{(0)}$置为全零张量。输出层将上采样输出的特征图与输入图像通过Concat(·)进行级联，并通过两次3×3卷积、批量归一化（Ioffe和Szegedy，2015）以及ReLU激活函数输出低分辨率预测特征（1通道的蒙版预测图、3通道的前景残差图及32通道的隐藏特征图）。

1.2 精细化网络

基准网络中联合上采样操作将导致输出图像的边缘模糊，根据 Wu 等人(2018)提出的快速引导滤波，通过设计高分辨率信息指导模块构建精细化网络提高人像抠图质量。高分辨率图像 $\boldsymbol{I}_h$ 通过下采样处理获得低分辨率图像 $\boldsymbol{I}_l$，并将 $\boldsymbol{I}_l$ 与低分辨率预测特征(蒙版图 $\boldsymbol{\alpha}_l$、前景残差图 $\boldsymbol{F}_c^R$ 和隐藏特征图 $\boldsymbol{H}_c$)作为精细化网络的输入，传递给高分辨率信息指导模块，如图 4 所示。

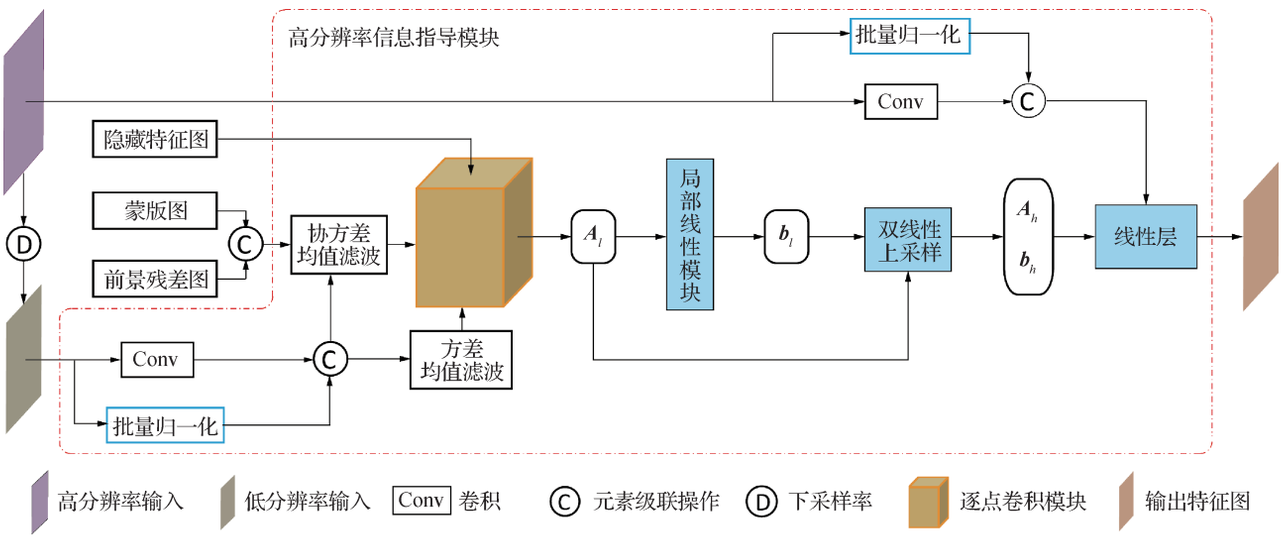

图 4 精细化网络

Fig. 4 Refined network

在高分辨率信息指导模块中，将 $\boldsymbol{I}_h$ 经卷积处理减小通道维度，并与批量归一化结果融合获得高分辨率特征图 $\boldsymbol{P}_h$，对 $\boldsymbol{I}_l$ 执行相同的操作获得低分辨率特征图 $\boldsymbol{P}_l$，并通过方差均值滤波计算 $\boldsymbol{P}_l$ 像素间的关联信息，进一步提取多尺度特征。

为充分提取图像的低分辨率特征，根据基准网络输出的蒙版图 $\boldsymbol{\alpha}_l$ 及前景残差图 $\boldsymbol{F}_c^R$ 提取中期低分辨率指导信息 $\boldsymbol{Q}_l$，具体为

$$\boldsymbol{Q}_l = C\left(\boldsymbol{\alpha}_l, \boldsymbol{F}_c^R\right) \tag{7}$$

式中，$C(\cdot)$ 为元素级联操作，采用 Concat(·)函数进行拼接。通过协方差均值滤波模块将 $\boldsymbol{P}_l$ 与 $\boldsymbol{Q}_l$ 融合，再将方差均值滤波模块输出的特征图和隐藏特征图共同输入到逐点卷积模块中，计算低分辨率预测值与输入图像间的重构误差，获得低分辨率线性指导参数 $\boldsymbol{A}_l$，该参数受到如下方程的约束，具体为

$$\boldsymbol{Q}_i = \boldsymbol{A}_l \otimes \boldsymbol{I}_i + \boldsymbol{b}_l, \ \boldsymbol{I}_i \in \boldsymbol{w}_k \tag{8}$$

式中，$\boldsymbol{w}_k$ 为 3 × 3 卷积滤波窗口，⊗为元素乘法，$\boldsymbol{I}_i$ 为卷积滤波窗口 $\boldsymbol{w}_k$ 内第 i 个像素，$\boldsymbol{Q}_i$ 为特征图内第 i 个像素。通过式(8)局部线性变换计算低分辨率图像偏移量 $\boldsymbol{b}_l$，将参数 $\boldsymbol{A}_l$ 和 $\boldsymbol{b}_l$ 通过双线性上采样获得高分辨率线性指导参数 $\boldsymbol{A}_h$ 及高分辨率图像偏移量 $\boldsymbol{b}_h$。为保留原始高分辨率图像信息，在高分辨率特征图 $\boldsymbol{P}_h$ 的指导下，通过线性层处理生成高分辨率蒙版图 $\boldsymbol{Q}_h$，具体为

$$\boldsymbol{Q}_h = \boldsymbol{A}_h \otimes \boldsymbol{P}_h + \boldsymbol{b}_h \tag{9}$$

1.2.1 协方差均值滤波模块

协方差均值滤波可以更充分地提取低分辨率特征信息，采用如图 5 所示的协方差均值滤波模块融合 $\boldsymbol{Q}_l$ 和 $\boldsymbol{P}_l$，计算式为

$$Cov(\boldsymbol{F}) = b(\boldsymbol{Q}_l \times \boldsymbol{I}_l) - b(\boldsymbol{Q}_l) \times b(\boldsymbol{I}_l) \tag{10}$$

式中，$b(\cdot)$ 为均值滤波卷积，卷积核大小均为 4 × 4、权值大小相同。对输入的 $\boldsymbol{Q}_l$ 和 $\boldsymbol{P}_l$ 进行元素乘积，通过均值滤波卷积剔除图像噪声点并保留细节，对 $\boldsymbol{Q}_l$ 和 $\boldsymbol{P}_l$ 分别经均值滤波卷积去噪，再进行元素乘积，通过减法操作计算协方差均值特征信息 $Cov(\boldsymbol{F})$。

1.2.2 方差均值滤波模块

方差均值滤波对特征图进行平滑处理，并充分提取低分辨率图像特征信息。将单一变量 $\boldsymbol{P}_l$ 作为输入，通过如图 6 所示的方差均值滤波模块计算方差均值特征信息 $Var(\boldsymbol{F})$，具体为

$$Var(\boldsymbol{F}) = b(\boldsymbol{P}_l \times \boldsymbol{P}_l) - b(\boldsymbol{P}_l) \times b(\boldsymbol{P}_l) \tag{11}$$

式中，$b(\cdot)$ 为均值滤波卷积操作，卷积核大小均为 4 × 4、权值大小均等，$Var(\boldsymbol{F})$ 将作为逐点卷积模块的输入。

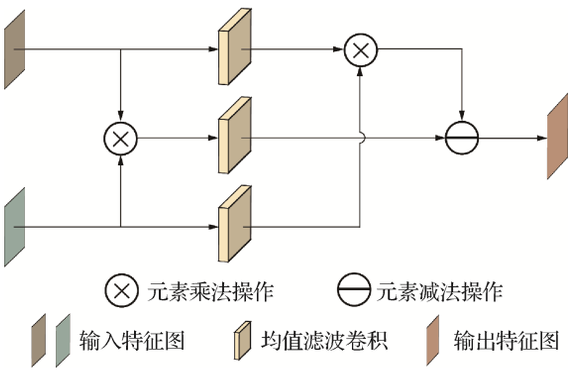

图5　协方差均值滤波模块

Fig. 5　Covariance mean filtering module

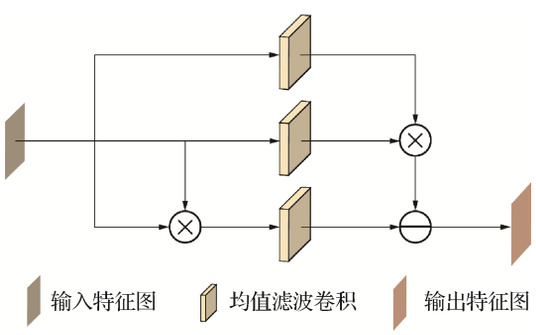

图6　方差均值滤波模块

Fig. 6　Variance mean filter module

1.2.3　逐点卷积模块

多阶段特征融合过程中，图像像素间各通道信息的相关性至关重要。逐点卷积操作可以通过跨通道的方式对特征进行整合，提高综合信息表达能力，采用如图7所示的逐点卷积模块提取低分辨率线性指导参数$\boldsymbol{A}_l$。

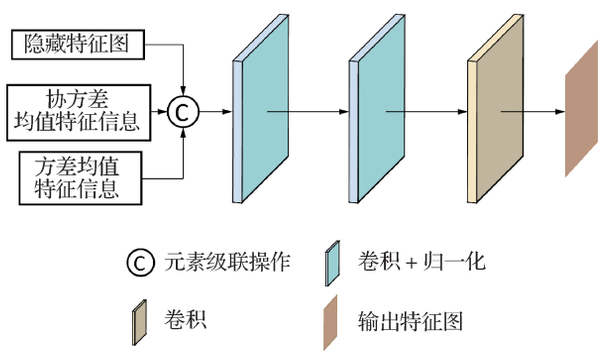

图7　逐点卷积模块

Fig. 7　Pointwise convolution block

采用隐藏特征图、协方差均值特征信息及方差均值特征信息相结合提取多阶段融合图$\boldsymbol{M}$，根据得到的多阶段融合图提取每位像素特征，计算逐像素特征图$\boldsymbol{F}_1$。具体为

$$\boldsymbol{M} = C\left(H_c, Var(\boldsymbol{F}), Cov(\boldsymbol{F})\right) \tag{12}$$

$$\boldsymbol{F}_0 = B\left(Conv^1_{1\times1}(\boldsymbol{M})\right) \tag{13}$$

$$\boldsymbol{F}_1 = B\left(Conv^1_{1\times1}(\boldsymbol{F}_0)\right) \tag{14}$$

式中，$C(\cdot)$为元素级联操作，采用Concat()函数进行级联，$Conv^1_{1\times1}(\cdot)$表示卷积核大小为1×1、步长为1的二维卷积，$B(\cdot)$为批量归一化(Ioffe和Szegedy，2015)。最后经过逐点卷积处理输出参数$\boldsymbol{A}_l$，具体为

$$\boldsymbol{A}_l = Conv^1_{1\times1}(\boldsymbol{F}_1) \tag{15}$$

1.3　损失函数

对输入的视频帧计算蒙版图和前景图的损失，为了加快模型的收敛速度，同时降低离群点的敏感度，采用标准$L1$损失函数L^{α}_1计算蒙版图$\boldsymbol{\alpha}$和真值$\hat{\boldsymbol{\alpha}}$间的损失，具体为

$$L^{\alpha}_1 = \left\| \boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}} \right\|_1 \tag{16}$$

同时，为了平衡视频帧局部信息与全局信息间的差异，引入拉普拉斯损失函数L^{α}_{lap}(Hou和Liu，2019)和时间相干损失函数L^{α}_{tc}(Sun等，2021a)，具体为

$$L^{\alpha}_{\text{lap}} = \sum_{i=1}^{5} 2^{i-1} \left\| L^i(\boldsymbol{\alpha}) - L^i(\hat{\boldsymbol{\alpha}}) \right\|_1 \tag{17}$$

$$L^{\alpha}_{\text{tc}} = \left\| \frac{\mathrm{d}\boldsymbol{\alpha}}{\mathrm{d}t} - \frac{\mathrm{d}\hat{\boldsymbol{\alpha}}}{\mathrm{d}t} \right\|_2 \tag{18}$$

式中，$L^i(\cdot)$为第i层拉普拉斯金字塔，t为时间。采用标准$L1$损失函数L^F_1及时间相干损失函数L^F_{tc}共同计算前景图$\boldsymbol{F}$与真值$\hat{\boldsymbol{F}}$之间的损失，具体为

$$L^F_1 = \left\| B(\hat{\boldsymbol{\alpha}} > \boldsymbol{0}) \cdot (\boldsymbol{F} - \hat{\boldsymbol{F}}) \right\|_1 \tag{19}$$

$$L^F_{\text{tc}} = \left\| B(\hat{\boldsymbol{\alpha}} > \boldsymbol{0}) \cdot \left(\frac{\mathrm{d}\boldsymbol{F}}{\mathrm{d}t} - \frac{\mathrm{d}\hat{\boldsymbol{F}}}{\mathrm{d}t} \right) \right\|_2 \tag{20}$$

式中，$B(\cdot)$表示布尔运算法则，则网络的总损失函数为

$$L_{\text{total}} = L^{\alpha}_1 + ML^{\alpha}_{\text{lap}} + N_1 L^{\alpha}_{\text{tc}} + L^F_1 + N_2 L^F_{\text{tc}} \tag{21}$$

式中，M、N_1、N_2为拉普拉斯损失函数L^{α}_{lap}和时间相干损失函数(L^{α}_{tc}，L^F_{tc})的权重。

1.4　训练方法

采用Pytorch架构与Adam优化器进行训练。训练基准网络时，将编码器、金字塔池化模块及循环解码器的初始学习率设置为{0.000 1，0.000 1，0.000 5}，批量大小为1。训练精细化网络时，将编码器、金字塔池化模块、循环解码器和高分辨率信息指导模块的初始学习率置为{0.000 05，0.000 05，0.000 1，

0.000 2}，批量大小为1，采用DeepLabV3预训练的官方权重初始化特征提取器。采用单张NVIDIA GTX 1080Ti GPU显卡在3种数据集上依次训练：在Video240K SD数据集上训练基准网络，输入帧序列为15，训练8轮后，在Video240K HD数据集上训练1轮精细化网络。此外，为提高模型在处理高分辨率视频上的鲁棒性，在Human2K数据集上继续训练精细化网络，令精细化网络的下采样率D为0.25，输入帧序列为2并进行50轮训练。由于AIM数据集精度较低，会导致模型训练精度下降，因此将AIM数据集仅用于测试。

1.5 数据集

VideoMatte240K数据集（Lin等，2021）给出一系列视频抠图数据。该数据集提供了484个视频片段，其中384个视频为4 K分辨率，100个为高清分辨率视频，通过AE（adobe after effects）软件生成两组不同分辨率下的同一图像数据集（高清数据集Video240K HD和标清数据集Video240K SD），包含240 709幅蒙版图和前景帧。分别采用479组视频帧用于训练，剩余5组视频帧用于验证。

Human2K数据集（Liu等，2021）提供了2 100幅高精度人体图像，平均分辨率为2 560×1 440像素。采用2 000幅图像作为训练集，100幅图像作为测试集。

AIM数据集（Xu等，2017）的训练集包含431幅图像，测试集包含50幅图像。从中选出269幅人类图像作为训练集，11幅人类图像作为测试集，图像平均分辨率为1 000×1 000像素。

背景数据集采用Lin等人（2021）给出的共享背景图，并在百度上抓取8 859幅背景图像，分辨率均在1 920×1 080像素以上，将其分别按8 832、200、20的比例构建训练集、验证集和测试集。为增强数据，引入多种噪声到数据集中，并对其作模糊处理，对所有视频帧采用随机裁剪、缩放和旋转等操作进行处理。对前景视频帧进行仿射变换、亮度、对比度和色调变化等操作，并在视频帧中加入帧率变换、随机翻转和跳帧取样等操作，以此增加数据的多样性。

2 实验评估

2.1 评估指标

与Xu等人（2017）评估方法相同，在蒙版图上采用绝对误差（sum of absolute difference，SAD）、均方误差（mean squared error，MSE）、梯度（gradient error，Grad）及连通性（connectivity error，Coon）进行客观评估。视频前景采用均方误差进行评估，此外将MSE、Grad和Coon分别缩放至10^3、10^{-3}和10^{-3}倍，以便更好地对实验结果进行对比评估。

2.2 评估结果

在3种测试数据集（Video240K SD、AIM和Human2K）上进行实验，表1给出了多种人像实时抠图结果。由表1可见，本文方法优于多种结合三分

表1 不同方法在多个数据集上的性能对比

Table 1 Comparison of different methods on multiple datasets

方法	Video240K SD					AIM					Human2K				
	蒙版图				前景图	蒙版图				前景图	蒙版图				前景图
	SAD	MSE	Grad	Coon	MSE	SAD	MSE	Grad	Coon	MSE	SAD	MSE	Grad	Coon	MSE
DIM†（Xu等，2017）	0.409	13.123	708	322	–	21.481	14.559	11 098	19 966	–	13.443	19.51	15 723	11 337	–
MGM†（Yu等，2021）	0.424	17.574	947	332	–	26.704	38.99	13 686	25 555	–	14.907	40.409	19 291	14 060	–
MF†（Park等，2022）	0.519	20.584	1 170	440	50.6	17.504	15.499	15 630	16 470	34.766	10.458	14.143	12 350	9 691	7.237
OTVM†（Seong等，2022）	0.368	11.112	508	288	8.03	17.894	22.667	32 167	15 783	37.974	20.578	36.091	32 695	19 936	21.224
AEM†（Liu等，2023）	0.314	8.28	418	225	–	**10.603**	**5.569**	**4 898**	**8 595**	–	6.557	6.617	4 965	5 421	–
BGM（Sengupta等，2020）	0.511	19.038	1 000	445	9.312	27.721	38.76	44 425	27 275	48.999	25.095	55.31	46 187	24 635	24.65
BGMv2(Lin等,2021)	0.526	23.223	1 226	445	122.524	15.408	12.828	12 944	13 640	**27.61**	8.89	10.736	9 690	7 879	8.616
本文	**0.232**	**4.083**	**180**	**136**	**7.683**	**11.612**	**6.918**	**6 783**	**10 582**	**30.192**	**5.32**	**4.023**	**2 940**	**4 285**	**5.14**

注：加粗字体表示各列最优结果，“–”表示该方法无法估计前景图像，†代表需要手动制作三分图的方法。DIM：deep image matting；MGM：mask guided matting；MF：matteformer；OTVM：one-trimap video matting；BGM：background matting；BGMv2：real-time high-resolution background matting。

图与背景的方法DIM（Xu等，2017）、MGM（Yu等，2021）、MF（Park等，2022）、OTVM（Seong等，2022）、AEM（Liu等，2023）、BGM（Sengupta等，2020）和BGMv2（Lin等，2021），在AIM数据集的测试上各种指标略低于AEM方法，然而AEM方法需要通过手动设置三分图作为辅助信息，本文方法只需设置一幅背景图像。

图8给出了本文方法与相关方法的可视化实验结果。第1组结果中，其他方法预测的舞者手部（红色方框）蒙版图较为模糊或为半透明状，而本文方法可有效给出手部清晰的边缘细节。第2组结果中，与其他的方法相比，本文方法在鲜花边缘处（红色方框）抠图结果更加精细准确。第3组实验中，其他方法在人像的发丝部分（红色方框）未给出清晰的抠图结果，而本文方法更多地保留了人物的发丝细节信息。最后一组多人抠图实验中，其他方法在人物手臂交叉处（红色方框）未给出正确抠图结果，本文方法可以清晰地预测出人物各自的手臂。由此可见，本文方法可以更好地提取图像细节信息，从而提升人像抠图质量。

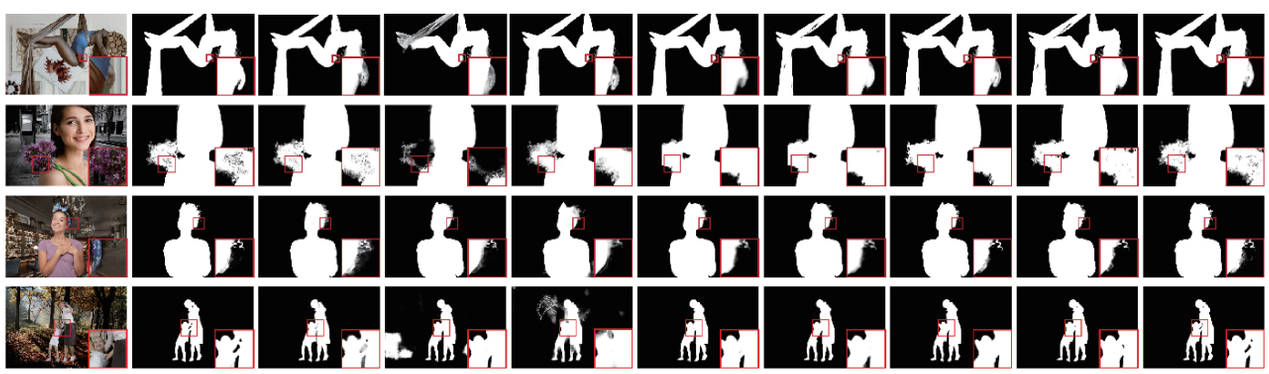

图8　可视化实验结果对比

Fig. 8　Comparison of visual experimental results((a)input images;(b)ground truth;(c) DIM (Xu et al., 2017);(d)MGM (Yu et al., 2021);(e)MF(Park et al.,2022);(f)OTVM (Seong et al., 2022);(g)AEM(Liu et al., 2023);(h)BGM(Sengupta et al.,2020);(i)BGMv2 (Lin et al., 2021);(j)ours)

2.3　性能评估

采用Vladislav Sovrasov测量模型的参数量（parameters）与乘加运算量（GMac）评估网络性能，结果如表2及表3所示。可以看出，与DIM、MGM、MF、OTVM、AEM、BGM及BGMv2方法相比，本文方法产生的模型参数量更小。处理多种分辨率（resolution）视频时，本文方法可以达到实时需求（见表3）。采用NVIDIA GTX 1080Ti GPU进行实验，本文方法在低分辨率（512 × 288像素）和中分辨率（1 024 × 576像素）视频上的处理速度分别为49帧/s和42.4帧/s，在HD（1 920 × 1 080像素）及4K（3 840 × 2 160像素）视频上的处理速度分别为43帧/s及26帧/s，与同类方法相比，本文方法获得了更佳的实时性。此外，若采用MobileNetV2作为编码器主干，执行速率将进一步提升，且参数量更小。

表2　模型参数量和大小对比

Table 2　Comparison of model parameter quantity and size

方法	Backbone	参数量/M	尺寸/MB
DIM（Xu等，2017）	-	25.58	307.1
MGM（Yu等，2021）	-	29.6	356.1
MF（Park等，2022）	-	44.9	513
OTVM（Seong等，2022）	-	58.6	282
AEM（Liu等，2023）	-	52	208
BGM（Sengupta等，2020）	-	72.23	275.53
BGMv2（Lin等，2021）	ResNet50	40.25	161.36
本文	ResNet50$^{\Delta}$	27.92	112.06
本文	ResNet101	46.92	188.35
本文	MobileNetV2	**3.30**	**13.48**

注：加粗字体表示各列最优结果，Δ代表本文的模型，“-”表示没有此选项。

此外，为探索拉普拉斯损失函数L_{lap}^{α}和时间相干损失函数（L_{tc}^{α}、L_{tc}^{F}）的权重M、N_1、N_2对网络结构的影响。本文进行了定量分析，依次改变权重大小观察其对指标SAD（绝对误差）的影响。如图9所示，当

表 3　不同方法间的性能比较

Table 3　Performance comparison between different methods

方法	分辨率/像素	下采样率	Backbone	处理速率(帧/s)	GMac
DIM(Xu等,2017)	HD	–	–	3.66	1 012.19
MGM(Yu等,2021)	HD	–	–	5.74	39.35
MF(Park等,2022)	HD	–	–	8.23	233.3
OTVM(Seong等,2022)	HD	–	–	1.04	245.8
AEM(Liu等,2023)	HD	–	–	7.25	295.4
BGM(Sengupta等,2020)	512 × 288	–	–	10.75	403.58
BGMv2(Lin等,2021)	HD	0.25	ResNet50	38.46	31.95
本文	512 × 288	1	ResNet50$^{\Delta}$	49	32.4
本文	512 × 288	1	ResNet101	38.2	43.3
本文	512 × 288	1	MobileNetV2	**70.9**	**11.5**
本文	1 024 × 576	0.5	ResNet50$^{\Delta}$	42.4	32.4
本文	1 024 × 576	0.5	ResNet101	35.8	43.3
本文	1 024 × 576	0.5	MobileNetV2	**62.4**	**11.6**
本文	HD	0.25	ResNet50$^{\Delta}$	43.4	28.67
本文	HD	0.25	ResNet101	35.7	38.37
本文	HD	0.25	MobileNetV2	**58.8**	**10.25**
本文	4 K	0.125	ResNet50$^{\Delta}$	26.3	28.78
本文	4 K	0.125	ResNet101	22.7	38.48
本文	4 K	0.125	MobileNetV2	**28.5**	**10.36**

注:加粗字体表示各列最优结果,Δ代表模型最终应用时的backbone模型,“–”表示没有此选项。

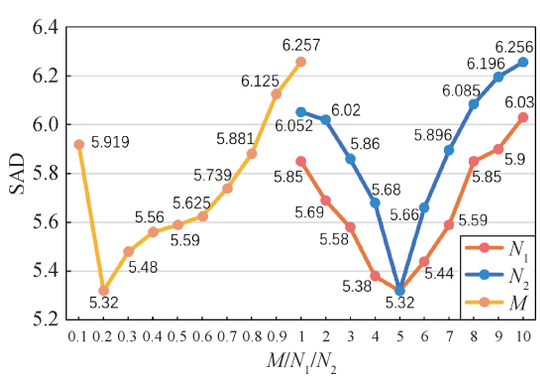

图 9　损失函数权重影响

Fig. 9　Loss function weight influence

M为0.2，N_1、N_2为5时获得的SAD最优。

2.4　消融实验

在消融实验中,针对改进的特征提取网络(编码器)、轻量级金字塔池化模块、循环解码器中的残差门控循环单元模块,以及精细化网络中的高分辨率信息指导模块进行消融实验,以验证本文方法的有效性。

2.4.1　编码器的作用

表4给出了分别采用改进的ResNet50(本文)、ResNet101和MobileNetV2网络作为编码器在Human2K测试集上的测试结果,与后两者相比,改进的ResNet50可以获取更精细的蒙版图,更适用于本文特征提取的任务。

2.4.2　金字塔池化模块与循环解码器模块的作用

通过在基准网络中增加相应模块,验证基准网

表 4　编码器的有效性

Table 4　Effectiveness of encoder

方法	蒙版图			
	SAD	MSE	Grad	Coon
MobileNetV2	8.95	9.481	7 389	8 361
ResNet101	6.735	6.299	4 798	5 766
本文	**5.320**	**4.023**	**2 940**	**4 285**

注:加粗字体表示各列最优结果。

络中金字塔池化模块(PPM)与循环解码器模块(R-ConvGRU)对最终结果产生的影响,表5为在Video240K SD测试集上的测试结果。其中,对于基准网络模型(baseline),网络中未含有PPM模块与R-ConvGRU模块;“+PPM”模型为在“baseline”模型中仅增加PPM模块;“+R-ConvGRU”模型(本文)为在“+PPM”模型基础上增加R-ConvGRU模块。采用相同的学习率和输入序列帧数训练每种模型,其中特征提取网络均采用改进的ResNet50结构。由表5可见,PPM模块可以提高蒙版图上的所有评估指标,而R-ConvGRU模块在指标上的提高更加显著,分别在SAD、MSE、Grad、Coon指标上降低了17.2%、27.3%、32.8%和23.7%。综上,实验结果验证了在PPM和R-ConvGRU模块的共同作用下,可以获得更为准确的蒙版图。

表5 PPM与R-ConvGRU模块的有效性

Table 5 Effectiveness module of PPM and R-ConvGRU

方法	蒙版图			
	SAD	MSE	Grad	Coon
baseline	0.373	9.112	372	295
+ PPM	0.353	8.616	359	274
本文	**0.292**	**6.262**	**241**	**209**

注:加粗字体表示各列最优结果。

2.4.3 精细化网络的作用

为验证精细化网络对高分辨率视频抠图的作用,在Human2K测试集上进行实验:一种采用不包含精细化网络的基准网络(base),另一种是包含精细化网络的总体网络(overall)。采用与2.4.2节相同的方法进行实验,结果如表6所示。由表6可见,总体网络分别在SAD、MSE、Grad及Coon指标上降低了10.1%、19%、12.4%和13.2%,因此验证了在精细化网络的作用下,总体网络可以更好地预测蒙版图。

表6 精细化网络的有效性

Table 6 Effectiveness of refined network

方法	蒙版图			
	SAD	MSE	Grad	Coon
基准网络(不包含精细化网络)	5.919	4.972	3 358	4 937
总体网络(包含精细化网络)	**5.320**	**4.023**	**2 940**	**4 285**

注:加粗字体表示各列最优结果。

3 结 论

本文给出一种实时高分辨率的视频抠图方法,采用基准网络中金字塔池化模块提取并融合视频帧的多尺度信息,通过残差门控循环单元模块聚合连续帧间时间信息。通过高分辨率信息指导模块捕获图像中的高分辨率信息,指导低分辨率图像计算高质量蒙版图与前景图。在高分辨率数据集Human2K上进行实验,结果表明,本文的网络明显优于对比的同类方法,不仅可以获得更精细的人像抠图结果,而且在处理速度上具备较高的实时性,能够为后续高级应用提供更好的支持。此外,本文方法仍存在一定的局限性,将背景作为辅助信息限制了网络在动态背景中的应用,未来将进一步探索动态背景环境下的高分辨率实时视频抠图方法。

参考文献(References)

Aksoy Y, Aydin T O and Pollefeys M. 2017. Designing effective inter-pixel information flow for natural image matting//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 228-236 [DOI: 10.1109/CVPR.2017.32]

Chen G W, Liu Y, Wang J, Peng J C, Hao Y Y, Chu L T, Tang S Y, Wu Z W, Chen Z Y, Yu Z L, Du Y N, Dang Q Q, Hu X G and Yu D H. 2022. PP-matting: high-accuracy natural image matting [EB/OL] [2023-04-03]. https://arxiv.org/pdf/2204.09433.pdf

Chen Q F, Li D Z Y and Tang C K. 2013a. KNN matting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(9): 2175-2188 [DOI: 10.1109/TPAMI.2013.18]

Chen X W, Zou D Q, Zhou S Z, Zhao Q P and Tan P. 2013b. Image matting with local and nonlocal smooth priors//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 1902-1907 [DOI: 10.1109/CVPR.2013.248]

Chuang Y Y, Curless B, Salesin D H and Szeliski R. 2001. A Bayesian approach to digital matting//Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, USA: IEEE: 1063-1069 [DOI: 10.1109/CVPR.2001.990970]

He K M, Rhemann C, Rother C, Tang X O and Sun J. 2011. A global sampling method for alpha matting//Proceedings of CVPR 2011. Colorado Springs, USA: IEEE: 2049-2056 [DOI: 10.1109/CVPR.

2011.5995495］

He K M, Sun J and Tang X O. 2010. Fast matting using large kernel matting Laplacian matrices//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE: 2165-2172［DOI: 10.1109/CVPR.2010.5539896］

Hong X, Yang Y Y and Wen S H. 2018. Improving comprehensive sampling sets matting using texture feature//Proceedings of 2018 IEEE 4th International Conference on Computer and Communications (ICCC). Chengdu, China: IEEE: 1617-1621［DOI: 10.1109/CompComm.2018.8780703］

Hou Q Q and Liu F. 2019. Context-aware image matting for simultaneous foreground and alpha estimation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 4129-4138［DOI: 10.1109/ICCV.2019.00423］

Ioffe S and Szegedy C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift［EB/OL］.［2015-03-02］. https://arxiv.org/1502.03167.pdf

Jin Y, Li Z X, Zhu D M, Shi M and Wang Z Q. 2022. Automatic and real-time green screen keying. The Visual Computer, 38(9): 3135-3147［DOI: 10.1007/s00371-022-02542-x］

Lee P and Wu Y. 2011. Nonlocal matting//Proceedings of CVPR 2011. Colorado Springs, USA: IEEE: 2193-2200［DOI: 10.1109/CVPR.2011.5995665］

Levin A, Rav-Acha A and Lischinski D. 2008. Spectral matting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(10): 1699-1712［DOI:10.1109/TPAMI.2008.168］

Lin S C, Ryabtsev A, Sengupta S, Curless B, Seitz S and Kemelmacher-Shlizerman I. 2021. Real-time high-resolution background matting//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 8758-8767［DOI: 10.1109/CVPR46437.2021.00865］

Lin S C, Yang L J, Saleemi I and Sengupta S. 2022. Robust high-resolution video matting with temporal guidance//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 3132-3141［DOI: 10.1109/WACV51458.2022.00319］

Liu T Y, Qiu J, He D and Liu C. 2022. Light field alpha matting based on spatial-angular consistency. Acta Optica Sinica, 42(16): #1612003 (刘天艺, 邱钧, 何迪, 刘畅. 2022. 基于空角一致性的光场抠图. 光学学报, 42(16): #1612003［DOI: 10.3788/AOS202242.1612003］

Liu Y H, Xie J K, Shi X, Qiao Y, Huang Y J, Tang Y and Yang X. 2021. Tripartite information mining and integration for image matting//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 7535-7544［DOI: 10.1109/ICCV48922.2021.00746］

Liu Q L, Zhang S P, Meng Q L, Li R, Zhong B N and Nie L Q. 2023. Rethinking context aggregation in natural image matting［EB/OL］.［2023-04-03］. https://arxiv.org/pdf/2304.01171.pdf

Park G, Son S, Yoo J, Kim S and Kwak N. 2022. MatteFormer: Transformer-based image matting via prior-tokens//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 11686-11696［DOI: 10.1109/CVPR52688.2022.01140］

Sengupta S, Jayaram V, Curless B, Seitz S M and Kemelmacher-Shlizerman I. 2020. Background matting: the world is your green screen//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 2288-2297［DOI: 10.1109/CVPR42600.2020.00236］

Seong H, Oh S W, Price B, Kim E and Lee J Y. 2022. One-Trimap video matting［EB/OL］［2023-04-03］. https://arxiv.org/pdf/2207.13353.pdf

Shahrian E, Rajan D, Price B and Cohen S. 2013. Improving image matting using comprehensive sampling sets//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 636-643［DOI: 10.1109/CVPR.2013.88］

Song S F. 2022. Attention-based Memory video portrait matting［EB/OL］［2023-04-03］. https://arxiv.org/pdf/2203.06890.pdf

Sun Y N, Tang C K and Tai Y W. 2021a. Semantic image matting//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 11115-11124［DOI: 10.1109/CVPR46437.2021.01097］

Sun Y N, Wang G Z, Gu Q, Tang C K and Tai Y W. 2021b. Deep video matting via spatio-temporal alignment and aggregation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 6971-6980［DOI: 10.1109/CVPR46437.2021.00690］

Wang J and Cohen M F. 2007. Optimized color sampling for robust matting//Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE: 1-8［DOI: 10.1109/CVPR.2007.383006］

Wu H K, Zheng S, Zhang J G and Huang K Q. 2018. Fast end-to-end trainable guided filter//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1838-1847［DOI: 10.1109/CVPR.2018.00197］

Wu Y E, He F Z, Zhang S L, Chen Z and Huang Z Y. 2010. A simple stroke-based iterative image matting approach. Journal of Image and Graphics, 15(12): 1769-1775 (吴玉娥, 何发智, 张胜龙, 陈钊, 黄志勇. 2010. 一种基于简单笔画交互的迭代图像抠图方法. 中国图象图形学报, 15(12): 1769-1775)［DOI: 10.11834/jig.20101206］

Xu N, Price B, Cohen S and Huang T. 2017. Deep image matting//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 311-320［DOI: 10.1109/CVPR.2017.41］

Xu Z B and Yang Y J. 2020. Fast portrait automatic matting based on multi-task deep learning. Engineering Journal of Wuhan University, 53(8): 740-745, 752 (许征波, 杨煜俊. 2020. 基于多任务深度学习的快速人像自动抠图. 武汉大学学报(工学版), 53(8): 740-745, 752) [DOI: 10.14188/j.1671-8844.2020-08-013]

Yu Q H, Zhang J M, Zhang H, Wang Y L, Lin Z, Xu N, Bai Y T and Yuille A. 2021. Mask guided matting via progressive refinement network//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 1154-1163 [DOI: 10.1109/CVPR46437.2021.00121]

Zhang Y K, Wang C, Cui M M, Ren P R, Xie X S, Hua X S, Bao H J, Huang Q X and Xu W W. 2021. Attention-guided temporally coherent video object matting//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: ACM: 5128-5137 [DOI: 10.1145/3474085.3475623]

Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 6230-6239 [DOI: 10.1109/CVPR.2017.660]

作者简介

彭泓,女,副教授,主要研究方向为智能感知、人工智能。
E-mail: penghong861@163.com
张家宝,通信作者,男,硕士研究生,主要研究方向为视频抠图。E-mail: zhang_jiabao316@163.com
贾迪,男,教授,主要研究方向为立体匹配与三维重建、摄影测量、视觉空间定位、视觉引导特种机械臂作业。
E-mail: lntu_jiadi@163.com
安彤,女,硕士研究生,主要研究方向为光流估计。
E-mail: 1319423118@qq.com
蔡鹏,男,硕士研究生,主要研究方向为立体匹配与三维重建。E-mail: pengcai980328@gmail.com
赵金源,男,硕士研究生,主要研究方向为人体姿态估计。
E-mail: lntu_zjy@163.com