文献引用格式：党雪云，王剑．融入多特征的篇章级新闻要素关系抽取[J]．电视技术，2022，46(6)：73-78.
DANG X Y，WANG J.Discourse-level news element relation extraction incorporating multi-features[J].Video Engineering，2022，46(6)：73-78.
中图分类号：TN931.3　　**文献标识码**：A　　**DOI**：10.16280/j.videoe.2022.06.016

融入多特征的篇章级新闻要素关系抽取

党雪云[1,2]，王　剑[1,2*]
（1.昆明理工大学 信息工程与自动化学院，云南　昆明　650500；2.云南省人工智能重点实验室，云南　昆明　650500）

摘要：随着互联网信息技术高速更新迭代，新闻文本信息在以指数级的速度增多。面对海量的新闻文本信息，如何自动提取长篇新闻文本中要素与要素之间的关系，成为研究的重点。篇章级新闻要素关系抽取是指从篇章级新闻文本中跨句子识别要素之间的关系信息，有助于加速人们对整篇新闻文本脉络的理解。本文以舆情新闻文本为例，提出融入多特征的篇章级新闻要素关系抽取方法，通过异构图模型将句子间的邻接关系、从属关系、句法依赖关系、要素间的多跳关系等多种特征进行融合，充分挖掘文本中潜在的上下文信息。在构建的篇章级舆情新闻要素关系数据集上的实验结果表明，融入的多种特征对要素关系抽取的性能均有明显的提升，F1值最高提升了4.09%，较目前主流方法取得了更好的效果。
关键词：舆情新闻文本信息；篇章级要素关系抽取；异构图模型

Discourse-Level News Element Relation Extraction Incorporating Multi-Features

DANG Xueyun[1,2],WANG Jian[1,2*]
(1.*Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650500, *China*;
2.*Yunnan Key Laboratory of Artificial Intelligence, Kunming* 650500, *China*)

Abstract: With the rapid update and iteration of Internet information technology, Internet news text information is also increasing at an exponential rate. In the face of massive news text information, how to automatically extract the relation between elements of long news texts has become the focus of research. The extraction of discourse-level news element relation refers to extracting the relation information between elements across sentences from news texts contained multiple sentences, which helps people to understand the context of the whole news text. Taking the extraction of element relation in the field of public opinion news as an example, proposes a method for extracting the relation between discourse-level news elements that incorporates multi-features, the adjacent relations, affiliation, syntactic dependency relation between sentences, and multi-hop relations between elements into a heterogeneous graph model, fully mine the potential context information of news. The experimental results on the discourse-level news element relation dataset show that multi-features can improve the extraction performance obviously, and the F1 value can be improved by 4.09% at most, which is better than the current mainstream method.
Keywords: discourse-level element relation extraction; heterogeneous graph; public opinion news text information

0　引　言

新闻要素关系抽取可以看作实体关系抽取（Relation Extraction，RE）任务。实体关系抽取是指抽取两个实体之间可能存在的语义关系，是信息抽取、构建问答系统的关键基础任务之一。以涉案舆情新闻为例，法院与人之间包含“审判”关系，人与罪名之间包含“涉嫌罪名”关系，原告和被告之间包含“涉事双方”关系等，从新闻中自动抽取这

基金项目：国家重点研发计划（No.2018YFC0830105；No.2018YFC0830101；No.2018YFC0830100）。
作者简介：党雪云（1994—），女，硕士，研究方向为自然语言处理、信息检索。
通信作者：王　剑（1976—），男，硕士，副教授，研究方向为软件演化、机器学习、智能计算等。E-mail：1542126163@qq.com。

些关系，对于人们快速理解舆情信息起着重要作用。当前，篇章级的要素关系抽取任务面临标注数据较少、任务复杂度更高的问题，导致抽取效果不佳，是一个值得研究的方向。现有的关系抽取方法主要侧重于从单个句子中抽取要素关系，通过对大量新闻文本进行分析会发现很多实体关系常常跨句子存在，如图 1 所示，通过整篇文本可分析出“品某良”和“张某雷”两者都是案件当事人，明显存在关系，但仅从其中某一个句子并不能抽取两者之间存在的关系，因为二者没有在同一句子同时出现过。通过对文章中多个句子中的要素关系进行分析，结合上下文语义，才能推断出两者之间存在的关系。因此，本文提出一种通过异构图模型融合多个句子的邻接关系、从属关系、句法依赖关系、要素间的多跳关系等多种特征的方法，通过挖掘篇章级文本中潜在的上下文信息，提升跨句子要素关系抽取的准确率和性能。

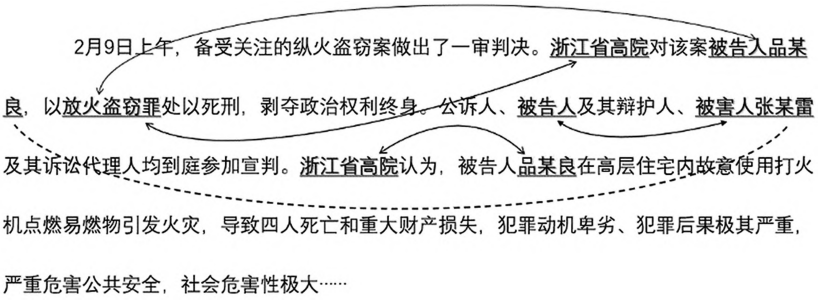

图 1　篇章级要素关系问题分析

1　相关工作

目前，按照训练文本的类型，关系抽取任务可以分为句子级关系抽取和篇章级关系抽取两大类，本文主要针对篇章级的要素关系抽取任务。篇章级关系抽取的目的主要是识别出整篇文章中要素之间的关系，包括单个句子中存在的实体关系，也包括跨多个句子存在的要素关系。根据输入文本的结构，可以将篇章级的实体关系抽取方法分为基于序列的篇章级实体关系抽取模型和基于图的篇章级实体关系抽取模型两类。

基于序列的篇章级实体关系抽取模型利用不同的序列编码获得词语表示，之后通过平均池化、注意力池化等各种池化操作计算实体关系的表示。ZENG 等人[1]使用卷积神经网络进行正则化的研究，利用外部知识资源如 WordNet、位置相关特征、词对信息以及词汇特征集等特征拼接为特征向量作为输入，进行关系分类；WANG 等人[2]在卷积神经网络架构上引入了一种新的多级注意力机制来捕获特定于要素的注意力和特定于目标关系的注意力，使其能够检测到更微妙的线索以自动学习与关系分类相关的部分；HE 等人[3]提出了一种带有注意力机制的长短期记忆（Long-Short Term Memory，LSTM）网络，该方法避免了标注数据存在误报，在提取过程中不采用人为设计的规则来提升效率，因此本研究利用词级别的注意特征提取关系，结合实例级别的注意机制处理数据中的误报问题；MIWA 等人[4]讨论的一种方法使用了双向 LSTM，将实体识别视为序列标注问题，模型嵌入层主要处理单词、依赖类型、词性标签及要素标签的嵌入，序列层主要用于单词在句子中的顺序信息，下一层通过一个神经网络，从左向右以一种贪心的策略分配要素标签，最后一个单词的标签用来预测当前的单词标签，最后一层提取预测到的要素之间的关系；GAO 等人[5]提出了神经雪球的方法，只需要使用少数的新关系样例，便可利用现有关系的先验知识从未标注数据中迭代地积累新的实例和事实，从而训练一个较好的神经关系分类器，实验结果进一步表明了其模型的效率和鲁棒性。

为了进一步捕获长期依赖关系，基于图的实体关系抽取模型被提出，通过构造图结构，距离较远的单词或者要素均可以成为相邻节点。相对序列编码器而言，图编码器可以聚合来自所有邻居节点的信息以捕获更长的依赖关系。ZENG[6] 等人为了更好地处理篇章级关系抽取任务，提出一种双图模型，引入了一种要素级别的异构图和一种图神经网络来模拟文章中不同要素之间的交互，他们还引入了要素级图并提出了一种新的路径推理机制，用于要素之间的关系推理；CHRISTOPOULOU[7] 等人提出一种新的面向边的图神经网络模型用于篇章级关系抽

取，该模型不同于现有模型，它专注于构建独特的节点和边，将信息编码为边表示而不是节点表示；ZHANG[8] 等人提出一种新颖的篇章级关系抽取模型，该模型构建双层异构图用于连续建模文章结构并实现关系推理；YANG[9] 等人针对关系抽取提出了两种树结构的图卷积神经网络的改进策略，一种策略是集成层次化注意力机制和主体、对象之间的相关性分析分别生成句子和要素向量，另一种策略合并命名实体识别子网络和图卷积网络树结构，以实现关系抽取和要素抽取的联合学习。

2 基于双层异构图的篇章级要素关系抽取方法

2.1 异构图的构建方法

本文将篇章级的实体关系抽取任务定义如下：给定一篇标注文章 $D=\{S_i\}_{i=1}^{n_s}$，实体集合为 $V=\{\varepsilon_i\}_{i=1}^{n_e}$，其中 $S_i=\{w_j\}_{i=1}^{n_w^i}$ 表示第 i 个句子中有 n_w^i 个单词，而 $\varepsilon_i=\{m_j\}_{i=1}^{n_m}$ 表示第 i 个实体中有 n_m^i 个单词，最终的目标是预测每个实体对之间的所有句内和句间关系 $R' \in R=\{r_i\}_{i=1}^{n_r}$。经过对长文本的大量分析发现，许多要素关系其实是跨多个句子存在的，所以篇章级的关系抽取任务比传统的句子级的关系抽取任务要更复杂，篇章级的关系抽取模型需要较强的语义建模能力和关系推理能力。

图 2 是本文关系抽取模型的系统架构图。该模型主要分为五层：输入层主要负责将输入的词进行向量化表征，文本编码层是任意的序列编码器，用于为每个单词生成上下文表示；结构化建模层负责建模文本中固有的结构信息，包括文本的邻接关系、从属关系以及句法依赖关系；关系推理层负责捕获文本中要素间的多跳关系，最后是输出层，负责输出可能存在的要素关系，相当于一个多标签分类层。

2.2 输入层

输入层负责对单词的语义信息、扩充信息进行编码并嵌入到单词的输入特征中。具体来说，就是先使用 d_w 维的词向量 $\boldsymbol{w}_i$ 来表征文本的上下文语义信息，再增加要素的类型表征 t_i 用于表征每个要素的类型信息；其次，增加指代特征 c_i 用于标记指代词所属的要素，帮助模型获取要素共指的信息；最后将这三种表征拼接起来构成输入特征 $x_i=[w_i;t_i;c_i] \in R^{d_x}$，其中 $[\cdot;\cdot]$ 表示向量拼接的操作，$d_x=d_w+d_t+d_c$。

2.3 文本编码层

文本编码层负责捕获单词的上下文信息。具体来讲，把整篇文章看作一个包含 n 个单词的长序列，然后使用序列编码器双向 LSTM 来编码长序列中每个单词的上下文信息。若将 LSTM 单元对 x_i 的操作表示为 $LSTM(x_i)$，则该单词的上下文语义信息可以表示为：

$$h_i = F\left[\overrightarrow{LSTM}\left(x_i\right);\overleftarrow{LSTM}\left(x_i\right)\right] \tag{1}$$

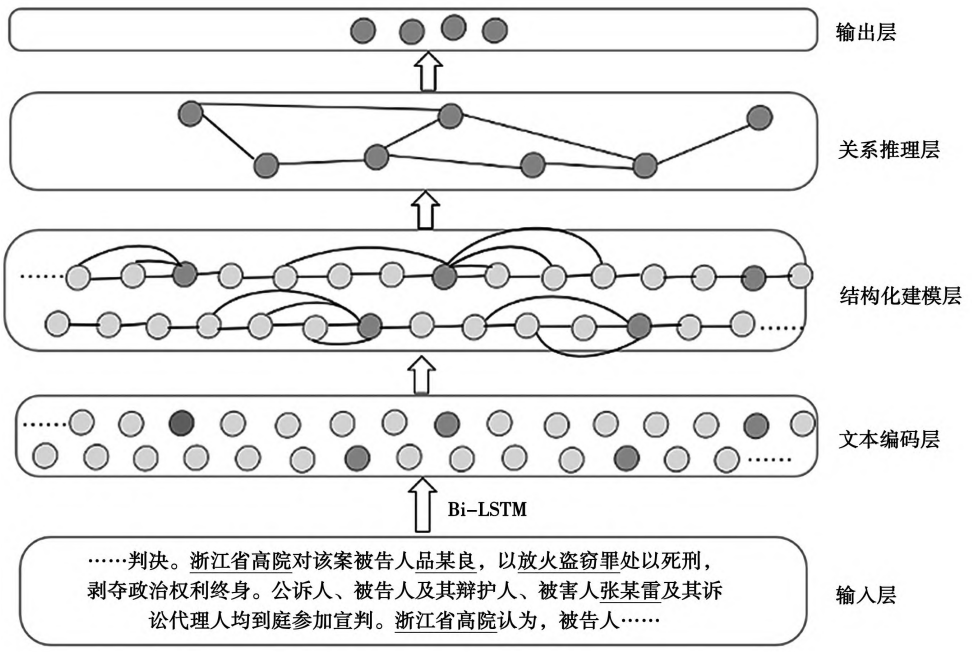

图 2 双层异构图模型

式中：h_i和F是一个线性函数，$h_i \in R^{d_h}$，$F:R^{2\times d_h}\rightarrow R^{d_h}$，$d_h$表示LSTM单元的隐藏层的维度。通过这种方式，可以捕获特定时间单词的前向状态和后向状态的特征表示，最后使用$\boldsymbol{HW}=\{h_1,h_2,\cdots,h_n\}$作为输入序列的表征向量。

2.4 结构化建模层

结构化建模层将文本序列的每一个句子、每一个单词均视为图中的一个节点。通常，一篇文章由多个句子组成，一个句子由多个单词组成，所以本文采用以下5种类型的边来建模文章内在结构信息：

（1）字－字邻接边，在文章每两个相邻的字节点之间建立一条边，以保持文章中每个字的自然顺序结构；

（2）句子－句子邻接边，在文章每两个相邻句子节点之间建立一条边，以保持文章中句子间的自然顺序结构；

（3）句子－句子补全边，将文章中没有相邻的句子节点之间连接一条边，以增强图结构的连通性；

（4）词－词依赖关系边，为了对语法结构进行编码，如果两个单词节点在句子级的依存关系树中相邻的话，则在它们之间连一条边；

（5）词－句子关联边，为了建模文章的层次结构，将单词节点和它们所在的句子节点之间连一条边。

结构化建模层直接利用文本编码层的输出作为单词节点的初始化特征，对每个句子中的所有单词节点进行最大池化操作得到句子节点的表示，即，最后$s=\max\{h_j\}_{i=1}^{n_w}$，然后利用图神经网络中常用的消息传播策略更新单词和句子节点的表示：

$$(\overline{H}_W,\overline{H}_S)=WR(H_W,H_S) \tag{2}$$

式中：$H_s=\{s_1,\cdots,s_{N_s}\}$指一篇文章中所有句子节点表征的集合，$H_W$是输入序列的词表征，$WR(\cdot)$表示图神经网络结构中的消息传播机制。最后对于每个单词节点，将其在WR之前和之后的特征拼接起来作为输出的表示$\hat{h}_i=F([h_i;\hat{h}_i])$，这种表示方式结合了词节点和句子节点的顺序特征和结构特征，为下一步推理提供基础。

2.5 关系推理层

关系推理层中将要素提及及要素当作图中的节点，建立如下4种类型的边：

（1）提及共现边，在同一句中的两个提及或要素之间建立一条边，用于表征句内关系；

（2）提及共指边，如果两个提及节点指向同一个实体，则在它们之间连一条边，用于表征句内关系；

（3）提及要素关联边，如果提及指向某要素，则在它们之间连一条边，用于传递提及层面的消息到要素层面；

（4）要素－要素互补边，将所有要素两两之间连一条边，用于防止出现不连通图，增强多跳关系。

具体来讲，对于文本中第s个单词到第t个单词组成的提及m，将其表征初始化为$m=1/[(s-t+1)\sum_{k=s}^{t}\hat{h}_k]$，则一个要素$e$的表征可以表示为其所有提及表征的平均值，即$e=(\sum_j m_j)/(nm)$，与结构化建模层中的消息传播机制类似，$(\overline{H}_E,\overline{H}_M)=WR(H_E,H_M)$，其中$H_M$和$H_E$分别指提及节点和要素节点的表征集合，经过$L$次的消息传递之后，便能得到所有节点的最终表征。

2.6 输出层

将关系预测看作一个多标签分类问题，对于每个要素对(e_i,e_j)，将这些要素特征和相对距离表征向量拼接起来，并使用一个双线性函数来计算每个关系的概率：

$$\begin{cases}\hat{e}_i=\left[\overline{e}_i;d_{ij}\right]\\ \hat{e}_j=\left[\overline{e}_j;d_{ji}\right]\end{cases} \tag{3}$$

$$y=sigmoid(\hat{e}_i^T W e_j+\boldsymbol{b}) \tag{4}$$

式中：$d_{ij}\in R^{dd}$和$d_{ji}\in R^{dd}$指文章中第一次提到的两个要素间的相对距离表征，$W\in R^{d\times nr\times d}$是双仿射学习参数，$\boldsymbol{b}\in R^d$指的是偏差向量，其中$d=d_h+d_d$，$y\in R^{nr}$表示所有关系的预测值。

3 实 验

3.1 数据集

本文使用的新闻要素关系语料集一共包含1 200篇新闻文本数据，共4类关系。其中，涉案人员－涉案人员关系共2 352组，涉案人员－受理法院关系共1 348组，涉案人员－涉嫌罪名关系共1 732组，受理法院－判处罪名关系共1 285组，具体信息如表1所示。

表1 涉案舆情篇章级要素关系抽取语料库信息

类型	数量
篇章级文本数	1 200篇
涉案人员－涉案人员关系	2 352组
涉案人员－受理法院关系	1 348组
涉案人员－涉嫌罪名关系	1 732组
受理法院－判处罪名关系	1 285组

3.2 实验设置

实验采用128维的词向量对输入文本进行初始化，得到其向量化表示。训练时，Dropout设置为0.8，学习率lr设置为0.01，训练轮次epoch设置为200，batch_size设置为10，优化器使用SGD。

3.3 评价标准

本文采用准确率（Precision，P）、召回率（Recall，R）和F1值（F1-Measure，F1）作为评价指标，其计算公式如下：

$$P=TP/(TP+FP) \tag{5}$$

$$R=TP/(TP+FN) \tag{6}$$

$$F1=(2\times P\times R)/(P+R) \tag{7}$$

式中：*TP*表示把正例预测为正的概率，*FP*表示把负例预测为正的概率，*FN*表示把正例预测成负的概率。

3.4 实验结果

本文采用如下3个基准模型：ME-CNN模型[10]使用具有语言特征的最大熵模型、具有多级语义特征的卷积神经网络分别用于提取句间要素关系和句内要素关系，并在训练阶段考虑要素之间的上位词关系以构建更精确的训练实例；RPCNN模型[11]提出一个将领域知识、注意力机制、分段池化以及多实例学习策略结合的篇章级循环分段卷积神经网络；GCNN模型[12]提出了一种使用图卷积神经网络来捕获本地和非本地依赖关系的句间关系抽取模型，在篇章级的图上构建了一个带标签边的图卷积神经网络，这也是在篇章级关系抽取中利用图神经网络的首次尝试。实验结果如表2所示，本文模型与其他模型相比，F1值有0.46 ~ 4.09个百分点的提升；对比ME-CNN和RPCNN，实验结果表明图神经网络模型的确具有一定的优越性；对比GCNN，结果表明了双层异构图在要素关系抽取任务上的多跳推理能力。

表2 本文模型与基准模型实验对比结果

模型	*P*/%	*R*/%	*F*1/%
ME-CNN	58.61	60.20	59.63
RPCNN	61.20	62.10	62.51
GCNN	63.15	59.82	62.26
本文模型	65.26	61.33	63.72

本文还进行了消融实验，逐一去掉模型中不同类型的边特征进行实验，实验结果如表3所示。

表3 消融实验结果

模型	*P*/%	*R*/%	*F*1/%
本文模型	65.26	61.33	63.72
去掉字字邻接边	62.56	60.63	61.24
去掉词句关联边	61.56	61.49	62.62
去掉句子去掉句子邻接边	63.24	60.56	62.12
去掉提及共现边	62.34	60.28	61.26
去掉提及实体关联边	61.36	60.84	61.69
去掉实体去掉实体互补边	60.47	60.82	62.26

分析表3的结果可知，字字邻接边表征了文章中每个字的自然顺序，对模型的准确率做出了较大贡献；词句关联边对模型的层次进行建模，有效提高了模型的性能；句子-句子邻接边保证了句子的顺序结构，也增强了模型的准确性；提及共现边捕捉了提及之间的全局关系；提及实体关联边传递提及和实体间的关系。这些边的建立提高了模型的整体性能。

4 结 语

本文针对新闻要素关系抽取任务，通过对新闻文本中的词、句子作为图节点建模，根据节点间的位置及语义关系精心设计多种边特征，捕获了文本的序列、语法、层次等固有结构信息，并利用图模型的多跳推理能力，对新闻文本上下文信息进行有效的表征，最终篇章级要素关系抽取的性能得到了较好的提升。

参考文献：

[1] ZENG D，LIU K，LAI S，et al.Relation classification via convolutional deep neural network[C]//The 25th International Conference on Computational Linguistics，2014.

[2] WANG L，CAO Z，DE M G，et al.Relation classification via multi-level attention cnns[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics，2016.

[3] HE D，ZHANG H，HAO W，et al.A customized attention-based long short-term memory network for distant supervised relation extraction[J].Neural Computation，2017，27(7)：1964-1985.

[4] MIWA M，BANSAL M.End-to-end relation extraction using lstms on sequences and tree structures[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics，2016.

[5] GAO T, HAN X, XIE R, et al.Neural snowball for few-shot relation learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[6] ZENG S, XU R, CHANG B, et al.Double graph based reasoning for document-level relation extraction[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

[7] CHRISTOPOULOU F, MIWA M, ANANIADOU S.Connecting the dots: document-level neural relation extraction with edge-oriented graphs[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

[8] ZHANG Z, YU B, SHU X, et al.Document-level relation extraction with dual-tier heterogeneous graph[C]//Proceedings of the 28th International Conference on Computational Linguistics, 2020.

[9] YANG S, GUO J.Improved strategies of relation extraction based on graph convolutional model on tree structure for web information processing[J].Journal of Industrial Information Integration, 2022, 25:100301.

[10] Gu J, Sun F, Qian L, et al.Chemical-induced disease relation extraction via convolutional neural network[J]. Database, 2017(1):1-12.

[11] LI H, YANG M, CHEN Q, et al.Chemical-induced disease extraction via recurrent piecewise convolutional neural networks[J].BMC medical informatics and decision making, 2018, 18(2):45-51.

[12] SAHU S K, CHRISTOPOULOU F, MIWA M, et al.Inter-sentence relation extraction with document-level graph convolutional neural network[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

编辑:张玉聪

(上接第 62 页)

[11] HU J, SHEN L, SUN G.Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[12] IOFFE S, SZEGEDY C.Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning, 2015.

[13] SZEGEDY C, VANHOUCKE V, IOFFE S, et al.Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[14] YU F, KOLTUN V.Multi-scale context aggregation by dilated convolutions[EB/OL].[2022-05-11].https://arxiv.org/pdf/1511.07122.pdf.

[15] YANG J, ZHU J, WANG H, et al.Dilated MultiResUNet: dilated multiresidual blocks network based on U-Net for biomedical image segmentation[J].Biomedical Signal Processing and Control, 2021, 68:102643.

[16] WANG J, LV P, WANG H, et al.SAR-U-Net: squeeze-and-excitation block and atrous spatial pyramid pooling based residual U-Net for automatic liver segmentation in computed tomography[J].Computer Methods and Programs in Biomedicine, 2021, 208:106268.

[17] LIU Z, SONG Y Q, SHENG V S, et al.Liver CT sequence segmentation based with improved U-Net and graph cut[J]. Expert Systems with Applications, 2019, 126:54-63.

[18] SEO H, HUANG C, BASSENNE M, et al.Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images[J].IEEE transactions on medical imaging, 2019, 39(5):1316-1325.

[19] LEI T, ZHOU W, ZHANG Y, et al.Lightweight V-Net for liver segmentation[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

编辑:张玉聪