智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

# 用于胎儿超声切面识别的知识蒸馏方法

张欣培， 周尧， 章毅

引用本文:
张欣培, 周尧, 章毅. 用于胎儿超声切面识别的知识蒸馏方法[J]. 智能系统学报, 2022, 17(1): 181–191.
ZHANG Xinpei, ZHOU Yao, ZHANG Yi. Knowledge distillation method for fetal ultrasound section identification[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(1): 181–191.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202105007

**DOI**: 10.11992/tis.202105007
**网络出版地址**: https://kns.cnki.net/kcms/detail/23.1538.TP.20211217.1240.002.html

# 用于胎儿超声切面识别的知识蒸馏方法

张欣培, 周尧, 章毅
(四川大学 计算机学院, 四川 成都 610065)

**摘　要**: 胎儿超声切面识别是产前超声检查的主要任务之一, 直接影响了产前超声检查的质量。近年来, 深度神经网络方法在临床超声辅助诊断方面取得了许多进展。然而, 已有研究大多应用预训练模型微调进行迁移学习, 这不仅容易导致参数冗余和过拟合问题, 而且限制了在实际应用中的实时分析能力。本文提出用于胎儿超声切面识别的知识蒸馏方法。第 1 阶段, 在学生教师网络模型中采用残差网络, 对二者隐藏层特征融入注意力机制, 提取隐藏层关键信息, 进行一次知识迁移, 使学生网络获得先验权重; 第 2 阶段, 使用教师网络模型指导学生网络模型进行知识蒸馏训练, 进一步从整体上提升知识迁移的性能。实验结果表明: 学生网络在提升各项性能的同时, 降低了模型复杂度, 有利于超声设备终端的部署和实时分析能力的提升。

**关键词**: 深度学习; 卷积神经网络; 残差网络; 产前检查; 胎儿超声; 计算机辅诊; 知识蒸馏; 模型压缩
**中图分类号**: TP30　　**文献标志码**: A　　**文章编号**: 1673−4785(2022)01−0181−11

**中文引用格式**: 张欣培, 周尧, 章毅. 用于胎儿超声切面识别的知识蒸馏方法 [J]. 智能系统学报, 2022, 17(1): 181–191.
**英文引用格式**: ZHANG Xinpei, ZHOU Yao, ZHANG Yi. Knowledge distillation method for fetal ultrasound section identification[J]. CAAI transactions on intelligent systems, 2022, 17(1): 181–191.

# Knowledge distillation method for fetal ultrasound section identification

ZHANG Xinpei, ZHOU Yao, ZHANG Yi
(College of Computer Science, Sichuan University, Chengdu 610065, China)

**Abstract**: Fetal ultrasound section recognition is one of the main tasks of prenatal ultrasonography, which directly affects the quality of prenatal ultrasonography. In recent years, by the Deep Neural Network method, we have made great advances in clinical ultrasound-assisted diagnosis. However, most of the existing studies have applied fine-tuned pre-trained model for migration learning, which not only easily leads to parameter redundancy and overfitting problems, but also limits the real-time analysis capability in practical applications. Therefore, this paper proposes a knowledge distillation method for fetal ultrasound section recognition. In the first stage, a residual network is used in the student and teacher network model to incorporate attention mechanisms for both hidden layer features, extract key information in the hidden layer, and perform one knowledge migration so that the student network can obtain a priori weight. In the second stage, the teacher network model is used to guide the student network model to perform knowledge distillation training, so as to further improve the performance of knowledge migration in an overall manner. The experimental results show that the student network reduces the model complexity while improving various performances, which is beneficial to the deployment of ultrasound device terminals and real-time analysis capability.

**Keywords**: deep learning; convolutional neural network; residual network; prenatal diagnosis; fetal ultrasound; computer aided diagnosis; knowledge distillation; model compression

产前超声检查是监测胎儿在母体内生长情况的重要步骤, 在传统产前超声检查的过程中, 临床医生利用超声设备获得胎儿各个部位的二维超声标准切面, 并在此基础上测量各种体征数据, 以评估胎儿在母体内的发育情况, 预测早产的风险。但产前超声检查用到的切面种类多、不同切面的主要结构和复杂程度都不一样, 使用传统方式手动获取切面会面临很多问题, 如: 1) 标准切面的获取难度大, 对超声医生的临床经验依赖度极高; 2) 因不同超声医生专业水平的差异, 获取到的标准切面结果可能不同, 切面图像的规范性

**收稿日期**: 2021−05−11.　**网络出版日期**: 2021−12−17.
**基金项目**: 国家自然科学基金项目(62006163).
**通信作者**: 章毅. E-mail: zhangyi@scu.edu.cn.

得不到保障；3) 临床工作效率低，易使孕妇受检时间过长，引起不良反应。近年来，随着深度神经网络在医学图像分析领域的发展与应用，为解决传统方法的弊端，研究人员逐渐将深度神经网络应用到胎儿超声切面图像的自动识别任务中，辅助医生进行诊断。

Maraci 等[1]采用动态纹理分析和支持向量机(SVM)算法[2]对产妇孕中期的超声检查视频的每帧图像进行标准切面识别。SVM 算法是一种利用高维映射来解决机器学习中线性不可分问题的算法，但在数据量过大时，其鲁棒性和准确率无法得到保证，所以 SVM 算法的性能是有限的。随着大数据和深度神经网络的发展，各种深度神经网络方法被应用在胎儿超声切面图像识别任务中。Baumgartner 等[3]首次提出了基于图像级别标签的弱监督方法，使用卷积神经网络对胎儿标准切面图像进行实时自动检测，其 $F_1$ 评价指标达到了 0.791 8，且在回溯帧检索中的准确率达到了 90.09%。Maraci 等[4]使用条件随机场模型从超声检查视频的每一帧图像对胎儿心脏切面进行检测。条件随机场模型[5]是一种判别式模型，在观测序列的基础上对目标序列进行建模，可以通过超声视频的每一帧及其前后帧所提供的序列化信息来检测胎儿心脏切面，但此方法在训练时的收敛速度极慢。Ryou 等[6]提出了一种基于随机森林的矢状面胎儿全域定位方法，利用卷积神经网络对胎儿头部、身体和非胎儿切面进行识别。Cheng 等[7]用基于卷积神经网络的迁移学习模型对胎儿腹部二维超声切面进行识别，分别使用两个卷积神经网络 CaffeNet[8]、VGGNet[9]进行对比实验，基于 CaffeNet 的迁移学习模型达到了平均 77.3% 的准确率，基于 VGGNet的迁移学习模型达到了 77.9% 的准确率。

近年来，越来越多的研究人员将深度神经网络应用于临床辅诊任务中。随着计算机硬件设备的不断发展，在图形处理器(GPU)上训练各种深度神经网络已不是一件难事。但庞大神经网络模型在训练过程中的计算资源占用量是巨大的，不可避免地耗费大量时间开销对输入数据进行处理，极大限制了实际应用时的运行效率。同时，在目前的研究和应用中，多使用预训练模型针对不同任务进行微调，该方式极易造成参数冗余的问题，增加不必要的时间开销，难以提高实时分析能力；且在实际部署时，深度神经网络模型占用大量内存，对终端设备的计算资源需求高。

针对以上问题，本文提出改进的两阶段知识蒸馏方法，在保留分类性能的同时提升模型的实时分析能力。首先，根据胎儿超声切面图像的特征，调研和使用几种主流分类模型进行实验，综合考量其计算资源占用量和分类性能，选择 Resnet8 和 Resnet101 分别作为学生网络和教师网络。再者，通过第 1 阶段，使用预训练好的教师网络的隐藏层信息初始化学生网络的中间层，将 Resnet101 模型的隐藏层输出作为 Resnet8 模型中间层训练的标签信息，使学生网络的中间层获得初始化的先验权重；最后，通过第 2 阶段进行知识蒸馏，将教师网络的负样本标签蕴含的软标签信息“蒸馏”，作为此阶段训练的监督信息。通过以上方法得到的学生网络模型，在分类性能的各项指标上超过教师网络模型，且其计算资源占用量大幅降低，模型被有效压缩，加快了实际应用时的分析速度。

## 1　胎儿超声切面分类与网络压缩

### 1.1　胎儿超声切面分类

针对医学超声图像分类任务的特点，分别选取 MobileNetV2、MobileNetV3small、Resnet8、VGG16、Resnet34 和 Resnet101 模型。前 3 个模型属于轻量级模型，适合从中选择合适的学生网络；后 3 个模型参数数量较多，适合作为教师网络的选项，其参数量对比如表 1 所示。由表 1 可得，前 3 个模型与后 3 个模型相比，参数数量更少，具有更轻量级的特征。基于此，设计对比实验从前 3 个模型中选择合适的学生网络，从后 3 个模型中选择合适的教师网络。MobileNetV1 是 Andrew 等[10]提出的一种神经网络结构，利用深度可分离卷积减少了参数数量，从而降低计算量，提高计算效率。这种神经网络模型适合部署到移动端或嵌入式系统中，但其不足之处在于该网络是一种较简单的单通道结构，在任务中的准确率等性能表现往往不能达到预期目标。随着 ResNet 和 DenseNet 等网络的提出，研究人员验证了卷积层输出的复用对提升网络性能的有效性，MobileNetV2[11]应运而生，引入具有线性瓶颈的逆残差结构模块，一定程度改善了原有 MobileNetV1 模型的不足。MobileNetV3[12]是该系列的最新版本，包含 MobileNetV3Small 和 MobileNetV3Large 两种模型，结合自动机器学习技术以及人工微调构建了更轻量级的模型。本文所述的 Resnet8 模型是对 Resnet18 模型进行改造而形成的层数和参数数量更少的轻量级模型，由 7 层卷积层与 1 层全连接层构成。由表 1 可知，在评价计算资源占用量指标的参数数量上，Resnet8 模型比其他两个轻量级模型具有一定优势。

**表 1　不同分类模型参数量对比**
**Table 1　Parameter comparison of different models**

| 模型名称 | 网络深度 | 参数数量/$10^6$ | 模型文件大小/MB |
|---|---|---|---|
| MobileNetV2 | 54 | 2.25 | 9.27 |
| MobileNetV3small | 49 | 1.24 | 5.12 |
| Resnet8 | 8 | 1.00 | 4.02 |
| VGG16 | 16 | 134.39 | 537 |
| Resnet34 | 34 | 21.33 | 85.5 |
| Resnet101 | 101 | 43.09 | 173 |

## 1.2　知识蒸馏方法

近年来, 随着大数据和深度神经网络的不断发展, 在强大的 GPU 上训练各种复杂的神经网络模型已不是一件难事。但在实际部署时, 因用户端终端设备的运算能力有限, 使得复杂模型的部署变得困难; 在实际应用方面, 深度神经网络需要巨大的时间开销对输入数据进行处理, 极大限制了实时分析能力。基于此, 研究人员逐渐将目光放在模型压缩领域的各项研究上来。知识蒸馏作为模型压缩的一大分支, 也不断取得各项进展。Bucilua 等[13] 首次提出模型压缩的概念, 这种方法能将有效信息从深度神经网络模型转移到训练浅层模型, 而不会显著降低原有模型的精度。Romero 等[14] 提出 FitNet, 不仅利用教师网络最后一层神经元的输出信息, 还利用了其中间层信息, 成功训练了较原有教师网络更深但更窄的学生网络。Hinton 等[15] 正式将这种学习模式定义为“知识蒸馏”, 并提出了带温度系数 $T$ 的 Softmax 函数。通过此函数将教师网络的负样本信息输出的概率分布“蒸馏”出来, 以对学生网络的训练提供额外的监督信息。他们在 MNIST 数据集上进行初步试验, 证明了带温度系数 $T$ 的 Softmax 函数对深度神经网络模型精度提高的有效性; 并分别在语音数据集和大型数据集 JFT 上进行对比实验, 证明了知识蒸馏对模型精度提高和模型压缩的有效性。受课程学习 (curriculum learning) 的启发, Jin 等[16] 发现由学生网络和教师网络间的结构差异而造成蒸馏失败的问题, 并针对此提出了路由约束提示学习方法。2019 年 Phuong 等 [17] 从理论上论述了知识蒸馏中学生网络具有快速收敛的泛化边界的原因, 解释了知识蒸馏的工作原理。2020 年 Ji 等[18] 分别从风险界、数据效率和不完美的老师 3 个角度进一步对广义神经网络上的知识蒸馏方法进行了理论解释。目前的知识蒸馏方法已扩展到师生学习[14]、相互学习[19]、辅助教学[20]、终身学习[21] 和自主学习[22] 等模式。通过知识蒸馏训练后的学生网络, 能保留甚至超过教师网络的性能, 网络结构比教师网络更简单, 减少了冗余参数, 能有效提高实时分析性能, 缓解终端部署和实际应用的困难。

虽然现有知识蒸馏方法已经取得了良好的效果, 但也具有一定局限性。神经网络的隐藏层特征表达往往蕴含了丰富的有用信息, 现有方法仅依托于神经网络最后一层神经元输出信息, 提供的监督信息是有限的。考虑到隐藏层特征表达和映射对深度神经网络模型的影响, 在传统知识蒸馏方法中融入隐藏层的特征表达, 将在一定程度上为学生网络提供更丰富的监督信息。首先, 在第 1 阶段, 通过对学生网络预先进行训练, 使其学到教师网络隐藏层丰富的特征表达, 获得优于原始学生网络中间层的权重信息; 第 2 阶段, 对已学习到教师网络隐藏层特征表达的学生网络进行知识蒸馏。同时, 考虑到在教师网络训练时会产生很多中间模型 (anchor points[16]), 应使用结构相似的神经网络模型作为学生网络, 以便于学生网络从其中间模型更好地进行特征学习, 从而提升知识蒸馏的效率。基于此, 本文使用师生学习模式, 提出改进的两阶段知识蒸馏方法。

### 1.2.1　知识蒸馏方法

众所周知, 基于深度神经网络的分类任务都具有共同的特征: 神经网络最后一层神经元的输出信息都会通过一个 Softmax 函数, 如式 (1) 所示, 将输出信息变成概率分布, 才能与标签信息求其极大似然值, 此种经过 Softmax 层直接输出的信息被称为硬标签信息。

$$q_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)} \tag{1}$$

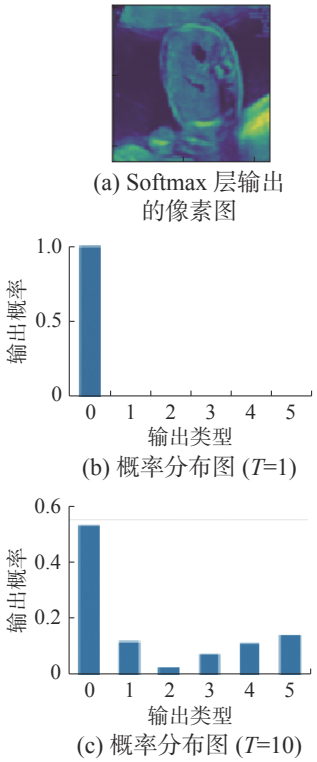
式中: $q_i$是教师网络输出每一类的概率分布; $z_i$是最后一层的神经元的输出信息。

但由于 Softmax 函数只输出概率分布的独热编码, 会均一化所有负样本标签的信息, 将负样本标签的概率都还原为 0, 弱化了负样本标签的概率信息对模型训练的影响。对此, Hinton 等[15] 提出带温度系数 $T$ 的 Softmax 函数, 如式 (2) 所示, 此种经过温度系数 $T$ 的输出信息被称为软标签信息。最后一层神经元的输出信息通过带温度系数 $T$ 的 Softmax 函数后, 能“蒸馏”出负样本标签的概率信息, 为学生网络的训练提供更为丰富的“暗知识”, 使学生网络不只接受正样本标签的监督训练。

$$q_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)} \tag{2}$$

式中: $q_i$是教师网络输出每一类的软标签; $z_i$是最后一层的神经元的输出信息; $T$ 为温度系数。根

据不同的温度蒸馏出的知识占比不同，需进行对比实验选出最适合的温度系数 $T$。实验数据图片在不同温度系数 $T$ 下的预测概率如图 1 所示。




(a) Softmax 层输出的像素图

(b) 概率分布图 ($T$=1)

(c) 概率分布图 ($T$=10)

**图 1　不同温度系数 $T$ 下的概率分布**

**Fig. 1　Probability distribution with different parameter $T$**

知识蒸馏方法基于软标签信息，在给定教师网络的条件下，使用教师网络最后一层神经元的输出信息经过带温度系数 $T$ 的 Softmax 函数，将其预测的所有类别的概率分布“蒸馏”，作为知识蒸馏的监督信息，指导学生网络进行训练。此方法为学生网络的训练提供了来自教师网络的先验知识，本质上是在学生网络的训练中加入一种新的正则化机制。具体流程图如图 2 所示。

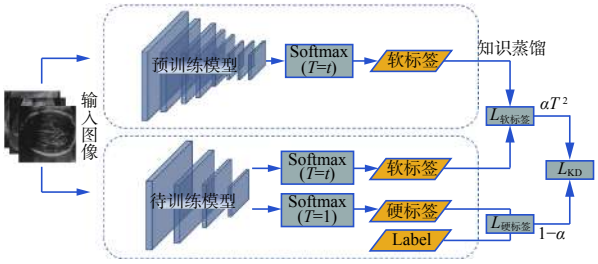



**图 2　知识蒸馏网络结构**

**Fig. 2　Network structure of knowledge distillation**

最终损失函数表达式为

$$L_{\mathrm{KD}}(\boldsymbol{W}_s)=\alpha T^2\varphi\left(\boldsymbol{W}_s^{\mathrm{T}},\boldsymbol{W}_t^{\mathrm{T}}\right)+(1-\alpha)\psi(\boldsymbol{W}_s,\boldsymbol{Y}_{\mathrm{label}}) \tag{3}$$

式中：$\alpha$为蒸馏强度；$T$ 为温度系数；$\varphi$为 KL 散度；$\boldsymbol{W}_s^{\mathrm{T}}$为学生网络经过带温度系数 $T$ 的 Softmax 层的权重矩阵；$\boldsymbol{W}_t^{\mathrm{T}}$为教师网络的权重矩阵；$\psi$为交叉熵；$\boldsymbol{W}_s$为学生网络的硬标签信息；$\boldsymbol{Y}_{\mathrm{label}}$为输入图像的标签信息。

### 1.2.2　改进的两阶段知识蒸馏方法

在第 1 阶段，旨在提取教师网络隐藏层的特征表达，将其作为此阶段训练的监督信息，以此来指导学生网络的中间层权重的初始化，使其学习到教师网络的隐藏层特征表达。第 1 阶段流程如图 3 所示。

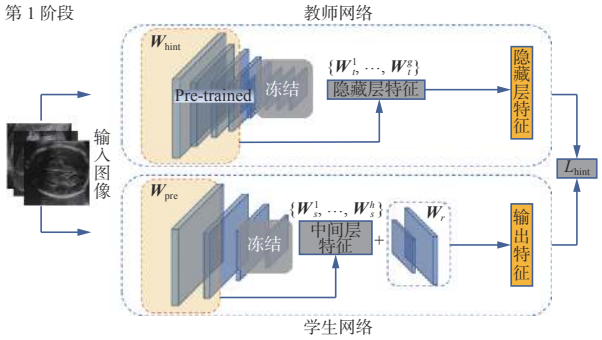



**图 3　第 1 阶段网络结构**

**Fig. 3　Network structure of the first stage**

为获得基于教师网络隐藏层特征表达的学生网络，$\boldsymbol{W}_s$需冻结学生网络的最后一层残差连接层、池化层以及全连接层，仅训练学生网络的第一层至中间层 $h$ 的权重矩阵。训练集$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_i,y_i)\}$，其中，$x_i\in x\subset\mathbf{R}^{s\times s\times c}$即通道数为 $c$ 的输入大小为$s\times s$的图像数据，$y_i\in(0,5)$即输入的图像数据的标签信息，在本文中代表属于编号为 0~5 的 6 类标签；$\boldsymbol{W}_{\mathrm{pre}}=\{\boldsymbol{W}_s^1,\boldsymbol{W}_s^2,\cdots,\boldsymbol{W}_s^h\}$即教师网络隐藏层前 $g$ 层的特征表达；$\boldsymbol{W}_{\mathrm{pre}}=\{\boldsymbol{W}_s^1,\boldsymbol{W}_s^2,\cdots,\boldsymbol{W}_s^h\}$即学生网络中间前 $h$ 层的特征表达。为解决教师网络前 $g$ 层输出特征与学生网络前 $h$ 层输出特征表达的维度不匹配问题，加入随机初始化权重的卷积回归层。最终通过最小化损失函数来优化和卷积回归层，其表达式为

$$L_{\mathrm{hint}}(\boldsymbol{W}_{\mathrm{pre}},\boldsymbol{W}_r)=\frac{1}{2}\left\|\mu(x;\boldsymbol{W}_{\mathrm{hint}})-r(\upsilon(x;\boldsymbol{W}_{\mathrm{pre}});\boldsymbol{W}_r)\right\|^2 \tag{4}$$

式中：$\mu$为教师网络隐藏层函数；$\upsilon$为学生网络中间层函数；$r$ 为卷积回归函数。为保证教师网络隐藏层输出特征和学生网络卷积回归层输出特征维度一致，$\mu$函数和 $r$ 函数应具有相同的非线性性质。

经过第 1 阶段训练的学生网络模型已经具有基于教师网络隐藏层特征表达的中间层权重信息，类比教师教授学生知识的环节，相当于学生已经在教师布置的预习任务中获得了一定量的知识储备，为接下来的教师教学打下基础，即为第 2 阶段的知识蒸馏训练做铺垫。在第 2 阶段，使用知识蒸馏方法再次对学生网络进行训练，通过最小化损失函数去优化学生网络模型，不断迭代，直至其损失函数值收敛。

#### 1.2.3 学生网络和教师网络结构

综合各种分类模型在胎儿超声切面数据集上的性能, 考虑到分类性能与计算资源占用量之间的平衡, 将 Resnet8 作为学生网络模型, 其层数浅、参数量少, 具体参数如表 2 所示; 将 ResNet101 作为教师网络模型, 其层数深、参数量大, 具体参数如表 3 所示。二者都具有良好的分类性能, 且二者具有相同的残差结构, 可以方便学生网络学习教师网络的特征表达。

**表 2　学生模型网络参数**

**Table 2　Network parameters of student module**

| 层名 | 卷积核尺寸及个数 | 步长 | 填充 |
|---|---|---|---|
| Conv1 | 3×3,16 | 1 | 1 |
| Layer1 | 3×3,16<br>3×3,16 | 1<br>1 | 1 |
| Layer2 | 3×3,32<br>3×3,32 | 2<br>1 | 1 |
| Layer3 | 3×3,64<br>3×3,64 | 2<br>1 | 1 |
| Avgpool | 8×8 | 1 | 0 |
| Fc | — | — | — |
| Softmax | — | — | — |

**表 3　教师模型网络参数**

**Table 3　Network parameters of teacher module**

| 层名 | 卷积核尺寸及个数 | 步长 | 填充 |
|---|---|---|---|
| Conv1 | 3×3 | 1 | 1 |
| Layer1 | $3\times\begin{cases}1\times1,64\\3\times3,64\\1\times1,256\end{cases}$ | 1 | 1 |
| Layer2 | $4\times\begin{cases}1\times1,128\\3\times3,128\\1\times1,512\end{cases}$ | 1 | 1 |
| Layer3 | $23\times\begin{cases}1\times1,256\\3\times3,256\\1\times1,1024\end{cases}$ | 1 | 1 |
| Layer4 | $3\times\begin{cases}1\times1,512\\3\times3,512\\1\times1,248\end{cases}$ | 1 | 1 |
| Avgpool | 8×8 | 1 | 0 |
| Fc | — | — | — |
| Softmax | — | — | — |

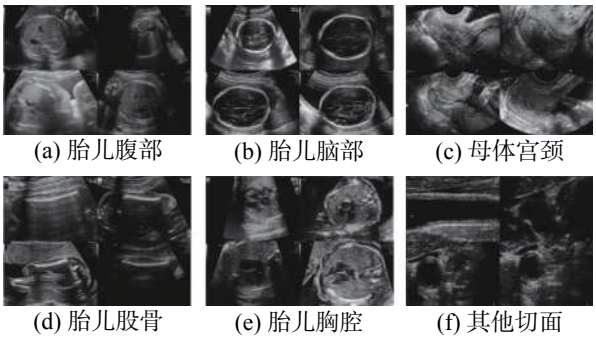
## 2　实验设计

### 2.1　实验平台

本实验在 GPU 深度神经网络集成计算平台上进行, 操作平台为 Ubuntu, 使用的 GPU 为 Nvidia GeForce RTX 3090Ti, 显存为 24 GB, 使用的深度神经网络框架为 PyTorch。

### 2.2　实验准备

实验采用的数据是胎儿超声切面数据集, 由 BCNatal 收集, 涵盖了来自两个医学中心共计 12 400 张胎儿超声切面图像, 图像格式为 PNG, 均做了匿名处理。此数据集包含了 6 类切面类型, 各类型切面图像概览如图 4 所示。胎儿超声切面图像作为产前检查的重要依据, 均由专业的超声科阅片医师进行手动标注, 每类切面的临床意义其数据分布情况如表 4 所示, 其中“其他类型”的存在可以提高模型对于不同类别在有干扰情况下的准确率。



(a) 胎儿腹部　(b) 胎儿脑部　(c) 母体宫颈

(d) 胎儿股骨　(e) 胎儿胸腔　(f) 其他切面

**图 4　胎儿超声切面概览**

**Fig. 4　Examples of fetal ultrasound section images**

**表 4　数据集分布情况**

**Table 4　Component distribution of datasets**

| 切面类型 | 临床用途 | 图像数量 |
|---|---|---|
| 胎儿腹部 | 胎儿形态学特征 | 711 |
| 胎儿头部 | 检测胎儿神经发育情况 | 3 092 |
| 胎儿股骨 | 估算胎儿体重 | 1 040 |
| 胎儿胸腔 | 检测胎儿心脏和肺部发育情况 | 1 718 |
| 母体宫颈 | 预测早产风险 | 1 626 |
| 其他类型 | | 4 213 |
| 总计 | | 12 400 |

本实验将此数据集划分为训练集和测试集, 其比例约为 4∶1, 具体分布情况如表 5 所示。为满足不同分类模型对输入图像大小的限制, 预先将图像进行了拉伸缩放的预处理方式, 将其调整为像素尺寸。同时, 为了提高彩超图像在基于 ImageNet 预训练模型上的泛化能力, 对原始超声图像进行归一化等预处理。

**表 5　实验数据集分布情况**

**Table 5　Experimental component distribution of datasets**

| 数据类型 | 胎儿腹部 | 胎儿头部 | 胎儿股骨 | 胎儿胸腔 | 胎儿宫颈 | 其他类型 | 总计 |
|---|---|---|---|---|---|---|---|
| 训练集 | 569 | 2 474 | 832 | 1 374 | 1 301 | 3 370 | 9 920 |
| 测试集 | 142 | 618 | 208 | 344 | 325 | 843 | 2 480 |
| 总计 | 711 | 3 092 | 1 040 | 1 718 | 1 626 | 4 213 | 12 400 |

### 2.3　实验步骤

#### 2.3.1 胎儿超声切面分类实验

针对本文所述的胎儿超声切面分类任务的特

点，使用不同深度神经网络分类模型进行实验，并评估各种模型在胎儿超声切面数据集上的准确率及其损失函数值。在 MobileNetV2、MobileNetV3Small、Resnet8、VGG16、Resnet34、Resnet101 模型上进行分类实验。此阶段的学习率为 $1\times10^{-6}$，并设置 Warmup 机制，首先使用较大的学习率进行训练，然后逐渐逼近实验设置的学习率；本实验中的损失函数使用交叉熵函数，优化方法采用自适应梯度下降法 (adam) 算法，此方法较随机梯度下降（SGD）算法能取得更优的效果。

2.3.2　改进的两阶段知识蒸馏实验

本方法对现有知识蒸馏方法进行改进，先进行第 1 阶段训练，将教师网络隐藏层的输出信息作为监督信息，将其迁移到学生网络的中间层，使学生网络的中间层获得教师网络的隐藏层特征表达作为监督信息训练的初始权重。在第 2 阶段，使用知识蒸馏方法对既得学生网络模型进行二次训练，整体训练流程为

1) 将实验所用数据集进行预处理和数据集划分，分别用于训练和测试；

2) 将训练集输入 Resnet101 模型，训练教师网络，使用测试集测试其分类性能，并保存性能最好的 Resnet101 模型作为教师网络；

3) 固定教师网络模型参数，将其隐藏层的输出信息作为学生网络中间层知识迁移的监督信息；

4) 冻结学生网络的最后 3 层参数，即全连接层、最后池化层、和最后一层残差网络层。为解决教师网络中间层输出特征和学生网络中间层输出特征维度不一致的问题，需在学生网络中间层的最后添加一个卷积回归层。

5) 在第 1 阶段，将训练集输入学生网络，使用步骤 3) 获得的教师网络隐藏层特征表达作为监督信息，训练学生网络中间层 $W_{\text{pre}}$ 和 $W_r$。使用 $L_{\text{hint}}$ 作为损失函数，通过反向传播算法不断迭代优化式 (4)，最小化其损失函数值，直到收敛。保存此阶段训练的学生网络模型。

6) 用知识蒸馏方法对步骤 5) 获得的学生网络模型进行二次训练。将学生网络直接训练的输出作为硬标签信息，结合教师网络最后一层神经元的输出经过带温度系数 $T$ 的 Softmax 层后的软标签信息，将二者加权求和作为监督信息，最小化 $L_{\text{KD}}$ 来优化学生网络的权重参数。通过反向传播算法迭代式 (3)，最小化损失值，直到收敛。同时计算各种性能指标，保存性能最佳的学生网络模型。

7) 用训练好的学生网络模型进行预测，测试其各项性能指标。

### 2.4　评价指标

针对本任务，使用多个评价指标，即准确率 (Acc)、宏精确率 (MacroPre)、宏召回率 (MacroRecall)、宏 $F_1$-score 值 (MacroF1) 和前向传播时的计算力 (FLOPs)。Acc 即预测正确的样本类别占总样本的比例，体现了模型的预测能力。精确率在二分类中即正确预测为该类别的占全部预测为该类别的比例，在多分类中，对每个标签分别计算其精确率，再对其取算数平均 (Macro)，得到 MacroPre；召回率在二分类中即正确预测为该类别的样本数占全部实际为该样本的比例，在多分类中，对每个标签分别计算其召回率，再对其取算数平均，得到 MacroRecall；$F_1$ 值在二分类中，即对精确率和召回率的评估，在多分类中，对于每个标签，分别计算其 $F_1$ 值，然后对其取算数平均，得到 MacroF1。以上参数数值越大，分类模型的性能越好。FLOPs(fLoating point operations)，即浮点运算数，衡量模型复杂度，体现了模型的运算能力。

分别计算每一类的 $\text{Pre}_i$、$\text{Recall}_i$、$F_1$ 的公式为

$$\text{Pre}_i=\frac{\text{TP}_i}{\text{TP}_i+\text{FP}_i}$$

$$\text{Recall}_i=\frac{\text{TP}_i}{\text{TP}_i+\text{FN}_i}$$

$$F_{1_i}=\frac{2\times\text{Pre}_i\times\text{Recall}_i}{\text{Pre}_i+\text{Recall}_i}$$

对既得的每一类的 Pre 和 Recall 以及 $F_1$，再使用 Macro 算法。先分别求出每个类别对应的值，再对其求算数平均值：

$$\text{宏精确率：MacroPre}=\frac{1}{N}\sum_{i=1}^{N}\text{Pre}_i$$

$$\text{宏召回率：MacroRecall}=\frac{1}{N}\sum_{i=1}^{N}\text{Recall}_i$$

$$\text{宏F}_1\text{值：MacroF1}=\frac{1}{N}\sum_{i=1}^{N}F_{1_i}$$

Acc 可计算为

$$\text{Acc}=\frac{\text{TP}_i+\text{TN}_i}{\text{TP}_i+\text{TN}_i+\text{FP}_i+\text{FN}_i}$$

式中：$\text{TP}_i$、$\text{TN}_i$、$\text{FP}_i$、$\text{FN}_i$ 分别代表第 $i$ 类别的正阳性、正阴性、假阳性和假阴性。

卷积核 FLOPs 的计算为

$$\text{FLOP}_s=2HW\left(C_{in}K^2+1\right)C_{\text{out}}$$

式中：$H$、$W$ 和 $C_{in}$ 分别是输入特征图的高度、宽度和通道数；$K$ 是卷积核宽度 (假定卷积核长宽相等)，$C_{\text{out}}$ 是输出通道数。全连接层 FLOPs 的计算为

$$\text{FLOP}_s=(2I-1)O$$

式中：$I$ 是输入维数；$O$ 是输出维数。

在本文中，为了凸显本方法的有效性，还关注各个模型的网络深度、显存占用量、GPU 占用率、损失值、模型文件大小等性能指标。

## 3　实验结果与分析

### 3.1　胎儿超声切面分类实验

使用不同的分类模型, 在相同的训练集和验证集上进行对比训练, 从其计算资源占用情况、准确率和损失值来衡量其分类性能。其中损失值为交叉熵函数的输出, 体现了分类模型的预测值和真实值之间的概率分布情况, 本实验中选用损失值低于 0.1 的指标作为分类器取得了好的效果的基准, 具体情况如表 6 所示。

**表 6　各神经网络模型性能对比**

**Table 6　Experimental results with different neural network methods**

| 模型名称 | Acc/% | 显存占用量/MB | GPU占用率/% | MFLOPs | 损失值 | 模型文件大小/MB |
|---|---|---|---|---|---|---|
| MobileNetV2 | 93.27 | 1 384 | 21 | 2 420 | 0.065 1 | 9.27 |
| MobileNetV3small | 89.35 | 116 | 8 | 64 | 0.093 9 | 5.12 |
| Resnet8 | 92.22 | 172 | 9 | 624 | 0.093 2 | 4.02 |
| VGG16 | 94.52 | 2 630 | 39 | 15 530 | 0.089 8 | 537 |
| Resnet34 | 93.15 | 1 970 | 27 | 56 940 | 0.075 7 | 85.5 |
| Resnet101 | 97.50 | 12 656 | 51 | 123 570 | 0.078 6 | 173 |

由表 6 可知, 在准确率性能的表现上, Resnet101 模型较 VGG16 模型提升了 2.98% 个百分点, 较 Resnet34 模型提升了 4.35%, 较 Resnet8 模型提升了 5.28%, 取得了最优的准确率性能表现。在计算资源占用量方面, Resnet101 模型的网络深度是 Resnet8 模型的近 12 倍, 相比其他两个较大的模型也增加了近 3~5 倍, 训练时的显存占用量和 GPU 占用率和 FLOPs 也是最高的。综上所述, 充分表明 Resnet101 模型具有最好的分类性能的同时, 其计算资源占用量也最庞大, 适合作为教师网络进行后续实验, 以验证知识蒸馏方法能否在保留其分类性能的情况下将模型压缩, 并达到提升实时性分析能力的目的。

在学生网络的选择方面, 应考虑模型本身的参数数量和计算资源占用情况, 尽量减少冗余参数; 同时, 学生网络本身的分类准确率也是重要指标之一, 不能为了压缩模型的大小, 使得分类性能得不到保证。由表 6 可得, 轻量级模型 MobileNetV2、MobileNetV3Small 和 Resnet8 模型都具有较好的基本分类性能, 但 MobileNetSmall 在准确率上的表现却不如其他两个模型。对比 MobileNetV2 和 Resnet8 的各项性能指标, 虽然前者在准确率性能指标上超过后者 1.05%, 但其显存占用量是后者的近 8 倍, 在 GPU 占用率和 FLOPs 等性能上也处于劣势。Resnet8 模型有较好的分类性能, 其在训练时的计算资源占用量是更轻量级的, 最终得到的模型文件大小较前两者也是最小的。基于此, 本文综合考虑准确率和计算资源占用量, 同时假设与教师网络模型具有相同残差结构的 Resnet8 模型, 能更好地学习到以 Resnet101 网络特征表达作为监督信息的“知识”, 选用 Resnet8 模型作为学生网络, 如表 7 所示。

综上所述, Resnet101 模型在胎儿超声切面分类任务中具有最优异的分类性能, 较 Resnet8 模型具有 5.28% 的准确率指标提升。综合各种分类模型在胎儿超声切面数据集上的性能, 考虑到分类性能与计算资源占用量之间的平衡, 将 Resnet8 作为学生网络, 将 ResNet101 作为教师网络, 此二者都具有良好的分类性能, 且具有相同的残差结构, 可以方便学生网络学习教师网络的隐藏层特征表达, 提高泛化能力。学生网络模型和教师网络模型在训练时的资源占用情况对比如表 8。

**表 7　学生网络和教师网络计算资源占用**

**Table 7　Occupation of computational resource of student and teacher models**

| 模型名称 | 参数量/$10^6$ | 显存占用量/MB | GPU占用率/% | MFLOPs | 模型文件大小/MB |
|---|---|---|---|---|---|
| Resnet8 | 1.00 | 172 | 9 | 624 | 4.02 |
| Resnet101 | 43.09 | 12 656 | 51 | 123 570 | 173 |

**表 8　不同温度系数 $T$ 的性能对比**

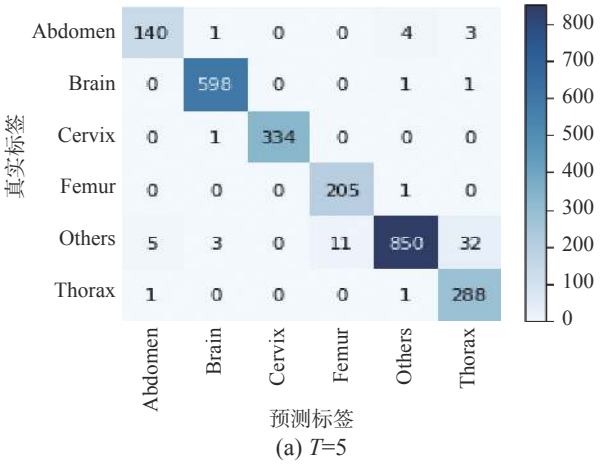
**Table 8　Experimental results with different parameter $T$**

| 温度系数$T$ | 准确率/% | 宏精确率/% | 宏召回率/% | 宏$F_1$值 |
|---|---|---|---|---|
| 5 | 97.38 | 97.85 | 96.34 | 0.970 3 |
| 10 | 96.05 | 94.83 | 95.98 | 0.953 8 |
| 20 | 95.60 | 94.39 | 95.20 | 0.947 8 |
| 30 | 95.56 | 94.42 | 95.06 | 0.947 2 |
| 40 | 95.48 | 94.46 | 94.79 | 0.946 1 |

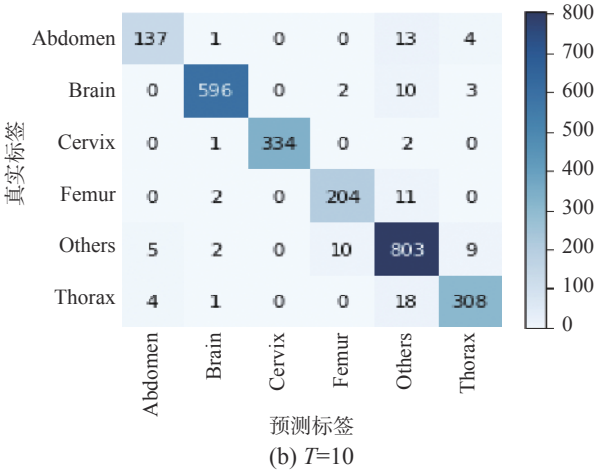
由表 8 可知, Resnet8 模型较 Resnet101 模型的训练参数量减少了近 4 210 万, 在训练时的显存占用和 GPU 使用率上也更具优势, 模型文件大小也缩小近 43 倍, 占用的计算资源不再冗余, FLOPs 缩小了近 198 倍, 提升了实际部署的可行性。

### 3.2　温度参数 $T$ 对比实验

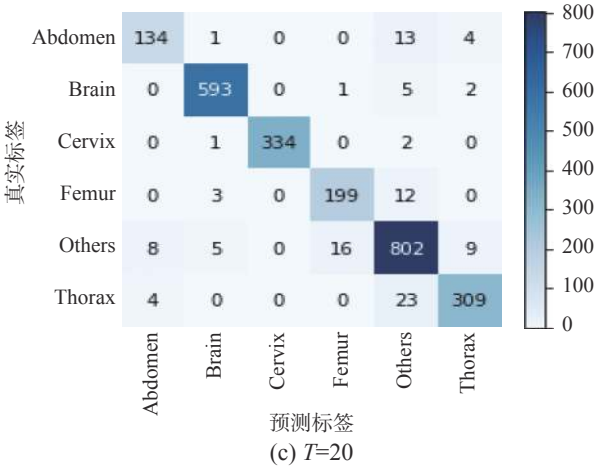
使用不同的温度系数 $T$ 进行对比实验, 选择 5、10、20、30、40 作为实验的温度参数 $T$, 其分类可视化混淆矩阵如图 5 所示。



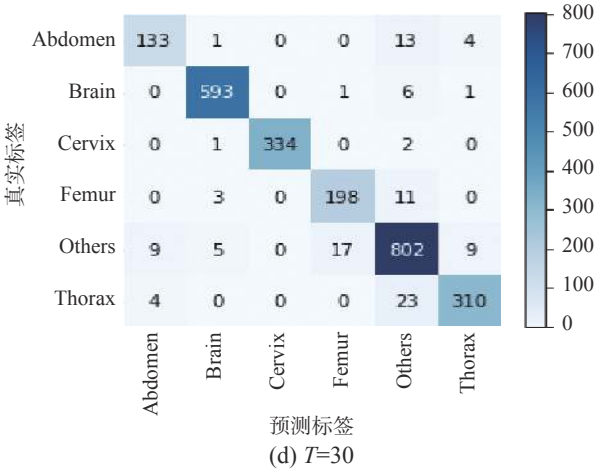

(a) $T$=5



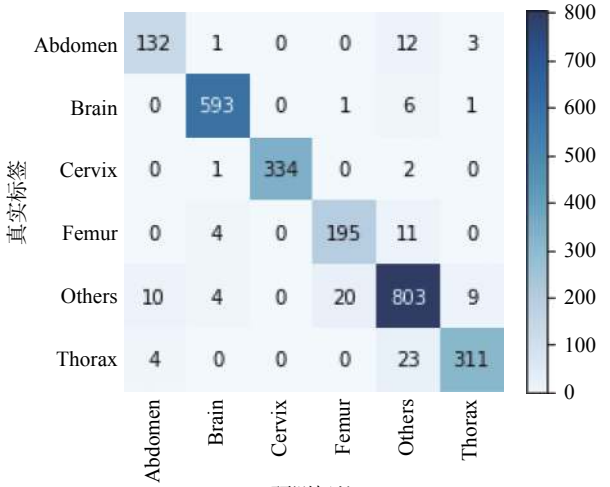

(b) $T$=10




(c) $T$=20




(d) $T$=30




(e) $T$=40

**图 5　不同温度系数 $T$ 的知识蒸馏分类混淆矩阵**

**Fig. 5　Confusion matrix of knowledge distillation classification with different parameter $T$**

比较学生网络在不同温度系数 $T$ 的训练结果，选择最合适的 $T$ 作为整个实验中的温度系数 $T$。由表 9 可得，当温度系数 $T$=5 时，学生网络在准确率、宏准确率、宏召回率、宏 $F_1$ 值等性能指标相比其他温度系数 $T$ 得到的模型是最具优势的。同时，在温度系数 $T$=5 的情况下，通过现有知识蒸馏方法训练的学生网络模型与学生网络单独训练时得到的模型的性能相比，各项性能都得到了提升，较原有学生网络的准确率提升 5.16%，并不断逼近教师网络模型的准确率，在宏精确率和宏 $F_1$ 值上都超过了教师网络模型，涨幅分别为 1.19% 和 0.07%，且其计算资源占用量远小于原始教师网络。基于此，选择 $T$=5 作为实验中的温度参数 $T$。

**表 9　不同优化方法的性能对比**

**Table 9　Performance comparison of different models**

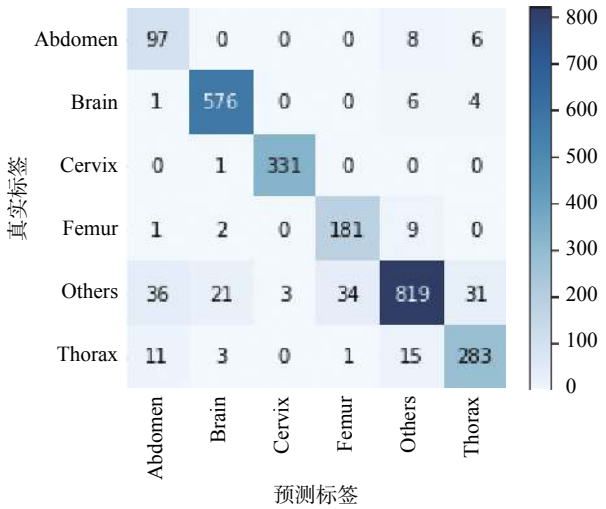
| 模型名称 | 准确率/% | 宏精确率/% | 宏召回率/% | 宏$F_1$值 |
|---|---|---|---|---|
| Resnet101 | 97.50 | 96.66 | 97.29 | 0.969 6 |
| Resnet8 | 92.22 | 92.69 | 87.95 | 0.900 0 |
| Resnet8 + stage1 | 93.75 | 92.62 | 91.87 | 0.922 0 |
| Resnet8+KD | 97.38 | 97.85 | 96.34 | 0.970 3 |
| Resnet8 + Hint | 98.59 | 98.25 | 98.17 | 0.982 1 |

### 3.3　改进的两阶段知识蒸馏方法实验

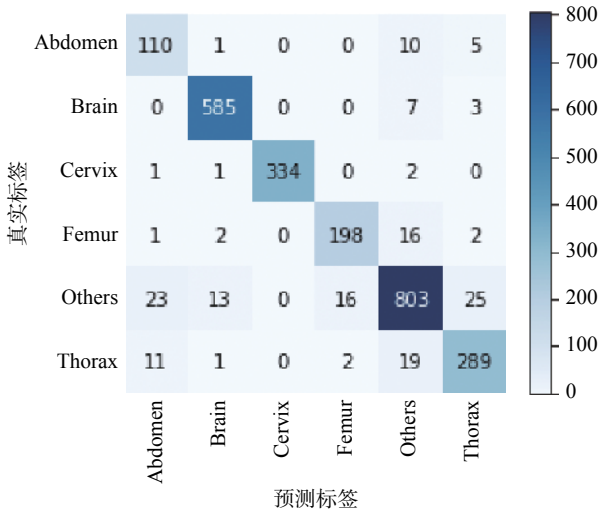
#### 3.3.1　第 1 阶段有效性实验

Resnet8 模型与 Resnet8+stage1 模型相比，前者是 Resnet8 模型直接训练得到的学生网络模型；而后者是经过改进的两阶段知识蒸馏方法第 1 阶段的 Resnet8 模型，再直接训练得到的学生网络模型。由图 6 可知，Resnet8+stage1 模型在除

“其他类型”切面图像外的各个分类的成功样本数较 Resnet8 模型增加 3~20 例不等。由表 9 可知, Resnet8+stage1 模型较 Resnet8 模型在准确率上提升了 1.53%, 宏召回率提高 3.92%, 宏 $F_1$ 值提高 2.2%, 仅在宏精确率上降低 0.07%。以上实验结果充分表明了改进的两阶段知识蒸馏方法的第 1 阶段训练的必要性和有效性, 具有相同残差结构的学生网络在第 1 阶段的训练中, 从教师网络的隐藏层特征表达能学习到有用的权重信息。




(a) Resnet8



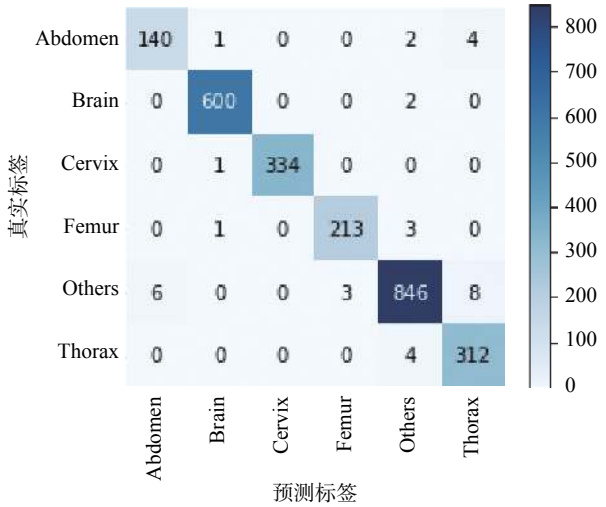

(b) Resnet8+Stage1

**图 6　混淆矩阵对比**

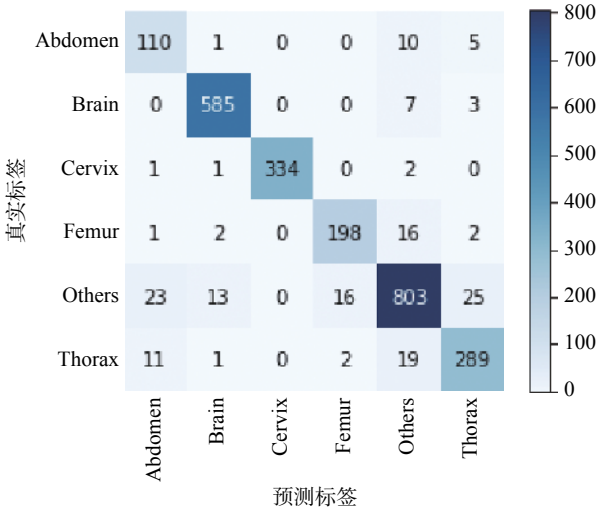
**Fig. 6　Confusion matrix with different methods**

### 3.3.2　第 2 阶段有效性实验

Resnet8+stage1 模型与 Resnet8+Hint 模型相比, 前者是经过改进的两阶段知识蒸馏方法第 1 阶段的 Resnet8 模型, 再直接训练得到的学生网络模型; 后者是经过改进的两阶段知识蒸馏方法的学生网络模型。由图 7 可知, Resnet8+Hint 模型较 Resnet8+stage1 模型, 在每个类别正确的分类样本数的最大增幅达到了 27.2%(腹部类切面图像); 由表 9 可知, Resnet8 + Hint 模型较 Resnet8+stage1 模型的各项指标性能有了大幅提升, 准确率提升 4.84%, 宏精确率提升 5.63%, 宏召回率提升 6.3%, 宏 $F_1$ 值提升 8.21%, 以上实验结果充分表明改进的两阶段知识蒸馏方法的第 2 阶段训练的有效性。




(a) Resnet8+Hint




(b) Resnet8+Stage1

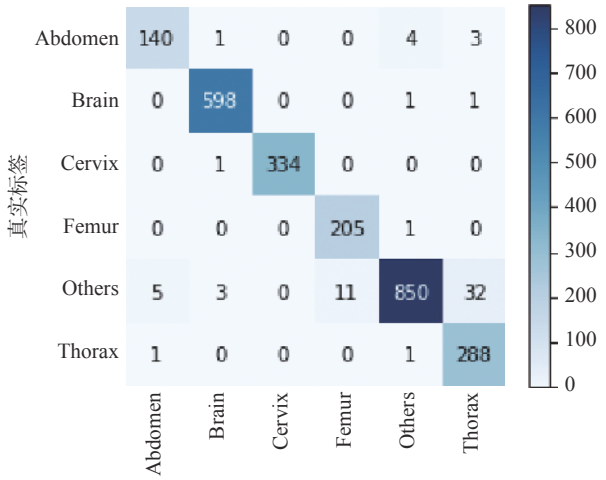
**图 7　混淆矩阵对比**

**Fig. 7　Confusion matrix with different methods**

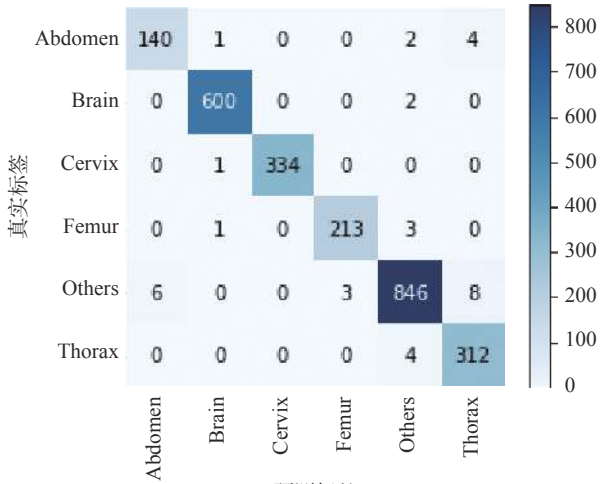
### 3.4　传统和改进的知识蒸馏方法对比

Resnet8+KD 模型与 Resnet8 + Hint 模型相比, 前者是经过传统知识蒸馏方法训练得到的学生网络模型, 后者是经过改进的两阶段知识蒸馏方法训练得到的学生网络模型。由图 8 可知, Resnet8+Hint 模型的主要提升在于“胎儿股骨”和“胎儿胸腔”切面的分类结果上, 而 Resnet8+KD 模型在这两类上的分类性能是次与前者的。由表 9 可知, Resnet8 + Hint 模型的准确率较 Resnet8+KD 模型提升 1.21%, 宏精确率提升了 0.4%, 宏召回率提升了 1.83%, 宏 $F_1$ 值提升了 1.18%, 以上各项性能指

标的提升都充分证明了改进的两阶段知识蒸馏方法的有效性。




(a) Resnet8+KD




(b) Resnet8+Hint

**图 8　混淆矩阵对比**

**Fig. 8　Confusion matrix with different methods**

经过改进的两阶段知识蒸馏方法的学生网络模型在各项分类指标都取得了大幅提升，较原有学生网络模型，准确率提升 6.37%，其他各项性能也得到了明显提升。较传统知识蒸馏方法训练的学生网络模型，准确率提升 1.21%，且在准确率指标上超过教师网络模型 1.09%。实验表明，在改进的两阶段知识蒸馏方法的第 1 阶段，与教师网络具有相同残差结构的学生网络能以教师网络的隐藏层特征表达作为监督信息，获得良好的中间层初始权重，为第 2 阶段知识蒸馏打下了良好基础。同时，使用层数浅、参数量较少的学生网络，可以有效避免模型因层数过深、参数量过大产生的过拟合问题，提升了模型的泛化能力，在保留分类性能的同时成功将模型参数量进行压缩。综上所述，充分表明了改进的两阶段知识蒸馏方法在提升学生网络模型各项性能的有效性。

## 4　结束语

针对医学图像的特点，考虑到深度神经网络模型在实际应用时的实时性能，本文提出了一种用于胎儿超声切面识别的改进的两阶段知识蒸馏方法。利用两种结构相似，但计算量相差较大的残差网络，即 Resnet8 作为学生网络，Resnet101 作为教师网络，通过现有知识蒸馏方法和改进的两阶段知识蒸馏方法在胎儿超声切面数据集上进行实验，分别达到 97.38% 和 98.59% 的准确率，后者在各项分类的性能指标上都取得了突破，由此可以得出改进的两阶段知识蒸馏方法优于现有知识蒸馏方法的结论。通过对比实验，表明改进的两阶段知识蒸馏方法的第 1 阶段，在具有相同残差结构的学生网络和教师网络之间进行隐藏层特征迁移的必要性和有效性。通过改进的两阶段知识蒸馏方法得到的学生网络模型 Resnet8+Hint 在准确率和各项性能上远超原有学生网络模型，在分类性能方面超过了教师网络模型，在计算资源占用量方面，大幅降低了对计算资源的需求，同时加快了实际应用时的分析速度，表明本文所述的改进的两阶段知识蒸馏方法的有效性。

## 参考文献：

[1] MARACI M A, NAPOLITANO R, PAPAGEORGHIOU A, et al. Searching for structures of interest in an ultrasound video sequence[C]//Proceedings of the 5th International Workshop on Machine Learning in Medical Imaging. Boston, USA, 2014: 133–140.

[2] PLATT J. Sequential minimal optimization: a fast algorithm for training support vector machines[J]. Advances in kernel methods-support vector learning, 1998, 7:1–22.

[3] BAUMGARTNER C F, KAMNITSAS K, MATTHEW J, et al. SonoNet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound[J]. IEEE transactions on medical imaging, 2017, 36(11): 2204–2215.

[4] MARACI M A, BRIDGE C P, NAPOLITANO R, et al. A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat[J]. Medical image analysis, 2017, 37: 22–36.

[5] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning.

San Francisco, USA, 2001: 282–289.

[6] RYOU H, YAQUB M, CAVALLARO A, et al. Automated 3D ultrasound biometry planes extraction for first trimester fetal assessment[C]//Proceedings of the 7th International Workshop on Machine Learning in Medical Imaging. Athens, Greece, 2016: 196–204.

[7] CHENG P M, MALHI M S. Transfer learning with convolutional neural networks for classification of abdominal ultrasound images[J]. Journal of digital imaging, 2017, 30(2): 234–243.

[8] DONAHUE J. CaffeNet(GitHubPage)[EB/OL]. (2016-05-6)[2021-05-11].https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet

[9] SIMONYAN K. VGG team ILSVRC-2014 model with 16 weight layers (GitHub Page)[EB/OL]. (2016-05-16)[2021-05-11]. https://gist.github.com/ksimonyan/211839e770f7b538e2d8

[10] HOWARD A G, ZHU Menglong, CHEN Bo, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2021-05-11]. https://arxiv.org/abs/1704.04861v1.

[11] SANDLER M, HOWARD A, ZHU Menglong, et al. MobileNetV2: inverted residuals and linear bottlenecks [C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4510–4520.

[12] ANDREW A, SANDLER M, GRACE C, et al. Searching for mobileNetV3[EB/OL]. (2019-05-06)[2021-05-11].https://arxiv.org/abs/1905.02244v3.

[13] BUCILUǍ C, CARUANA R, NICULESCU-MIZIL A. Model compression[C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 535–541.

[14] ROMERO A, BALLAS N, KAHOU S E, et al. FitNets: hints for thin deep nets[J]. (2014-12-19)[2021-05-11]. https://arxiv.org/abs/1412.6550v2.

[15] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09)[2021-05-11].https://arxiv.org/abs/1503.02531.

[16] JIN Xiao, PENG Baoyun, WU Yichao, et al. Knowledge distillation via route constrained optimization [C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision . Seoul, Korea , 2019: 1345–1354.

[17] PHUONG M H, LAMPERT C H. Towards understanding knowledge distillation[C]//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019.

[18] JI Guangda, ZHU Zhanxing. Knowledge distillation in wide neural networks: risk bound, data efficiency and imperfect teacher[EB/OL]. (2020-10-20)[2021-05-11].https://arxiv.org/abs/2010.10090.

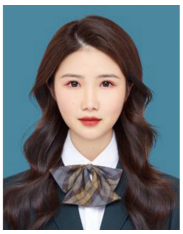
[19] ZHANG Ying, XIANG Tao, HOSPEDALES T M, et al. Deep mutual learning[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4320–4328.

[20] MOBAHI H, FARAJTABAR M, BARTLETT P L. Self-distillation amplifies regularization in Hilbert space [EB/OL]. (2020-02-13)[2021-05-11].https://arxiv.org/abs/2002.05715v1.

[21] ZHAI Mengyao, CHEN Lei, TUNG F, et al. Lifelong GAN: continual learning for conditional image generation[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision . Seoul, South Korea, 2019: 2759–2768.

[22] YUAN Li, TAY F E H , LI Guilin, et al. Revisit knowledge distillation: a teacher-free framework[C]//Proceedings of the International Conference on Learning Representation. Addis Ababa, Ethiopia, 2020.

## 作者简介：



张欣培，硕士研究生，主要研究方向为神经网络、医学图像处理。

周尧，助理研究员，主要研究方向为神经网络、进化计算。参与科技部科技创新 2030–“新一代人工智能”重大项目 1 项，主持国家自然科学基金项目 1 项。发表学术论文 7 篇。

章毅，教授，博士生导师，IEEE Fellow，主要研究方向为人工智能与智能医学。获国家自然科学二等奖、教育部自然科学一等奖、四川省科技进步一等奖。主持科技部科技创新 2030–“新一代人工智能”重大项目。发表学术论文 500 余篇，出版英文学术专著 3 部。