

面向机器视觉任务的多尺度图像特征压缩算法

张云飞, 贺丽君, 王子溪, 李凡

(西安交通大学电子与信息学部, 710049, 西安)

摘要: 针对协作智能框架中机器视觉任务的中间特征数据量大、难以高效传输等问题,提出了一种多尺度图像特征融合压缩算法。首先,根据边缘设备上深度学习模型输出的多尺度特征,设计级联残差变换模块,通过不同尺寸特征的逐级相减,消除多尺度特征的相关冗余,并将残差特征压缩到统一尺寸;然后,设计自编码器,用算术编码来消除紧凑特征的统计冗余;接着,在云端根据解码的紧凑特征设计预测重建模块生成预测特征,并与残差特征相结合实现多尺度特征的精确重建;最后,通过构建联合优化函数对残差变换、自编码和预测重建3个模块开展协同优化,实现传输码率与信息表征能力的最优权衡。仿真实验结果表明:所提算法在空间压缩时不仅特征压缩率最大,而且重建特征最完整;在传输码率为0.1 bit/像素时,与图像编码算法(VVC)和特征压缩算法(MS-FC)相比,所提算法的模型精度分别提高了8.57%和3.87%。该研究可为机器视觉编码框架提供技术支撑,具有一定的工程应用价值。

关键词: 特征压缩;图像编码;深度学习;多尺度特征;自编码器

中图分类号: TP391 **文献标志码:** A

DOI: 10.7652/xjtuxb202312001 **文章编号:** 0253-987X(2023)12-0001-10

Image Multi-Scale Feature Compression Algorithm for Machine Vision Tasks

ZHANG Yunfei, HE Lijun, WANG Zixi, LI Fan

(Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: For the problems that, in the collaborative intelligence framework, the intermediate feature data of machine vision tasks is large and difficult to transmit efficiently, a multi-scale image feature fusion compression algorithm was proposed. Firstly, a cascaded residual transformation module was designed according to the multi-scale features output by the deep learning model on the edge device, the redundancy of multi-scale features was eliminated by stepwise subtraction of features of different sizes, and the residual features were compressed to a unified size. Then, an autoencoder was designed to eliminate the statistical redundancy of compact features by arithmetic coding. Next, a prediction and reconstruction module was designed on the cloud according to the compact features of decoding to generate the prediction features, which were combined with the residual features to accurately reconstruct the multi-scale features. Finally, a joint optimization function was built for the collaborative optimization of the modules including residual transformation, autoencoder, and prediction reconstruction, thus achieving the optimal

trade-off between transmission bit rate and information representation ability. The simulation results show that the proposed algorithm has not only the largest feature compression ratio, but also the most complete reconstructed features in the space compression, and that when the transmission bit rate is 0.1bpp, the model accuracy of the proposed algorithm is improved by 8.57% and 3.87%, respectively, compared with the image coding algorithm VVC and the feature compression algorithm MSFC. This study can provide technical support for the coding framework of machine vision, and has certain value in engineering application.

Keywords: feature compression; image coding; deep learning; multi-scale feature; autoencoder

近年来,随着深度神经网络(deep neural networks, DNN)的发展,庞大的计算量和显存开销等问题日益凸显。当前边缘设备应用 DNN 的常见算法是将数据传输到云端执行视觉任务,但纯云端的数据传输方式延时大,通信开销高。Kang 等^[1]提出了协作智能(collaborative intelligence, CI)框架,将 DNN 的前半段放在边缘端,传输中间特征到云端完成推理任务,实现了边缘设备和云端的负载均衡^[2]。由于中间特征扩展了图像的维度,造成数据量较大,因此,特征的压缩传输效率是影响 CI 框架性能的关键因素。机器视觉编码(video coding for machine, VCM)^[3]技术应运而生,该编码技术致力于弥合视觉特征压缩与经典视频编码之间的鸿沟,将视觉特征作为桥梁,以可扩展方式连接信号级和任务级的紧凑表示^[4]。由于难点在于现有的编码器主要面向图像、视频设计,因此,将其应用于中间特征编码并不高效。此外,由于中间特征具有稀疏的数据分布,增强中间特征的紧凑性也是提高应用效率的关键。

目前,已有的特征压缩算法可分为编码压缩、空间压缩、特征优化和多尺度特征压缩 4 类。编码压缩利用传统或非传统的编码器将特征量化,消除统计冗余后编码为二进制码流。元辉等^[5]提出基于局部亮度补偿的视频编码算法,提高了视频的编码效率。类比可知,特征也能去除冗余信息,从而进行压缩传输。汪维等^[6]对残差网络(ResNet)^[7]各层输出开展了有损编码与无损编码的对比实验,证明无损压缩效率较低。Choi 等^[8]研究了基于高效视频编码(HEVC)^[9]的特征编解码方案,对每个通道的特征图进行量化处理,使用 HEVC 编码器将特征编码进行传输。Suzuki 等^[10]研究了中间特征各通道之间的伪时间关联性,通过对特征进行时空排列,使用视频编码器完成了编码传输。然而,上述研究仅通过量化编码去除统计冗余,并没有考虑特征的空

间冗余。

空间压缩从特征本身的稀疏性出发,通过调整特征的空间尺寸进一步压缩特征数据量。Eshrati-far 等^[11]提出了一种名为瓶颈单元的可学习压缩单元和有损压缩感知训练算法,分别在 ResNet-50 和 VGG-19^[12]网络上实现了特征压缩。Ikusan 等^[13]考虑特征之间的相似性,设计了分层聚类特征选择策略。由于上述算法依然采用传统编码器对紧凑特征编码进行传输,码流中仍存在较多的统计冗余。

通过可训练的编码模块,特征优化能够将特征压缩与参考网络联合优化。宋蓓蓓等^[14]提出了结合码率预测和率失真优化的码率控制算法,实现了对图像码率的高精度控制。Alvar 等^[15]设计了一个损失函数衡量特征的压缩特性,用来约束模型的训练过程。Chamain 等^[16]利用压缩模块和任务算法组成链路,联合微调编解码器和任务网络显著提高了任务精度。然而,端到端的训练令压缩传输的特征与任务强相关,限制了特征的泛化能力。

多尺度特征压缩是适用于复杂网络的特征压缩算法。为了保证对多尺寸目标的检测效果,主流的检测、分割网络如 Faster RCNN^[17]、YOLO^[18]、Mask RCNN^[19]等,需要在主干网络中引出多条支路分别完成推理任务。Liu 等^[20]提出将图像转换为多个子带的金字塔,通过对变换模块进行训练,发现特征对多种机器视觉任务有用。Zhang 等^[21]提出了一种端到端的多尺度特征压缩框架,将多尺度特征融合为单尺度特征进行编码传输。但是,以上这些算法既没有探索不同尺度特征之间的关联,也没有针对中间特征进行编码优化。

为了实现多尺度特征的高效压缩传输,本文对比了不同尺度特征的轮廓和细节,发现多尺度特征之间存在着强相关性,可以通过残差变换去除相关性,然后使用预测重建的方式恢复特征,消除传输过程的空间冗余,同时采用自编码器消除传输过程的

统计冗余。在此基础上,提出了一种基于残差变换模块、自编码器和预测重建模块联合优化的多尺度特征压缩算法。首先,设计了预测重建模块,将多尺度特征逐层作差压缩为紧凑特征;其次,采用面向特征的自编码器,将紧凑特征编码传输;然后,设计预测重建模块,将预测特征和重建的残差特征相结合,实现了多尺度特征的精确重建,并采用 3 个可训练的算法模块,使用码流与保真度构建损失函数进行联合优化;最后,通过仿真实验对比了多个特征压缩算法码率-精度的性能,证明了本文算法的优越性。

1 多尺度特征联合压缩算法

1.1 算法总述

本文以 Faster RCNN 作为骨干网络,提出了一种多尺度特征联合压缩算法,算法总览如图 1 所示。该算法由残差变换模块、面向特征的自编解码器和预测重建模块 3 部分组成,通过 3 个模块的联合优化,不仅实现了多尺度特征的融合压缩编码,还降低了 VCM 算法的传输延时与能耗,提高了协作智能的应用效率。

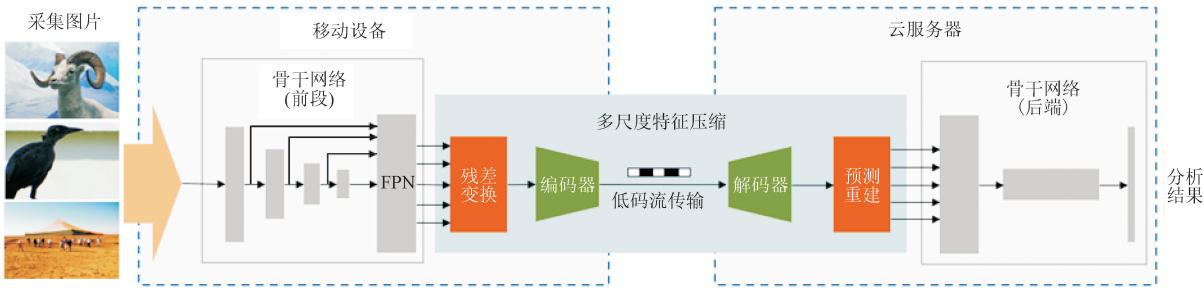


图 1 面向多尺度特征的联合压缩算法总览

Fig. 1 Overview of joint compression algorithms for multi-scale features

残差变换模块和自编码器部署在边缘设备,解码器和预测重建模块部署在云端。在边缘端使用骨干网络的前端提取多尺度特征,采用残差变换模块消除多尺度特征的相关性,并转化为统一尺寸的规范特征,然后使用自编码器编码为比特流上传到云服务器。解码器对接收到的数据进行解码处理,经过预测重建模块完成多尺度特征的高质量恢复,最后通过骨干网络的后端完成推理任务。

1.2 残差变换模块

在设计多尺度特征压缩模块之前,先探索了不同尺度特征之间的关联性。提取 Faster RCNN 网络模型中特征金字塔(feature pyramid network, FPN)的不同尺寸特征 p_2 、 p_3 、 p_4 、 p_5 、 p_6 ,其中 p_6 由 p_5 通过池化操作获得,由于不包含 p_5 以外的信息,所以在进行特征压缩研究时不予考虑。图 2 给出了经过伪彩图处理的 p_2 、 p_3 、 p_4 、 p_5 层特征的可视化结果。

统一尺寸后发现,不同层的特征具有相似的轮廓,即多尺度特征之间存在明显的相关性。若将这种相关性应用于特征压缩中,可以极大地节省中间特征压缩传输的数据量。但是,不同尺度特征之间的细节差异也不能忽略,因为这些细节信息赋予了 DNN 网络对多尺度物体的辨识能力。因此,本文设

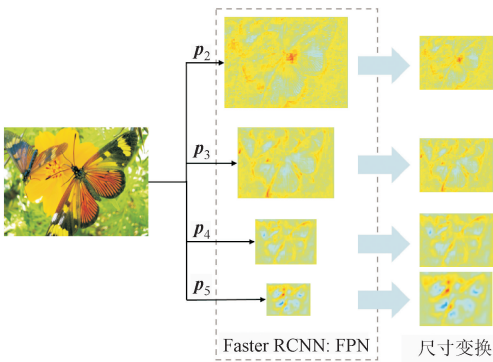


图 2 Faster RCNN 网络模型中多尺度特征可视化结果

Fig. 2 Visualization of intermediate multi-scale features in Faster RCNN net

计了针对多尺度特征的差异信息进行压缩编码的残差变换模块,在保留不同特征差异信息的同时,通过残差处理高效压缩多尺度特征。

多尺度特征在空间上呈现为金字塔结构,可通过逐层作差的方式消除层级相关性。文献 [21] 的实验结果表明, p_5 层对冻结模型推理的影响程度最大。因此在进行残差计算时,以 p_5 层作为基本特征,后续处理只进行数值变换。值得注意的是,不同特征的空间尺寸、数据分布都不相同,不能直接进行残差计算,而是需要加入反卷积层^[22] 充当空间变换的桥

梁,从而扩展低级特征的尺寸。残差变换模块的具体网络结构见图 3,其中 f_2 、 f_3 、 f_4 、 f_5 分别表示被压缩的 p_2 、 p_3 、 p_4 、 p_5 层的残差特征, f 表示残差变换模块输出的统一格式的单尺度特征, k 、 s 、 c 分别表示卷积层和反卷积层的尺寸、步长和输出通道数。

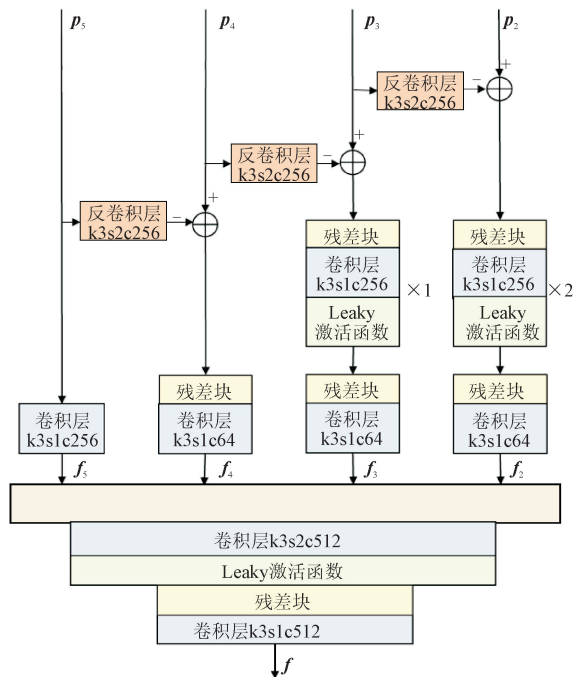


图 3 残差变换模块的网络结构

Fig. 3 Network of the residual transformation module

该模块由卷积层、残差块、反卷积层和通道拼接操作组成,根据多尺度特征的金字塔特性,逐层作差消除不同尺寸特征的相关信息,然后对残差特征进行空间变换,与 p_5 的尺寸保持一致。与此同时,对 p_5 特征采用尺寸为 3×3 的卷积进行数值变换,尽量消除与残差特征数值分布上的差异。最后对所有特征按通道进行拼接,经过卷积层与残差块的处理,输出具有统一格式的单尺度特征。压缩过程可表示为

$$f_5 = \text{trans}(p_5) \quad (1)$$

$$f_4 = \text{trans}(p_4 - \text{sizeAlign}(p_5, p_4)) \quad (2)$$

$$f_3 = \text{trans}(p_3 - \text{sizeAlign}(p_4, p_3)) \quad (3)$$

$$f_2 = \text{trans}(p_2 - \text{sizeAlign}(p_3, p_2)) \quad (4)$$

$$f = \text{trans}(\text{concat}(f_5, f_4, f_3, f_2)) \quad (5)$$

式中: $\text{trans}(\cdot)$ 操作为对残差特征进行尺寸变换和数值变换; $\text{sizeAlign}(a, b)$ 操作为将张量 a 的空间尺寸变换到和张量 b 一致; $\text{concat}(\cdot)$ 操作为将特征按照通道进行拼接。

残差块是由主分支和快捷连接组成的双分支结构,输入输出均为特征图。主分支由两个尺寸为

3×3 的卷积和广义分裂归一化(GDN)层^[23]构成,第一个卷积步长设置为 2,在完成输入特征空间尺寸的变换后,采用 GDN 层代替批归一化(BN)层,以便后续多尺度特征的重建。快捷连接限制输入特征的内容损失,使用尺寸为 1×1 的卷积实现两条支路输出特征空间尺寸的匹配。上述具体结构见图 4,其中, C 为输入特征图的通道数目, H 为特征高度, W 为特征宽度。

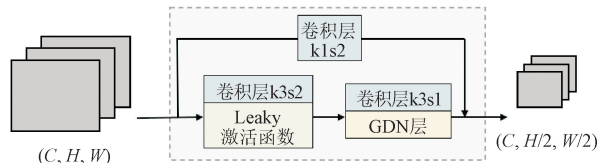


图 4 残差块的基本结构

Fig. 4 Basic structure of the residual block

综上所述,残差变换模块通过消除多尺度特征之间的空间冗余,实现了对多尺度特征的高效压缩。将残差块与卷积层相组合进行特征变换,在实现残差特征压缩的同时,保留了中间特征的表达能力,让压缩后的特征在后期能够得到高质量的恢复。

1.3 面向特征的自编码器

近年来,深度图像编码算法得到深入发展。与 HEVC 算法相比,在相同的码率下,最新的 Cheng2020 算法^[24]获得的图片重建性能更高。此算法通过深度编码器扩展图片特征空间,经过量化编码传输,采用对称的孪生网络重建图片,其本质为特征的压缩传输。残差变换模块输出的特征已经过压缩处理,能够较方便地进行编码传输,之后再通过对称的解码结构生成重建特征。自编码器的网络结构见图 5,其中, f 表示残差变换模块输出的压缩特征, f' 表示自编码器的重建特征。

由骨干网络前端提取到的该模块输入的中间特征,经过残差变换模块的压缩已经能满足紧凑性的要求,不需要进一步再压缩。对于超先验编码器,该模块输入的压缩特征尺寸较小,最多能进行两次下采样,否则包含的信息不足以指导算术编码器。解码器的目标是最大程度地还原压缩信息,所以采用对称设计。此时,使用逆函数 iGDN 层替代解码器的 GDN 层,将下采样的卷积层改为上采样的反卷积层,并对称设计输入、输出的通道数。

采用面向特征的自编码器去除编码冗余,可实现特征的高效编码。该模块是可训练的,与残差变换模块进行联合优化,能够实现传输码率与表现力的最佳权衡。

不同的数值。

2.2 数据集

实验使用 COCO 数据集^[25]训练,在 OpenImage 数据集^[26]上进行性能评估。COCO 数据集是图像检测领域的经典数据集,OpenImage 数据集是可商用的、规模较大且标注精确的图像检测分割数据集。

2.2.1 COCO 数据集

此数据集主要从复杂场景中截取,图像中的目标通过精确分割进行位置标定。面向检测任务,训练集中有 118 287 张图片,测试集中有 5 000 张图片,各划分为 80 个类别。实验中,将 COCO 数据集中的图片输入到冻结 DNN 网络的前端以获取多尺度特征,用于特征压缩算法的训练优化。

2.2.2 OpenImage 数据集

此数据集包含 900 万张训练图片、41 620 张验证图片和 125 456 张测试图片,能够提供图像层标签以及分割掩码,可用于目标检测和语义分割任务的训练及测试。

2.3 测试流程和评估指标

通过 Detection2 中的检测模型 Faster RCNN 和分割模型 Mask RCNN,对 OpenImage 数据集的子集进行测试。对于原始图像,使用图像编码算法(VVC)进行压缩传输,重建图像后转换成特定格式送入神经网络,得到不同机器视觉任务的推理结果。对于 Faster RCNN 网络模型中 FPN 输出的多尺度特征,采用 VVC 算法进行压缩传输,重建特征以完成后续机器视觉任务,得到不同任务的推理结果。最终比较时,以 VVC 算法压缩传输的原始图像和多尺度特征作为参考,通过码率-精度曲线证明本文算法的优越性。

对于目标检测和语义分割任务,使用传输码率和 DNN 网络推理精度对提出的特征编码算法进行评估。对于图像数据集,传输码率 R 定义为压缩码流比特数与原始图片像素数之比,以每个像素占用的比特数($\text{bit} \cdot \text{像素}^{-1}$)作为衡量单位。

对于目标检测和语义分割,采用平均精度(mean average precision, mAP)来衡量 DNN 网络的性能。其中,给定阈值的召回率定义为该阈值对应的所有真正样本与假负样本中所有真正样本的比例,推理精度定义为来自所有正例与所有真正例的比例。实验中测试 mAP 时,阈值设置为 0.5。

2.4 实验结果评估

2.4.1 特征融合与空间性能评估

对残差变换模块进行特征融合,需要完成空间

压缩。因此,首先验证经过空间压缩的特征能否被精确恢复。为了进行比较,分别采用 MSFC 算法、Cheng2020 算法和本文算法测试了特征压缩的推理精度。由于残差变换模块没有编码压缩能力,因而仅计算原始码率。

原始图像的宽、高、通道数分别定义为 w 、 h 、 c ,由于图像处于像素空间, c 取 3,因此原始图像的空间信息可以表示为 $(3, h, w)$ 。由于每个元素都是 uint8 的非负整数,数值深度为 8 bit,可知原始图像的每像素比特为 24,下面分别说明各算法的空间压缩效率。对于 Cheng2020 算法,在骨干网络前端对图像提取特征的基础上,将输入张量的通道数与特征通道数保持一致,对特征逐个进行压缩编码,其中 p_2 空间尺寸过大,采用采样算法将空间尺寸变为原来的一半, p_2 、 p_3 、 p_4 、 p_5 压缩后的空间信息分别为 $(256, h/32, w/32)$ 、 $(256, h/32, w/32)$ 、 $(256, h/64, w/64)$ 、 $(256, h/128, w/128)$ 。

对于 MSFC 算法,多尺度特征进行了融合压缩,压缩后的融合特征表示为 $(64, h/32, w/32)$ 。对于本文算法,多尺度特征进行残差处理再经过融合压缩,可表示为 $(192, h/64, w/64)$ 。特征图元素均为 Float32 浮点数,数值深度为 32 bit,计算每种算法的码率并测试 DNN 网络的推理精度,结果见表 1。其中,特征压缩率定义为每种算法的码流与多尺度特征码率之比,用以说明各算法空间压缩的效果。

表 1 不同压缩传输方法空间压缩性能对比(未编码)

Table 1 Comparison of feature compression performance based on VCM test conditions (uncoded)

传输方法	码率/ ($\text{bit} \cdot \text{像素}^{-1}$)	压缩率/ %	精度/ %
原始图像	24.00	10.15	79.28
多尺度特征	236.46	100.00	79.28
Cheng2020 ^[24]	12.59	5.32	79.08
MSFC ^[21]	2.78	1.18	79.18
本文算法	2.16	0.91	79.23

从表 1 可以看出,DNN 网络对特征的提取扩展了图片的表达空间,多尺度特征的数据量最大。3 种特征压缩算法中,Cheng2020 由于面向图像编解码设计,特征信息损失最多。本文算法的特征压缩比最大,但 DNN 网络的推理效果最好,表明本文算法重建的多尺度特征表达能力最完整。

2.4.2 特征压缩算法性能评估

本文通过计算码率-精度(rate-accuracy, RA)

性能评估特征的传输效率,分别测试了 VVC 传输图片、VVC 传输多尺度特征、Cheng2020 编码、MSFC编码及本文算法编码的传输特征。对于传统编解码算法,通过调整量化参数实现对码率的控制;对于深度学习算法,可通过调整损失函数中的权重参数实现对码率的约束;对于本文算法,通过设置不同的权重参数 $\lambda=\{15,25,50,100,1\,000,5\,000\}$,得到不同的传输码率。采用不同压缩传输算法得到的 RA 曲线如图 6 所示。

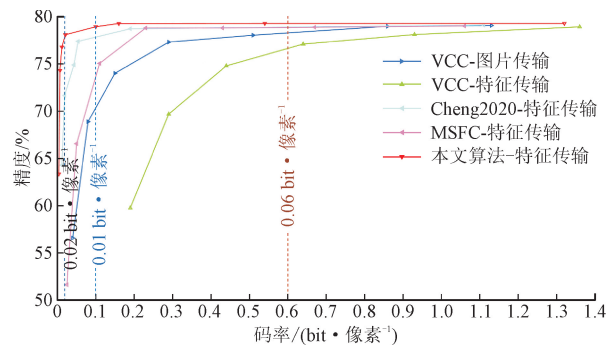


图 6 不同压缩传输算法码率-精度性能对比(已编码)
Fig. 6 Comparison of RA curves for different compression methods (encoded)

由图可见,本文算法的特征压缩效果最优。在传输码率小于 $0.1\text{ bit} \cdot \text{像素}^{-1}$ 的极低码率环境下,各种压缩传输算法均不能维持特征信息的完整性,导致推理精度急剧下降。然而,随着传输码率的降低,本文算法退化最慢,可以在极低码率下获得最优的推理性能。在传输码率约为 $0.02\text{ bit} \cdot \text{像素}^{-1}$ 时,DNN 网络的推理精度损失不超过 2%,在码率为 $0.004\text{ bit} \cdot \text{像素}^{-1}$ 时,本文算法依然能够维持 63.30% 的推理精度。上述结果表明,本文算法的压缩特征是有用的,能够提高 VCM 框架的传输效率。

为了直观地比较各种压缩传输算法的差异,将压缩特征重建并可视化。由于码率为 $0.1\text{ bit} \cdot \text{像素}^{-1}$ 时,各算法重建特征差异最大,因而给出此码率下可视化 p_2 层的特征,如图 7 所示。可以看出,VVC-图片传输和本文算法更多地保留了原始特征的轮廓信息。但 VVC 算法在低码率传输图片时,由于图片信息丢失,导致中间特征的数值分布和原始图像特征存在较大差异,推理效果并不理想。而本文算法在保留特征轮廓信息的同时,以 MSE 函数作为约束,尽可能在细节上与原始特征保持一致,因此特征重建效果最好。

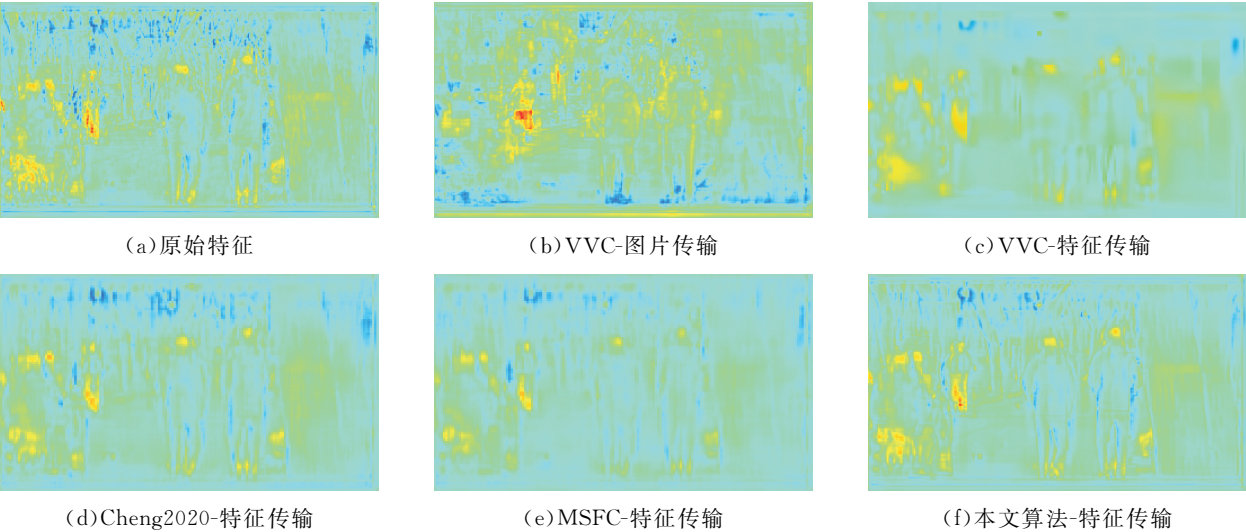


图 7 不同压缩传输算法 p_2 层特征可视化对比
Fig. 7 Visual comparison of p_2 features in various compression transmission methods

2.5 消融实验

为了证明本文算法的有效性,将残差变换模块替换为直接下采样开展对比实验。多尺度特征可以压缩为单尺度特征进行编码传输,其中最简单的方式就是用双线性插值下采样将多尺度特征尺寸规范化,然后按通道拼接,再做编码传输。

为了便于对比,将残差变换模块直接替换成双线性插值下采样,用双线性插值上采样代替对应的预测重建模块,算法的其他部分保持不变,计算得到的结果见表 2。可以看到,在传输码率基本一致时,残差变换模块的加入令 CI 框架表现出更强的检测能力。

表 2 直接下采样与残差变换模块特征压缩效果对比

Table 2 Comparison of feature compression effect between direct down sampling and residual transform module

残差变换		下采样	
码率/ (bit · 像素 ⁻¹)	精度/ %	码率/ (bit · 像素 ⁻¹)	精度/ %
0.160	79.22	0.150	76.29
0.021	78.04	0.018	74.51
0.012	76.73	0.013	73.23
0.007	74.25	0.006	69.45
0.004	63.30	0.004	52.39

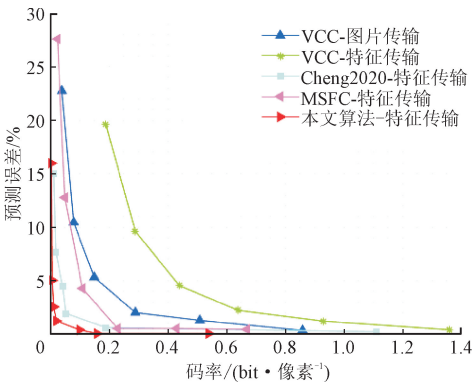
接着,开展消融实验进行复杂度分析,分别测试了 CI 框架的推理时间和运算量,结果见表 3。虽然在 CI 框架中插入残差变换模块导致整体计算量有所增加,但在相似的码率下,残差变换模块对协作推理性能的提升超过了 3%。因此,相比于压缩性能的提升,残差变换模块带来的额外计算量可以接受。表 3 插入下采样与残差变换模块 CI 框架复杂度对比

Table 3 Complexity comparison of CI framework after inserting down-sampling and residual transformation modules

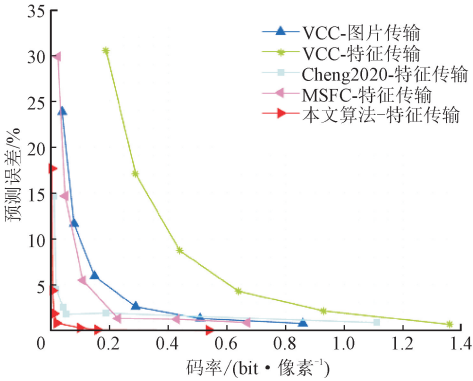
变换方法	推理时间/ ms	浮点运算量/ 10 ⁹	网络参数量/ 10 ⁶
下采样	104.5	409.35	106.19
残差变换	120.4	436.58	114.77

2.6 泛用性验证

本文算法针对中间特征进行编码传输,可以移植到各种多任务网络,根据压缩前后的多尺度特征差异进行联合优化,具有泛用性。因此,将其应用于广泛使用的多任务网络模型 Mask RCNN,分别开展目标检测和语义分割性能测试,完成泛用性验证。实验设置与 Faster RCNN 网络模型相同,为了观察多任务网络经过特征压缩后推理能力的衰减情况,定义预测误差为未压缩预测结果和压缩预测结果的平均精度之差,得到的结果如图 8 所示。以上实验结果表明,本文算法可以用于多任务网络,在多个任务上都具备良好的压缩性能。在低码率场景下,本文算法的信息损失更慢,表明其在多任务网络中也可以进行特征压缩。



(a)目标检测任务的预测误差



(b)语义分割任务的预测误差

图 8 面向多任务网络的压缩传输性能对比

Fig. 8 Comparison of feature compression performance for multitask network

3 结 论

针对协作智能框架复杂网络中间特征数据量大、难以高效传输的问题,提出了一种面向机器视觉任务的多尺度图像特征压缩算法,将残差变换模块与自编码器的联合压缩优化,实现了对多尺度中间特征的高效压缩。通过仿真实验,得到以下结论。(1)在相同的特征压缩比下,本文算法重建的多尺度特征表达能力最完整。当压缩中间特征的数据量为原始特征数据量的 0.91% 时,依然能够正常完成 DNN 网络的推理任务。(2)相较于已有的传统图像编码和深度特征编码算法,当传输码率为 0.02 bit · 像素⁻¹时,本文算法 DNN 网络的推理精度损失不超过 2%;当码率为 0.1 bit · 像素⁻¹时,不同压缩算法下特征可视化重建的对比结果表明,本文算法在细节上能与原始特征保持一致,重建效果最好。在云边协作智能框架中,对中间特征的高效压缩和传输是提高框架效率的核心问题,本文算法具有良好的特征压缩和编码传输性能,能够提

高云边协作智能框架的应用效率。此外,开展的泛化性实验表明,本文算法解码特征具有一定的泛化能力。

参考文献:

- [1] KANG Yiping, HAUSWALD J, GAO Cao, et al. Neurosurgeon: collaborative intelligence between the cloud and mobile edge [J]. *ACM SIGARCH Computer Architecture News*, 2017, 52(4): 615-629.
- [2] 陈玉平, 刘波, 林伟伟, 等. 云边协同综述 [J]. *计算机科学*, 2021, 48(3): 259-268.
CHEN Yuping, LIU Bo, LIN Weiwei, et al. Survey of cloud-edge collaboration [J]. *Computer Science*, 2021, 48(3): 259-268.
- [3] AKBARI M, LIANG Jie, HAN Jingning. DSSLIC: deep semantic segmentation-based layered image compression [C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ, USA: IEEE, 2019: 2042-2046.
- [4] XIA Sifeng, LIANG Kunchangtai, YANG Wenhao, et al. An emerging coding paradigm VCM: a scalable coding approach beyond feature and signal [C]//2020 IEEE International Conference on Multimedia and Expo (ICME). Piscataway, NJ, USA: IEEE, 2020: 1-6.
- [5] 元辉, 常义林, 卢朝阳, 等. 一种应用于光照变化场景的视频编码方法 [J]. *西安交通大学学报*, 2010, 44(10): 62-66.
YUAN Hui, CHANG Yilin, LU Zhaoyang, et al. A novel video coding method for luminance variation scenes [J]. *Journal of Xi'an Jiaotong University*, 2010, 44(10): 62-66.
- [6] 汪维, 徐龙, 陈卓. 深度学习模型中间层特征压缩技术综述 [J]. *计算机应用研究*, 2023, 40(5): 1281-1291.
WANG Wei, XU Long, CHEN Zhuo. Review of intermediate deep feature compression [J]. *Application Research of Computers*, 2023, 40(5): 1281-1291.
- [7] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 770-778.
- [8] CHOI H, BAJIC I V. Deep feature compression for collaborative object detection [C]//2018 25th IEEE International Conference on Image Processing (ICIP). Piscataway, NJ, USA: IEEE, 2018: 3743-3747.
- [9] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding (HEVC) standard [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2012, 22(12): 1649-1668.
- [10] SUZUKI S, TAKEDA S, TAKAGI M, et al. Deep feature compression using spatio-temporal arrangement toward collaborative intelligent world [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, 32(6): 3934-3946.
- [11] ESHRATIFAR A E, ESMAILI A, PEDRAM M. BottleNet: a deep learning architecture for intelligent mobile cloud computing services [C]//2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED). Piscataway, NJ, USA: IEEE, 2019: 1-6.
- [12] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10) [2022-10-16]. <https://arxiv.org/abs/1409.1556>.
- [13] IKUSAN A, DAI Rui. Rate-distortion optimized hierarchical deep feature compression [C]//2021 IEEE International Conference on Multimedia and Expo (ICME). Piscataway, NJ, USA: IEEE, 2021: 1-6.
- [14] 宋蓓蓓, 何帆, 马穗娜, 等. 一种高精度的自适应码率控制图像压缩算法 [J]. *西安交通大学学报*, 2022, 56(2): 198-206.
SONG Beibei, HE Fan, MA Suina, et al. An image compression algorithm with high-precision adaptive rate control [J]. *Journal of Xi'an Jiaotong University*, 2022, 56(2): 198-206.
- [15] ALVAR S R, BAJIC I V. Bit allocation for multi-task collaborative intelligence [C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ, USA: IEEE, 2020: 4342-4346.
- [16] CHAMAIN L D, RACAPÉ F, BÉGAINT J, et al. End-to-end optimized image compression for machines, a study [C]//2021 Data Compression Conference (DCC). Piscataway, NJ, USA: IEEE, 2021: 163-172.
- [17] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [C]//Advances in Neural Information Processing Systems. San Francisco, CA, USA: Curran Associates Inc., 2015: 28.
- [18] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 779-788.

- [19] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN [C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA; IEEE, 2017: 2980-2988.
- [20] LIU Kang, LIU Dong, LI Li, et al. Semantics-to-signal scalable image compression with learned revertible representations [J]. International Journal of Computer Vision, 2021, 129(9): 2605-2621.
- [21] ZHANG Zhicong, WANG Mengyang, MA Mengyao, et al. MSFC: deep feature compression in multi-task network [C]//2021 IEEE International Conference on Multimedia and Expo (ICME). Piscataway, NJ, USA; IEEE, 2021: 1-6.
- [22] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks [C]//Computer Vision-ECCV 2014. Cham, Germany: Springer International Publishing, 2014: 818-833.
- [23] BALLÉ J, LAPARRA V, SIMONCELLI E P. Density modeling of images using a generalized normalization transformation [EB/OL]. (2016-02-29) [2023-01-10]. <https://arxiv.org/abs/1511.06281>.
- [24] CHENG Zhengxue, SUN Heming, TAKEUCHI M, et al. Learned image compression with discretized Gaussian mixture likelihoods and attention modules [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA; IEEE, 2020: 7936-7945.
- [25] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [C]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.
- [26] KUZNETSOVA A, ROM H, ALLDRIN N, et al. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale [J]. International Journal of Computer Vision, 2020, 128(7): 1956-1981.

(编辑 李慧敏 刘杨)