

基于局部短接单向融合网络的骨架检测*

乔杨, 肖士湘, 刘悦, 焦建彬[†]

(中国科学院大学电子电气与通信工程学院, 北京 100049)

(2021 年 4 月 21 日收稿; 2021 年 5 月 30 日收修改稿)

Qiao Y, Xiao S X, Liu Y, et al. Skeleton detection based on local short-connection unidirectional fusion networks[J]. Journal of University of Chinese Academy of Sciences, 2023, 40(2): 250-257. DOI: 10. 7523/j.ucas. 2021. 0048.

摘要 近年来, 基于侧输出网络的骨架检测方法获得了显著的性能提升。但是, 现有方法仍无法解决侧输出结构中高倍上采样和下采样带来的图像失真问题, 固定的感受野大小也限制了其视觉特征表达能力。为解决这些问题, 提出一种基于侧输出连接的局部短接单向融合网络。该网络由特征提取网络和侧输出网络组成。特征提取网络为深度卷积神经网络, 主要用于多层次视觉特征提取。侧输出网络包含局部短接网络和单向融合网络 2 个模块, 其中局部短接网络通过整合感受野邻近特征逐步构建起连续的大感受野特征, 而多尺度特征从深到浅的单向融合则实现了对目标从粗糙到精细的刻画。在 4 种常用骨架检测数据集上的实验结果验证了所提方法的有效性。

关键词 骨架检测; 局部短接单向融合网络; 侧输出网络

中图分类号: TP391 文献标志码: A DOI: 10. 7523/j.ucas. 2021. 0048

Skeleton detection based on local short-connection unidirectional fusion networks

QIAO Yang, XIAO Shixiang, LIU Yue, JIAO Jianbin

(School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract In recent years, skeleton detection based on side-output network has shown significant effectiveness. However, the existing methods are still unable to tackle the problem of image distortion in side-output structure caused by high-multiplier up/down-sampling, and the fixed receptive field limits the feature expression of the networks. To solve these problems, this paper proposes a local short-connection unidirectional fusion network based on side-output connection, which includes a feature extraction network and a side-output connection network. The feature extraction network is a deep convolutional neural network, which is used for multi-layer feature extraction. The side-output connection network consists of a local short-connection unit and a unidirectional fusion network. The local short-connection gradually constructs the continuous large receptive field features by integrating the adjacent features of receptive field, while the unidirectional fusion of multi-scale features from deep to shallow can achieve the characterization of the object from

* 国家自然科学基金(61771447)资助

[†] 通信作者, E-mail: jiaojb@ucas.ac.cn

rough to fine. Experimental results on four commonly used skeleton detection datasets demonstrate the effectiveness of the proposed method.

Keywords skeleton detection; local short-connection unidirectional fusion network; side-output network

骨架是一种对目标形状和形状间的拓扑连接进行抽象化描绘的特征,包含着物体丰富的拓扑信息和几何形状描述。目标骨架可作为其他视觉任务的输入,如人体姿态估计^[1]、手势识别^[2]、文本检测^[3]和区域提取^[4]等;又可作为视觉任务中的空间约束,进行模型的可解释性研究。对骨架检测的研究和应用,有效地推动着计算机视觉算法的发展。

图像中目标的骨架又被称为对称轴,需满足以下要求:1)左右两侧纹理相近;2)左右两侧到达目标边缘距离相似。在这样的约束下,骨架能够比较有效地反映出目标的结构。特别地,人体、手掌等目标的骨架可描述目标的自身形态。基于监督学习的骨架检测是一种像素级二分类问题^[5]:对输入图像中每一个像素点进行分类,判定其为骨架或背景。

深度卷积神经网络^[6]能够显著增强视觉特征表达能力,基于深度卷积神经网络的骨架检测算法自然地成为骨架检测的一种有效途径。深度卷积神经网络具有天然的特征金字塔特性:在特征提取过程中,随着特征图在网络中的逐层传递,网络先后提取到尺度较小、语义较浅的特征和尺度较大、语义较深的特征。将不同尺度、不同语义深度的特征进行融合,进行监督学习,即是深度学习时代骨架检测的范式。

最早的基于深度学习的像素级二分类模型是由 Xie 和 Tu^[7]提出的整体嵌套边缘检测网络(holistically-nested edge detection, HED)。该模型在边缘检测问题上展示出良好的性能,取得了相对非深度学习方法(如 MIL^[8]等)的显著性能提升。HED 通过截取深度卷积神经网络中不同尺度阶段的特征图输出作为侧输出加以融合,同时对各侧输出和融合输出进行监督学习。由于深度卷积神经网络中多尺度感受野的天然特性,HED 获得了良好的多尺度特征融合能力。多尺度特征侧输出融合的结构也成为后续骨架检测研究的雏形,在 HED 的基础上,多种骨架检测器^[9-15]先后出现,分别以不同的方式进行多尺度特征融合,提升了检测器在像素级二分类任务上的性能表现。

侧输出残差网络^[9](side-output residual network, SRN)基于侧输出范式,在融合阶段加入残差单元,每个单元输出一个侧输出的同时向上传递残差,各单元从深到浅进行嵌套。SRN 从深到浅的特征融合有效地提升了模型性能。线性延展网络^[10](linear span network, LSN)基于 SRN,在残差传递模块中加入密集连接(dense connection),在不增加模型参数的同时获得了性能提升。

现有的基于侧输出网络的方法虽然在骨架检测上获得了显著的性能提升,然而,其结构仍存在一些不足之处:

1)侧输出的融合过程中会对较深层的小尺度的特征图进行多至 16 倍甚至更高的上采样,导致高倍上采样后的输出特征图存在较大的失真,无法真实地反映骨架的特征,使得具有高层语义信息的深层特征无法得到充分利用,进而严重影响融合后多尺度特征的表达能力。为解决这一问题,DSS^[11]丢弃了最深层的侧输出,Hi-Fi^[14]和 FSDS^[15]则提出尺度相关的侧输出思想,但是引入了更为复杂的标注过程。我们希望在无添加任何复杂处理的条件下,通过直接改变网络结构的方式减少深层特征信息的失真。

2)侧输出网络中每层特征的感受野大小由特征提取网络的卷积核大小和池化层确定,是一组固定的数值。对于一组经多尺度特征融合的输出特征,其感受野大小由每一层的感受野决定,也是固定的,而固定的感受野大小限制了深层特征表达。

为解决以上问题,受神经科学中人类视觉认知的 glance-and-focus 模型^[16-17]的启发,本文提出一种基于侧输出连接的局部短接单向融合网络(local short-connection unidirectional fusion, LoSUF)。根据 glance-and-focus 模型,本文将骨架检测的过程描述为扫视定位和聚焦观察的过程:

首先,在大感受野上对图像进行全局浏览(glance),定位出目标的大致位置;然后,对定位到的目标区域进行小感受野切换,聚焦目标骨架进行细致观察(focus)。在这个 2 步学习的过程

中,大感受野、高层语义的深度特征如位置信息等被用于过滤和修正浅层特征,提升模型的特征表达能力。实现方法上,本文采用类似 SRN 结构的从深到浅的单向特征融合网络进行多层特征学习。

为解决融合特征感受野固定、变化不够丰富的问题,在网络中加入了局部短接结构,将特征提取网络每一阶段的输出都与相邻阶段的输出进行整合,并利用可学习的方式设计局部短接的权重,从而得到对输入自适应的感受野大小。同时,为解决深层特征高倍上采样和下采样造成的失真问题,在融合过程中抛弃所有侧输出,仅采用单向结构进行特征融合。该设计能够有效地提升网络性能。

1 基于局部短接单向融合网络的骨架检测

1.1 基本思想和框架

整个检测网络 LoSUF 由特征提取网络和局部短接单向融合侧输出网络构成,如图 1 所示。其中,特征提取网络(又称为主干网络,backbone)包含结构复杂的深度卷积神经网络,其具有强大的非线性拟合能力,在整个检测网络中起到特征提取的作用。侧输出网络则由局部短接单元和特征

单向融合网络组成。在该网络中,多尺度特征先后经过局部短接整合和从深到浅的单向融合,以整合大感受野深层特征和浅层的细致纹理特征。

1.2 特征提取网络

本文选用 VGG-16^[18] 卷积神经网络作为特征提取网络,即主干网络 Backbone,如图 1 中(1)所示。VGG-16 共有 13 个卷积层和多个用于下采样的池化层,池化层将各卷积层分开为 5 个阶段,各阶段之间的下采样倍数为 2。本文选取各阶段最深层的卷积层输出作为侧输出网络的输入,并将其记为 Conv1~Conv5,其输出记为 $C_1 \sim C_5$,其中 $C_i = \{C_{i1}, C_{i2}, \dots, C_{id}\}$, $d = \text{depth}(C_i)$ 表示通道数, C_{ij} 表示特征矩阵。

1.3 局部短接单向融合网络

在侧输出网络中,不同尺度的特征融合过程分为 2 步:局部短接融合(local short-connection fusion)和从深到浅的单向融合(unidirectional fusion)。整体构成局部短接单向融合网络,其结构如图 1 所示。

局部短接融合网络通过将不同阶段的特征进行相邻跳接融合以调节各阶段感受野大小,如图 1 中(2-a)所示。本文中,局部短接融合网络由 5 个局部短接融合单元 lsfuse₁~lsfuse₅ 构成。其中第 1 个局部短接单元的输入是 C_1 和 C_2 、第 5 个局

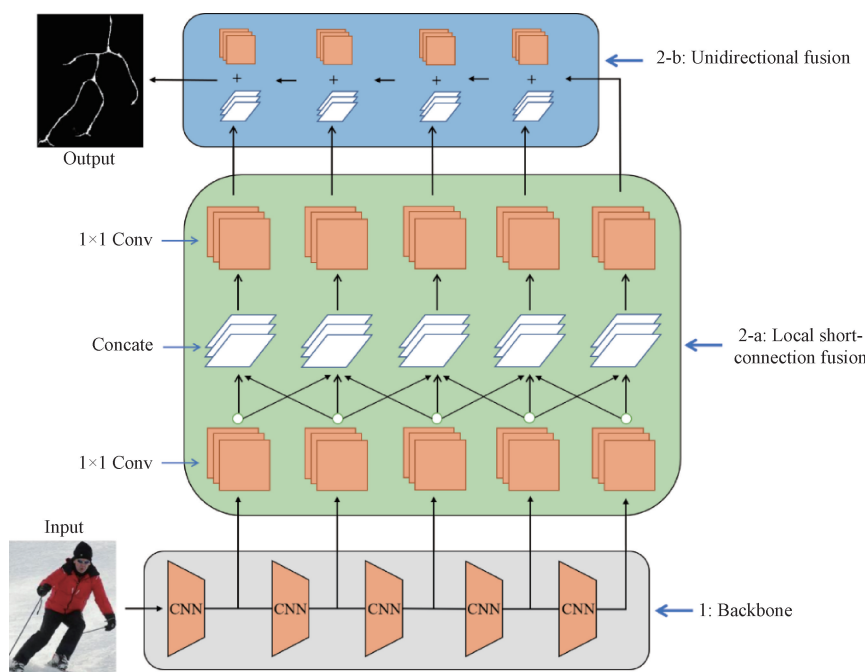


图 1 LoSUF 网络的整体结构图

Fig. 1 The architecture of LoSUF

部短接单元的输入是 C_4 和 C_5 , 其余局部短接单元 lsfuse_i 的输入为 C_{i-1} 、 C_i 和 C_{i+1} , 并记其局部短接融合输出 (local short-connection output) 为 $LSOP_1 \sim LSOP_5$ 。在每个局部短接单元中, 输入特征先通过逐点卷积 (1×1 Conv) 进行特征的降维和筛选, 再上/下采样至单元输出相应的尺度, 最后对各组处理后的特征进行堆叠, 共同经过逐点卷积进行加权和作为输出。其计算过程可表示为

$$LSOP_i = W_f(D(W_{i-1,i}^{(i)} C_{i-1}), W_{ii}^{(i)} C_i, U(W_{i+1,i}^{(i)} C_{i+1})). \quad (1)$$

其中: D (downsample) 表示下采样操作, U (upsample) 表示上采样操作。 W 表示各逐点卷积的系数矩阵。

单向融合过程按照从深到浅的过程进行, 其网络结构为由堆叠和逐点卷积所组成的单向融合单元的自底向上的嵌套, 如图 1 中 (2-b) 所示。最底层的单向融合单元以 $LSOP_5$ 作为输入, 记其单向融合输出 (unidirectional output) 为 UOP_5 ; 而其余的单向融合单元的输入为 $LSOP_i$ 和 UOP_{i+1} 。除最底层的单向融合单元, 其余的单向融合单元以残差连接的方式从深到浅地融合特征。本文将最上层的残差单元输出 UOP_1 传入 sigmoid 函数进行最终的网路输出。

1.4 网络训练

本文采用监督学习的方法进行 LoSUF 网络训练。将数据集划分为训练集和测试集, 假设整个数据集满足独立同分布。给定含有 N 对训练样本的训练集 $S = \{(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)}), \dots, (X^{(N)}, Y^{(N)})\}$, $X^{(k)} (k = 1, 2, \dots, N) = \{X_1^{(k)}, X_2^{(k)}, X_3^{(k)}\}$ 为输入的 3 通道 RGB 图像, $Y^{(k)} = \{y_{ij} = 1 \text{ or } 0\}$ 为对应图像的骨架真值标注, 表示为一个二值矩阵: 图像中被表示为骨架的像素点处的取值为 1, 反之为 0。在训练过程中, 输入图像 X 在经过网络结构后得到的输出为估计矩阵 $\hat{Y} = \{p_{ij} \mid p_{ij} = P\{y_{ij} = 1 \mid \Theta\}\}$, 表示在当前网络参数 Θ 下, 各个像素被预测为骨架的概率。

骨架检测的过程可看作对二维矩阵上的像素点进行二分类的过程, 因此一张训练图像的损失采用二维二值交叉熵 (2-dimension binary cross entropy) 计算:

$$L(Y, \hat{Y} \mid \Theta) = -\beta \sum_{y=1} \log(p) -$$

$$(1 - \beta) \sum_{y=0} \log(1 - p). \quad (2)$$

其中 $1 - \beta$ 为真值标注矩阵的均值, 即

$$\beta = 1 - \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W y_{ij}. \quad (3)$$

公式 (2) 的第 1 项为对骨架真值错分情况的惩罚: $-\log(p)$ 在 p 越接近于 0 时越大, 即当 y 为正例时, 概率预测值越小, 损失越大; 第 2 项同理, 为对背景错分情况的惩罚。加入系数 β 可有效解决真值标注矩阵中绝大部分像素为背景导致的正负例不均衡的问题。

在训练过程中, 使用训练集对模型进行端到端监督学习; 对损失通过链式法则和反向传播逐层计算权重参数的梯度, 对参数使用基于梯度下降的优化算法进行优化; 在进行一定次数的迭代之后, 保存优化参数并使用这一组参数在测试集上进行测试以评估模型。

2 实验结果与分析

2.1 实验数据与预处理

本文的实验使用到 4 种骨架检测领域的常用数据集: WH-SYMMAX^[19]、SK-SMALL (SK-506)^[15]、SK-LARGE^[20] 和 SYM-PASCAL^[9]。4 种数据集的样本图像和真值标注如图 2 所示。

WH-SYMMAX 数据集包含 328 张马匹样本的图像, 其中 228 张为训练数据, 100 张为测试数

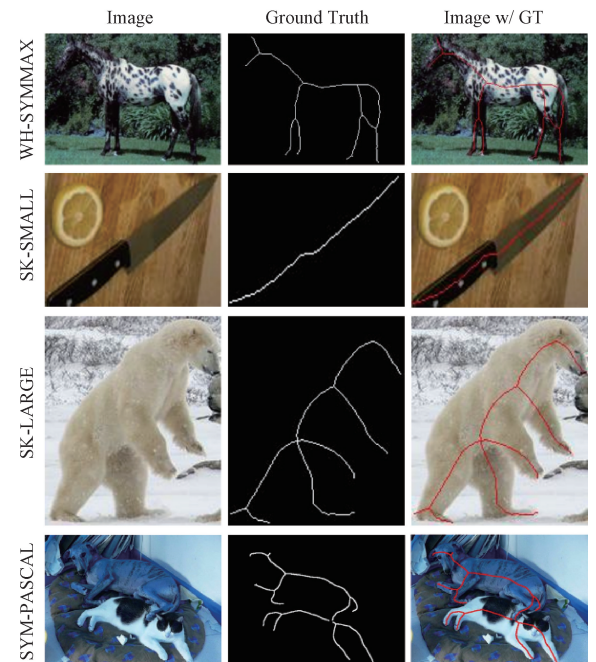


图 2 4 种数据集的样本图像和真值标注

Fig. 2 Examples of the four datasets

据。该数据集中只有一种前景样本类别,且其拍摄角度均为侧面,对模型性能的考验不大。

SK-LARGE 和 SK-SMALL 数据集均从 MSCOCO^[21] 图像实例分割数据集中裁剪提取得到,骨架的标注较为精细。SK-LARGE 包含 16 类目标样本的图片和骨架标注,每张图片上只有一个目标。SK-LARGE 共有 746 张训练数据和 745 张测试数据,本文在实验中将测试数据进一步划分为包含 245 张图片的验证集和包含 500 张图片的评估集。SK-SMALL 数据集同样包含 16 类样本,但是只有 300 张训练数据和 206 张测试数据。

SYM-PASCAL 数据集对 PASCAL-VOC2012 数据集^[22] 中的图像不经裁剪直接进行骨架的标注得到,这使得该数据集的图片间尺度差异较大,且在一张图片中可能会有多个目标。不仅如此,在 SYM-PASCAL 数据集中还出现了目标间的遮挡、背景遮挡目标、小尺度目标等情况,这对模型的性能提出了较为严苛的挑战。该数据集共包括 648 张训练图像和 787 张测试图像。

在网络训练时,为使模型具有更好的泛化能力,还对各个数据集的训练样本进行了数据的预处理,以扩充训练集。初始训练图像的预处理方式包括 3 类:1) 图像的尺度放缩(0.8 倍、1 倍、1.2 倍);2) 图像的旋转(0°、90°、180°、270°);3) 图像的翻转(水平、垂直)。扩充后的训练集在样本容量上扩大为原数据集的 36 倍。

2.2 超参数设置与模型初始化

训练过程使用 Adam 算法进行参数优化,使用到的超参数如下:初始学习率为 $5e-5$;权重衰减为 0.000 2; β 为(0.9, 0.999)。在训练过程中,每次前馈过程网络输入一张图片,每进行 10 次前馈过程进行一次参数更新,作为一组迭代。根据经验,在 SK-LARGE 和 SYM-PASCAL 数据集的训练中,共进行 50 000 次迭代,在进行到第 40 000 次迭代时将初始学习率缩小为初始值的 1/10;在 SK-SMALL 和 WH-SYMMAX 的训练中,总迭代次数为 25 000,在第 20 000 次迭代时将学习率缩小为初始值的 1/10。

使用在 ImageNet^[23] 进行预训练后的 VGG-16 网络权值作为特征提取网络的初值。对于侧输出网络中的权值,采用如下策略:将最上层的残差融合单元的逐点卷积层初始权重设为 0.1,并将其余单向融合单元的初值设定为 0.01,保证在训练开始时梯度在反向传递时不会出现梯度爆炸,亦

不会出现启动过慢的情况;将局部短接融合网络中的逐点卷积层采用 Xavier-Init^[24] 方法进行正态初始化;由于在该网络中不存在 Relu 函数,该方法在本文模型中比 Kaiming-Init^[25] 方法更好。

2.3 测试和评价指标

在模型测试阶段,对测试样本在骨架检测器上的输出概率矩阵 \hat{Y} 采用非极大值抑制(non-maximal suppression, NMS)进行骨架的细化,再将其结果由概率矩阵以阈值分割的方式转化为二值矩阵,最后提取宽度为 1 的像素的骨架。将 NMS 处理后的矩阵作为测试输出与真值标注矩阵进行比对,计算精确度(precision, P)、召回率(recall, R)和 F 度量(F -measure, 综合评价指标)。精确度和召回率的计算如下:

$$P = \frac{T_p}{T_p + F_p}, \quad (4)$$

$$R = \frac{T_p}{T_p + F_N}. \quad (5)$$

其中: T_p 表示一张输出矩阵中正确预测为正例(骨架)的像素数, F_p 表示错误预测为正例像素数, F_N 表示错误预测为反例(背景)的像素数。二值化阈值越高,精确度通常会越高,而召回率通常会降低。 F 度量表示为精确度和召回率的调和平均数:

$$F = \frac{2PR}{P + R}. \quad (6)$$

本文在模型评估时将对每张图遍历阈值,对每张图寻找最佳阈值获得单张输出的最优 F 度量,并对整个数据集中每张图片的最优 F 度量求取均值,作为该数据集的 F 度量评估结果。

2.4 实验结果及分析

将 LoSUF 模型在 4 种数据集中进行训练和测试。对比模型有基于多实例学习的模型 MIL^[8]、基于单一标注侧输出网络的模型 HED^[7]、SRN^[9] 和 LSN^[10], 和基于尺度相关标注的 FSDS^[15]。各深度学习模型均使用 VGG-16 作为特征提取网络。

本文使用 F 度量进行模型性能量化比对,结果如表 1 所示。

F 度量的对比结果显示,在 WH-SYMMAX^[19]、SK-SMALL (SK-506)^[15]、SK-LARGE^[20] 和 SYM-PASCAL^[9] 数据集上,本文模型 LoSUF 的性能均优于对比算法。在 SYM-

表 1 测试模型在 4 种数据集上的性能表现

Table 1 Performance on the four datasets

Methods	WH-SYMMAX	SK-SMALL	SK-LARGE	SYM-PASCAL
MIL ^[8]	0.365	0.392	0.353	0.174
HED ^[7]	0.732	0.541	0.487	0.369
FSDS ^[15]	0.769	0.623	0.633	0.418
SRN ^[9]	0.780	0.632	0.658	0.443
LSN ^[10]	0.797	0.633	0.668	0.425
LoSUF	0.830	0.663	0.683	0.432

PASCAL 数据集上,LoSUF 算法性能略低于 SRN,可能有如下 2 个原因:1)SYM-PASCAL 数据集含有大量的小目标和多目标数据,训练时需要重新设置优化算法的初始学习率等参数;2)SYM-PASCAL 数据集中有一部分样本的挑选并不合理,同时存在一部分标注并不准确的情况,本文算法和现有其他算法均无法较好地对方形目标(如门、屏幕等)的骨架检测。

从表 1 还可以看出,所有算法在 SYM-PASCAL 上的性能都比较低,而在 WH-SYMMAX 上的性能都比较高,这和数据集自身的图像复杂度有关。WH-SYMMAX 中图像目标全部为马匹,比较单一。而 SYM-PASCAL 中图像包含多目标和小目标,图像复杂度更高。

可视化对比结果如图 3 所示。其中第 1 列为样例输入图像,第 2、3、4 列分别为对比方法 HED、SRN、LSN 的结果展示,第 5 列为 LoSUF 的骨架检测结果,最后 1 列为标注真值 Ground Truth 骨架展示。第 1 行样例来自于 WH-SYMMAX 数据集,第 2 行样例来自 SK-SMALL 数据集,第 3 和第 4 行样例来自 SK-LARGE 数据集,第 5 和第 6 行样例来自 SYM-PASCAL 数据集。可以看出,相比对比算法,本文模型获得了更加准确的、更加接近 Ground Truth 的骨架检测结果。

2.5 消融实验

为展示 LoSUF 中子模块对模型性能的影响,以进一步验证本文模型的有效性,基于 SK-LARGE 数据集设计了 2 组对比实验。

1)保留/舍去侧输出 新模型从深到浅的特征融合结构与 SRN 结构的最大不同在于,SRN 结构中每个残差侧输出单元均有 2 个输出:向上传播的残差和送入融合单元的侧输出;而 LoSUF 仅保留了向上传播的输出。我们认为深层的、小尺度的特征图存在由于上采样造成的失真问题,直接对其进行侧输出融合会导致模型整体性能的弱

化。模型保留或舍去侧输出的对比实验结果如表 2 所示。

对比实验显示,去掉侧输出后,LoSUF 的性能取得了 5 个百分点的 *F* 度量性能提升。该结果验证了单向融合结构的有效性。

2)短接的范围选择 局部短接能够有效保持单向融合结构输入的尺度差异性,去掉较高上/下采样率的输入能保证特征融合的鲁棒性。只保留 2 倍上/下采样的局部短接(LoSUF)与其他采样率的对比实验结果如表 3 所示。

实验结果表明,本文最终选择的局部短接方式(2 倍)优于其他采样率(4 倍、8 倍和 16 倍)下的短接方式。

表 2 保留/舍去侧输出的 LoSUF 模型对比

Table 2 Comparison of LoSUF with/without side-output	
Methods	<i>F</i> 度量
LoSUF without side-output	0.683
LoSUF with side-output	0.632

表 3 不同短接范围选择下 LoSUF 模型对比

Table 3 Comparison of LoSUF models with various short-range choices	
Methods	<i>F</i> 度量
LoSUF	0.683
LoSUF 4x	0.670
LoSUF 8x	0.663
LoSUF 16x	0.659

3 结论

受神经科学研究中人眼辨认识事物的 glance-and-focus 模型的启发,提出一种新的骨架检测模型:基于侧输出连接的局部短接单向融合网络 LoSUF。该模型中,图像多尺度特征先后经过局部短接整合和从深到浅的单向融合,实现了大感受野深层特征和浅层的细致纹理特征的学习。实

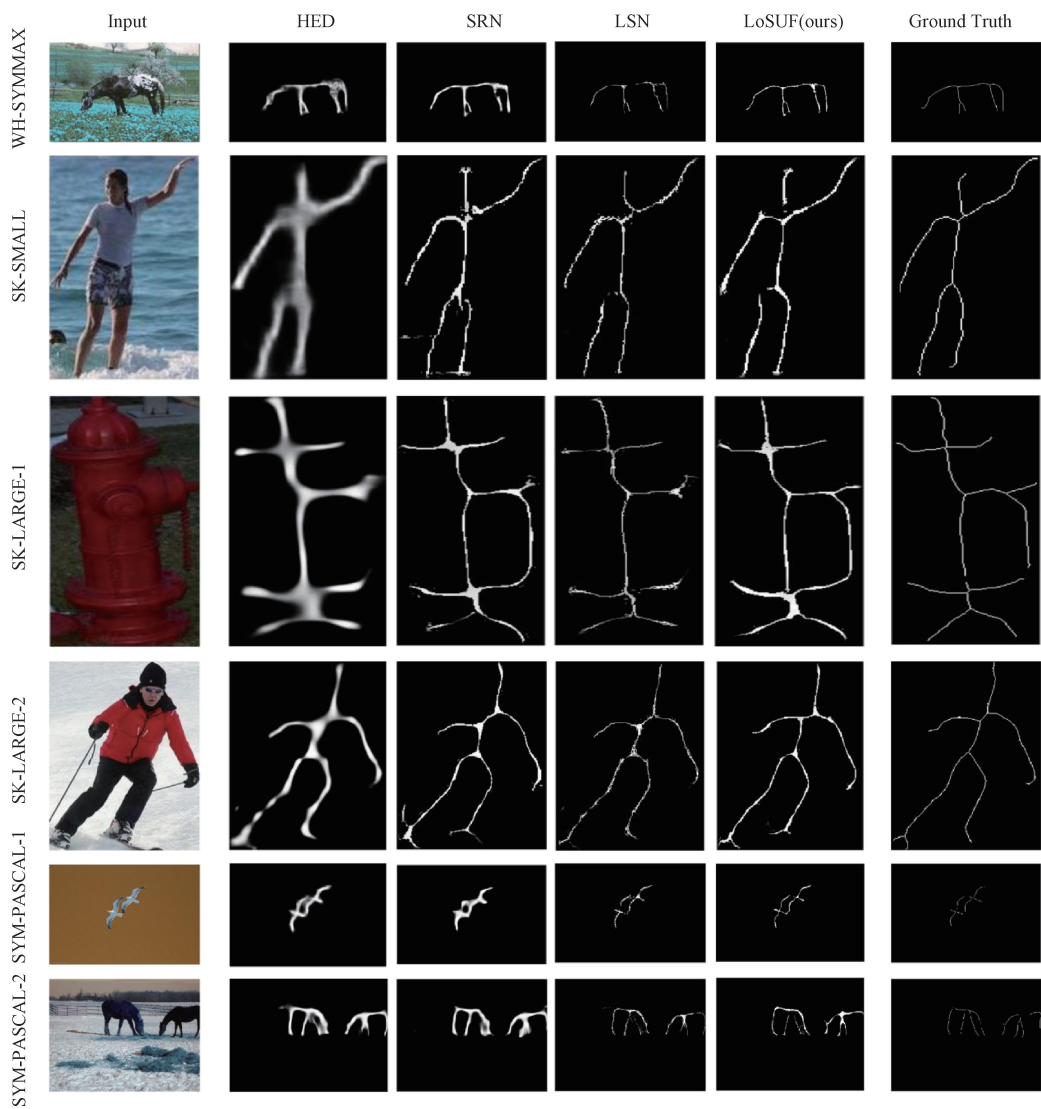


图 3 实验测得输出和真值标注对比

Fig. 3 Visualizations for experimental results

验结果表明,LoSUF 的性能优于现有的单一标注骨架检测模型。消融实验验证了所提出的局部短接和单向融合模块的有效性。

参考文献

[1] Wei S H, Ramakrishna V, Kanade T, et al. Convolutional pose machines [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27 - 30, 2016, Las Vegas, NV, USA. IEEE, 2016: 4724-4732.

[2] Teo C L, Fermüller C, Aloimonos Y. Detection and segmentation of 2D curved reflection symmetric structures [C] // 2015 IEEE International Conference on Computer Vision (ICCV). December 7 - 13, 2015, Santiago, Chile. IEEE, 2015: 1644-1652.

[3] Zhang Z, Shen W, Yao C, et al. Symmetry-based text line detection in natural scenes [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 7 - 12, 2015, Boston, MA, USA. IEEE, 2015: 2558-2567.

[4] Lee T, Fidler S, Dickinson S. Learning to combine mid-level cues for object proposal generation [C] // 2015 IEEE International Conference on Computer Vision (ICCV). December 7 - 13, 2015, Santiago, Chile. IEEE, 2015: 1680-1688.

[5] Lam L, Lee S W, Suen C Y. Thinning methodologies-a comprehensive survey [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(9): 869-885.

[6] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[7] Xie S N, Tu Z W. Holistically-nested edge detection [C] // 2015 IEEE International Conference on Computer Vision (ICCV). December 7 - 13, 2015, Santiago, Chile. IEEE, 2015: 1395-1403.

[8] Tsogkas S, Kokkinos I. Learning-based symmetry detection in natural images [M] // Computer Vision - ECCV 2012. Berlin,

- Heidelberg: Springer, 2012: 41-54.
- [9] Ke W, Chen J, Jiao J B, et al. SRN: side-output residual network for object symmetry detection in the wild[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). July 21-26, 2017, Honolulu, HI, USA. IEEE, 2017: 302-310.
- [10] Liu C, Ke W, Qin F, et al. Linear span network for object skeleton detection[M]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 136-151.
- [11] Hou Q B, Cheng M M, Hu X W, et al. Deeply supervised salient object detection with short connections[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(4): 815-828.
- [12] Wang Y K, Xu Y C, Tsogkas S, et al. DeepFlux for skeletons in the wild[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 15-20, 2019, Long Beach, CA, USA. IEEE, 2019: 5282-5291.
- [13] Qiao Y, Tian Y J, Liu Y, et al. Genetic feature fusion for object skeleton detection[J]. Security and Communication Networks, 2021, 2021: 1-9.
- [14] Zhao K, Shen W, Gao S H, et al. Hi-fi: hierarchical feature integration for skeleton detection[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. July 13 - 19, 2018. Stockholm, Sweden. California: International Joint Conferences on Artificial Intelligence Organization, 2018: 1191-1197.
- [15] Shen W, Zhao K, Jiang Y, et al. Object skeleton extraction in natural images by fusing scale-associated deep side outputs[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA. IEEE, 2016: 222-230.
- [16] Blake R, Logothetis N K. Visual competition[J]. Nature Reviews Neuroscience, 2002, 3(1): 13-21.
- [17] Wang Y L, Lv K C, Huang R, et al. Glance and focus: a dynamic approach to reducing spatial redundancy in image classification[EB/OL]. arXiv: 2010.05300. (2020-10-11) [2021-06-01]. <https://arxiv.org/abs/2010.05300>.
- [18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. arXiv: 1409.1556. (2015-04-10) [2021-06-01]. <https://arxiv.org/abs/1409.1556>.
- [19] Shen W, Bai X, Hu Z H, et al. Multiple instance subspace learning via partial random projection tree for local reflection symmetry in natural images[J]. Pattern Recognition, 2016, 52: 306-316.
- [20] Shen W, Zhao K, Jiang Y, et al. DeepSkeleton: learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images[J]. IEEE Transactions on Image Processing, 2017, 26(11): 5298-5311.
- [21] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Computer Vision - ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.
- [22] Everingham M, Gool L, Williams C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.
- [23] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 20-25, 2009. Miami, FL, USA. IEEE, 2009: 248-255.
- [24] Hubara I, Courbariaux M, Soudry D, et al. Quantized neural networks: training neural networks with low precision weights and activations[EB/OL]. arXiv: 1609.07061. (2016-09-22) [2021-06-01]. <https://arxiv.org/abs/1609.07061>.
- [25] He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//2015 IEEE International Conference on Computer Vision (ICCV). December 7-13, 2015, Santiago, Chile. IEEE, 2015: 1026-1034.