基于二次生成对抗的人体姿态估计

张显坤，张荣芬，刘宇红*
贵州大学大数据与信息工程学院大数据与智能技术重点实验室，贵州 贵阳 550025

摘要 针对人体姿态估计中因肢体、环境复杂性导致的估计结果不精确问题，提出了一种基于二次生成对抗的人体姿态估计方法，通过两个阶段对堆叠沙漏网络(SHN)进行生成对抗训练。首先将 SHN 作为第一个生成对抗网络模型的判别器，通过在线对抗数据加强训练，以提升 SHN 的估计性能；然后将 SHN 作为第二个生成对抗网络模型的生成器，将肢体几何约束作为判别器，通过第二次对抗训练再一次提升 SHN 的估计性能，得到最终的 SHN。在公开数据集 LSP 和 MPII 上对本方法进行测试，结果表明，该方法能有效提升 SHN 的估计精确度。

关键词 人体姿态估计；生成对抗网络；模型再训练；肢体几何约束
中图分类号 O436 **文献标志码** A **doi**：10.3788/LOP57.201509

Human Pose Estimation Based on Secondary Generation Adversary

Zhang Xiankun, Zhang Rongfen, Liu Yuhong*
Key Laboratory of Big Data and Intelligent Technology, College of Big Data and Information Engineering, Guizhou University, Guiyang, Guizhou 550025, *China*

Abstract Aiming at the problem of inaccurate estimation results caused by the complexity of limbs and environment in human pose estimation, a human pose estimation method based on secondary generation adversary is proposed in this work. The stacked hourglass network (SHN) is trained for generation adversary through two stages. First, the SHN is used as a discriminator in the first generation adversarial network model, and the on-line adversarial data is used to strengthen training to improve the estimation performance of the SHN. Then, the SHN acts as a generator in the second generation adversarial network model, and the limb geometric constraints are used as the discriminator. The estimation performance of the SHN is improved again through the second adversarial training, and the final SHN is obtained. The proposed method is tested on the public data sets LSP and MPII, and the results show that it can effectively improve the estimation accuracy of the SHN.

Key words human pose estimation; generation adversarial network; model retraining; limb geometric constraints
OCIS codes 150.4065; 150.1135; 100.4996

1 引 言

对图像进行识别并估计出其中的二维(2D)人体姿态，是机器视觉研究领域的基础性工作。如人体跟踪、动作认知、人机交互应用以及三维(3D)人体姿态研究中，都需要精确地估计人体姿态。由于图像中肢体的复杂性、摄像机角度等客观原因，导致图像中的人体存在不同程度的扭曲和遮挡，使姿态估计变成一项极具挑战性的任务。对于机器视觉而言，标准的深度神经网络(DNN)观察并学习肢体结构需要进行大量训练，在面对复杂的人体姿态时，估计精度不高，是人体姿态估计中的难点。

Andriluka 等[1-2]通过图像结构模型或随机部件推理模型进行姿态估计，但均采用人工提取特征的方式，缺乏有效的功能表征，不能充分利用图像信息，受制于图像的视角、外观及几何模糊性。DNN 具有自主提取特征以及理解上下文特征关联的能力，使基于 DNN 的姿态估计得到了深入研究。Toshev 等[3]提出的 DeepPose 方法将 DNN 应用到人体姿态估计中，采用多阶段回归思路设计卷积神

收稿日期：2020-01-14；**修回日期**：2020-03-01；**录用日期**：2020-03-09
基金项目：贵州省科技计划项目(黔科合平台人才[2016]5707)
* **E-mail**：1693623574@qq.com

经网络(CNN),直接回归人体骨骼关节点的二维坐标,但该方法缺乏关节间的结构信息,对于多尺度的姿态估计泛化性能较差。Newell[4]提出的堆叠沙漏网络(SHN)通过级联沙漏网络结构和中继监督训练,以热图(heatmap)检测的方式,学习人体关节点的图像特征及关节点间的结构信息,推理出整个图像的检测结果。但当人体图像的肢体存在重叠或遮挡时,SHN 对人体的姿态估计不合理。

Luo 等[5]提出了生成对抗网络(GAN),通过生成器和判别器之间的对抗训练得到想要的模型。刘坤等[6]使用半监督 GAN 实现 X 光图像的分类;杨晓莉等[7]使用 GAN 在动态平衡中实现图像融合;张清博等[8]基于改进的 GAN 改善了水下激光图像的去噪和照明。Wang 等[9]提出的快速卷积网络(A-Fast-RCNN)使用 GAN 生成变形输入以达到数据增强的目的,从而进行鲁棒性更强的目标检测。Peng 等[10]将姿态估计网络作为判别器,并新建一个神经网络作为生成器,引入奖励、惩罚机制,与姿态估计网络进行对抗训练,实现了在线数据增强,提升了姿态估计网络的泛化能力和精确度;Chou 等[11]将 SHN 作为生成器进行姿态估计,并新建一个神经网络作为判别器,生成器和判别器进行对抗训练,提高了 SHN 的姿态估计性能;Chen 等[12]建立了两个神经网络作为判别器,分别对多任务生成器得到的姿态关节 heatmap 置信度和关节点定位偏差进行判别并反馈给生成器,达到对抗训练的目的。

本文以 SHN 为姿态估计网络(Target Net),融合了两种基于 GAN 的姿态估计模型,并添加了肢体几何约束,将姿态估计网络经过两次生成对抗训练,以提升 SHN 的姿态估计精确度。

2 模型分析

本方法的结构如图 1 所示,Target Net 在第一个 GAN(GAN1)中作为判别器(D_1),新建神经网络作为生成器(G_1)。生成器生成的增强图像与随机增强的图像输出到判别器,并加入奖励、惩罚权衡机制,使生成器进行在线数据增强,实现对抗训练,以提升 SHN 姿态估计模型的性能;该姿态估计模型在第二个 GAN(GAN2)中再次进行生成对抗训练,其在 GAN2 中作为生成器(G_2),新建一个 SHN 作为判别器(D_2),并在 D_2 中加入肢体几何约束(body geometric constraints)[12]。将生成器生成的姿态估计 heatmap 特征图和带有正确 heatmap 标记的数据输入判别器,并加入边界参数权衡机制,通过第二次生成对抗训练得到最终的 SHN 姿态估计网络。

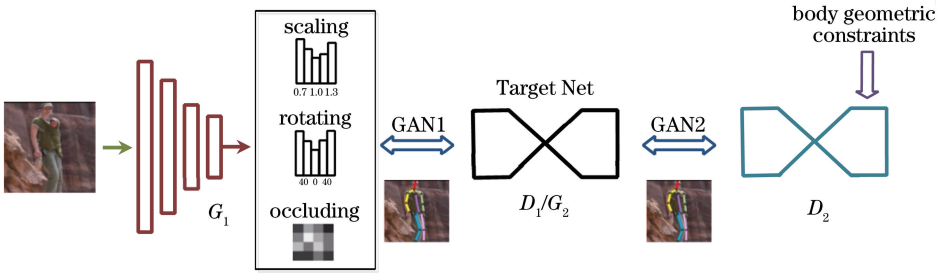

图 1 本方法的结构示意图

Fig. 1 Structure schematic diagram of our method

2.1 堆叠沙漏模型

模型中的目标优化网络:SHN[4]通过串联多个沙漏(hourglass)网络估计人体姿态,用多阶段分辨率 heatmap 学习姿态关节点的坐标,对图像中每个像素对应的概率值进行估算,像素点位置越靠近关节点,其对应的概率值就越接近 1,反之则越接近 0;将训练后 heatmap 形式的 feature map 结果映射到原图像,得到对应关节点的坐标,从而估计人体姿态,如图 2 所示。

SHN 结构前后对称,形似沙漏,如图 3 所示。模型中的卷积模块采用残差模块(residual module)[13],通过残差连接避免了网络层数过多产生的网络过拟合等退化问题,并利用 1×1 卷积核减少参数数量。其中,每个方框均为一个残差模块,在沙漏网络的前半部通过卷积及降采样(max pooling)操作得到分辨率逐渐降低的 heatmap,并向沙漏网络后半部传递,逐步扩大感受野(receptive field);在沙漏网络中心得到低分辨率以及最大感受野的特征图;同时,沙漏网络上面的分支结构(shortcut)经逐步特征提取向沙漏网络后半部传递分辨率特征,然后将 shortcut 上的特征与主干路的低分辨率特征进行最近邻上采样(nearest neighbor

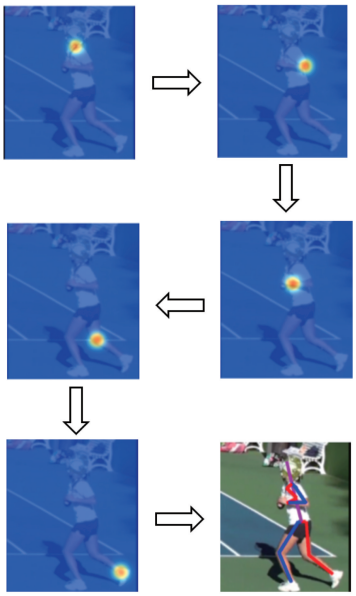

图 2 heatmap 示意图

Fig. 2 Schematic diagram of heatmap

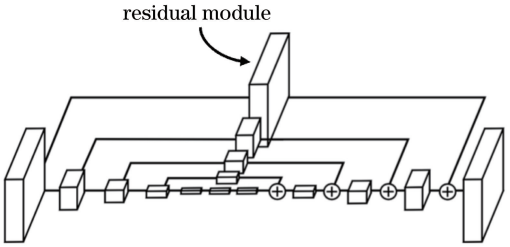

图 3 沙漏网络结构

Fig. 3 Structure of hourglass network

upsampling)[14]，逐步恢复出一个高分辨率的 heatmap。

将多个沙漏网络级联，每个沙漏网络作为一个姿态估计阶段，在每个阶段建立残差连接，避免网络退化。其输出的混合特征经过一个 1×1 全卷积网络分支输出混合特征和 heatmap，两者合并后传输到下一个阶段，实现多阶段姿态估计，从而得到 SHN，如图 4 所示。

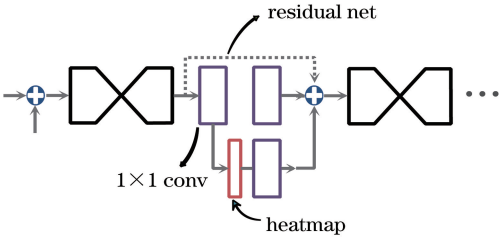

图 4 SHN 的级联结构图

Fig. 4 Cascade structure diagram of SHN

SHN 通过级联结构的沙漏网络和多阶段中间监督对整个图像的初始特征和检测结果进行评估，通过隐式学习关节点特征之间的结构关系，在最后一个沙漏网络输出最终的姿态估计结果。每个沙漏网络输出的估计结果包含 M 个关节点，可采用均方误差损失 $\mathscr{L}_{\mathrm{MSE}}$ 对预测的 heatmap($\widetilde{C}$)与正确标记 heatmap(C)进行对比，可表示为

$$\mathscr{L}_{\mathrm{MSE}}=\sum_{j=1}^{M}(C_j-\widetilde{C}_j)^2, \tag{1}$$

式中，j 为第 j 个关节点。SHN 根据姿态估计得到所有输出 heatmap 的 $\mathscr{L}_{\mathrm{MSE}}$ 损失添加中间监督，从而在多个阶段调整姿态的预测精确度，如图 5 所示。

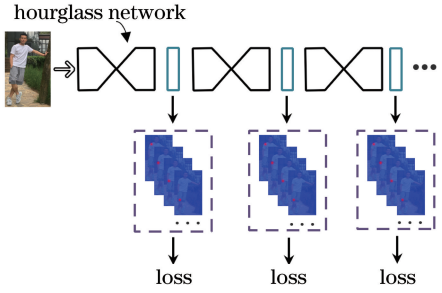

图 5 中间监督结构

Fig. 5 Structure of intermediate supervision

2.2 生成对抗模型

GAN 由 Luo 等[5]提出，可同时训练生成器(G)和判别器(D)两个网络模型。G 接收随机噪声 z 生成数据分布 $G(z)$，D 用于评估并判别输入的数据分布，如数据 x 来自真实的数据分布 $\bar{x}$ 还是由 G 生成的数据分布 $G(z)$。训练过程中，G 的目标是生成接近真实数据的分布以欺骗 D，而 D 的目标是区分 G 生成的数据分布与真实的数据分布，并生成判别输出 $D(\cdot)$传回 G，输入判别器的数据越逼近真实数据，$D(\cdot)$越大。两者由多层感知机建立，通过反向传播(BP)机制，形成动态二元 minimax 博弈过程，优化函数可表示为

$$\min_G\max_D V(D,G)=E_{\bar{x}\in p_{\mathrm{data}}(\bar{x})}[\log_e D(\bar{x})]+E_{z\in p_z(z)}\{\log_e\{1-D[G(z)]\}\}, \tag{2}$$

式中，E 为期望，$p_{\mathrm{data}}(\bar{x})$为真实数据的分布范围，$p_z(z)$为噪声的分布范围，通过最大化 $\log_e D(\bar{x})$和 $\log_e\{1-D[G(z)]\}$的期望值优化判别器 D，通过最小化 $\log_e\{1-D[G(z)]\}$的期望值优化生成器 G。

2.2.1 生成对抗模型 1

实验中第一个生成对抗模型 GAN1 采用在线数据增强对抗网络，以 SHN 作为判别器 D_1，从生成器 G_1 生成的增强数据中采样，在有限数据集下提升其姿态估计性能。G_1 使用 SHN 分支结构上的桥特征作为 D_1 的输入，生成使 D_1 损失增加的 hard 数据，优化函数可表示为

$$\max_{\theta_{G_1}} E_{x\in\Omega}\left\{E_{\substack{\tau_r\in\Gamma\\ \tau_a\in G_1(x,\theta_{D_1})}}\mathscr{L}_{\mathrm{MSE}}\{D[\tau_a(x),y]\}-\mathscr{L}_{\mathrm{MSE}}\{D[\tau_r(x),y]\}\right\},\tag{3}$$

式中，θ_{G_1} 和 θ_{D_1} 分别为 G_1 和 D_1 中的一系列参数及变量，$D(\cdot)$为判别器 D_1 以不同增强方式和正确标注图像作为输入时的输出，Ω 为训练图像集，x 为输入 G_1 的图像，y 为正确标注的图像，Γ 为随机增强空间，$G_1(x,\theta_{D_1})$为增强网络 G_1、x 和 D_1 的函数。数据增强分为对抗增强 $\tau_a(x)$和随机增强 $\tau_r(x)$，G_1 使对抗增强产生的损失大于随机增强产生的损失，从而让判别器获得更 hard 的数据。D_1 在 GAN1 中一方面根据(1)式评估 G_1 生成样本的质量，另一方面从对抗增强的样本中训练网络，可表示为

$$\min_{\theta_{D_1}} E_{x\in\Omega}\left\{E_{\tau_a\in G(x,\theta_{D_1})}\mathscr{L}_{\mathrm{MSE}}\{D[\tau_a(x),y]\}\right\}。\tag{4}$$

G_1 网络有两种对抗增强训练方式：对抗缩放及旋转(ASR)和对抗多特征图遮挡(AHO)。在 ASR 增强中，选取 m 种缩放尺度和 n 种旋转尺度，将一个 batch 图像输入 G_1。计算 $m\times n$ 种增强方式的均方误差损失，并对损失进行归一化处理，生成缩放和旋转分别对应的两个归一化分布 P^{s} 和 P^{r} 作为 ground truth；同时，G_1 会预测出 $\tilde{P}^{\mathrm{s}}$ 和 $\tilde{P}^{\mathrm{r}}$ 两种分布，通过损失函数对 G_1 进行训练，可表示为

$$\mathscr{L}_{\mathrm{ASR}}=\sum_{i=1}^{m}P_i^{\mathrm{s}}\log_e\frac{P_i^{\mathrm{s}}}{\tilde{P}_i^{\mathrm{s}}}+\sum_{i=1}^{n}P_i^{\mathrm{r}}\log_e\frac{P_i^{\mathrm{r}}}{\tilde{P}_i^{\mathrm{r}}}。\tag{5}$$

训练结果通过 m 和 n 种高斯分布进行采样后生成增强样本，如图 6 所示。

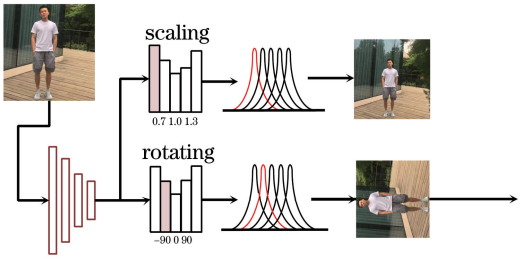

图 6 ASR 流程

Fig. 6 Procedure of ASR

在 AHO 增强中，G_1 以 D_1 的上分支结构特征作为输入，在最低像素(尺寸为 4 pixel×4 pixel)的特征图下生成遮挡，将图像分为 $w\times h$ 个网格($w=h=4$)，在像素增大过程中，统计一个 batch 图像中每个关键点落在遮挡部分的概率，得到最大概率。以一系列最大概率生成的遮挡概率分布特征 $P_{i,j}^{\circ}$ 作为 ground truth，其中，$P_{i,j}^{\circ}$ 为第 i 行、第 j 列的分布特征，如图 7 所示，通过损失函数对 G_1 进行训练，可表示为

$$\mathscr{L}_{\mathrm{AHO}}=\sum_{i=1}^{h}\sum_{j=1}^{w}P_{i,j}^{\circ}\log_e\frac{P_{i,j}^{\circ}}{\tilde{P}_{i,j}^{\circ}}。\tag{6}$$

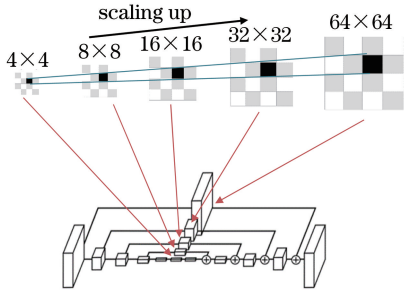

图 7 AHO 流程

Fig. 7 Procedure of AHO

对 G_1、D_1 两个网络进行联合对抗训练时，为了将 D_1 的训练状态反向传播到 G_1，同时避免 GAN 训练时由标签缺失、无法收敛等原因导致的训练崩溃问题，引入了奖励、惩罚机制。计算由对抗增强 τ_a 和随机增强 τ_r 造成的 D_1 损失，若 $\mathscr{L}\{D[\tau_a(x),y]\}-\mathscr{L}\{D[\tau_r(x),y]\}$增大，证明对抗增强有效。增大该采样的概率作为奖励，同时减少对其他采样的概率，可表示为

$$P_m=\tilde{P}_m+\alpha\tilde{P}_m,\ P_n=\tilde{P}_n-\frac{\alpha\tilde{P}_n}{k-1},\ \forall\, m\neq n,\tag{7}$$

式中，P 为由采样获得的 ground truth 分布，$\tilde{P}$ 为预测的分布，α 为超参数，k 为 ASR 或 AHO 的增强方式数量，m、n 分别为第 m、n 种采样。(7)式以增加 P 的方式增加对该采样的概率(P_m)，相对减少对其他采样的概率(P_n)。通过表 1 中 G_1 和 D_1 网络的批次图像分类训练和表 2 中的单图像训练，由交替迭代训练的方式实现第一次生成对抗训练，在有限数据集下提升 SHN 姿态估计网络的性能。其中，$\check{x}$ 为对抗增强后的采样数据，$\hat{x}$ 为随机增强后的数据，$\check{\mathscr{L}}_{\mathrm{MSE}}$、$\hat{\mathscr{L}}_{\mathrm{MSE}}$ 分别为由 $\check{x}$、$\hat{x}$ 计算的均方误差，$\check{\mathscr{L}}$、$\hat{\mathscr{L}}$分别为由 $\check{x}$、$\hat{x}$ 计算得到的 ASR 和 AHO 损失。

表 1 批次图像训练流程

Table 1 Training process of batch images

Input: a mini-batch training image set X
1. X is randomly and equally divided into X_1、X_2、X_3;
2. Train D_1 using X_1;
3. Train G_1、D_1 using X_2 with table 2 on ASR;
4. Train G_1、D_1 using X_3 with table 2 on AHO.

2.2.2 生成对抗模型 2

实验中的第二个生成对抗模型GAN2采用对

表 2　单图像训练流程

Table 2　Training process of single image

Input: image x
1. Get shortcut features from D_1;
2. Get distribution P from shortcut features in G_1;
3. Sample an adversarial augmentation data $\check{x}$ from P;
4. Compute the loss of D_1: $\check{\mathscr{L}}_{\mathrm{MSE}}$ with $\check{x}$;
5. Random augment x to get $\hat{x}$;
6. Compute the loss of D_1: $\hat{\mathscr{L}}_{\mathrm{MSE}}$ with $\hat{x}$;
7. Compare $\check{\mathscr{L}}$ and $\hat{\mathscr{L}}$ with formula (5) and formula (6) to update G_1;
8. Update D_1.

称堆叠沙漏生成对抗网络，以第一次训练得到的 SHN 作为生成器 G_2，以另一个 SHN 作为判别器 D_2，第二次对抗训练以目标网络 G_2 作为姿态估计网络。生成器 G_2 的目标是从 RGB(Red, Green, Blue)图像中学习并生成关节点 heatmap 热图的映射。引入判别器 D_2 后，将 G_2 生成的 heatmap 与带有正确标签的 heatmap 之间的误差反向传播到 G_2，使 G_2 在学习图像中人体特征及上下文依赖关系的同时，能生成更合理的人体姿态。

对于 G_2，采用 $\mathscr{L}_{\mathrm{MSE}}$ 对堆叠的沙漏网络进行中间监督，并定义对抗损失函数进行训练。假设 G_2 包含 N 个堆叠的沙漏网络，将原始沙漏网络输出的混合特征 heatmap 从多维矩阵中提取出来，输出 M 个单关节点 heatmap，每个 heatmap 都在第 l 个 ground truth 的关节点定位下达到高斯峰值，其 $\mathscr{L}_{\mathrm{MSE}}$ 可表示为

$$\mathscr{L}_{\mathrm{MSE}}=\sum_{i=k}^{N}\sum_{j=l}^{M}(C_{kl}-\widetilde{C}_{kl})^2, \tag{8}$$

式中，C_{kl} 为第 k 个沙漏网络第 l 个关节点定位的正确标记 heatmap，$\widetilde{C}_{kl}$ 为预测的 heatmap。对抗损失函数 $\mathscr{L}_{\mathrm{adv}}$ 可表示为

$$\mathscr{L}_{\mathrm{adv}}=\sum_{l=1}^{M}[\widetilde{C}_l-D(\widetilde{C}_l,x)]^2, \tag{9}$$

式中，$\widetilde{C}_l$ 为 G_2 最后一个沙漏网络预测的第 l 个关节 heatmap，$D(\cdot)$为判别器 D_2 的输出，x 为输入 G_2 的图像，可根据(9)式计算 G_2 预测的 heatmap 和 D_2 重建的 heatmap 的损失。生成器 G_2 的总损失函数可表示为

$$\mathscr{L}_{\mathrm{G}}=\mathscr{L}_{\mathrm{MSE}}+\lambda_{\mathrm{G}}\mathscr{L}_{\mathrm{adv}}, \tag{10}$$

式中，λ_{G} 为对抗损失权重控制的超参数。以 G_2 预测的 heatmap 和带有正确标签的 heatmap 作为 D_2 的输入，然后重建对应的两组 heatmap，如图 8 所示。定义损失函数 $\mathscr{L}_{\mathrm{real}}$、$\mathscr{L}_{\mathrm{fake}}$，输入包含正确标签的 heatmap($C_j$)时，$D_2$ 重建的 heatmap 应尽量缩小两个 heatmap 之间的误差 $\mathscr{L}_{\mathrm{real}}$；输入包含预测 heatmap($\widetilde{C}_j$)时，判别器重建的 heatmap 应尽量扩大两者间误差 $\mathscr{L}_{\mathrm{fake}}$，可表示为

$$\begin{aligned}\mathscr{L}_{\mathrm{real}}&=\sum_{j=1}^{M}[C_j-D(C_j,x)]^2,\\ \mathscr{L}_{\mathrm{fake}}&=\sum_{j=1}^{M}[\widetilde{C}_j-D(\widetilde{C}_j,x)]^2,\\ \mathscr{L}_{\mathrm{D}}&=\mathscr{L}_{\mathrm{real}}-k_t\mathscr{L}_{\mathrm{fake}},\end{aligned} \tag{11}$$

式中，k_t 为权衡参数，t 为第 t 次迭代训练，$\mathscr{L}_{\mathrm{D}}$ 为 D_2 的损失函数，D_2 通过 $\mathscr{L}_{\mathrm{D}}$ 计算输入 heatmap 和重建 heatmap 的像素，从而对 D_2 进行优化。

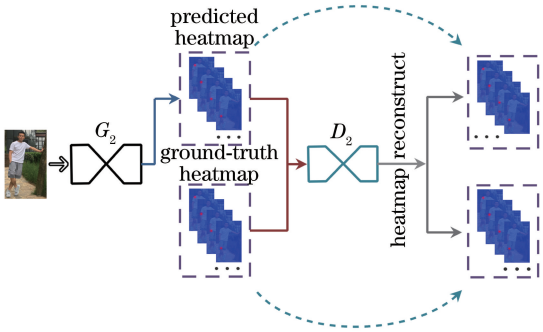

图 8　heatmap 重建

Fig. 8　Reconstruction of heatmap

为了使 G_2 网络在第二次 GAN 训练中生成更合理的姿态估计结果，在 G_2 网络中添加肢体几何约束。在 SHN 姿态估计结果中，可将人体姿态理解为 M 个关节点的定位及连接形成的肢体结构，要使预测的人体姿态更接近真实的姿态，则关节点定位应更接近真实的关节点定位，定义 $p_{\widetilde{C}}^{i}$ 为

$$p_{\widetilde{C}}^{i}=\begin{cases}1 & d_i<\delta\\ 0 & d_i\geqslant\delta\end{cases}, \tag{12}$$

式中，δ 为阈值参数，d_i 为预测的第 i 个关节点与重建 heatmap 关节点间的归一化距离，$\boldsymbol{p}_{\widetilde{C}}$ 为由 1 和 0 组成的 M 维向量。为了使 G_2 网络更好地学习逼近 D_2 重建的 $D(\widetilde{C},x)$热图，将 $\boldsymbol{p}_{\widetilde{C}}$ 的 2 范数引入 $\mathscr{L}_{\mathrm{D}}$ 损失，可表示为

$$\mathscr{L}'_{D}=\mathscr{L}_{\mathrm{real}}+\|\boldsymbol{p}_{\widetilde{C}}\|_2-k_t\mathscr{L}_{\mathrm{fake}}。 \tag{13}$$

为了避免对抗训练中 G_2 或 D_2 太好或太坏而导致的不稳定问题，采用文献[15]中的边界参数均衡思想，用参数 k_t 权衡 G_2 和 D_2 间的对抗训练。在第 $t+1$ 次迭代中，参数 k_t 的更新可表示为

$$k_{t+1}=k_t+\lambda_k[\xi(\mathscr{L}_{\text{real}}+\|\boldsymbol{p}_{\widetilde{C}}\|_2)-\mathscr{L}_{\text{fake}}], \tag{14}$$

式中，ξ、λ_k 为权衡自变量，k_t 为 $\mathscr{L}_{\text{fake}}$ 的相关权重参数。当 G_2 生成的姿态估计结果能使 $\mathscr{L}_{\text{adv}}$ 足够小，表明 G_2 的表现比 D_2 更好，$\mathscr{L}_{\text{fake}}$ 比 $\xi(\mathscr{L}_{\text{real}}+\|\boldsymbol{p}_{\widetilde{C}}\|_2)$更小，第 $t+1$ 次迭代中 k_t 就会增大为 k_{t+1}；反之，当 D_2 的表现比 G_2 更好时，k_t 就会减小，避免了对抗训练中 $\mathscr{L}'_D$ 收敛过慢或过快导致的崩塌问题。通过表 3 中交替迭代的训练实现第二次生成对抗训练，得到最终的 SHN 姿态估计网络。

表 3　第二次生成对抗的训练流程

Table 3　Training process of the secondary generation adversary

Input: image x; ground truth heatmap C
1. D_2 reconstructs heatmap: $D(C,x)$;
2. Compute $\mathscr{L}_{\text{real}}$ with formula (11);
3. G_2 generates predictive heatmap: $\widetilde{C}=G(x)$;
4. Compute $\mathscr{L}_{\text{MSE}}$ with formula (8);
5. D_2 reconstructs heatmap: $D(\widetilde{C},x)$;
6. Compute $\boldsymbol{p}_{\widetilde{C}}$;
7. Compute $\mathscr{L}_{\text{fake}}$、$\mathscr{L}'_D$ with formula (11)、formula (12);
8. Update D_2;
9. Compute $\mathscr{L}_{\text{adv}}$、$\mathscr{L}_G$ with formula (9)、formula (10);
10. Update G_2.

3　实验及结果分析

3.1　数据集

通过 LSP[16] 和 MPII 数据集[17] 对本方法进行测试，LSP 数据集包括 11000 张和 1000 张从运动场景中截取的人体姿态图像，分别用于训练和测试，每张图像有 14 个注释的关节点；MPII 数据集包括 30000 张和 10000 张从 YouTube 网站视频中截取的人体姿态图像，分别用于训练和测试，每张图像有 16 个注释的关节点。

3.2　实验设置

因为部分训练图像中存在多个人体姿态，因此需对部分样本进行预处理，以图像中主要人体的髋部为中心进行裁剪，同时将图像的分辨率统一为 256 pixel×256 pixel。

在生成对抗训练中，堆叠沙漏网络为 4 个沙漏网络级联的结构，以步长为 2 的 7×7 卷积层开始，输入分辨率为 256 pixel×256 pixel 的图像，后接沙漏网络中的残差模块均为 BN-ReLU-conv(1×1)-BN-ReLU-conv(3×3)-BN-ReLU-conv(1×1)的 bottleneck 结构[18]，其中，BN 为批归一化，ReLU 为线性整流函数，括号内为卷积层的尺寸。两个 1×1 卷积层分别用于降维和升维，可在保持 bottleneck 输入/输出维度不变的情况下减少卷积核的参数数量。同时结合最大池化层在网络前半部不断降低图像分辨率，并在后半部通过上采样恢复分辨率，最终输出分辨率为 64 pixel×64 pixel 的 heatmap。

在第一次生成对抗训练中，使用 RMSProp[19] 优化器优化网络，训练时首先将学习率设置为 2.5×10^{-4} 对 D_1 单独训练，然后保持 D_1 不变，使用相同的学习率训练 G_1 网络的 ASR 和 AHO 两种图像增强方式，最后将学习率降低为 5×10^{-5} 对 G_1 和 D_1 进行联合对抗训练；在第二次生成对抗训练中，同样使用 RMSProp 优化器，设置学习率为 2.5×10^{-4} 对 G_2 和 D_2 进行第二次联合对抗训练。

3.3　评估指标

采用关键点正确估计百分比(PCK)评价本方法对 LSP 数据集[20] 的估计结果，PCK 以躯干直径作为归一化参考，计算检测的关键点 $\widetilde{\boldsymbol{y}}_i$ 与对应 ground truth 关键点 $\boldsymbol{y}_i$ 间的归一化距离小于设定阈值的比例，可表示为

$$\frac{\|\boldsymbol{y}_i-\widetilde{\boldsymbol{y}}_i\|}{\|\boldsymbol{y}_{\text{lhip}}-\boldsymbol{y}_{\text{rshoulder}}\|}\leqslant r, \tag{15}$$

式中，$\boldsymbol{y}_{\text{lhip}}$ 和 $\boldsymbol{y}_{\text{rshoulder}}$ 分别为左髋部和右肩部的 ground truth 坐标，r 为大小在 0 到 1 之间的阈值。对 MPII 数据集采用正确关键点的头部归一化概率(PCKh)[17] 作为实验评估指标，与 PCK 不同的是，PCKh 以头部长度作为归一化参考。

3.4　实验结果分析

表 4 和表 5 分别为在数据集 LSP 和 MPII 的测试集上，阈值 r 分别设置为 0.2 和 0.5 时，本方法与其他方法在 7 个主要关节点(头部 head、肩部 shoulder、肘部 elbow、腕部 wrist、髋部 hip、膝部 knee、踝部 ankle)的 PCK 和 PCKh(腕部等对称关节点取两者均值)。其中，文献[4]使用 8 个沙漏网络堆叠的 SHN，该网络训练后的姿态估计精度较高，在 LSP 和 MPII 测试集的平均 PCK 和 PCKh 分别为 93.0%和 90.9%；文献[10]和文献[11]分别为本方法二次生成对抗模型中 GAN1 和 GAN2 未改进前的模型；相比其他方法，本方法将两种对抗训练模式融合并添加肢体几何约束后，使 SHN 的姿态估计性能得到了一定的提升，在 LSP 和 MPII 测试集的平均 PCK 和 PCKh 分别为 94.8%和 92.2%。

表 4 不同方法在 LSP 数据集的 PCK

Table 4 PCK of different methods in LSP data set unit: %

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Ref. [21]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Ref. [4]	98.2	94.0	91.2	87.2	93.5	94.5	92.6	93.0
Ref. [12]	98.5	94.0	89.8	87.5	93.9	94.1	93.0	93.1
Ref. [10]	98.6	95.3	**92.8**	90.0	94.8	95.3	94.5	94.5
Ref. [11]	98.2	94.9	92.2	89.5	94.2	95.0	94.1	94.0
Ours	**98.8**	**95.7**	92.6	**90.8**	**94.8**	**96.1**	**95.0**	**94.8**

表 5 不同方法在 MPII 数据集的 PCKh

Table 5 PCKh of different methods in the MPII data set unit: %

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Ref. [21]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Ref. [4]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Ref. [12]	**98.6**	96.2	90.9	86.7	89.8	87.0	83.2	90.6
Ref. [10]	98.1	96.6	92.5	88.4	90.7	87.7	83.5	91.5
Ref. [11]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Ours	98.4	**97.1**	**93.4**	**88.7**	**92.5**	**90.3**	**85.2**	**92.2**

将文献[4]、文献[10]、文献[11]以及本方法在 MPII 测试集上测试的一种 heatmap 估计结果进行可视化比较，从四种 SHN 输出的混合特征 heatmap 中提取出单特征关节点 heatmap，并对腕部、肘部、肩部和膝部的关节点 heatmap 进行可视化，如图 9 所示。可以发现，人体的腕部、肘部、肩部和膝部均

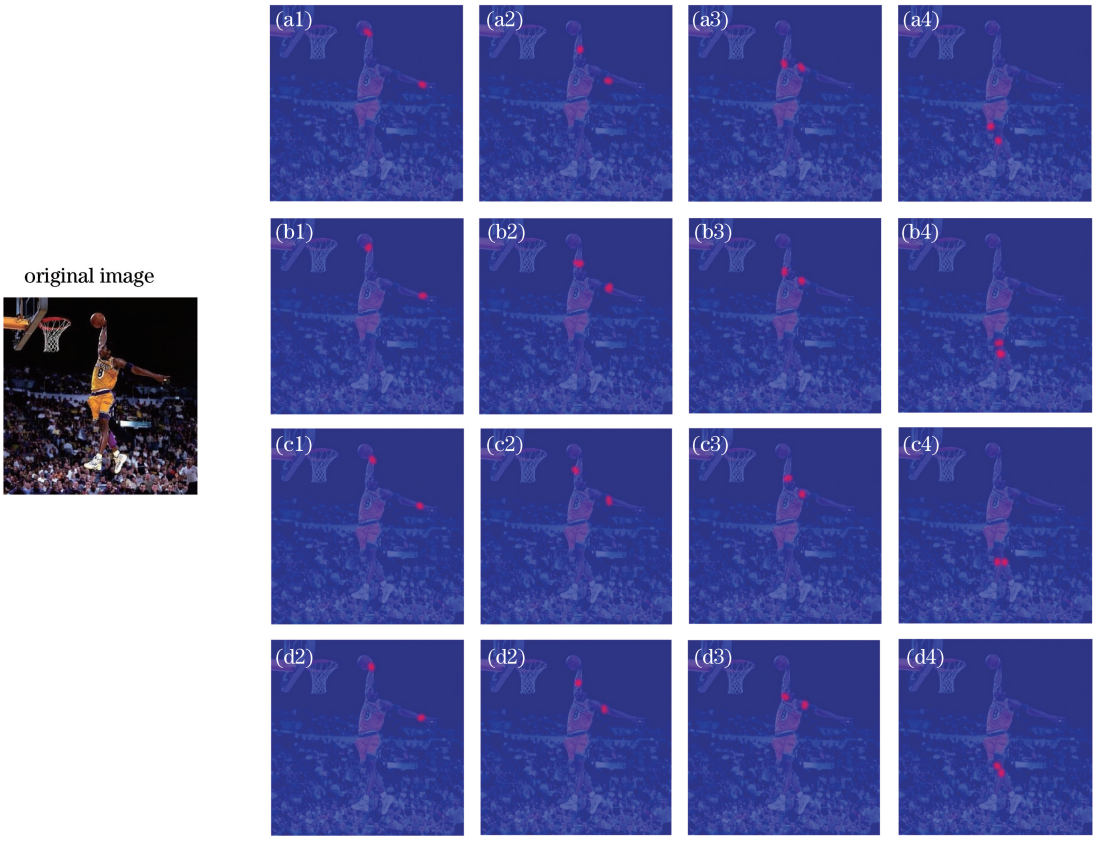

图 9 不同方法得到的 heatmaps。(a)文献[4]；(b)文献[10]；(c)文献[11]；(d)本方法

Fig. 9 Heatmaps obtained by different methods. (a) Ref. [4]; (b) Ref. [10]; (c) Ref. [11]; (d) ours

存在不同程度的遮挡。将图 9 中四种方法的关节点估计归一化误差(Normalized error)进行对比，如图 10 所示。可以发现，在面对复杂肢体结构时，本方法中 SHN 输出的 heatmap 关节点定位误差更小，估计精确性更高。

表 6 为本方法与文献[11]中方法在同一 GPU 上将等损失函数收敛到同一比例标准时模型的迭代次数、对 MPII 测试集的平均处理时间、模型浮点运算次数(GFLOPs)以及模型参数数量。可以发现，相比文献[11]中的方法，本方法的模型使用效率有一定程度的降低，但复杂度有所增加。

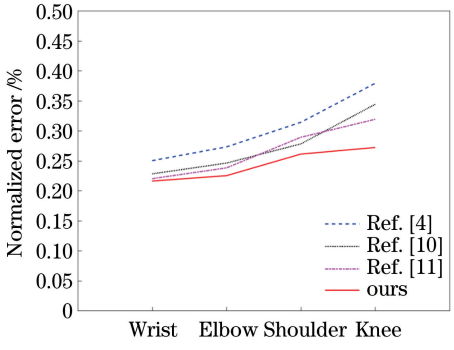

图 10 关节点估计误差对比

Fig. 10 Comparison of joint estimation errors

表 6 模型使用效率的对比

Table 6 Comparison of model efficiency

Method	Convergence iteration times	Average processing time /s	GFLOPs /(10^9 times)	Number of parameters /10^7
Ref. [11]	19500	0.48	10.820	5.495
Ours	26600	0.73	13.702	6.738

4 结 论

以 SHN 作为优化目标，将 SHN 与两种不同思路的 GAN 进行融合，并将肢体几何约束加入训练模型中。实验结果表明，相比其他的人体姿态估计网络，本方法经两次训练得到的人体姿态估计精度有一定程度的提高，在 LSP 和 MPII 测试集的平均 PCK 和 PCKh 均有所提升，估计误差有所降低。但如何在保持估计精度的前提下降低模型的大小和复杂度，提升模型的使用效率，还需进一步研究。

参 考 文 献

[1] Andriluka M, Roth S, Schiele B. Pictorial structures revisited: people detection and articulated pose estimation[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE, 2009: 1014-1021.

[2] Ladick L, Torr P H S, Zisserman A. Human pose estimation using a joint pixel-wise and part-wise formulation[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE, 2013: 3578-3585.

[3] Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks[J]. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1653-1660.

[4] Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation[M]//Leibe B, Matas J, Sebe N, et al. Computer Vision-ECCV 2016. Lecture Notes in Computer Science. Cham: Springer, 2016, 9912: 483-499.

[5] Luo J Y, Xu Y, Tang C W, et al. Learning inverse mapping by autoencoder based generative adversarial nets[M]// Liu D, Xie S, Li Y, et al. Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science. Cham: Springer, 2017, 10635: 207-216.

[6] Liu K, Wang D, Rong M X. X-ray image classification algorithm based on semi-supervised generative adversarial networks[J]. Acta Optica Sinica, 2019, 39(8): 0810003.
刘坤, 王典, 荣梦学. 基于半监督生成对抗网络 X 光图像分类算法[J]. 光学学报, 2019, 39(8): 0810003.

[7] Yang X L, Lin S Z, Lu X F, et al. Multimodal image fusion based on generative adversarial networks[J]. Laser & Optoelectronics Progress, 2019, 56(16): 161004.
杨晓莉, 蔺素珍, 禄晓飞, 等. 基于生成对抗网络的多模态图像融合[J]. 激光与光电子学进展, 2019, 56(16): 161004.

[8] Zhang Q B, Zhang X H, Han H W. Backscattered light repairing method for underwater laser image based on improved generative adversarial network[J]. Laser & Optoelectronics Progress, 2019, 56(4): 041004.
张清博, 张晓晖, 韩宏伟. 基于改进生成对抗网络的

水下激光图像后向散射光修复方法[J]. 激光与光电子学进展, 2019, 56(4): 041004.

[9] Wang X L, Shrivastava A, Gupta A. A-fast-RCNN: hard positive generation via adversary for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 3039-3048.

[10] Peng X, Tang Z Q, Yang F, et al. Jointly optimize data augmentation and network training: adversarial data augmentation in human pose estimation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 2226-2234.

[11] Chou C J, Chien J T, Chen H T. Self adversarial training for human pose estimation[EB/OL]. [2020-01-02]. https://arxiv.org/abs/1707.02439.

[12] Chen Y, Shen C H, Wei X S, et al. Adversarial PoseNet: a structure-aware convolutional network for human pose estimation[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 1221-1230.

[13] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[EB/OL]. [2019-12-28]. http://arxiv.org/abs/1602.07261.

[14] Tompson J, Jain A, Lecun Y, et al. Joint training of a convolutional network and a graphical model for human pose estimation[EB/OL]. [2020-01-01]. https://arxiv.org/abs/1406.2984.

[15] Berthelot D, Schumm T, Metz L. BEGAN: boundary equilibrium generative adversarial networks [EB/OL]. [2019-12-30]. https://www.arxiv.org/abs/1703.10717.

[16] Johnson S, Everingham M. Clustered pose and nonlinear appearance models for human pose estimation[C]//Procedings of the British Machine Vision Conference, BMVC 2010, August 31-September 3, 2010, Aberystwyth, UK. UK: BMVA, 2010: 1-11.

[17] Andriluka M, Pishchulin L, Gehler P, et al. 2D human pose estimation: new benchmark and state of the art analysis[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 3686-3693.

[18] Sze V, Chen Y H, Yang T J, et al. Efficient processing of deep neural networks: a tutorial and survey[J]. Proceedings of the IEEE, 2017, 105(12): 2295-2329.

[19] Tieleman T, Hinton G. Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude[J]. COURSERA: Neural Networks for Machine Learning, 2012, 4(2): 26-31.

[20] Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts[C]//CVPR 2011, June 20-25, 2011, Providence, RI, USA. New York: IEEE, 2011: 1385-1392.

[21] 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. New York: IEEE, 2016: 4724-4732.