E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP18;TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)12-3870-14

论文引用格式:Jiang L, Liu Z C, Xiong Y, Wu W and Li K G. 2025. Adaptive ground-truth heatmap generation for bottom-up human pose estimation. Journal of Image and Graphics, 30(12):3870-3883(江玲, 刘卓程, 熊源, 吴威, 李凯歌. 2025. 面向自下而上人体姿态估计的自适应真值热力图生成方法. 中国图象图形学报, 30(12):3870-3883)[DOI:10.11834/jig.240615]

面向自下而上人体姿态估计的自适应真值热力图生成方法

江玲[1],刘卓程[2],熊源[2],吴威[2],李凯歌[2*]

1. 安徽理工大学计算机科学与工程学院,淮南　232001; 2. 北京航空航天大学虚拟现实技术与系统全国重点实验室,北京　100191

摘　要:**目的**　热力图回归方法因能够提供丰富的空间信息,在人体姿态估计领域受到广泛关注。然而,由于传统真值热力图通常由固定标准差的2D高斯核覆盖标注点位置生成,当人体尺度变化较大时,固定的高斯核覆盖范围可能与关键点的实际语义区域不匹配,导致模型对关键点定位的模糊性和语义不确定性。对此,提出面向自下而上人体姿态估计的自适应真值热力图生成方法。**方法**　首先设计一种自适应真值热力图生成模块,通过学习图像中关键点的固有尺度信息以及近邻关键点之间的几何关系生成自适应尺度因子,为图像定制尺度自适应的真值热力图。另外,由于现有方法使用的热力图损失函数未能有效捕捉局部结构的相关性,导致其对关键点位置偏差不敏感。为此,提出局部概率一致性损失函数,通过在热力图的局部区域上计算结构相似性,提升模型对局部结构的学习和理解,同时引入动态权重来平衡样本的贡献,进一步引导模型优化方向,提高模型鲁棒性。**结果**　在两个公开数据集MS COCO(Microsoft common objects in context)和CrowdPose上进行实验评估,实验结果表明所提方法相较对比工作,关键点检测平均准确率分别提高1.6%与6.5%,达到72.1%和74.1%,验证了所提方法的有效性。此外,所提方法在拥挤场景的CrowPose数据集上也能带来显著的性能提升,这进一步表明其能够有效缓解复杂场景中的人体尺度变化带来的问题。同时消融实验验证了所提方法的有效性。**结论**　提出的面向自下而上人体姿态估计的自适应真值热力图生成方法,通过学习图像中关键点的固有尺度信息以及近邻关键点之间的几何关系生成自适应热力图作为真值,结合局部概率一致性损失函数来处理图像中尺度变化问题,有效提高了人体姿态估计准确率。

关键词:人体姿态估计;自适应尺度;自下而上;热力图回归;动态权重

Adaptive ground-truth heatmap generation for bottom-up human pose estimation

Jiang Ling[1], Liu Zhuocheng[2], Xiong Yuan[2], Wu Wei[2], Li Kaige[2*]

1. School of Computer Science and Engineering, Anhui University of Science & Technology, Huainan　232001, China;

2. State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing　100191, China

Abstract: **Objective** Human pose estimation aims to locate skeletal keypoints of individuals in a given image. As a fundamental task in computer vision, human pose estimation has wide applications in human activity recognition, person re-identification, pose tracking, and related fields. Two main approaches for human pose estimation are available: top-

收稿日期:2024-10-21;**修回日期**:2025-05-22;**预印本日期**:2025-05-29

*通信作者:李凯歌　likg@mail.sysu.edu.cn

基金项目:国家自然科学基金项目(62272018);海南省交通科技项目(HNJTT-KXC-2024-3-22-02)

Supported by: National Natural Science Foundation of China (62272018); Transportation Science and Technology Program of Hainan, China (HNJTT-KXC-2024-3-22-02)

down and bottom-up. Top-down methods first detect human bodies in the image, crop out each person, and then estimate the keypoint coordinates. While effective, these methods perform poorly in cases of occlusion, and their computation cost increases with the number of people in the image. In contrast, bottom-up methods detect all identity-independent keypoints simultaneously and then group them into individual poses. These methods are typically lightweight and fast but must handle varying human scales. Bottom-up human pose estimation methods commonly use 2D Gaussian kernels to generate keypoint heatmaps as regression targets because they provide rich spatial information. However, conventional approaches apply Gaussian kernels with a fixed variance across all keypoints, resulting in uniform heatmap structures. This uniformity is problematic given the existing scale variability in bottom-up methods. On the one hand, different keypoints cover different pixel areas in images, and using large Gaussian kernels may introduce semantic ambiguity, particularly for small joints. On the other hand, differences in keypoint scale imply different levels of annotation uncertainty, which the heatmap variance should ideally reflect. The variance of the Gaussian kernel represents uncertainty; thus, it should be proportional to the scale and ambiguity associated with each keypoint. Aiming to address these issues, an adaptive heatmap generation network (AHGNet) for bottom-up human pose estimation is proposed. AHGNet estimates the appropriate radius of the Gaussian kernel for each keypoint by integrating inherent scale information and geometric relationships. Through formula derivation, the relationship between the radius and the Gaussian kernel variance is established, enabling the creation of customized, scale-adaptive ground-truth heatmaps. This approach improves localization accuracy by effectively aligning the heatmap structure with the spatial characteristics of each keypoint. **Method** First, an adaptive heatmap generation module is introduced. This module combines the inherent scale information from image features and the geometric relationship between adjacent keypoints to constrain the coverage areas of kernels. Keypoint scale is defined by semantic coverage areas in images. However, in the actual scene, accurately allowing pixel areas to occupy keypoints is almost impossible, and determining the potential relationship between Gaussian kernels and coverage areas is difficult. Interestingly, the areas occupied by keypoints are found to be related to geometric distance from adjacent keypoints. Therefore, an adaptive heatmap generation module is introduced to generate kernel scale maps of keypoints. This module combine the geometric relationship between adjacent keypoints and inherent scale information from image features. Second, local probabilistic consistency loss is presented to define the distance between the predicted and ground truth heatmaps globally and locally. Most methods based on heatmap regression use L_2 loss for supervised learning. However, as the loss function for heatmap regression, L_2 loss assumes that each pixel point is independent and overlooks the local structural correlation, making it difficult to describe the probability distribution of heatmaps. A keypoint heatmap is a probability distribution that describes pixels belonging to a certain joint. Thus, KL Divergence must be added to describe local probability consistency. Moreover, samples with large prediction errors are difficult to predict; thus, the weight of difficult samples should be increased. Similarly, the weight of easily detected samples should be reduced. Therefore, the dynamic weight is added to balance the contribution of different samples. Inspired by focal loss, which allows the model to actively focus on hard-to-detect samples, this paper utilizes dynamic weights to reduce the contribution of easily detected samples while enhancing the contribution of hard-to-detect samples. **Result** HrHRNet is used as the baseline to establish AHGNet for bottom-up human pose estimation. The model is tested on two public datasets: MS COCO and CrowdPose. Experimental results reveal that AHGNet surpasses HrHRNet in terms of average precision (AP), achieving 72.1% AP and 74.1% AP on COCO test-dev and CrowdPose dataset, providing improvements of +1.6% AP and +6.5% AP, respectively. In addition, the substantial improvement on the CrowdPose dataset with crowded scenes indicates that AHGNet helps alleviate the problem of human scale changes in complex crowded scenes. Simultaneously, the ablation experiments verified the effectiveness of the proposed method. **Conclusion** AHGNet leverages geometric features between adjacent keypoints and inherent scale information within the image to generate adaptive heatmaps as groundtruth. This network further employs a local probability consistency loss function to address the challenges posed by various human scales, effectively improving the accuracy of bottom-up human pose estimation. AHGNet provides a new paradigm for optimizing supervision signals in bottom-up pose estimation. By dynamically adjusting the Gaussian kernel scale and enforcing local probability constraints, it effectively reduces multi-scale ambiguity in complex scenarios.

Key words: human pose estimation; adaptive scale; bottom-up; heatmap regression; dynamic weight

0 引 言

人体姿态估计旨在从给定图像中定位人体的骨骼关键点。人体姿态估计作为计算机视觉的基本任务之一，广泛应用于安防监控、虚拟现实等领域（喻莉 等，2024；贾伟 等，2024）。一般而言，人体姿态估计方法可分为两种：自上而下和自下而上。自上而下的方法（Chen 等，2018；Newell 等，2016；Yang 等，2017）首先利用人体检测器裁剪出图像中的人物，然后回归关键点的坐标。这种方法容易受到人体检测器结果影响，效率与性能在拥挤场景中通常会严重下降（Chen 等，2018）。自下而上的方法（Li 等，2020；Qi 等，2019；Luo 等，2021）直接检测所有身份独立的关键点，然后对它们进行分组。这种方法更加轻量快捷，但需要处理多尺度人体带来的误检漏检问题（Cheng 等，2020）。

自下而上的方法通常使用关键点坐标或热力图作为回归目标。然而人体关节很难通过像素位置准确定义，并且人工精确标注成本很高。此外，标注点附近的区域也可能带有人体关节语义，直接将它们标记为负样本，可能会对网络的训练产生负面影响。与坐标回归（Sekii，2018；Sun 等，2018）相比，基于热力图的方法（Cheng 等，2020；Sun 等，2019；Newell 等，2016；Cai 等，2020）通过以关键点为中心的高斯核生成概率图进行监督学习。热力图中像素值表示当前像素点属于该关键点的概率。越接近标注点，像素值越大，使得网络可以快速学习到收敛方向。此外，热力图不仅保留了关键点中的空间关系，还捕捉了前景（关键点）与背景的对比关系来引导网络学习。因此热力图回归受到人体姿态估计领域的广泛关注。

传统热力图通常使用具有固定标准差的 2D 高斯核作为姿态估计网络的回归目标（Newell 等，2016）。这意味着不同尺度的关键点使用相同结构的高斯核进行监督学习。当尺度变化较大时，这种方法容易导致模型在关键点定位上出现模糊性和语义不确定性。一方面，关键点在图像中占据的语义区域应与高斯核的覆盖范围一致。不同尺度的关键点在图像中占据不同大小的语义区域，使用相同结构的高斯核真值，可能导致高斯核的覆盖范围与关键点的实际语义区域不匹配。对于小尺度关键点，高斯核覆盖面积过大会导致语义歧义。如图 1 所示，其中，图 1（a）是原始图像，图 1（b1）（c1）是局部放大图，图 1（b2）（c2）是关键点语义区域，图 1（b3）（c3）是标准真值热力图，图 1（b4）（c4）是自适应真值热力图。可以看出，图 1（b1）—（b4）肘部关键点的固有尺度大于腕部，本文方法增大了右肘的高斯核区域。图 1（c1）—（c4）由于图像中两个关键点距离近，导致左脚踝处像素面积小。本文方法减小了左踝的高斯核区域。另一方面，关键点在图像中占据的语义区域应与高斯核的标准差相关。关键点的语义区域面积越大，人工标注偏差越大，也就意味着对应高斯分布的不确定性越大，即高斯核标准差越大。因此关键点的高斯标准差应该与其尺度成正比。综合以上分析，本文方法通过探索关键点的语义区域，调整高斯核的标准差，优化高斯核覆盖面积，避免语义歧义。

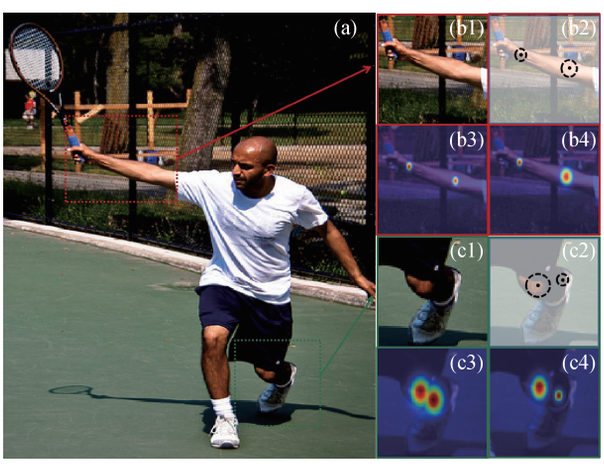


图 1 不同尺度关键点的高斯核覆盖面积对比

Fig. 1 Comparisons of Gaussian kernel coverage area for keypoints at different scales

综合以上分析，本文提出面向自下而上人体姿态估计的自适应真值热力图生成方法，设计了一个自适应热力图生成网络模型（adaptive heatmap generation network，AHGNet）。具体而言，本文的主要贡献如下：1）针对图像中由于透视引起的多尺度问题，设计自适应真值热力图生成模块（adaptive ground-truth heatmap generation module，AHGM）。通过结合关键点固有的尺度信息以及近邻关键点之间的几何关系来生成关键点尺度因子，调整高斯核标准差来定制真值热力图作为姿态估计网络的回归目标。2）提出局部概率一致性损失函数（local probabilistic consistency loss，LPCLoss）。通过在热力图的

局部区域上计算结构相似性，提升模型对局部结构的学习和理解，增强损失函数对于位置偏差的敏感性，同时引入动态权重平衡样本的贡献，进一步引导模型优化方向。3）在两个公开数据集 MS COCO（Microsoft common objects in context）和 CrowdPose 上进行实验评估。实验结果表明，AHGNet 相较对比工作，关键点检测平均准确率分别提高 1.5% 与 6.4%。此外，AHGNet 在拥挤场景的 CrowPose 数据集上带来的显著性能提升，表明其能够有效缓解复杂场景中的人体尺度变化。

1 相关工作

相较于坐标回归（Sun 等，2018；Toshev 和 Szegedy，2014；Li 等，2021），热力图具有丰富的空间信息广泛用于语义特征点定位。真值热力图通过在关键点上放置 2D 高斯核构建，像素值表示属于关键点的概率。该方法实现简便且可达到像素级精度。然而，当前方法通常使用统一结构的高斯核覆盖所有关键点。这种方法适用于将人物调整至统一尺度的自上而下方法（Cheng 等，2020；Sun 等，2019；Newell 等，2016；Cai 等，2020）。但在自下而上的方法（Li 等，2020；Qi 等，2019；Luo 等，2021）中，网络需处理整幅图像的透视引起的尺度变化问题。为此，部分研究利用多分辨率特征融合增强网络对尺度变化的适应性，通过跨层连接保留高分辨率特征。然而，固定高斯核仍限制了关键点语义区域的精确建模，尤其在小尺度关键点检测中。大多数方法通过多尺度感知网络或改进损失函数减轻尺度变化的影响。

1.1 多尺度感知网络

一些方法（Ding 等，2022；Jiang 等，2022a；Mao，2021）通过设计多尺度网络结构来对不同尺度的关键点进行检测。堆叠的沙漏网络（Newell 等，2016）通过设计从高到低、从低到高的对称结构来获取和捕捉关键点之间的空间关系。金字塔残差网络（Yang 等，2017）通过不同比例的亚采样获取多尺度的特征。多尺度结构感知神经网络（Ke 等，2018）对现有的深度卷积沙漏模型进行改进，有效地利用多尺度特征定位关键点。Li 等人（2019b）提出一种多阶段姿态估计网络来融合各阶段的特征信息。Qi 等人（2019）提出一种用于人体姿态估计的空间捷径网络，使多尺度信息在空间上共享。Ke 等人（2018）提出深度高分辨率表示学习网络 HRNet（high-resolution network），将多尺度特征由传统的串行转变为并行。在此基础上进行了改进与扩展，HrHRNet（HigherHRNet）（Cheng 等，2020）进一步优化特征融合效率，通过并行分支维持高分辨率表示以提升小尺度关键点的检测精度，它从由两个 3 × 3 的卷积组成的主干开始，将分辨率降低到 1/4，并进一步使用反卷积进行多尺度融合，然后结合使用关联嵌入作为聚类方法，成为目前最先进的自下而上姿态估计方法。

1.2 损失函数

一些方法在损失函数中添加尺度约束（Kreiss 等，2019；Li 等，2019a；Li 等，2020）来处理多尺度人体问题。PifPaf（part intensity field and part association field）（Kreiss 等，2019）提出部件强度场和部件关联场，并添加表示关节大小的组件进行监督学习，同时在损失函数中融入尺度信息，使网络回归不同关键点的多尺度响应区域。CrowdPose（Li 等，2019a）在损失函数中加入衰减因子，以加强目标关节的响应，抑制干扰关节的响应。SimplePose（Li 等，2020）提出使用 Focal Loss（Lin 等，2020）来处理难检测和易检测样本的不平衡问题。然而，这些方法均忽略了传统真值热力图使用统一尺度高斯核对关键点定位的影响。

2 AHGNet 网络框架

针对图像中由于透视引起的多尺度问题，本文提出一种面向自下而上人体姿态估计的自适应真值热力图生成方法，设计了一种自适应真值热力图生成网络（AHGNet），如图 2 所示。AHGNet 由 3 部分组成：姿态估计网络、自适应真值热力图生成模块和多尺度热力图融合模块。姿态估计网络选用自下而上的人体姿态估计先进方法 HrHRNet（Cheng 等，2020）。自适应真值热力图生成模块以相邻关键点间的向量为输入，学习其几何关系，并结合图像特征的固有尺度信息生成尺度因子，从而定制真值热力图作为姿态估计网络的回归目标。多尺度热力图融合模块重组多尺度预测热力图并恢复至原始图像分辨率，通过提取预测热力图的局部最大值获取关键点坐标，分组后输出完整的人体姿态估计结果。

姿态估计网络是一种基于热力图回归的自下而

上的姿态估计方法。姿态估计网络将图像作为输入，从图像特征中预测多分辨率关键点热力图。本文主要基于HrHRNet（Cheng等，2020）展开实验，该模型是一个带有反卷积模块的HRNet（Sun等，2019）。它通过两个3×3卷积将分辨率降低到1/4，最后通过一个反卷积模块，输出两种分辨率（$H_{1/4}$和$H_{1/2}$）关键点热力图。在测试过程中，两种分辨率的预测热力图融合以形成最终预测。

自适应真值热力图生成模块使用标注的关键点图及其到相邻关键点的向量作为输入以学习关键点之间的几何关系，结合从姿态估计网络中提取的图像特征（$\boldsymbol{F}_{\text{img}}$）来学习关键点固有尺度信息，生成自适应尺度因子，作用于不同关键点高斯核来定制真值热力图，作为姿态估计网络的回归目标。

多尺度热力图融合模块通过对预测的不同分辨率热力图进行重组，实现分辨率恢复。本文独立训练多尺度融合模块来恢复热力图分辨率，将不同分辨率的预测热力图结合通过反卷积恢复到原始图像分辨率，得到最终的预测热力图，最后通过提取局部最大值定位关键点的位置坐标。

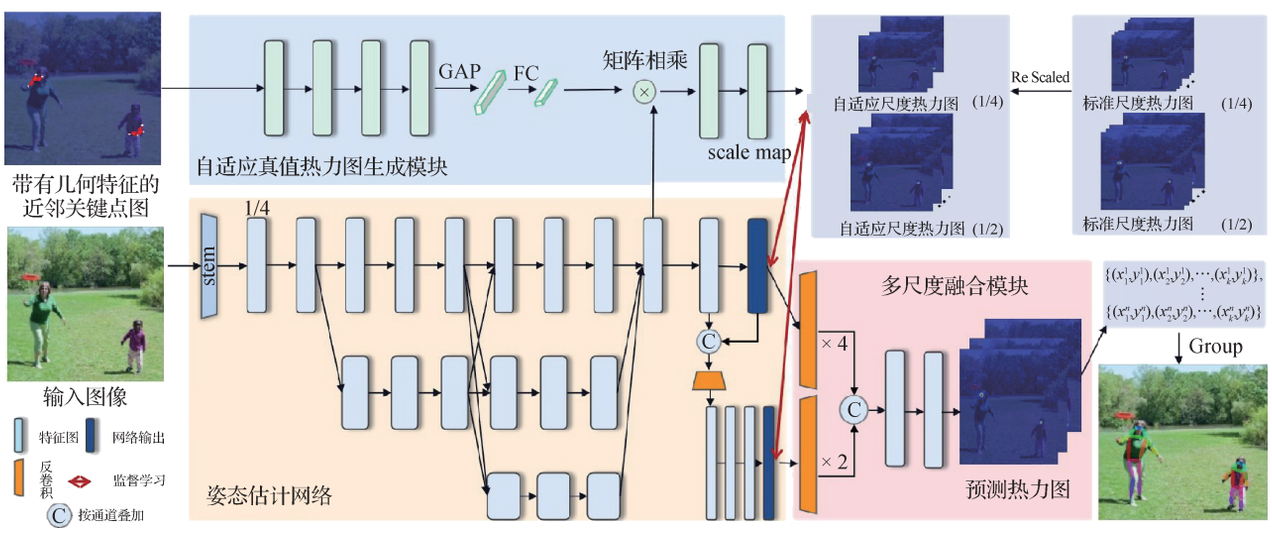


图2　AHGNet框架图

Fig. 2　AHGNet framework

3　自适应真值热力图生成模块

3.1　标准真值热力图生成

关键点热力图可以认为是以标注点为中心的高斯核生成的概率图。像素值表示属于目标关键点的概率。在生成真值热力图时，将所有人体的相同类型关键点的高斯核作为一个通道，真值热力图中的通道数等于关键点类型的数量。

标准真值热力图通常由固定结构的高斯核放置在标注点位置生成。这里以左肩关键点为例详细说明标准真值热力图生成过程，具体为

$$H(\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma})=\sum_{j=1}^{N}\lambda\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\tag{1}$$

式中，N表示图像中人体数量，j表示第j个人，$\boldsymbol{x}$表示像素位置，$\boldsymbol{\mu}$表示均值向量。λ的定义为

$$\lambda=\frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}}\tag{2}$$

式中，$\boldsymbol{\Sigma}$表示协方差，σ表示标准差，一般使用固定值（$\sigma=2$）。$\boldsymbol{\Sigma}$的计算式为

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{bmatrix}\tag{3}$$

高斯分布对于远离标注点的像素仍然有概率值。然而，这些像素已经没有关键点相关的语义信息，可能导致模型学习的语义歧义。为此，通过一个阈值（θ）来截断高斯核可以计算出一个合理的区间来确定高斯核的边界，得到二维高斯核的截断半径（R），具体为

$$H(x,\boldsymbol{\mu},R)=\begin{cases}H(x;\boldsymbol{\mu},\boldsymbol{\Sigma}) & d(x,\boldsymbol{\mu})\leqslant R\\ 0 & \text{其他}\end{cases}\tag{4}$$

式中，R表示高斯核的截断半径，d表示切比雪夫距离，定义为

$$d(i,j)=\max(|i_x-j_x|,|i_y-j_y|) \tag{5}$$

基于PauTa准则，为确保在截断半径内的点的概率足够高(99.7%)，通常将R设为3σ。

3.2 自适应真值热力图生成

标准真值热力图采用固定标准差的高斯核生成，导致不同尺度关键点的高斯核结构一致，易引发语义歧义。首先，不同类型关键点具有固有尺度差异，例如膝盖关键点的固有尺度通常大于鼻部关键点。其次，在拥挤场景中，关键点可能受到周围关键点的干扰，仅考虑固有尺度不足以准确建模，还需结合图像中相邻关键点间的几何关系动态调整高斯核大小。此外，固定标准差无法适应透视变形引起的尺度变化，进一步加剧了关键点定位的模糊性，尤其在复杂背景中表现明显。因此，本文提出一种面向自下而上人体姿态估计的自适应真值热力图生成方法，通过融合图像特征中的固有尺度信息和相邻关键点的几何关系，调整高斯核的截断面积，增加真值热力图的尺度信息，旨在提升多尺度场景下的鲁棒性，为自下而上姿态估计提供更精确的监督信号。

不同关键点应具有差异化的高斯核覆盖面积，为此本文为每个关键点学习一个尺度因子，以生成定制的真值热力图。以一维高斯分布为例，通过引入阈值和截断半径，在固定阈值条件下，可推导出高斯核截断半径与标准差的关系，具体为

$$\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{R^2}{2\sigma^2}\right)=\theta \tag{6}$$

$$\exp\left(-\frac{R^2}{2\sigma^2}\right)=\sqrt{2\pi}\,\sigma\theta \tag{7}$$

$$\frac{R^2}{2\sigma^2}=\ln((\sqrt{2\pi}\,\sigma\theta)^{-1}) \tag{8}$$

$$R=\sqrt{2\sigma^2\ln((2\pi)^{\frac{1}{2}}\sigma\theta)^{-1}} \tag{9}$$

基于PauTa准则，高斯核的截断半径(R)可以由标准差以及固定的阈值(θ)确定，如式(9)所示。由此可知，截断半径与标准差成正比，而标准差反映了关键点的不确定性，进一步验证了本文假设。

为进一步验证该假设，通过调整标准差值改变高斯核及其覆盖区域。如图3所示，标准差越大，高斯核截断面积越大，关键点的语义区域相应增大。

自适应真值热力图生成模块使用标注的点和相邻关键点的向量作为输入，以学习图像中相邻关键点之间的几何关系，结合从图像特征中学习到的关键点的固有尺度信息，学习关键点的尺度因子(s)，用来优化高斯核的标准差，具体为

$$\boldsymbol{\Sigma}_s=\begin{bmatrix}\sigma_s^2 & 0\\ 0 & \sigma_s^2\end{bmatrix}\ \text{s.t.}\ \sigma_s=\sigma\times s \tag{10}$$

式中，σ_s表示自适应尺度标准差。热力图中的每个高斯核在尺度因子图中对应一个比例因子s。自适应真值热力图生成模块为每个关键点回归一个比例因子，通过该因子调整高斯核标准差，生成定制化真值热力图，从而有效避免语义歧义。

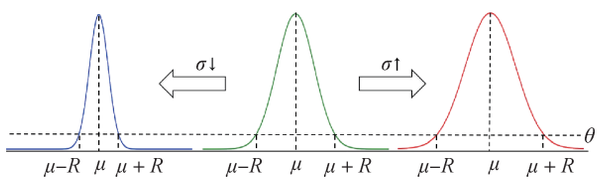


图3 高斯核的标准差与截断面积的关系图

Fig. 3 The relationship between the standard deviation and truncation area of Gaussian kernel

3.3 模块设计

为生成自适应真值热力图，本文以标注的关键点图及其至相邻关键点的向量作为输入，通过学习关键点之间的几何关系，生成各关键点的尺度因子，从而调整高斯核的覆盖面积，如图4所示。具体而言，对于每个标注点，计算其与相邻关键点的距离$(d_1,d_2,\cdots,d_m)$以及平均距离$\bar{d}$，然后通过式(9)计算得到尺度因子和平均差。对于某一类型的关键点，将所有人体的该关键点整合为一个通道，并将相邻关键点的信息作为其他通道。本文设定$m=3$作为相邻关键点的数量。因此，自适应真值热力图生成模块的输入$\boldsymbol{P}_{\text{map}}$具有17×2个通道(其中17表示关键点的总数，2表示坐标维度)。

自适应真值热力图生成模块通过4个卷积层(C^k是一个具有k个特征和核大小3×3的卷积层)、1个全局平均池化(global average pooling，GAP)和2个全连接层(f^l是一个具有l个节点的全连接层)来学习几何关系特征(geometric relationship feature，GRF)，具体为

$$\boldsymbol{f}_{\text{GRF}}(\boldsymbol{P}_{\text{map}})=f^{128}(\text{GAP}(C^{64}(C^{128}(C^{128}(C^{64}(\boldsymbol{P}_{\text{map}}))))))) \tag{11}$$

与全连接层相比，自适应真值热力图生成模块中采用了GAP，可以垂直聚合相邻关键点之间的空间信息和几何关系。

在生成尺度自适应的自定义真值热力图时，除了考虑图像中相邻关键点之间的几何关系外，还需

(a) 当前关键点到近邻关键点的几何向量　(b) 自适应真值热力图　(c) 标准真值热力图

图4　利用相邻关键点之间的几何关系调整高斯核截断半径获取自适应真值热力图

Fig. 4　Adjust the truncation radius of the Gaussian kernel using the geometric relationship between adjacent keypoints to obtain an adaptive ground-truth heatmap ((a) geometric vectors; (b) adaptive ground-truth heatmaps; (c) standard ground-truth heatmaps)

要考虑图像特征中关键点的固有尺度信息。因此，本文将图像特征($\boldsymbol{F}_{\text{img}}$)集成到自适应真值热力图生成模块中，以生成尺度因子图($\boldsymbol{S}$)，具体为

$$\boldsymbol{S} = C^{17}(C^{64}(f_{\text{GRF}}(\boldsymbol{P}_{\text{map}}) \otimes \boldsymbol{F}_{\text{img}})) \tag{12}$$

通过调整高斯核的尺度，定制自适应尺度真值热力图，作为姿态估计网络的回归目标进行训练。

多尺度热力图融合模块主要负责对预测热力图进行重组，以实现分辨率的恢复。为降低计算成本，现有方法通常采用低分辨率热力图作为监督学习的真值。然而，在测试阶段，需将热力图恢复至原始图像分辨率，大多数方法依赖双线性插值进行分辨率恢复(Cheng等，2020；Luo等，2021)，这常导致定位偏差。为此，本文独立训练了一个多尺度融合模块以恢复热力图分辨率。该模块以不同尺度的预测热力图($\boldsymbol{H}_{1/4}$和$\boldsymbol{H}_{1/2}$)作为输入，通过反卷积层(D)统一提升分辨率，随后进行连接(Z_{concate})。经过两个3×3卷积层处理后，生成与原始图像分辨率一致的预测热力图，具体为

$$\boldsymbol{H}_{\text{final}} = C^{17}(C^{64}(Z_{\text{concate}}(D_1(\boldsymbol{H}_{1/4}), D_2(\boldsymbol{H}_{1/2})))) \tag{13}$$

最后，根据预测热力图的局部最大值提取关键点位置坐标。

4　局部概率一致性损失函数

大多数基于热力图回归的方法采用L_2损失进行监督学习。然而，L_2损失假设每个像素点相互独立，忽视了局部结构间的相关性，导致难以准确刻画热力图的概率分布。因此，本文提出一种局部概率一致性损失函数(LPCLoss)，通过计算局部结构相似性，增强模型对局部结构的学习与理解，同时引入动态权重以平衡不同样本的贡献，进一步优化模型的训练方向。

L_2损失函数定义了全局热力图的像素级误差，具体为

$$L_2 = \left\| \boldsymbol{H}^{\text{GT}} - \boldsymbol{H}^{\text{Pred}} \right\| \tag{14}$$

式中，$\boldsymbol{H}^{\text{Pred}}$和$\boldsymbol{H}^{\text{GT}}$分别表示预测与真值热力图。$L_2$损失并不能充分表示两个高斯分布之间的差异。如图5所示，即使L_2损失大幅减少，两个高斯分布的中心点也可能保持不变。这说明L_2损失对位置偏差和标准差的变化不敏感。关键点热力图是描述属于某个关节的周围像素的概率分布，需要更精确地描述两个概率分布的一致性。

为此，本文提出使用KL散度(Kullback-Leibler divergence)来描述局部概率一致性。KL散度用于度量两个概率分布$P(x)$和$Q(x)$的差异。对于随机变量$x \in \boldsymbol{X}$，KL散度计算在x处两个分布之间差异，计算式为

$$D_{\mathrm{KL}}[P(x) \parallel Q(x)] = \sum_{x \in \mathrm{X}} P(x) \log \frac{P(x)}{Q(x)} \tag{15}$$

对于真值热力图中的每个高斯核，计算以标注点为中心的9×9框的KL损失，计算式为

$$L_{\mathrm{prob}} = \frac{1}{\boldsymbol{H}^{\mathrm{total}}} \sum_{h \in \boldsymbol{H}^{\mathrm{total}}} D_{\mathrm{KL}}[h^{\mathrm{GT}} \parallel h^{\mathrm{Pred}}] \tag{16}$$

式中，$\boldsymbol{H}^{\mathrm{total}}$是真值热力图中非空高斯核集合。$h^{\mathrm{GT}}$和$h^{\mathrm{Pred}}$分别表示真值与预测热力图中的高斯核。模型训练容易受到大量易检测的关键点的影响。受到Focal Loss的启发，为了解决难检测样本和易检测样本的不平衡问题，让模型主动关注难检测样本，本文利用动态权重来降低易检测样本的贡献，增强难检测样本的贡献，具体为

$$L_{\mathrm{pixel}} = W \times \left\| \boldsymbol{H}^{\mathrm{GT}} - \boldsymbol{H}^{\mathrm{Pred}} \right\| \tag{17}$$

$$W = \left| \boldsymbol{H}^{\mathrm{GT}} - \boldsymbol{H}^{\mathrm{Pred}} \right|^{\gamma} \tag{18}$$

此外，为了稳定尺度因子的回归方向，本文同样添加了一个正则化器损失，计算式为

$$L_{\mathrm{scale}} = \left\| 1/s - 1 \right\|^2 \tag{19}$$

综上，本文将最终的目标损失函数L定义为

$$L = \alpha L_{\mathrm{pixel}} + \beta L_{\mathrm{prob}} + \rho L_{\mathrm{scale}} + \varphi L_{\mathrm{group}} \tag{20}$$

式中，分组损失L_{group}采用与HrHRNet(Cheng等，2020)一样的方式，对提取的关键点进行分组以构建完整的人体姿态。采用L_2损失量化热力图像素间的差异，并引入KL散度定义局部概率一致性损失，以增强局部结构的学习，同时通过轻量化权重设计平衡样本贡献。

式(20)中的超参数$\alpha = 0.01$，$\beta = 0.01$，$\rho = 0.1$，$\varphi = 0.02$。上述超参数经在COCO数据集上进行多轮实验并结合性能评估结果优化而得，以实现模型性能的最优配置。

5 实验及结果分析

5.1 实验设置

5.1.1 数据集

本文使用两个公开数据集COCO和CrowdPose评估所提出的AHGNet模型。COCO数据集(Lin等，2014)大多来自日常活动场景，包含训练集(75 k幅图像)、验证集(5 k幅图像)和测试集(20 k幅图像)。CrowdPose数据集(Li等，2019a)收集了较难的拥挤场景中的活动图像，包含20 k幅图像，包含大约80 000人。按照5∶1∶4比例划分为训练集、验证集和测试集。

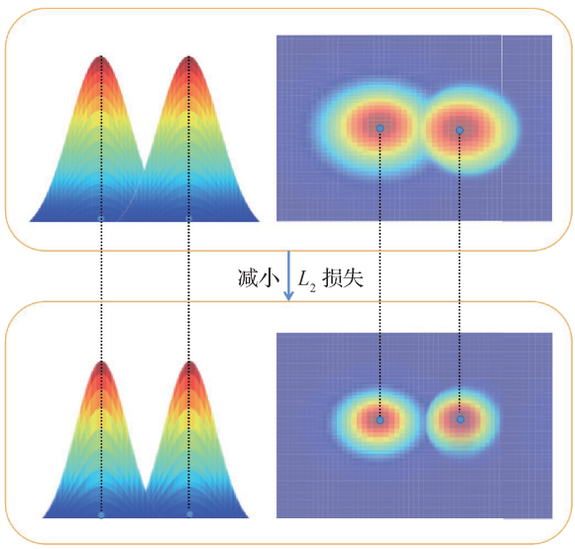


图5 L_2损失减小但中心点不变的示例

Fig. 5 An example of L_2 loss reduction but center unchanged

5.1.2 评价指标

为了比较不同算法的性能，使用数据集的评价方法，使用OKS(object keypoint similarity)来测试预测关键点与真值的相似性，具体为

$$OKS = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\sigma(v_i > 0)}{\sum_i \sigma(v_i > 0)} \tag{21}$$

式中，d_i表示预测关键点与真值之间的欧几里得距离，v_i表示标注的关键点的可见性，s表示目标的尺度，而k_i表示控制衰减的常数。在两个数据集上计算平均准确率(average precision，AP)，包括AP^{50}($OKS = 0.5$)和AP^{75}($OKS = 0.75$)，AP($OKS = 0.5$，0.55，…，0.9，0.95得分的平均值)。除此之外，COCO数据集中添加了额外的评价：中等尺度的人体姿态估计准确率AP^{M}和大尺度的人体姿态估计准确率AP^{L}。CrowdPose数据集在相对简单的样本(AP^{E})、中等难度样本(AP^{M})和困难样本(AP^{H})上增加了额外的评价。

5.1.3 模型训练

所有实验均在带有Ubuntun系统的电脑上完

成，配置设备包括6核心3.2 GHz的中央处理器（central processing unit，CPU），32 GB内存以及两块NVIDIA GeForce RTX 2080Ti显卡。实验通过随机旋转[−30，30]、随机缩放[0.75，1.5]、随机平移[−40，40]和随机水平翻转来扩充训练数据。输入图像被裁剪为512像素（或640像素）的分辨率。采用Adam优化器，初始学习率为10^{-3}，学习率将线性衰减。

5.1.4　模型测试

模型测试时，将测试图像的短边调整为512像素（或640像素），并保持图像的长宽比。在多尺度测试中，本文使用3个尺度（0.5、1、2），并通过将所有预测热力图调整为相同的大小来计算最终的预测热力图。模型测试时，不需要自适应真值热力图生成模块，直接使用姿态估计网络进行热力图预测，通过识别预测热力图中的局部最大值来获得关键点的位置坐标，并组合成完整的人体姿态。

5.2　实验结果

本文以HrHRNet（Cheng等，2020）为姿态估计网络，构建了基于自适应真值热力图的人体姿态估计网络AHGNet。为了验证本文方法的性能，将其与人体姿态估计中的主流模型进行对比。

5.2.1　COCO数据集实验结果

不同方法在COCO测试集实验结果对比如表1所示。可以看出，在没有多尺度测试情况下，相比基线方法，AHGNet-W48准确率提高2.5%，达到70.9%；引入多尺度测试后，相比基线方法，AHGNet-W48准确率提高1.6%，达到72.1%，验证了本文方法的有效性。

表1　不同方法在COCO测试集实验结果对比

Table 1　Comparisons with other methods on COCO test set

/%

	方法	输入/像素	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
非多尺度实验	CenterNet-DLA（Zhang等，2016）	512	57.9	84.7	63.1	52.5	67.4
	CenterNet-HG（Zhang等，2016）	512	63.0	86.8	69.6	58.9	70.4
	OpenPose（Cao等，2017）	368	61.8	84.9	67.5	57.1	68.2
	PersonLab（Papandreou等，2018）	1 401	66.5	88.0	72.6	62.4	72.3
	PifPaf（Kreiss等，2019）	641	66.7	–	–	62.4	72.9
	HrHRNet-W32（Cheng等，2020）	512	66.4	87.5	72.8	61.2	74.2
	AHGNet-W32（本文）	512	68.7	89.2	75.4	63.1	76.5
	SWAHR-W32（Luo等，2021）	512	67.9	88.9	74.5	62.4	75.5
	SWAHR-W48（Luo等，2021）	640	70.2	89.9	76.9	65.2	77.0
	ED-Pose（Yang等，2023）	–	69.8	90.2	77.2	64.3	77.4
	GroupPose（Liu等，2023）	–	70.2	90.5	77.8	64.7	78.0
	HrHRNet-W48（Cheng等，2020）	640	68.4	88.2	75.1	64.4	74.2
	AHGNet-W48（本文）	640	**70.9**	89.9	77.0	65.9	76.6
多尺度实验	PersonLab（Papandreou等，2018）	1 401	68.7	89.0	75.4	64.1	75.5
	HrHRNet-W32（Cheng等，2020）	512	69.0	89.0	75.8	64.4	75.2
	CKG（Brasó等，2021）	640	71.1	90.5	77.5	66.9	76.7
	PETR（Shi等，2022）	–	71.2	91.4	77.6	66.9	78.0
	PoseTrans（Jiang等，2022b）	512	69.9	89.3	77.0	65.2	76.2
	KAPAO-M（McNally等，2022）	1 024	70.3	91.2	77.8	66.3	76.8
	HrHRNet-W48（Cheng等，2020）	640	70.5	89.3	77.2	66.6	75.8
	AHGNet-W48（本文）	640	**72.1**	90.6	78.9	68.1	77.3

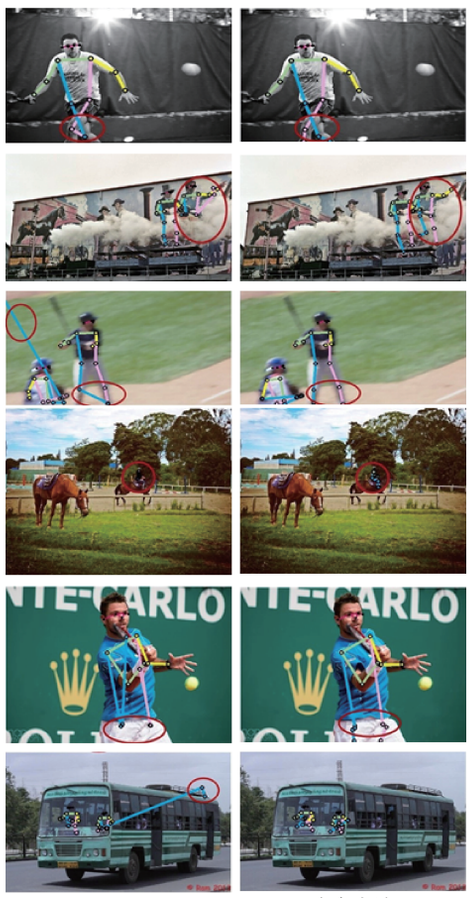
注：加粗字体表示最优结果，“–”表示论文未提供数据。

5.2.2　COCO数据集实验结果可视化

本文选择不同的测试图像进行可视化，图6展示了HrHRNet和本文方法的可视化结果。可以看出，对于尺度差异显著的图像，特别是在包含小尺度关键点的场景中，HRNet易出现误检和漏检的情况，而本文方法显著改善了预测性能，尤其在小尺度关键点的检测上表现优异。可视化结果表明，本文方法能够准确检测尺度变化较大的图像中的关键点位置，提升了对尺度较小的关键点的检测准确率。

5.2.3　CrowdPose数据集实验结果

如表2所示，相比基线方法，在非多尺度与多尺度测试下，AHGNet准确率分别提升6.7%和6.5%，达到72.6%和74.1%。结果表明，AHGNet能够有效应对CrowdPose数据集中的密集关键点分布，尤其是在多人体重叠和遮挡复杂的情况下，展现出稳定的检测能力。此外，AHGNet的鲁棒性在不同测试配置下保持一致，进一步验证了本文方法在复杂拥挤场景中的有效性。对比表1和表2可以看出，AHGNet在CrowdPose数据集准确率的提升比在COCO数据集更显著。CrowdPose中，拥挤场景下的关键点检测面临更大挑战，仅考虑关键点固有的尺度信息是不够的，还需考虑密集区域的关键点所占的实际语义区域。AHGNet创新性地融合了相邻关键点之间的几何关系和其固有尺度信息，动态生成关键点的尺度因子，并据此调整高斯核的截断面积，有效缓解了因语义歧义导致的误检和漏检问题，从而提高人体姿态估计模型的准确率。



HrHRNet　　本文方法

图6　COCO数据集上的可视化结果

Fig. 6　Visualization results on COCO dataset

表2　不同方法在CrowdPose测试集实验结果对比

Table 2　Comparisons with other methods on CrowdPose test set

/%

	方法	AP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
非多尺度实验	HrHRNet-W48(Cheng等,2020)	65.9	86.4	706	73.3	66.5	57.9
	SWAHR-W48(Luo等,2021)	71.6	88.5	77.6	78.9	72.4	63.0
	AHGNet-W48(本文)	**72.6**	90.6	78.5	79.6	73.3	64.0
多尺度实验	HrHRNet-W48(Cheng等,2020)	67.6	87.4	72.6	75.8	68.1	58.9
	SWAHR-W48(Luo等,2021)	73.8	90.5	79.9	81.2	74.7	64.7
	AHGNet-W48(本文)	**74.1**	91.2	79.8	81.3	74.8	65.3

注：加粗字体表示最优结果。

5.3　消融实验

为了验证各模块的有效性，本文以HrHRNet-W32作为基线，在COCO val2017数据集上进行消融实验。所有消融实验均未采用多尺度测试，以确保结果的独立性和可比性。

5.3.1　高斯核标准差大小对实验结果的影响

如表3所示，随着标准差σ的增大，模型在中等尺度人体上的性能有所下降，而在大尺度人体上的

表现显著提升。实验结果表明，较大的σ值更适用于体型较大的人体，这与本文的假设相符：大尺度关键点通常对应更大的语义区域和更高的标注歧义，需要采用标准差较大的高斯核进行真值热力图的构建。本文通过生成尺度自适应真值热力图，有效缓解了多尺度变化人体带来的语义模糊问题。

表3　高斯核标准差σ对模型精度的影响

Table 3　The impact of Gaussian kernel standard deviation σ on model accuracy

/%

σ	AP	AP^M	AP^L
1.0	64.9	61.2	71.0
1.5	65.1	61.2	72.3
2.0	66.6	61.3	75.0
2.5	66.1	60.1	75.2
3.0	65.4	58.3	75.4

5.3.2　AHGM、KL散度和动态权重对实验结果的影响

如表4所示，本文对自适应真值热力图生成模块（AHGM）、KL散度和动态权重进行消融实验。可以看出，三者均提升了模型的准确率，三者结合使用，AHGNet的准确率提升了2.6%。

1）自适应真值热力图生成模块。从表4第2、3行结果对比可以看出，通过生成尺度自适应真值热力图，模型准确率提高1.3%。通过表4第2、8行结果对比，结合LPCLoss，准确率达到69.7%，也验证了自适应真值热力图生成模块的有效性。

2）KL散度。从表4第2、4行结果对比可以看出，使用KL散度定义局部概率一致性损失提升了模型1.2%的精度，表明KL散度可以用来描述概率损失，结合KL散度训练的模型比仅使用L_2损失进行训练的模型准确率更高。

3）动态权重。从表4第2、5行结果对比可以看出，动态权重的加入提升了模型0.9%的准确率。并且从图6可以看出，AHGNet可以准确检测小尺度关键点，表明动态权重可以平衡难易样本的贡献，从而提升模型的准确率。

5.3.3　KL散度计算框大小选择对实验结果的影响

本文研究了KL散度计算框大小选择对实验结果的影响。如表5所示，结果显示，9×9计算框的模型结果最优。这说明过小的KL散度计算框不足以包含有用的信息，过大的KL散度计算框可能会掺杂噪声，影响模型准确率。

表4　不同模块的有效性实验结果

Table 4　Experimental results on the effectiveness of different modules

/%

AHGM	KL散度	动态权重	AP	AP^{50}	AP^{75}	AP^M	AP^L
–	–	–	67.1	86.2	73.0	61.5	76.1
√	–	–	68.4	76.4	74.3	61.5	77.7
–	√	–	68.3	87.3	74.1	62.0	77.2
–	–	√	68.0	87.5	74.1	62.3	76.8
√	√	–	69.2	**88.2**	74.6	63.1	77.9
√	–	√	68.8	87.7	74.8	62.6	77.7
√	√	√	**69.7**	88.1	**75.0**	**63.3**	**78.6**

注：加粗字体表示各列最优结果，“√”表示加入此模块，“–”表示未加入。

5.3.4　动态权重指数对于模型的影响

本文研究了动态权重指数对于模型的影响。本文通过式（17）和式（18）来平衡困难和容易的样本。W越大，样本贡献越大，反之W越小，样本贡献越小，进而调整难易样本的平衡。如表6所示，模型在γ=1时的准确率达到69.7%。

表5　不同KL散度计算框大小的实验结果

Table 5　Experimental results of different KL divergence calculation box sizes

KL散度计算框大小	AP/%
3×3	68.7
5×5	69.1
7×7	69.1
9×9	**69.7**
11×11	69.2
13×13	69.1

注：加粗字体表示最优结果。

5.3.5　多尺度热力图融合模块对于模型的影响

本文通过消融实验来证明多尺度热力图融合的有效性，结果如表7所示。第2行展示了未采用多尺度热力图融合的基线结果，第3行则呈现了引入多尺度热力图融合后的测试结果。实验结果表明，多

表 6　不同动态权重指数的实验结果

Table 6　Experimental results of different dynamic weight indices

γ	AP/%
0.5	68.8
1	**69.7**
2	65.8
3	58.6

注：加粗字体表示最优结果。

尺度热力图融合模型能够有效纠正因人工比例放大导致的关键点位置偏差。通过融合不同尺度热力图的信息，模型的准确率提升了 0.4%，验证了多尺度融合策略的有效性。

表 7　多尺度融合模块有效性实验结果

Table 7　Experimental results on the effectiveness of the multi-scale fusion module

/%

多尺度融合模块	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
未采用	69.7	**88.1**	75.0	63.3	78.6
采用	**70.1**	88.0	**76.2**	**66.9**	**78.7**

注：加粗字体表示各列最优结果。

6　结　论

本文提出一种面向自下而上人体姿态估计的自适应真值热力图生成方法，为该任务提供了一种新的监督信号优化范式。该方法通过动态调整高斯核尺度并结合局部概率约束，有效缓解了复杂场景中的多尺度歧义问题。此外，针对现有方法热力图损失函数难以有效捕捉局部结构的相关性，导致其对关键点位置偏差不敏感的问题，提出局部概率一致性损失函数。该损失函数在热力图的局部区域计算结构相似性，提升模型对局部结构的建模能力，并通过引入动态权重机制来平衡样本贡献，进一步优化模型的学习方向，提高其鲁棒性。实验结果表明，所提出的 AHGNet 通过学习生成的自适应真值热力图，并结合局部一致性损失函数，有效提升了人体姿态估计的准确率。

参考文献（References）

Brasó G, Kister N and Leal-Taixé L. 2021. The center of attention: center-keypoint grouping via attention for multi-person pose estimation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 11833-11843 [DOI: 10.1109/ICCV48922.2021.01164]

Cai Y H, Wang Z C, Luo Z X, Yin B Y, Du A G, Wang H Q, Zhang X Y, Zhou X Y, Zhou E J and Sun J. 2020. Learning delicate local representations for multi-person pose estimation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 455-472 [DOI: 10.1007/978-3-030-58580-8_27]

Cao Z, Simon T, Wei S E and Sheikh Y. 2017. Realtime multi-person 2D pose estimation using part affinity fields//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1302-1310 [DOI: 10.1109/CVPR.2017.143]

Chen Y L, Wang Z C, Peng Y X, Zhang Z Q, Yu G and Sun J. 2018. Cascaded pyramid network for multi-person pose estimation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7103-7112 [DOI: 10.1109/CVPR.2018.00742]

Cheng B W, Xiao B, Wang J D, Shi H H, Huang T S and Zhang L. 2020. HigheRHRnet: scale-aware representation learning for bottom-up human pose estimation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5385-5394 [DOI: 10.1109/CVPR42600.2020.00543]

Ding Y, Zhang Z L, Zhao X F, Hong D F, Cai W, Yu C G, Yang N J and Cai W W. 2022. Multi-feature fusion: graph neural network and CNN combining for hyperspectral image classification. Neurocomputing, 501: 246-257 [DOI: 10.1016/j.neucom.2022.06.031]

Jia W, Li J, Li S J, Zhao Y and Min H. 2024. A high-order graph convolutional network for homomorphic and heterogeneous skeletal motion retargeting. Journal of Image and Graphics, 29(12): 3712-3726 (贾伟, 李骏, 李书杰, 赵洋, 闵海. 2024. 面向同胚异构骨骼运动重定向的高阶图卷积网络. 中国图象图形学报, 29(12): 3712-3726) [DOI: 10.11834/jig.230909]

Jiang J J, He Z X, Zhao X Y, Zhang S Y, Wu C R and Wang Y. 2022a. MLFNet: monocular lifting fusion network for 6DoF texture-less object pose estimation. Neurocomputing, 504: 16-29 [DOI: 10.1016/j.neucom.2022.06.096]

Jiang W T, Jin S, Liu W T, Qian C, Luo P and Liu S. 2022b. PoseTrans: a simple yet effective pose transformation augmentation for human pose estimation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 643-659 [DOI: 10.1007/978-3-031-20065-6_37]

Ke L P, Chang M C, Qi H G and Lyu S W. 2018. Multi-scale structure-aware network for human pose estimation//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 731-746 [DOI: 10.1007/978-3-030-01216-8_44]

Kreiss S, Bertoni L and Alahi A. 2019. PifPaf: composite fields for human pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 11969-11978 [DOI: 10.1109/CVPR.2019.01225]

Li J, Su W and Wang Z F. 2020. Simple pose: rethinking and improving a bottom-up approach for multi-person pose estimation//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 11354-11361 [DOI: 10.1609/aaai.v34i07.6797]

Li J F, Wang C, Zhu H, Mao Y H, Fang H S and Lu C W. 2019a. CrowdPose: efficient crowded scenes pose estimation and a new benchmark//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10855-10864 [DOI: 10.1109/CVPR.2019.01112]

Li K, Wang S J, Zhang X, Xu Y F, Xu W J and Tu Z W. 2021. Pose recognition with cascade transformers//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1944-1953 [DOI: 10.1109/CVPR46437.2021.00198]

Li W B, Wang Z C, Yin B Y, Peng Q X, Du Y M, Xiao T Z, Yu G, Lu H T, Wei Y C and Sun J. 2019b. Rethinking on multi-stage networks for human pose estimation [EB/OL]. [2024-10-21]. https://arxiv.org/pdf/1901.00148.pdf

Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2020. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 318-327 [DOI: 10.1109/TPAMI.2018.2858826]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu H, Chen Q, Tan Z C, Liu J J, Wan, J, Su X B, Li X L, Yao K, Han J Y, Ding E R, Zhao Y and Wang J D. 2023. Group pose: a simple baseline for end-to-end multi-person pose estimation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 14983-14992 [DOI: 10.1109/ICCV51070.2023.01380]

Luo Z X, Wang Z C, Huang Y, Wang L, Tan T N and Zhou E J. 2021. Rethinking the heatmap regression for bottom-up human pose estimation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13259-13268 [DOI: 10.1109/CVPR46437.2021.01306]

Mao W A, Ge Y T, Shen C H, Tian Z, Wang X L and Wang Z B. 2021. TFPose: direct human pose estimation with transformers [EB/OL]. [2024-10-21]. https://arxiv.org/pdf/2103.15320.pdf

McNally W, Vats K, Wong A and McPhee J. 2022. Rethinking keypoint representations: modeling keypoints and poses as objects for multi-person human pose estimation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 37-54 [DOI: 10.1007/978-3-031-20068-7_3]

Newell A, Yang K Y and Deng J. 2016. Stacked hourglass networks for human pose estimation//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 483-499 [DOI: 10.1007/978-3-319-46484-8_29]

Papandreou G, Zhu T, Chen L C, Gidaris S, Tompson J and Murphy K. 2018. PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 282-299 [DOI: 10.1007/978-3-030-01264-9_17]

Qi T, Bayramli B, Ali U, Zhang Q C and Lu H T. 2019. Spatial shortcut network for human pose estimation [EB/OL]. [2024-10-21]. https://arxiv.org/pdf/1904.03141.pdf

Sekii T. 2018. Pose proposal networks//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 350-366 [DOI: 10.1007/978-3-030-01261-8_21]

Shi D H, Wei X, Li L Q, Ren Y and Tan W M. 2022. End-to-end multi-person pose estimation with transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11059-11068 [DOI: 10.1109/CVPR52688.2022.01079]

Sun K, Xiao B, Liu D and Wang J D. 2019. Deep high-resolution representation learning for human pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5686-5696 [DOI: 10.1109/CVPR.2019.00584]

Sun X, Xiao B, Wei F Y, Liang S and Wei Y C. 2018. Integral human pose regression//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 536-553 [DOI: 10.1007/978-3-030-01231-1_33]

Toshev A and Szegedy C. 2014. DeepPose: human pose estimation via deep neural networks//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 1653-1660 [DOI: 10.1109/CVPR.2014.214]

Yang J, Zeng A L, Liu S L, Li F, Zhang R M and Zhang L. 2023. Explicit box detection unifies end-to-end multi-person pose estimation//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR

Yang W, Li S, Ouyang W L, Li H S and Wang X G. 2017. Learning feature pyramids for human pose estimation//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1290-1299 [DOI: 10.1109/ICCV.2017.144]

Yu L, Du C J, Yan Z Q, Zhao H J and He S J. 2024. Review of 2D

human pose encoding and decoding methods: from the perspective of ambiguity mitigation. Journal of Image and Graphics, 29(11): 3319-3344 (喻莉, 杜聪炬, 闫增强, 赵慧娟, 何双江. 2024. 二维人体姿态编解码方法综述: 从解决歧义性问题的角度出发. 中国图象图形学报, 29(11): 3319-3344) [DOI: 10.11834/jig.230648]

Zhang Y Y, Zhou D S, Chen S Q, Gao S H and Ma Y. 2016. Single-image crowd counting via multi-column convolutional neural network//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 589-597 [DOI: 10.1109/CVPR.2016.70]

作者简介

江玲，女，讲师，主要研究方向为计算机视觉和图像分析与理解。E-mail: jiangjiang@aust.edu.cn

李凯歌，通信作者，男，博士后，主要研究方向为计算机视觉、图像分析与理解和虚拟现实。E-mail: likg@mail.sysu.edu.cn

刘卓程，男，硕士研究生，主要研究方向为计算机视觉和图像分析与理解。E-mail: zcliu@buaa.edu.cn

熊源，男，博士研究生，主要研究方向为计算机视觉和三维重建。E-mail: xiongyuanxy@buaa.edu.cn

吴威，男，教授，博士生导师，主要研究方向为计算机视觉、图像分析与理解和虚拟现实。E-mail: wuwei@buaa.edu.cn