## 二、研究方案及技术途径

详细论述研究思路、研究方案、技术途径。

## 总体方案

本项目需要完成对多运动人员关重部位的精确识别与跟踪，为精确击伤提供重要技术支撑，从而大幅提升智能化作Z效能。本项目总体方案将涉及数据获取与生成、检测模型、追踪算法等方面进行宏观考虑。

在深度学习中，数据集是训练深度学习模型的基础。通过将大量的数据馈送给模型，模型能够学习到数据中的模式和规律，从而提高其性能。模型的性能很大程度上取决于用于训练的数据的多样性和数量。足够丰富、多样的数据集可以帮助模型学习到更好的特征和表示，提高模型的泛化能力。一个丰富、多样的数据集可以帮助模型提高泛化能力。模型在训练过程中接触到的不同样本和场景，使其能够更好地适应各种不同的情况，而不仅仅是记住训练集中的特定样本。受限于当前所公开的人体姿态估计数据集样本及数量的丰富性，仅使用公开数据集不能够支撑起本项目拟对运动人员关重部位识别与跟踪达到的高鲁棒性与高精确性。因此本项目提出采用：“虚实结合”与数据增强两种主要方式扩展训练数据，增加训练数据量。

在人体姿势估计（关键点检测）方面，研究人员拟针对该项目构建人体姿势估计网络。网络主要有三部分构成，即骨干网络、解耦头、输出层。该网络使用CSPDarknet53 作为其骨干网络。CSPDarknet53是一种卷积神经网络（CNN），其用于从输入图像中提取特征。特征提取是目标检测的基础，因为它捕获了图像中的语义信息。通过使用解耦的头结构，将不同层次的特征融合在一起，以提高对不同尺度目标的检测性能。这有助于处理不同尺寸的目标，并提高模型的泛化能力。网络采用双检测头模型进行输出。其中，Detect head识别人体姿势的种类，Pose head对人体关键部位的特征点位置进行输出。

在追踪算法部分，本项目拟采用一种高效的检测结果关联方法，该算法根据检测目标置信度与追踪轨迹关联，采用卡尔曼滤波预测目标的新位置，并采用匈牙利算法进行相似性匹配。

## 技术途径

1.数据扩增算法研究

为满足实际场景应用下的准确性与鲁棒性需求，本项目调研了当前所公开的人体姿势估计数据集，但是由于公开数据集样本及数量丰富性缺乏，研究人员拟采用虚实结合的数据扩增方法。在公开数据集的基础上，研究人员使用无人机拍摄了真实人体姿势图片，并在训练过程中加入数据增强策略，提高模型的准确性与泛化性。

本项目采用Unreal Engine仿真模型训练所需数据，具体过程为：通过导入FBX、OBJ、ABC等格式的3D模型文件来创建模型。导入模型时，可以选择将模型转换为静态或动态物体，也可以选择将模型设置为可移动或可交互，通过创建场景来编辑场景中的物体、材质、光照等属性。场景编辑包括创建场景、添加物体、设置物体属性、添加材质、设置材质属性、添加光照、设置光照属性等，通过拖拽和旋转物体来布置场景中的物体。可以使用不同的工具来布置物体，如平移工具、旋转工具、缩放工具等，通过创建动画来使场景中的物体可运动，用UE的动画编辑器来创建动画，包括平移、旋转、缩放、颜色等属性的动画，通过导出场景和模型来将其保存为文件，可以选择将场景和模型导出为FBX、OBJ、ABC等格式的文件。通过仿真可生成大量高质量图像数据，该方法大幅提高了模型检测的准确率及泛化性。

本项目拟使用无人机俯拍真实场景中的人体姿势模型。无人机真实拍摄的数据集提供了真实世界中的场景和目标行为数据。这些数据可以更好地反映真实世界中的变化、复杂性和不确定性，并能提供真实的视觉信息，包括光照、纹理、背景等。从而更好地模拟真实世界的视觉输入，提供更具挑战性的训练和评估环境。这些优势可以帮助深度学习模型更好地适应真实世界的变化和复杂性，提高其性能和泛化能力。

此外，本项目同时拟使用COCO数据集与真实拍摄数据集。COCO数据集是一个广泛应用于目标检测和图像分割的数据集，其中也包含了人体姿态估计的标注信息。该数据集包含超过20万张图像，并包含超过250000个人体姿态的标注，COCO数据集提供了丰富的姿态变化和复杂场景的图像，是人体姿态估计领域的重要基准数据集；在真实场景拍摄中，研究人员共拍摄8000余张不同角度下的人体姿势图像，由于实际拍摄的复杂性与随机性，一些不包含目标主体或者目标主体占比过于小的图像经过剔除，实际可使用的数量达6000余张。

2.多姿态运动人员关重部位识别技术研究

本项目采用的人体姿势估计网络架构其主要由三部分组成，分别是骨干网络、解耦头、输出层，模型采用SGD算法优化参数。

2.1 多尺度多姿态人员鲁棒特征提取技术

在深度学习中，主干网络是整个深度学习模型的主要特征提取部分。它负责从输入图像中提取特征，并逐渐减小特征图的分辨率。主干网络通常由卷积层、池化层和激活函数等组成，用于捕获不同层次的图像特征。结合本项目的实际应用场景，为了更好的提取多尺度多姿态人员鲁棒特征，设计骨干网络如下。

骨干网络的结构：图4显示了骨干网络的整体架构。该网络使用 CSPDarknet53 作为其骨干网络。CSPDarknet53是一种卷积神经网络（CNN），其用于从输入图像中提取特征。特征提取是目标检测的基础，因为它捕获了图像中的语义信息。

骨干网络通过一系列卷积运算，有效地学习提取有意义的特征。它采用不同大小和深度的滤波器来捕捉空间模式、边缘、纹理和其他有助于检测物体的相关视觉线索。虽然骨干网络在特征提取时可能会导致特征损失，但后续的网络在没有骨干网络的情况下无法学习真实世界图像中存在的丰富特征，特别是在网络较浅的情况下。物体检测算法可以从骨干网络提取和编码丰富特征的能力中受益。上述增强的特征可以使图像中的物体得到更准确的分类和精确的定位。骨干网络提取低级和高级特征的能力有助于提高物体检测系统的准确性和鲁棒性。根据实际应用需要，骨干网络必须表现出双重特性：轻量化以满足快速计算要求，鲁棒的特征提取能力以确保高精确度。CSP架构已经被证明可以提高卷积神经网络的学习能力，同时保持准确性，减少计算限制，最小化内存消耗和最小化网络重量，因此，本项目将CSPDarkNet选作模型的骨干网络。在CSP方法中，网络块的输入在处理前被分成两部分，其中一部分由原始块处理，而另一部分则直接以跳跃连接方式输出。随后，这两部分被结合起来，形成了该块的最终输出。

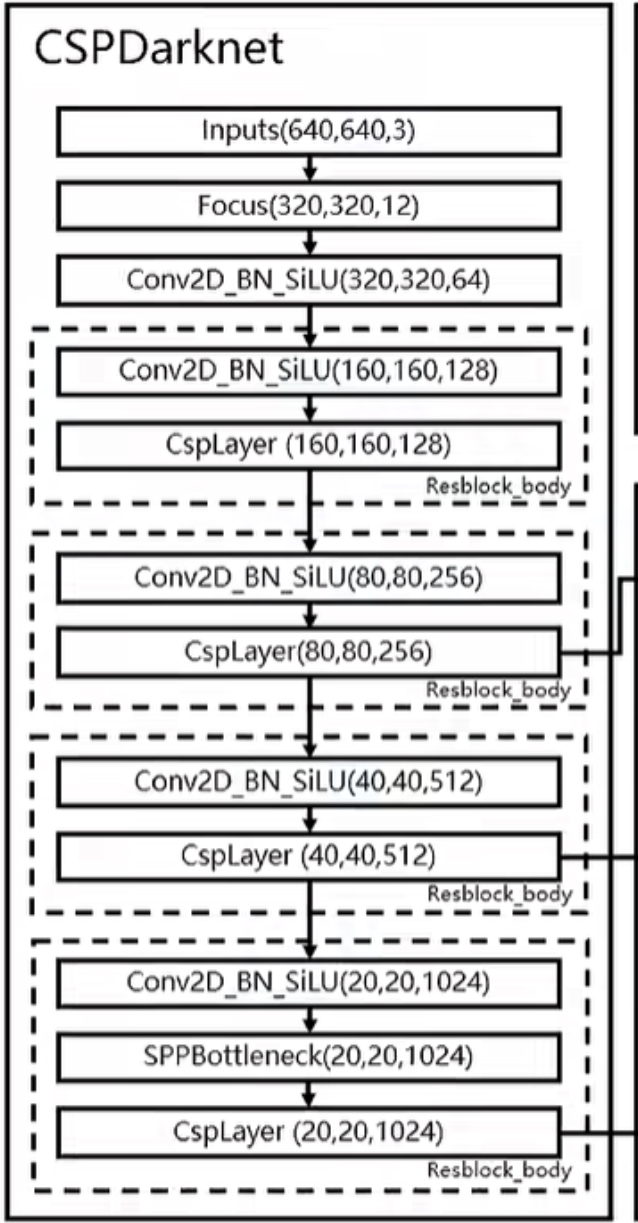


图 4检测模型骨干网络结构

在网络结构中，通过使用解耦的头结构，将不同层次的特征融

合在一起，以提高对不同尺度目标的检测性能。这有助于处理不同尺寸的目标，并提高模型的泛化能力。以80类检测任务为例，每个Anchor将产生85个通道的预测结果，其中目标背景分类占一个通道，目标坐标回归占四个通道，目标分类占80个通道。本项目检测器使用1×1卷积将不同通道数的特征图统一为256，然后同时使用3×3卷积的两个并行分支，在其分支上增加IoU分支。该结构明显提高了收敛速度，同时明显改善了AP。

2.2 基于拓扑关系的人员关重部位表征识别技术

多人2D姿态估计，是理解图像中人体的一种任务。输入一张图片，其目标是检测每一个人和定位到他们对应的关节点。推测图片中多人的姿态是有挑战的，由于图像中存在人体数量的变化，尺度变化，身体部位遮挡，人体非刚性等因素。

当前对于姿态估计的方法主要分为两大类：自下而上和自上而下的。自上而下或者双阶段的方法是目前最领先的方案。他们首先采用一个强大的人体检测器，然后针对每个人体做单人的姿态估计。自上而下的方法的复杂度随着图像中人体数量而线性增加。多数实时应用需要限制运行时间，不倾向于采用自上而下的方法，由于其复杂度较高。反而，自下而上的方法提供了恒定的运行时间，因为他们在单次检测中依赖热力图检测所有关键点，然后进行复杂的后处理将他们分组成个体。后处理可能涉及到如像素级别的NMS，线积分，细化，分组等步骤。坐标调整和细化减少了下采样热力图的量化误差，其中NMS用于寻找热力图中的局部最大值。即使经过后处理，热力图也许不够清晰，去区分同一类型的两个靠得很近的关键点。同样，自下而上的方法也不能端到端训练，因此后处理步骤是不可差分的，因此在卷积网络之外发生。他们在方法上存在很大差异，从线性规划到各种启发式算法。很难用CNN加速器去加速他们，因此速度也很慢。有单次推理的方法，虽然避免了分组任务，但不能与自下而上的方法相提并论。他们依赖额外的后处理来提升性能。

本项目算法拟解决不使用热力图进行姿态估计，而与目标检测保持一致，因为目标检测中的挑战与姿态估计类似，例如尺度多样性，遮挡，人体非刚性等。因此，如果一个人体检测网络能够处理这些问题，它也能处理姿态估计。调研发现，最近的目标检测框架尝试解决尺度变化的问题，通过在多个尺度上做预测。这里，研究人员拟采用相同的策略去针对每个检测去预测多尺度的人体姿态。

对于当前应用场景下的人体姿态估计，它归结为一个三种类别的人体姿态检测问题，每个人有6个相关的关键点，每个关键点输出一个位置和可信度：。因此，与一个锚点关联的6个关键点总共有18个元素。因此，对于每个锚点，关键点头预测18个元素。在神经网络的训练过程中，关键点置信度是基于该关键点的可见性标志进行训练的。如果一个关键点是可见的或者被遮挡的，那么地面真实置信度将设置为1在视场之外，置信度则设置为零。在推理过程中，我们保留了可信度大于0.5的关键点。预测的关键点置信度不用于评估。然而，由于网络预测了每个检测的所有6个关键点，我们需要过滤掉视场之外的关键点。否则，就会有摆动的关键点，导致输出是骨骼变形的。

网络采用双检测头模型进行输出。其中，Detect head识别人

体姿势的种类（分为三种），Pose head对人体关键部位的特征点位置进行输出。

3.人员目标关重部位并行稳定跟踪技术研究

目标跟踪是计算机视觉中非常重要的任务之一，它刚好在目标检测之后出现。为了完成目标跟踪任务，首先需要将目标定位在一帧中，然后给每个目标分配一个单独的唯一id。然后连续帧中的每个相同目标将生成轨迹。在这里，一个目标可以是任何类别，比如行人、车辆、运动中的运动员、天空中的鸟等。在项目的应用场景中，拟实现对项目中多运动人员的关重部位并行稳定跟踪。

多目标跟踪中不仅要面对在单目标跟踪中存在的遮挡、变形、运动模糊、拥挤场景、快速运动、光照变化、尺度变化等挑战，还要面对如轨迹的初始化与终止、相似目标间的相互干扰等复杂问题。多目标跟踪任务主要存在以下研究难点：

1.准确的对象检测的问题是未能检测到对象或者为检测到的对象分配错误的类别标签或错误地定位已识别的对象；

2.ID Switching发生在两个相似的物体重叠或混合时，导致身份切换；因此，很难跟踪对象 ID；

3.背景失真：复杂的背景使得在物体检测过程中难以检测到小物体；

4.遮挡：对象被另一个对象隐藏或遮挡时会产生这个问题；

5.多个空间空间、变形或对象旋转；

6.由于运动模糊而在相机上捕获的视觉条纹或拖尾。

近五六年来，随着目标检测的性能得到了飞跃式进步，也诞生了基于检测进行跟踪的方案，并迅速成为当前多目标跟踪的主流框架，极大地推动了多目标检测任务的前进。同时，近期也出现了基于检测和跟踪联合框架以及基于注意力机制的框架。基于Tracking-by-detaction框架的MOT算法是先对视频序列的每一帧进行目标检测，根据包围框对目标进行裁剪，得到图像中的所有目标。然后，转化为前后两帧之间的目标关联问题，通过IoU、外观等构建相似度矩阵，并通过匈牙利算法、贪婪算法等方法进行求解。

本项目采用了一种有效和通用的数据关联方法，保留了所有的检测框，并将其分为高置信度和低置信度检测框，首先将高置信度检测框与追踪轨道关联，有些追踪轨迹未与适当的高置信度检测框相匹配，这通常发生在遮挡、运动模糊或有较大尺度变化的情况下，将低置信度检测框和不匹配的追踪轨迹关联，以恢复低置信度检测框中的物体，同时抑制背景。追踪算法的输入是一个视频序列以及目标检测器，检测器设置阈值。追踪算法的输出是视频的轨迹，每条轨迹都包含每一帧物体的检测框和序号。对于视频中的每一帧，使用检测器预测检测框位置及置信度，根据阈值将检测框分为高置信度及低置信度检测框，对于置信度高于阈值的检测框，将其放入高置信度检测框队列中，对于置信度低于阈值的检测框，将其放入低置信度检测框队列中。在分离低置信度检测框和高置信度检测框后，采用卡尔曼滤波算法来预测目标在当前帧中的新位置。被追踪目标的相似性可通过IoU或Re-ID特征距离计算，即高置信度检测框和预测的距离。然后，采用匈牙利算法完成相似性匹配。追踪算法的伪代码如图 5所示。

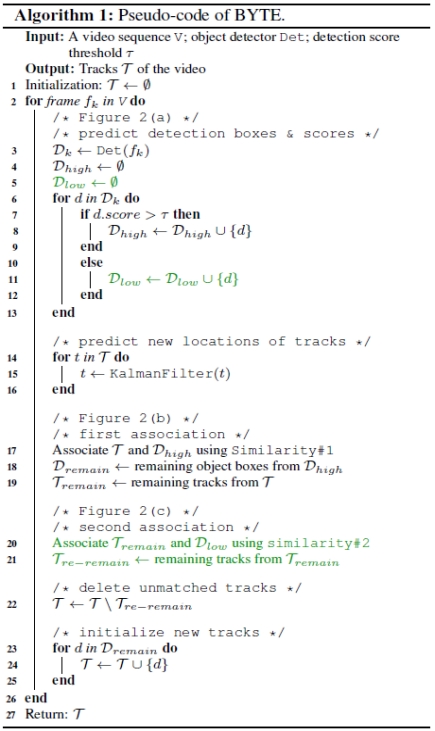


图 5追踪算法伪代码

具体来说，追踪算法的主要流程为:

1. 获取原始视频帧
2. 利用目标检测器对视频帧中的目标进行检测
3. 将检测到的目标的框中的特征提取出来，该特征包括表观特征（方便特征对比避免ID switch）和运动特征（运动特征方便卡尔曼滤波对其进行预测）
4. 计算前后两帧目标之前的匹配程度（利用匈牙利算法和级联匹配），为每个追踪到的目标分配ID。

3.1 基于卡尔曼滤波的多目标并行运动预测技术

卡尔曼滤波，也被称为线性二次估计，是一种使用一系列随时间观察到的测量结果，包括统计噪声和其他不准确因素，估计每个时间段的变量的联合概率分布，产生未知变量的估计值的算法，这些估计值往往比仅基于单一测量的变量更准确。这种数字滤波器有时被称为Stratonovich-Kalman-Bucy滤波器，该算法通过一个两阶段的过程工作。在预测阶段，卡尔曼滤波器产生对当前状态变量的估计，以及它们的不确定性。一旦观察到下一次测量的结果，这些估计值就会使用加权平均法进行更新，具有更大确定性的估计值会被赋予更多的权重。该算法是递归的，可以实时运行，只使用目前的输入测量值和之前计算的状态及其不确定性矩阵，不需要额外的过去信息。卡尔曼滤波的优化假定误差具有高斯分布。不管高斯性如何，如果过程和测量协方差是已知的，卡尔曼滤波器是最小均方误差意义上的最佳线性估计器。一个常见的误解是，除非所有的噪声过程被假定为高斯，否则卡尔曼滤波不能被严格地应用，其计算过程如图 6所示。

多目标跟踪算法中，卡尔曼滤波的主要作用就是当前的一系列运动变量去预测下一时刻的运动变量，但是第一次的检测结果用来初始化卡尔曼滤波的运动变量。

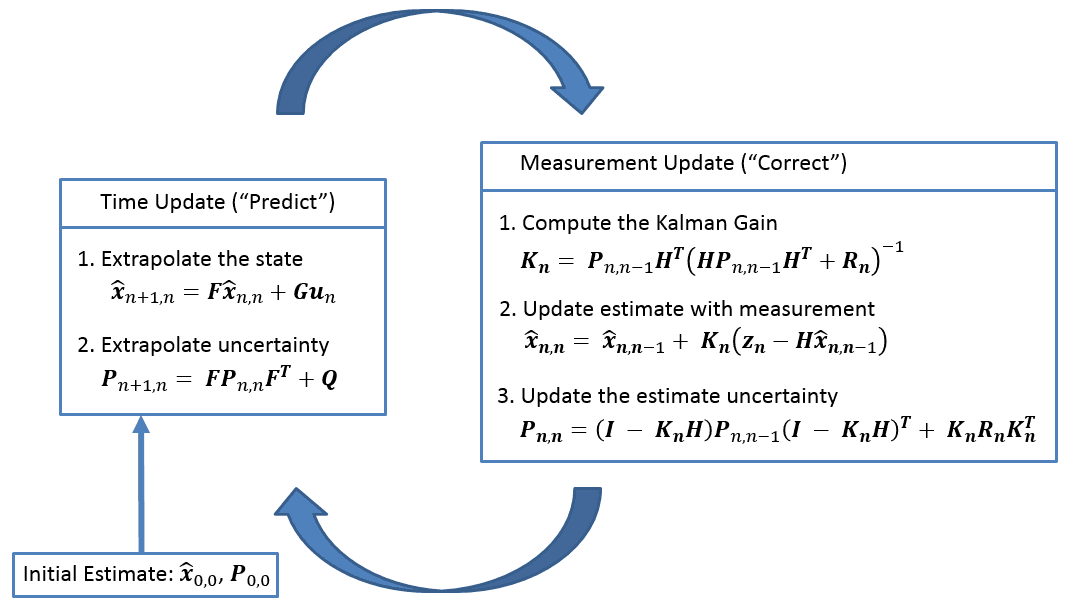


图 6 卡尔曼滤波器计算流程

3.2 基于匈牙利算法的多目标并行关联匹配技术

匈牙利匹配算法是一种组合优化算法，可在多项式时间内解决分配线性规划问题。解决最大成本分配问题可以转换为解决最小成本分配问题，反之亦然。假设成本矩阵是C，求解成本矩阵的最大成本分配问题C相当于求解成本矩阵的最小成本分配问题-C。成本矩阵通常不是正方形的。需要将成本矩阵变成正方形，并应用匈牙利算法求解最优成本分配问题，如图 7所示。匈牙利算法的计算过程如下：

1) 从该行中的所有项中减去每行中的最小值，使该行中的最小值等于0。

2) 从该列中的所有项中减去每列中的最小值，使该列中的最小值等于0。

3) 在包含0的行和列中绘制线条，使绘制的线条尽可能少。

4) 如果绘制了n条线，则可以进行最佳的零分配，并且算法完成。如果行数小于n，则尚未达到最佳的零分配，转到下一步。

5) 找到未被任何行覆盖的最小条目，从未被划掉的每一行中减去此条目，然后将其添加到被划掉的每一列中，返回步骤 3。

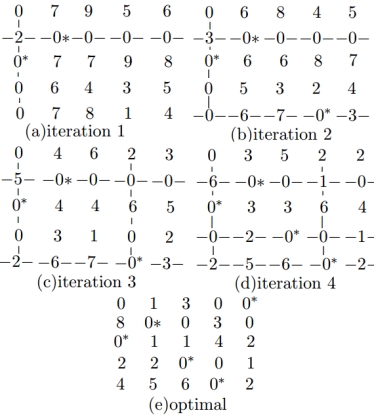


图 7匈牙利算法示例

多目标跟踪算法中，匈牙利算法的作用简单来讲即为解决分配问题，把图像帧中的检测框和卡尔曼预测的框做分配，让卡尔曼预测的框找到和自己最匹配的检测框，达到追踪的效果。

## （三）创新点

概括本项目的创新点，并简要说明。

1. 数据扩增算法研究

由于公开数据集样本及数量丰富性缺乏，研究人员拟采用虚实结合的数据扩增方法。在公开数据集的基础上，研究人员使用无人机拍摄了真实人体姿势图片，并在训练过程中加入数据增强策略，提高模型的准确性与泛化性。在公开数据集及仿真数据训练模型的基础上，利用实际拍摄的数据进行微调，可极大提高算法及模型对真实场景的适应能力，提高模型在多视角复杂姿态变化人员关重部位的精确识别能力。

1. 数据增强策略在模型训练过程中的运用

深度神经网络需要大量的训练数据才能获得良好的结果并防止过拟合。但是，通常很难获得足够的训练样本。多种原因可能使收集足够的数据变得非常困难甚至不可能。要创建训练数据集，需要先获取图像，并对其进行标记。对于该项目，需要标注薄弱部位的识别框，此过程需要人工，标记训练数据的成本非常高。对于该项目所面对场景，收集训练图像非常困难，因此，该项目采用数据增强扩增训练数据，提高模型的准确率及鲁棒性。数据增强是从现有样本创建新的训练样本的过程，通过变换原始图像获得新的数据。例如，通过水平翻转、裁剪、模糊、改变对比度、改变色调、改变饱和度、改变亮度、进行Gamma变换等方式得到新的图像，对一副输入图像做数据扩展的效果如图 2所示。



图 2 数据扩增示例

1. 多姿态运动人员关重部位识别算法研究

在设计出人员关重部位识别网络并运用数据集进行训练后，经过测试模型在单帧图像上推理时间达70+ ms，为了满足算法模型在轻量化开发板上实时性能的要求，研究人员开展了对检测算法的优化。

在识别网络中，网络最初引入了三种loss，即box\_loss（边界框损失）、cls\_loss(分类损失)、dfl\_loss(动态特征损失)。

其中，box\_loss是损失函数用于计算预测边界框与真实边界框之间的差异。识别网络使用IOU（Intersection over Union）作为度量，来衡量两个边界框之间的重叠程度。box\_loss通过计算预测框与真实框之间的IOU，来衡量预测框的位置准确度，并将其转化为一个损失值。通过最小化box\_loss，模型可以学习到更准确的边界框位置；cls\_loss（分类损失）用于计算预测类别与真实类别之间的差异。识别网络使用交叉熵损失（Cross Entropy Loss）来衡量分类准确度。cls\_loss通过比较预测类别分布与真实类别标签之间的差异，来计算分类的损失值。通过最小化cls\_loss，模型可以学习到更准确的类别分类；dfl\_loss（特征点损失）是识别网络中引入的自定义损失函数。dfl\_loss 引入了动态权重，以更好地平衡网络对小目标和大目标的重视程度。通常情况下，小目标由于其较小的尺寸和较低的像素值，容易被忽略或难以准确检测。dfl\_loss通过调整损失函数的权重，使得网络在训练过程中更关注小目标的检测，从而提高对小目标的检测性能。

在深度神经网络中，不同的任务可能需要不同的损失函数和模型结构。在某些情况下，为了特定的任务需求，则需要调整损失函数或模型结构，以取得更好的性能和效率。

在人体姿势估计中， 分类损失用于衡量模型对人体关节点所属类别的预测准确性。在人体姿势估计中，关节点的类别通常表示身体的不同部分，例如头、手、腿等。 准确的分类损失确保模型能够正确地识别人体的不同部分，为后续的关键点定位提供正确的上下文信息。错误的分类可能导致关节定位的混淆，影响最终的姿势估计准确性；边界框损失用于衡量模型对每个关节点边界框位置的预测准确性。在人体姿势估计中，每个关节点通常由一个边界框表示。精准的边界框损失确保模型能够准确地定位关节点的位置，提供了对人体姿势的空间信息。准确的位置信息是关键点定位和姿势估计的基础。动态特征损失（DFL Loss）引入了动态权重，以更好地平衡网络对小目标和大目标的重视程度。在目标检测中，这对于处理小目标非常重要。在人体姿势估计中，关节点通常并不算为小目标。因此，dfl\_loss对于处理小目标的优化作用相对有限。

研究人员针对实际应用场景，进行充分的实验以及损失函数调整后，确定在模型推理过程中，采用的损失函数组合为box\_loss与cls\_loss。经实验验证，模型最后在单张图片上推理时间为30ms，满足实时性的要求。