Tracking People by Predicting 3D Appearance, Location & Pose
通过预测 3d 外观、位置和姿势来跟踪人

Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Jitendra Malik
Jathushan Rajasegaran，Georgios Pavlakos，Angjoo Kanazawa，Jitendra Malik
UC Berkeley
加州伯克利分校

Abstract
摘要

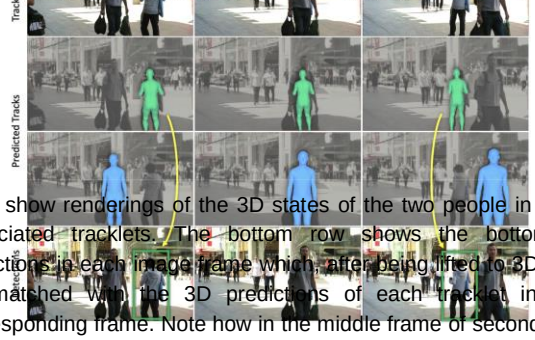
In this paper, we present an approach for tracking people in monocular videos, by predicting their future 3D repre-sentations. To achieve this, we first lift people to 3D from a single frame in a robust way. This lifting includes informa-tion about the 3D pose of the person, his or her location in the 3D space, and the 3D appearance. As we track a person, we collect 3D observations over time in a tracklet represen-tation. Given the 3D nature of our observations, we build temporal models for each one of the previous attributes. We use these models to predict the future state of the tracklet, including 3D location, 3D appearance, and 3D pose. For a future frame, we compute the similarity between the pre-dicted state of a tracklet and the single frame observations in a probabilistic manner. Association is solved with simple Hungarian matching, and the matches are used to update the respective tracklets. We evaluate our approach on vari-ous benchmarks and report state-of-the-art results.

在本文中，我们提出了一种通过预测人们未来的 3d 表象来跟踪单眼视频中的人们的方法。为了实现这一目标，我们首先以一种稳健的方式将人们从单一画面提升到 3d。这种提升包括人的 3d 姿势、他或她在 3d 空间中的位置以及 3d 外观的信息。当我们跟踪一个人的时候，我们通过轨迹表示法收集随着时间的推移而来的 3d 观察结果。鉴于我们观察的三维性质，我们为前面的每一个属性建立时间模型。我们使用这些模型来预测小径的未来状态，包括 3d 位置，3D 外观和 3d 姿态。对于未来的帧，我们以概率方式计算轨迹的预测状态与单帧观测之间的相似性。关联用简单的匈牙利匹配解决，匹配用于更新各自的小径。我们在各种基准上评估我们的方法，并报告最新的结果。

1. Introduction
1. 简介

When we watch a video, we can segment out individ-ual people, cars, or other objects and track them over time. The corresponding task in computer vision has been stud-ied for several decades now, with a fundamental choice be-ing whether to do the tracking in 2D in the image plane, or of 3D objects in the world. The former seems simpler be-cause it obviates the need for inferring 3D, but if we do take the step of back-projecting from the image to the world, other aspects such as dealing with occlusion become eas-ier. In the 3D world the tracked object doesn't disappear, and even young infants are aware of its persistence behind the occluder. A recent paper, Rajasegaran et al. [32] argues convincingly on the 3D side of this debate for people track-ing, and presents experimental evidence that indeed per-formance is better with 3D representations. In this paper, we will take this as granted, and proceed to develop a sys-tem in the 3D setting of the problem. While our approach broadly applies to any object category where parameterized 3D models are available and can be inferred from images, we will limit ourselves in this paper to studying people, the

当我们观看视频时，我们可以分割出个人，汽车或其他物体，并随着时间的推移跟踪他们。计算机视觉中的相应任务已经研究了几十年，其中一个基本的选择是在图像平面上进行二维跟踪，还是在世界上进行三维物体的跟踪。前者看起来更简单，因为它避免了推断 3d 的需要，但是如果我们真的采取从图像到世界的反向投影的步骤，其他方面，如处理遮挡变得更容易。在 3d 世界中，被跟踪的物体不会消失，甚至小婴儿都能意识到它在遮挡物后面的持久性。最近的一篇论文，Rajasegaran 等[32]令人信服地论证了这场关于人们跟踪的争论的三维方面，并提出了实验证据，证明三维表现确实更好。在这篇论文中，我们将认为这是理所当然的，并继续在问题的三维环境中开发一个系统。虽然我们的方法广泛适用于任何对象类别，参数化的三维模型是可用的，可以推断从图像，我们将在本文限制我们自己研究人，



rows show renderings of the 3D states of the two people in their associated tracklets. The bottom row shows the bottom-up detections in each image frame which, after being lifted to 3D, will be matched with the 3D predictions of each tracklet in the corresponding frame. Note how in the middle frame of second row, the 3D representation of the person persists even though he is occluded in the image. More videos at project site.

图 1。通过 3d 预测和匹配来跟踪人们。上面一行显示了我们在三个不同帧的跟踪结果。结果通过一个彩色的头部面具可视化，以获得独一无二的身份。第二行和第三行显示了两个人在相关小道中的 3d 状态的渲染图。下面一行显示了每个图像帧的自下而上的检测结果，在提升到 3d 后，将与相应帧中每个小径的 3d 预测结果相匹配。请注意，在第二行的中间帧中，即使人被遮挡在图像中，他的 3d 表示仍然存在。更多视频在项目网站。

Figure 1. Tracking people by predicting and matching in 3D. The top row shows our tracking results at three different frames. The results are visualized by a colored head-mask for unique identities. The second and third

most important case in practice.

实践中最重要的案例。

Once we have accepted the philosophy that we are tracking 3D objects in a 3D world, but from 2D images as raw data, it is natural to adopt the vocabulary from control the-ory and estimation theory going back to the 1960s. We are interested in the “state” of objects in 3D, but all we have access to are “observations” which are RGB pixels in 2D. In an online setting, we observe a person across multiple time frames, and keep recursively updating our estimate of the person’s state — his or her appearance, location in the world, and pose (configuration of joint angles). Since we have a dynamic model (a “tracklet”), we can also predict states at future times. When the next image frame comes in, we detect the people in it, lift them to 3D, and in that set-ting solve the association problem between these bottom-up detections and the top-down predictions of the different tracklets for this frame. Once the observations have been associated with the tracklets, the state of each person is re-estimated and the process continues. Fig. 1 shows this pro-

一旦我们接受了在三维世界中追踪三维物体，而将二维图像作为原始数据的哲学，就很自然地采用了可追溯到 20 世纪 60 年代的控制理论和参数估测理论的词汇。我们对 3d 物体的“状态”很感兴趣，但是我们所能获得的只是 2d 中的 RGB 像素的“观察”。在线环境中，我们观察一个人跨越多个时间框架，并不断递归地更新我们对该人状态的估计——他或她的外貌、世界位置和姿势(关节角度的配置)。由于我们有一个动态模型(“轨迹”)，我们也可以预测未来的状态。当下一个图像帧进来时，我们检测其中的人，将他们提升到 3d，在这个设置中解决了这些自下而上的检测和自上而下的预测之间的关联问题。一旦观察结果与踪迹相关联，每个人的状态就会被重新估计，这个过程就会继续下去。图 1 显示了这种支持

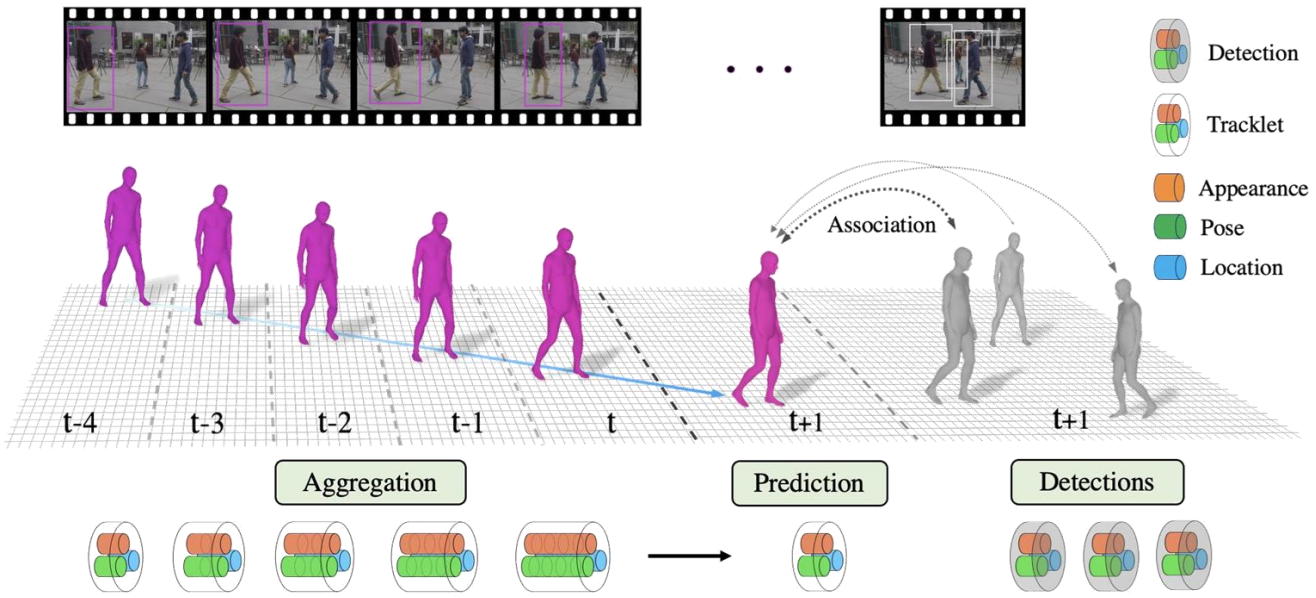


Figure 2. PHALP: Predicting Human Appearance, Location and Pose for Tracking: We perform tracking of humans in 3D from monocular video. For every input bounding box, we estimate a 3D representation based on the 3D appearance, 3D pose and 3D location of the person. During tracking, these are integrated to form corresponding tracklet-based representations. We perform tracking by predicting the future representation of each person and using it to solve for association given the detected bounding boxes of a future frame.

图 2。PHALP: 预测人类的外貌、位置和姿势进行跟踪: 我们通过单目视频在 3d 中进行人类的跟踪。对于每个输入边界框，我们基于人的 3d 外观，3D 姿势和 3d 位置估计一个 3d 表示。在跟踪期间，这些被整合以形成相应的基于轨迹的表示。我们通过预测每个人的未来表现来执行跟踪，并利用它来解决给定未来帧的检测到的边界框的关联。

cess at work on a real video. Note that during a period of occlusion of a tracklet, while no new observations are com-ing in, the state of the person keeps evolving following his or her dynamics. It is not the case that “Out of sight, out of mind”!

在一个真正的视频上工作。请注意，在一段时间的闭塞，一个小径，虽然没有新的观察到来，人的状态不断演变跟随他或她的动态。这并不是说“眼不见，心不烦”！

In an abstract form, the procedure sketched in the previous paragraph is basically the same as that followed in multiple computer vision papers from the 1980s and 1990s. The difference is that in 2021 we can actually make it work thanks to the advances brought about by deep learning and big data, that enable consistent and reliable lifting of peo-ple to 3D. For this initial lifting, we rely on the HMAR model [32]. This is applied on every detected bounding box of the input video and provides us with their initial, single frame, observations for 3D pose, appearance as well as location of the person in the 3D space.

以抽象的形式，上一段所描述的程序基本上与 1980 年代和 1990 年代的多篇计算机视觉论文所遵循的程序相同。不同之处在于，在 2021 年，由于深度学习和大数据带来的进步，我们实际上可以让它发挥作用，这些进步使人们能够一致而可靠地提升到 3d。对于这个初始的提升，我们依赖于 HMAR 模型[32]。这是应用于每个检测到的边界框的输入视频，并提供了他们的初始，单帧，三维姿态的观察，外观和位置的人在三维空间。

As we link individual detections into tracklets, these representations are aggregated across each tracklet, allowing us to form temporal models, i.e., functions for the aggre-gation and prediction of each representation separately (see left side of Fig. 2). More specifically, for appearance, we use the canonical UV map of the SMPL model to aggregate appearance, and employ its most recent version as a prediction of a person’s appearance. For pose, we aggregate information using a modification of the HMMR model [15], where through its “movie strip” representation, we can pro-duce 3D pose predictions. Finally, for 3D location, we use linear regression to predict the future location of the person.

当我们将单个检测连接到小径时，这些表示在每个小径上聚合，允许我们形成时间模型，也就是说，分别用于聚合和预测每个表示的函数(见图 2 的左侧)。更具体地说，对于外观，我们使用 SMPL 模型的规范 UV 映射来聚合外观，并使用其最新版本来预测一个人的外观。对于姿态，我们使用 HMMR 模型[15]的修正来聚合信息，通过它的“电影带”表示，我们可以产生三维姿态预测。最后，对于三维位置，我们使用线性回归来预测人的未来位置。

This modeling enables us to develop our tracking system, PHALP (Predicting Human Appearance, Location and Pose for tracking), which aggregates information over time, uses it to predict future states, and then associates the predictions with the detections. First, we predict the 3D location, 3D pose and 3D appearance for each tracklet for a short period of time (right side of Fig. 2). For a future frame, these pre-dictions need to be associated with the detected people of the frame. To measure similarity, we adopt a probabilistic interpretation and compute the posterior probabilities of ev-ery detection belonging to each one of the tracklets, based on the three basic attributes. With the appropriate similarity metric, association is then easily resolved by means of the Hungarian algorithm. The newly linked detections can now update the temporal model of the corresponding tracklets for 3D pose, 3D appearance and 3D location in an online manner and we continue the procedure by rolling-out fur-ther prediction steps. The final output is an identity label for each detected bounding box in the video. Notably, this approach can also be applied on videos with shot changes, e.g., movies [10], with minor modifications. Effectively, we only modify our similarity to include only appearance and 3D pose information for these transitions, since they (unlike location) are not affected by the shot boundary.

这个模型使我们能够开发我们的跟踪系统，PHALP (预测人类外观，位置和姿势进行跟踪)，它随着时间聚集信息，使用它来预测未来的状态，然后将预测与检测联系起来。首先，我们在短时间内预测每个小径的 3d 位置，3D 姿势和 3d 外观(图 2 右侧)。对于未来的框架，这些预测需要与框架的检测人员相关联。为了度量相似性，我们采用了一种概率解释方法，并基于三个基本属性计算了每个小径的每个检测的后验概率。使用适当的相似度量，然后通过匈牙利算法轻松解决关联。新的连接检测方法可以在线更新相应轨迹的三维姿态、三维外观和三维定位的时间模型，并通过进一步的预测步骤来继续该过程。最终的输出是视频中每个检测到的边界框的标识标签。值得注意的是，这种方法也可以应用于镜头改变的视频，例如电影[10]，稍作修改。有效地，我们只修改我们的相似性只包括外观和三维姿态信息为这些过渡，因为他们(不同于位置)不受镜头边界的影响。

2. Related work
2. 相关工作

Tracking. Object tracking is studied in various settings such as single object tracking, multi-object tracking for hu-

跟踪目标跟踪研究在不同的设置，如单个目标跟踪，多个目标跟踪

mans, and multi-object tracking for vehicles etc. The track-ing literature is vast and we refer readers to [6, 7, 42] for a comprehensive summary. In general tracking can be de-signed for any generic category, however, in this section we discuss the methods that focus on tracking humans. These approaches mostly work in a tracking by detection setting, where 2D location, key-point features [9,34,38] and 2D ap-pearance [4, 29, 40, 41] is used to associate detections over time. Quality of the detection plays a key role in tracking-by-detection setting and many works jointly learn or fine-tune their own detection models [4, 29]. In this work, we are interested in the effectiveness of 3D representations for tracking and thus assume that detection bounding boxes are provided, which we associate through our representations. On the other hand, tracking by regression predicts future lo-cations using the knowledge of the past detections. While this alleviate the requirement for good quality detections, most of the works regress in the image plane. The pro-jection from 3D world to the image plane makes it hard to make this prediction, therefore these methods need to learn non-linear motion models [2, 4, 45]. Compared to these methods, PHALP predicts short-term location in 3D coor-dinates, by simple linear regression. Additionally, we also predict appearance and pose features for better association.

人，多目标跟踪车辆等。跟踪文献是庞大的，我们提请读者参阅[6,7,42]以获得全面的总结。一般来说，跟踪可以针对任何通用类别进行设计，然而，在这一节中，我们讨论了关注于跟踪人的方法。这些方法主要工作在通过检测设置跟踪，其中 2d 位置，关键点特征[9,34,38]和 2d 外观[4,29,40,41]是用来关联检测随着时间的推移。检测的质量在检测跟踪设置中起着关键作用，许多工作共同学习或微调自己的检测模型[4,29]。在这项工作中，我们感兴趣的是有效的三维表示的跟踪，因此假设检测包围盒提供，我们关联通过我们的表示。另一方面，回归跟踪利用过去探测的知识预测未来的位置。虽然这减轻了对高质量探测的需求，但大多数工作在图像平面上是倒退的。从三维世界到图像平面的投影使得这种预测变得困难，因此这些方法需要学习非线性运动模型[2,4,45]。与这些方法相比，PHALP 通过简单线性回归预测三维坐标中的短期位置。此外，我们还预测外观和姿势特征以更好地关联。

Finally, there are methods that incorporate 3D informa-tion in tracking, however these approaches assume multiple input cameras [24, 44] or 3D point cloud observation from lidar data [37]. In this paper we focus on the setting where the input is a monocular video. Some recent works tracks occluded people based on the object permanence [17, 35]. These methods learn complex functions to predict the lo-cations of occluded people. However, by placing humans in 3D space and predicting their location, pose and appear-ance, object permanence is already built into our system.

最后，还有一些方法将三维信息纳入跟踪，但这些方法假设多个输入相机[24,44]或从激光雷达数据进行三维点云观测[37]。在本文中，我们关注输入是单目视频的设置。最近的一些工作轨迹基于物体永久性遮挡了人们[17,35]。这些方法学习复杂的函数来预测被遮挡人的位置。然而，通过将人类置于三维空间并预测他们的位置、姿势和外观，物体的持久性已经被植入我们的系统。

Monocular 3D human reconstruction. Although there is a long history of methods for 3D human reconstruction from monocular images, e.g., [5, 11], here we focus on more recent works. Many of the relevant approaches rely on the SMPL model [25], which offers a low dimensional parameterization of the human body. HMR [14] has been one of the most notable ones, using a neural network to regress the parameters of a SMPL body from a single image. Follow-up works have improved the robustness of the origi-nal model [19, 22], or added additional features like estima-tion of camera parameters [20], or probabilistic estimation of pose [23]. Recently, Rajasegaran et al. [32] introduced HMAR, by extending the model with an appearance head. Other works have focused on extending HMR to the tempo-ral dimension, e.g., HMMR [15], VIBE [18], MEVA [27] and more [31]. In this work, we make use of a modifica-tion of the HMAR model [32] as the main feature backbone, while also employing a model that follows the HMMR prin-

单目三维人体重建。虽然从单目图像进行三维人体重建的方法有着悠久的历史，例如，[5,11] ，但在这里我们关注更多的近期作品。许多相关的方法依赖于 SMPL 模型[25] ，它提供了人体的低维参量化。HMR [14]是最值得注意的一个，使用神经网络从单个图像回归 SMPL 身体的参数。后续工作提高了原始模型的稳健性[19,22] ，或增加了额外的特征，如摄像机参数的估计[20] ，或姿态的概率估计[23]。最近，Rajasegaran 等[32]通过扩展模型和外观头来引入 HMAR。其他工作集中于将 HMR 扩展到节奏维度，例如 HMMR [15]、 VIBE [18]、 MEVA [27]等等。在这项工作中，我们使用了一个修改的 HMAR 模型[32]作为主要的特征骨干，同时也采用了一个模型，遵循 HMMR 印刷

ciples [15] for temporal pose prediction, but instead, using a transformer [36] to aggregate pose information over time. Regarding human motion prediction, Kanazawa et al. [15], regress future poses from the temporal pose representation of HMMR, the “movie-strip”. Zhang et al. [43] extend this to PHD, employing autoregressive prediction of human mo-tion. Aksan et al. [1] also regress future human motion in an autoregressive manner, using a transformer.

Ciples [15]用于时间姿态预测，而是使用变压器[36]随时间聚集姿态信息。关于人体运动预测，Kanazawa 等[15]，从 HMMR (“电影带”)的时间姿态表示回归未来的姿态。Zhang 等[43]将其扩展到 PHD，采用人体运动的自回归预测。Aksan 等[1]也使用变压器以自回归的方式回归未来的人类运动。

3. Method
3. 方法

Tracking humans using 3D representations has significant advantages, including that appearance is independent of pose variations and the ability to have amodal completion for humans during partial occlusion. Our tracking algorithm accumulates these 3D representations over time, to achieve better association with the detections. PHALP has three main stages: 1) lifting humans into 3D representations in each frame, 2) aggregating single frame representations over time and predicting future representations, 3) associ-ating tracks with detections using predicted representations in a probabilistic framework. We explain each stage in the next sections.

使用三维表示追踪人类有显著的优势，包括外观独立于姿势变化和能力有模态完成的人类在部分遮挡。我们的跟踪算法随着时间的推移积累这些 3d 表示，以实现与检测更好的关联。PHALP 有三个主要阶段: 1)将人类提升为每帧中的三维表示; 2)随着时间的推移聚合单帧表示并预测未来表示; 3)在概率框架中使用预测表示将轨迹与检测联系起来。我们在下一节中解释每个阶段。

3.1. Single-frame processing
3.1 单帧处理

The input to our system is a set of person detections along with their estimated segmentation masks, provided by conventional detection networks, like Mask-RCNN [12]. Each detection is processed by our feature extraction back-bone that computes the basic representations for pose, ap-pearance and location on a single-frame basis. For this feature extraction we use a modification of the HMAR model [32]. HMAR returns a feature representation for the 3D pose p, for appearance a, while it can recover an esti-mate for the 3D location l for the person.

我们的系统的输入是一组人的检测和他们的估计分割掩码，由传统的检测网络，如 Mask-RCNN [12]提供。每一个检测是由我们的特征提取骨干处理，计算基本表示的姿态，外观和位置在一个单一的帧的基础上。对于这个特征提取，我们使用 HMAR 模型的修改[32]。HMAR 为三维姿态 p 返回一个特征表示，为外观 a，而它可以恢复一个人的三维位置 l 的估计。

The standard HMAR model takes as input the pixels in the bounding box corresponding to a detected person. This means that in a crowded, multi-person scenario, the input will contain pixels corresponding to more than one person in the bounding box, potentially confusing the network. To deal with this problem, we modify HMAR to take as additional input, the pixel level mask of the person of inter-est (this is readily available as part of the output of Mask R-CNN) and re-train HMAR. Obviously, we cannot expect this step to be perfect, since there can be inaccuracies in the bounding box detections or mask segmentations. However, we observed that the model gives more robust results in the case of close person-person interactions, which are common in natural videos.

标准的 HMAR 模型将对应于被检测人的边界框中的像素作为输入。这意味着在拥挤的多人场景中，输入将包含对应于边界框中多个人的像素，从而可能混淆网络。为了解决这个问题，我们修改了 HMAR 作为额外的输入，利益相关者的像素级掩码(这很容易作为 Mask R-CNN 输出的一部分)和重新训练 HMAR。显然，我们不能期望这个步骤是完美的，因为在边界框检测或蒙版分割中可能会有不准确的地方。然而，我们观察到这个模型在亲密的人与人的互动中给出了更强有力的结果，这在自然视频中是很常见的。

3.2. 3D tracklet prediction
3.2.3D 轨迹预测

The 3D estimates for each detection provide a rich and expressive representation for each bounding box. However,

每个检测的三维估计为每个包围盒提供了一个丰富而富有表现力的表示,

they are only the result of single-frame processing. Dur-ing tracking, as we expand each tracklet, we have access to more information that is representative of the state of the tracklet along the whole trajectory. To properly lever-age this information, our tracking algorithm builds a track-let representation during every step of its online processing, which allows us to also predict the future states for each tracklet. In this section we describe how we build this track-let representation, and more importantly, how we use it to predict the future state of each tracklet.

它们只是单帧处理的结果。在跟踪过程中，当我们扩展每个小径时，我们可以获得更多的信息，这些信息代表了小径沿着整个轨道的状态。为了正确地利用这些信息，我们的跟踪算法在其在线处理的每一个步骤中都建立了一个 track-let 表示，这使得我们能够预测每个 track-let 的未来状态。在本节中，我们将描述如何构建这个 track-let 表示，更重要的是，我们如何使用它来预测每个 track-let 的未来状态。

Appearance: The appearance pathway is used to inte-grate appearance information for each person over multi-ple frames. The single frame appearance representation for the person i at time step t, A^i_t, is taken from the HMAR model by combining the UV image of that per-son T^i_t 2 R^3 256 256 and the corresponding visibility map V_t^i 2 R^1 256 256 at time step t:

外观: 外观路径用于在多帧中整合每个人的外观信息。利用 HMAR 模型，结合 ti2r3256256 的紫外图像和相应的时间步长 t 的能见度图 vti2r1256256，得到了人在时间步长 t 的单帧外观表示:

$$A^i_t = [T^i_t; V_t^i] \ 2 \ R^4 \ 256 \ 256$$

Ait = [Tit; Vti]2 r4256256

Note that the visibility mask V_t^i 2 [0; 1] indicates whether a pixel in the UV image is visible or not, based on the esti-mated mask from Mask-RCNN. Now, if we assume that we have established the identity of this person in neighboring frames, we can integrate the partial appearance information coming from the independent frames to an overall tracklet appearance for the person. Using the set of single frame ap-pearance representations A^i = fA^i_t; $A^i_{t\ 1}$; $A^i_{t\ 2}$; :::g, after every new detection we create a singe per-tracklet appear-ance representation:

注意，可见性掩模 vti2[0; 1]基于 Mask-RCNN 估计的掩模，指示紫外图像中的像素是否可见。现在，如果我们假设我们已经在相邻帧中建立了这个人的身份，我们可以将来自独立帧的部分外观信息整合成这个人的整体轨迹外观。使用一组单帧外观表示 Ai = fAit; Ait 1; Ait 2; : : : g，在每次新的检测之后，我们创建一个单一的每个小径外观表示:

$$A^i_t = A(A^i_{t\ 1}; A^i_t) = (1\ ) \ A^i_{t\ 1} + \ A^i_t$$

i = a (Ati 1; Ati) = (1) Ati 1 + Ati

b 80b; if $V_{t\ 1}$ = 1 and = 1

0b; 如 Vt = 1 Vbt = 1

还有

Vbt

0; if = 1 and V = 0:

<0; 如 V = 1 和 v = 0:

bt

Bt 1 t

: b

Here, A is the appearance aggregation function, which takes a weighted sum of the previous tracklet appearance representation and the new detection appearance represen-tation. Note that, at the start of the tracklet we simply assign the initial single-frame representation to the tracklet repre-sentation ($Ab^i_0 = A^i_0$). With this definition of A, we can aggregate appearance information over time, while allowing the representation to change slowly to account for slight ap-pearance changes of the person during a video. Moreover, the UV image provides appearance of each point on the body surface independently of body pose and shape which enables the simple summation operation on the pixel space, without any learnable components. Figure 3 shows how the UV image of the person is aggregated over time and used for association of new detections.

在这里，a 是外观聚合函数，它采用以前的跟踪外观表示和新的检测外观表示的加权和。注意，在跟踪程序的开始，我们只是简单地将初始的单帧表示赋值给跟踪程序的表示(Abi0 = Ai0)。根据 a 的定义，我们可以随着时间的推移聚合外观信息，同时允许表示慢慢改变，以解释视频中人的轻微外观变化。此外，紫外图像提供了身体表面上的每个点的外观独立于身体姿态和形状，使得像素空间上的简单求和操作，没有任何可学习的组成部分。图 3 显示了这个人的紫外线图像是如何随着时间的推移聚合并用于新的检测的关联。

For appearance prediction, we make the realistic as-sumption that human appearance will not change rapidly

对于外貌预测，我们假设人类的外貌不会迅速改变

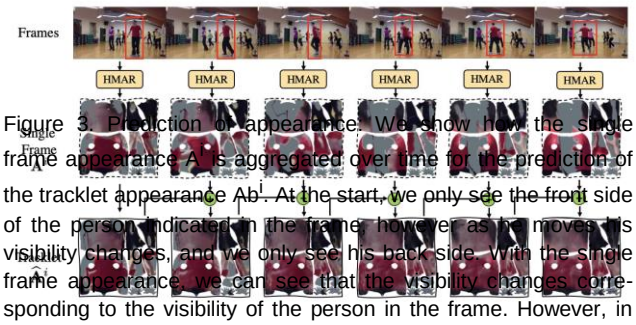


Figure 3. Prediction of appearance. We show how the single frame appearance A_i is aggregated over time for the prediction of the tracklet appearance Ab^i. At the start, we only see the front side of the person indicated in the frame, however as he moves his visibility changes, and we only see his back side. With the single frame appearance, we can see that the visibility changes corresponding to the visibility of the person in the frame. However, in the tracklet, the appearance is accumulated over time, and even if the front side is not visible in the last frame, we can see that the tracklet has predicted these regions using the past frames.

图 3。外观预测: 我们展示了单帧外观 Ai 如何随着时间的推移聚合来预测小径外观 Abi。一开始，我们只能看到画面中人物的正面，但是随着他移动的视野改变，我们只能看到他的背面。通过单帧的外观，我们可以看到能见度的变化与画面中人物的能见度相对应。然而，在跟踪程序中，外观是随着时间累积的，即使在最后一帧中正面不可见，我们也可以看到跟踪程序使用过去的帧预测了这些区域。

over time. Then, the appearance of the tracklet Ab^i_t can func-tion as a reasonable prediction for the future appearance of the person. Therefore, we use Ab^i_t as the prediction for ap-pearance and use it to measure similarity against a detection in the future frames.

随着时间的推移。然后，小径的出现可以作为一个人未来外貌的合理预测。因此，我们使用 bit 作为外貌的预测，并使用它来衡量相似性与未来帧中的检测。

Location: Lifting humans from pixels into the 3D space allows us to place them in the global 3D location. Let us as-sume that a person i at time t has an estimated 3D location l^i_t. Although, we can get an estimate for the location of the person in the global camera frame, this tends to be noisy, particularly along the z-axis. To avoid any instabilities when it comes to predicting future location, instead of per-forming our prediction on the Euclidean $(X; Y; Z)^T$ space, we express our locations in an equivalent $l^i_t = (x; y; n)^T$ space where (x; y) is the location of the root of the person in the pixel space and n is nearness, defined as log inverse depth n = log(1=z). Nearness is a natural parameterization of depth in multiple vision settings, e.g., [21], because of the 1=z scaling of perspective projection. In our case it cor-responds to the scale of the human figures that we estimate directly from images. We independently linearly regress the location predictions for x; y and n. This is somewhat like the Constant Velocity Assumption (CVA) used in past track-ing literature, but there is a subtlety here because constant velocity in 3D need not give rise to constant velocity in 2D (a person would appear to speed up as she approaches the camera). But local linearization is always a reasonable approximation to make, which is what we do.

位置: 提升人类从像素到 3d 空间允许我们将他们放置在全球 3d 位置。让我们假设一个人在时间 t 上有一个估计的 3d 位置点亮。虽然我们可以估计出人在全球相机框架中的位置，但这往往是噪声，特别是在 z 轴方向。为了避免在预测未来位置时出现任何不稳定性，我们不在 Euclidean (x; y; z) t 空间上进行预测，而是在等价的 lit = (x; y; n) t 空间中表示我们的位置，其中(x; y)是人在像素空间中的根的位置，n 是接近度，定义为对数反向深度 n = log (1 = z)。由于透视投影的 1 = z 比例，贴近度是多种视觉设置中深度的自然参量化，例如[21]。在我们的情况下，它对应于我们直接从图像估计的人物的尺度。我们独立地线性回归 x; y 和 n 的位置预测。这有点像过去跟踪文献中使用的恒定速度假设(CVA)，但这里有一个微妙之处，因为 3d 中的恒定速度不一定会导致 2d 中的恒定速度(一个人在接近摄像机时似乎会加速)。但是局部线性化总是一个合理的近似值，这就是我们所做的。

Let us assume that a tracklet has a set of past locations $L^i = fl^i_t; l^i_{t\,1}; l^i_{t\,2}; :::g$. Then, the prediction of the loca-

让我们假设一个跟踪具有一组过去的位置 Li = flit; lit 1; lit 2; : : : g

tion for time step t + 1 is given by:
T + 1 的时间步骤由以下人员给出:

lti+1 = (xti+1; yti+1; nti+1)T
Lti + 1 = (xti + 1; yti + 1; nti + 1) t

where, 在哪里, bt+1 Bt + 1 1 t 2 一平 f t t F t t 二 w g W g t ; :::; xi

xi = (xi; xi ; xi * * * ; t + 1):
Xi = (xi; xi) ; xi xi ; t + 1):

Here, 给你, L b L b is the location aggregation function and we use a 是位置聚合函数，我们使用

simple linear regression for prediction in our tracking algo-
简单线性回归预测在我们的跟踪算法
rithm. yti+1 and nti +1 are also predicted in a similar fashion.
Yti + 1 和 nti + 1 也以类似的方式被预测。

L takes the last w observations to fit a line by least squares regress the future location for x; y and n independently.
L 用最后的 w 观测值拟合一条直线，用最小二乘法独立地回归 x; y 和 n 的未来位置。

From the standard theory of linear regression, the prediction interval for x at a time step t^0 is given by the equation
从标准线性回归理论出发，利用该方程给出了 x 在时间步长 t0 上的预测区间
below: 下面:

x(t0) = t(1 =2) s MSE 1 + w + (t0 t)2 / P (t
X (t0) = t (1 = 2) s MSE 1 + w + (t0 t)2 1 (t)

Here, t(1 =2) is the Student's t distribution with confidence and degree of freedom w 2. M SE is the mean
在这里，t (1 = 2)是带有置信度和自由度 w 2 的 Student's t 分布。M SE 是平均值
squared error on the predicted locations and t is the mean of the time stamps for the previous observations. In a similar manner, we can compute prediction intervals y; n for y and n respectively.
预测位置的平方误差，t 是以前观测的时间戳的平均值。以类似的方式，我们可以分别计算 y 和 n 的预测区间 y;。

Pose: For the pose pathway, we need to integrate pose information across the tracklet and be able to predict future poses for the near future. To do this, we borrow ideas from the HMMR architecture [15]. Effectively, we learn a function P that takes as input a series of pose embeddings of a person Pi = fpit; pit 1; pit 2; :::g and com-putes a temporal pose embedding pt. We train this tem- poral pose aggregation function P to smooth the pose the future). We use a transformer [36] to compute . This choice allows for some additional flexibility, since some-times we are not able to detect an identity in some frames (e.g., due to occlusions), which can be handled gracefully by the transformer, by masking out the attention for the representation of the corresponding frame.
姿态: 对于姿态路径，我们需要整合穿过小径的姿态信息，并能够预测不久的将来的未来姿态。为了做到这一点，我们借鉴了 HMMR 架构的想法[15]。有效地，我们学习了一个函数 p，它以一个人 Pi = fpit; pit 1; pit 2; : : g 的一系列姿态嵌入为输入，并计算了一个时间姿态嵌入 pt。我们训练这个临时姿态聚合函数 p 来平滑未来的姿态)。我们使用变压器[36]来计算。这种选择允许一些额外的灵活性，因为有时我们无法在某些帧中检测到一个身份(例如，由于遮挡)，这些身份可以通过掩盖对相应帧的表示的注意而被变压器优雅地处理。

at frame bt, and regress the future pose
在帧，并回归 未来的姿态
fpti+1; pti +2; :::; pti +cg (typically for up to c = 12 frames in
Fpti + 1; pti + 2; : : ; pti + cg (通常最多 c = 12 帧

3.3. Tracking with predicted 3D representations
3.3. 用预测的 3d 表示进行跟踪

Given the bounding boxes and their single-frame 3D representations, our tracking algorithm associates identities across frames in an online manner. At every frame, we make future predictions for each tracklet and we measure the similarity with the detected single-frame representation. More specifically, let us assume that we have a tracklet T_i, which has been tracked for a sequence of frames and has information for appearance, pose and location. The tracklet predicts its appearance Ab, location bland pose pb for the next frame, and we need to measure a similarity score between these predictions of the track T_i and a detection D_j to make an association. Our tracklet representation has three differ-ent attributes (appearance, location and pose), so, directly

给出了包围盒和它们的单帧三维表示，我们的跟踪算法关联的身份跨帧在线方式。在每个帧中，我们对每个跟踪进行未来预测，并用检测到的单帧表示来测量相似性。更具体地说，让我们假设我们有一个跟踪 Ti，它已经被跟踪了一系列的帧，并且有外观、姿势和位置的信息。轨迹预测它的外观 Ab，位置平淡的姿态 pb 为下一帧，我们需要测量这些预测之间的相似性评分轨迹 Ti 和检测 Dj 作出关联。我们的小径表示有三个不同的属性(外观，位置和姿势)，所以，直接

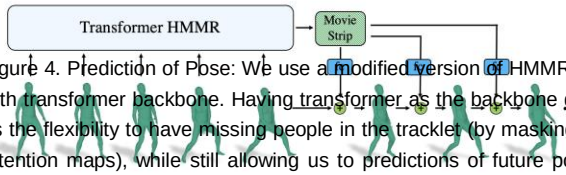


Figure 4. Prediction of Pose: We use a modified version of HMMR [15] with transformer backbone. Having transformer as the backbone gives us the flexibility to have missing people in the tracklet (by masking the attention maps), while still allowing us to predictions of future poses. Finally, the transformer give us a movie-strip representation and that is used to regress future poses.

图 4。姿态预测：我们使用带有变压器主干的 HMMR [15]的修改版本。有了变压器作为骨干，我们就可以灵活地把失踪的人放在轨迹中(通过屏蔽注意力地图)，同时仍然允许我们预测未来的姿势。最后，变压器给我们一个电影带的表示，用来回归未来的姿势。

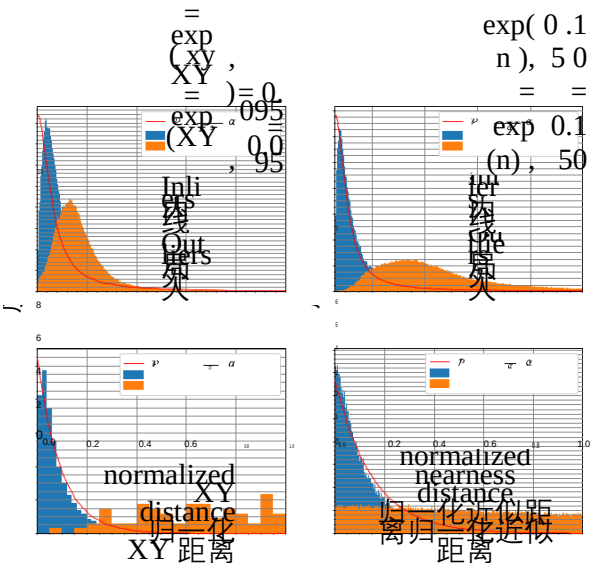
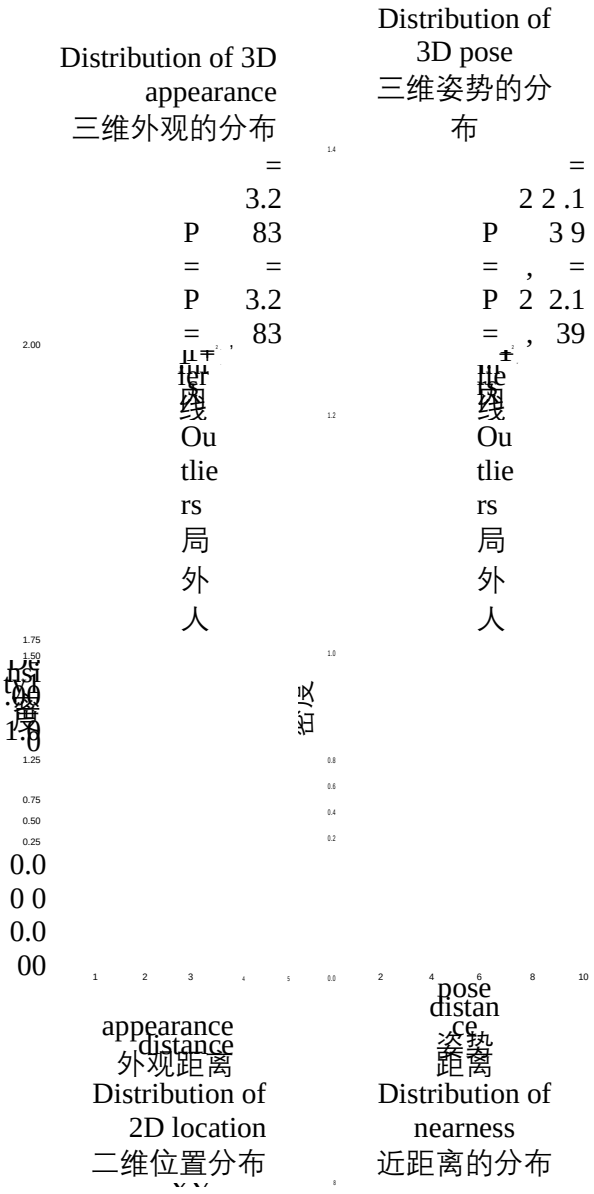


Figure 5. Conditional distributions of the attribute distances: We plot the data for the distances between the tracklet prediction and the single frame detection using the ground truth data from PoseTrack [3]. The curves show how the correct matches (inliers) and the incorrect matches (outliers) are distributed. Note that, for 2D location and nearness we plot the distances normalized by the prediction interval.

图 5。属性距离的条件分布：我们使用来自 PoseTrack [3]的地面真实数据绘制跟踪预测和单帧检测之间的距离数据。曲线显示了正确匹配(内部值)和不正确匹配(异常值)是如何分布的。注意，对于二维位置和接近度，我们绘制由预期区间归一化的距离。



combining their similarities/distances would not be ideal, since, each attribute has different characteristics. Instead, we investigate the conditional distributions of inliers and outliers of the attributes. Figure 5 presents the corresponding probability distributions for the PoseTrack dataset [3]. The characteristics of these distributions motivate our design decisions for our further modeling.

将它们的相似性/距离结合起来并不理想，因为每个属性都有不同的特征。相反，我们调查的条件分布的内因和异常值的属性。图 5 显示了 PoseTrack 数据集的相应概率分布[3]。这些分布的特征激发了我们进一步建模的设计决策。

Assume that tracklet T_i has an appearance representa-tion Ab^i_t. On the detection side, the detection D_j has a single-frame appearance representation A^j_t. Both of these representations are in the pixel space, therefore we first encode them into an embedding space using the HMAR appearance-encoder network. This gives us an appearance embedding ai and aj for the prediction of the tracklet Ti

假设小径 Ti 具有外观表示位。在检测方面，检测 Dj 有一个单帧外观表示 Ajt。这两种表示都在像素空间中，因此我们首先使用 HMAR 外观编码器网络将它们编码到嵌入空间中。这给我们提供了一个外观嵌入 ai 和 aj 来预测小径 Ti

b_t t

不行

and detection D_j, respectively. We are interested in estimat-ing the posterior probability of the event where the detection D_j belongs to the track T_i, given some distance measure of
和检测 Dj，分别。我们感兴趣的是估计事件的后验概率，其中检测 Dj 属于轨道 Ti，给出了一些距离测度的轨道 Ti

the appearance feature (a). Assuming that the appearance
外观特征(a)假设外观

distance is $a = jja^{i}_{t} \quad a^{j}_{tjj}{}^{2}{}_{2}$, then the posterior probability is proportional to the conditional probability of the appearance distances, given correct assignments based on Bayes rule. We model this conditional probability as a Cauchy distribution, based on the observations from the inlier dis-tribution of appearance distances (see also Fig 5):
距离为 a = jjait ajtjj22，则后验概率与出现距离的条件概率成正比，给出了基于贝叶斯规则的正确分配。我们将这种条件概率建模为柯西分布，基于对外观距离内在分布的观察(另见图 5)：

$$P_{A}(D_{j}\ 2\ T_{i}jd = \quad) / \frac{1}{1 + \quad a \quad a}$$

PA (Dj 2 Tijda = a)/1 + a

The distribution has one scaling hyper-parameter a. Similarly, for pose, we use Cauchy distribution to model
该分布有一个标度超参数 a。同样，对于姿态，我们使用柯西分布来建模

the conditional probability of inlier distances. We measure
内部距离的条件概率。我们测量

pose distance $p = jjp^{i}_{t} \quad p^{j}_{tjj}{}^{2}{}_{2}$ between the predicted pose representation pit from the track Ti and the pose represen-
姿态距离 p = jjpit pjtjj22 在来自轨道 Ti 的预测姿态表示坑和姿态表示坑之间

tation p_{t} of detection D_{j}. The posterior probability that the detection belongs to the track, given the pose distance is:
鉴于位置距离，检测属于轨道的后验概率是：

$$PP(Dj\ 2\ Tijdp = p) / \frac{1}{1 + p\ p}$$

PP (Dj 2 Tijdp = p)/ 1 + p p

Here, p = jjpti ptjjj22 and p is the scaling factor.
Here，p = jjpti 这里，p = jjpti Ptjjj22 和 p 是比例因子。

For location, let us assume the track Ti has predicted a location b(xti; yti; nti)T with a set of prediction lti = B (xti; yti; nti) t 与一组预测之间

对于位置，让我们假设轨道 Ti 预测了一个位置

$$f\ xjb \quad gj \quad y\ j\ N$$

$$j \quad F\ xjbY\ jgj_{T}.b\ b \quad j$$

$$t \quad t \quad t \quad \text{核} \quad B\ b$$

vals; ; , and the detection D is at a 3D location 瓦尔斯；探测 d 在一个 3d 位置

l = (x ; y ; n) We treat the 3D coordinates x; y and l = (x; y; n) 我们处理三维坐标 x; y 和

the nearness term n coordinates independently, and compute the posterior probabilities of the detection belongs to the tracklet given the location distance. We model the conditional probability distribution as an exponential distribu-tion, based on the findings from the empirical data. The Fig 5 shows the distribution of 2D distance and nearness distance, scaled by the confidence interval, of inliers ap-proximately follow the exponential distribution.
近似项 n 独立坐标，并计算给定位置距离的后验概率属于小径。我们根据经验数据的发现，将条件概率分布建模为指数分布。图 5 显示了二维距离和接近距离的分布，以置信区间为尺度，指数分布近似服从指数分布。

$$PXY(Dj\ 2\ Tijdxy = \ xy) / \ xy \quad \text{exp 译者} \quad xy\ xy$$

PXY (Dj 2 Tijdxy = xy)/xy 注: Xy xy

$$\frac{1}{\quad} \qquad \frac{\text{xy}}{: xy}$$

Here, xy is a scaling parameter for the exponential distri-bution, xy is the 2D pixel distance between the predicted
在这里，xy 是指数分布的比例参数，xy 是预测值之间的 2d 像素距离

track and the detection and $xy = \sqrt{x^{2} + y^{2}}$ is the predic-tion interval for the 2D location prediction. We use a similar form of exponential distribution for the posterior probability for nearness P_{N} :
跟踪和检测，xy = x2 + y2 是 2d 位置预测的预测区间。我们对接近 PN 的后验概率使用类似形式的指数分布：

$$PN(Dj\ 2\ Tijdn = n) / n \quad \text{exp 译者} \quad nn$$

PN (Dj 2 Tijdn = n)/n 注: 没有

$$\frac{1}{\quad} \qquad \frac{\text{n}}{\quad}$$

Here, n is the scaling parameter for the exponential distri-bution, n is the confidence interval for the nearness predic-tion, and n is the L_{1} distance between the nearness of the tracklet prediction and the detection.
在这里，n 是指数分布的尺度参数，n 是贴近度预测的置信区间，n 是轨迹预测与检测之间的 l1 距离。

Now that we have computed the conditional probabili-ties of the detection belonging to a track conditioned on the
现在，我们已经计算了条件概率的检测属于一个轨道的条件概率

Algorithm 1 Tracking Algorithm

算法 1 跟踪算法

1: procedure PHALP TRACKING
程序 PHALP 跟踪

2: Require: All active tracklets T , all detections and their single frame 3D representations at time t, D and maximum age of a track t_{max}.
要求: 所有活动轨迹 t，所有检测和他们的单帧 3d 表示时间 t，d 和最大年龄的轨道 tmax。

3: for T_j 2 T do
为了 tj2t do

4: # predict all attributes for the next frame.
预测下一帧的所有属性。

1;
Atj 2; :::g)

5: Atj A(fAtj 1; 2: : :

5: Atj A (fAtj Atj g)

6:

第一分名

pj (pj ; pj ; :::)
: Pj (pj ; ; :::)
b

7: ljtL(fljt 1; ljt 2; :::g)
ljtL (fljt1; ljt2; : : : g)

8: # Compute the cost between tracks and detections.
计算轨道和探测器之间的成本

9: $C_{i;j}$ C (Di; Tj) for all Di 2 D and Tj 2 T
9: 所有 Di 2 d 和 Tj 2 t 的 Ci; j c (Di; Tj)

10: # Hungarian to assign detections to tracklets.
匈牙利分配探测轨迹。

11: M; Tu; Du Assignment(C)
M; Tu; Du 分配(c)

12: # Update the matched tracks.
更新匹配的音轨。

13: T fTj(Di); 8(i; j) 2 Mg
T fTj (Di) ; 8(i; j)2 Mg

14: # Increase the age of unmatched tracks.
增加不匹配音轨的年龄。

15: T fTj(age)+ = 1; 8(j) 2 Tug
T fTj (年龄) + = 1; 8(j)2 Tug

16: # Make new tracks with unmatched detections.
用无与伦比的发现创造新的轨迹。

17: T fTj(Di); 8(i) 2 Du; j = jT + 1jg
T fTj (Di) ; 8(i)2 Du; j = jT + 1jg

18: Kill the tracks with age t_{max}.
关闭年龄 tmax 的轨道。

19: return Tracklets T
返回 Tracklets t

individual cues of appearance, location and pose, we can compute the overall conditional probability of the detection D_j belonging to the track T_i, given all the cues together (assumed to be independent):

外观、位置和姿势的个别线索，我们可以计算出属于音轨 Ti 的检测 Dj 的整体条件概率，给定所有的线索放在一起(假设是独立的) :

$$P(D_j \ 2 \ T_{ij} \ a; \ p; \ xy; \ n) \ / \ P_A P_P \ P_{XY} \ P_N$$
$$P \ (Dj \ 2 \ Tij \ a; \ p; \ xy; \ n)/PAPP \ PXY \ PN$$

This allow us to estimate how probable an association is based on various attribute distances. Finally, we map the similarity measures (probability values up to a scale), to cost values, for solving association. The cost function be-tween the detection representations and a predicted repre-sentations of the tracklet is defined as:
这使我们能够根据不同的属性距离估计一个关联的可能性有多大。最后，我们将相似性度量(一定范围内的概率值)映射到成本值，以解决关联问题。检测表示和预测表示之间的成本函数被定义为:

$$C \ (D_j; \ T_i) = \log(P(D_j \ 2 \ T_i))$$
$$C \ (Dj; \ Ti) = \log \ (p \ (Dj \ 2ti))$$
$$= \log(P_A) \ \log(P_P \ ) \ \log(P_{XY} \ ) \ \log(P_N \ );$$
日志(PA)日志(PP)日志(PXY)日志(PN) ;

where the second equality is up to an additive constant. Once the cost between all the tracks and the detection is computed, we simply pass it to the Hungarian algorithm for solving the association.
其中第二等式取决于加法常数。一旦所有轨道和检测之间的代价被计算出来，我们只需要把它传递给匈牙利算法来解决关联。

Estimating the parameters of the cost function: The cost function C has 4 parameters (a; p; xy and n). Ad-ditionally, the Hungarian algorithm has one parameter th to decide whether the track is not a match to the detec-tion. Therefore, overall we have five parameters for the whole association part of our tracking system. Now, we treat this as an empirical risk minimization problem and op-timize the values based on a loss function. We initialize
估计成本函数的参数: 成本函数 c 有 4 个参数(a; p; xy 和 n)。另外，匈牙利算法有一个参数 th 来决定轨道是否与检测结果不匹配。因此，总的来说，我们的跟踪系统的整个关联部分有五个参数。现在，我们将其视为一个经验风险最小化问题，并基于损失函数优化其值。我们初始化

a; p; xy and n with the values from the estimated density functions and use frame level association error as a loss function for the optimization. We use the Nelder–Mead [30] algorithm for this optimization. Finally, the optimized values are used for the cost function across all the datasets, and a simple tracking algorithm is used to associate detec-tions with tracklet predictions. The sketch of the tracking algorithm is shown in Algorithm 1.

A; p; xy 和 n 与估计密度函数的值，并利用帧级关联误差作为损失函数进行优化。我们使用 Nelder-Mead [30]算法进行优化。最后，优化值用于所有数据集的成本函数，并使用一个简单的跟踪算法将检测与跟踪预测关联起来。算法 1 显示了跟踪算法的示意图。

3.4. Extension to shot changes
3.4. 镜头改变的延伸

Our framework can easily be extended to also handle shot changes, which are common in edited media, like movies, TV shows, but also sports. Since shot changes can be detected relatively reliably, we use an external shot de-tector [13] to identify frames that indicate shot changes. In-formed by the detection of this boundary, during tracking, we update the distance metric accordingly. More specif-ically, since appearance and 3D pose are invariant to the viewpoint, we keep these factors in the distance computa-tion, while we drop the location distance from the distance metric, because of the change in the camera location. Then, the association is computed based on this updated metric. We use the AVA dataset [10] to demonstrate this utility of our tracking system and present results in Section 4.

我们的框架可以很容易地扩展到处理镜头变化，这在编辑媒体中很常见，比如电影，电视节目，还有体育节目。由于镜头变化可以相对可靠地检测到，我们使用外部镜头检测器[13]来识别指示镜头变化的帧。在跟踪过程中，通过检测这个边界形成的 in-form，我们相应地更新距离度量。更具体地说，由于外观和三维姿态对视点是不变的，因此我们将这些因素保留在距离计算中，而由于摄像机位置的变化，我们将位置距离从距离度量中去掉。然后，根据这个更新的度量来计算关联度。我们使用 AVA 数据集[10]来演示我们的跟踪系统的这个实用程序，并在第 4 节中呈现结果。

4. Experiments
4. 实验

In this section, we present the experimental evaluation of our approach. We report results on three datasets: Pose-Track [3], MuPoTS [28] and AVA [10], which capture a di-verse set of sequences, including sports, casual interactions and movies. Our method operates on detections and masks coming from an off-the-shelf Mask-RCNN network [12], and returns the identity label for each one of them. There-fore, the metrics we use to report results also focus on identity tracking at the level of the bounding box. More specifically, we report results using Identity switches (IDs), Multi-Object Tracking Accuracy (MOTA) [16], ID F1 score (IDF1) [33] and HOTA [26]. In all cases, we adopt the pro-tocols of Rajasegaran et al. [32] for evaluation.

在这一部分，我们介绍了我们的方法的实验评估。我们报告了三个数据集的结果: Pose-Track [3] ，MuPoTS [28]和 AVA [10] ，它们捕捉了一组不同的序列，包括运动，休闲互动和电影。我们的方法对来自现成的 Mask-RCNN 网络的检测和掩码进行操作[12] ，并返回其中每一个的身份标签。因此，我们用于报告结果的指标也集中在边界框水平上的身份跟踪。更具体地说，我们使用身份开关(ID)、多目标跟踪精度(MOTA)[16]、ID f1 得分(IDF1)[33]和 HOTA [26]来报告结果。在所有情况下，我们都采用 Rajasegaran 等[32]的方案进行评估。

First, we ablate the main components of our approach. Specifically, we investigate the effect of each one of the tracking cues we employ, i.e., appearance, 3D location and 3D pose, and how they affect the overall tracking pipeline. For this comparison, we report results on the Posetrack dataset [3]. The full results are presented in Table 1. As we can see, removing each one of the main cues leads to degradation in the performance of the system, where 3D lo-cation seems to have the largest effect on the performance, followed by appearance and 3D pose. Moreover, this abla-tion also highlights the importance of having the nearness term in the cost function, a feature that is not available to purely 2D tracking methods.

首先，我们消除我们方法的主要组成部分。具体来说，我们研究了我们使用的每一个跟踪线索的影响，即外观，三维位置和三维姿势，以及它们如何影响整个跟踪流水线。对于这个比较，我们在 Posetrack 数据集上报告结果[3]。表 1 列出了完整的结果。正如我们所看到的，删除每一个主要线索导致系统的性能下降，其中三维定位似乎对性能有最大的影响，其次是外观和三维姿势。此外，这种消融还强调了在成本函数中具有近似项的重要性，这是纯二维跟踪方法所不具备的特性。

Method 方法	PoseTrack 波塞特罗克		
	IDs# Id #	MOTA" MOTA“	IDF1" IDF1”
w/o 3D appearance 3D 外观 w/o 3D pose	632	58.4	74.9
W/o 3D pose w/o location	558	58.9	76.2
W/o 位置 w/o nearness	948	57.3	71.6
近在咫尺	622	58.5	74.8
Full system 全系统	541	58.9	76.4

Table 1. Ablation of the main components of PHALP on Pose-Track [3]. Removing each tracking cue (3D appearance, 3D pose or 3D location) leads to degradation in the performance.
表 1。在 Pose-Track 上消除 PHALP 的主要成分[3]。去除每个跟踪线索(3D 外观，3D 姿势或 3d 位置)导致性能下降。

Next, we evaluate our approach in comparison with the state-of-the-art methods. The results are presented in Ta-ble 2. We report results on PoseTrack [3], MuPoTS [28] and AVA [10]. Our method outperforms the previous baselines, as well as the state-of-the-art approach of Rajasegaran [32]. The gains are significant across all metrics. Our method also outperforms the other approaches in the HOTA metric.

接下来，我们将评估我们的方法，并与最先进的方法进行比较。结果见表 2。我们报告 PoseTrack [3] ，MuPoTS [28]和 AVA [10]的结果。我们的方法优于以前的基线，以及 Rajasegaran 的最新方法[32]。所有指标的收益都是显着的。我们的方法也优于 HOTA 指标中的其他方法。

Finally, we also show qualitative results of our method on multiple datasets in Fig 6. These results show that our method performs reliably even in very hard occlusion cases, while it is able to recover the correct identity over multiple successive occlusions. Fig 6 also shows the robustness of our method in complex motion sequences, shot changes and long trajectories.

最后，我们还在图 6 中显示了我们的方法在多个数据集上的定性结果。这些结果表明，即使在非常困难的闭塞情况下，我们的方法也能够可靠地执行，同时它能够恢复多个连续闭塞的正确身份。图 6 还显示了我们的方法在复杂运动序列，镜头变化和长轨迹中的稳健性。

5. Discussion
5. 讨论

We presented PHALP, an approach for monocular peo-ple tracking, by predicting appearance, location and pose in 3D. Our method relies on a powerful backbone for 3D human mesh recovery, modeling on the tracklet level for collecting information across the tracklet's detections, and eventually predicting the future states of the tracklet. One of the main benefits of PHALP is that the association as-pect requires tuning of only five parameters, which makes it very friendly for training on multi-object tracking datasets, where annotating the identity of every person in a video can be expensive. We should note that our approach can be naturally extended to make use of more attributes, e.g., a face embedding, which could be useful for cases with close-ups, like movies. The main assumptions for PHALP are that we have access to a good object detector for the initial bounding box/mask detection, and a strong HMAR network for single-frame lifting of people to 3D. If the performance of these components is not satisfactory, it can also affect PHALP. Regarding societal impact, tracking systems have often been used for human surveillance. We do not condone such use. Instead, we believe that a tracking system will be valuable for studying social-human interactions.

我们提出了 PHALP，一种单眼人跟踪的方法，通过预测外观，位置和姿势在 3d。我们的方法依赖于一个强大的骨干三维人体网格恢复，建模的轨迹水平收集信息，通过轨迹的检测，并最终预测轨迹的未来状态。PHALP 的主要优点之一是，作为方面的关联只需要调整五个参数，这使得它非常适合于在多目标跟踪数据集上进行培训，在这种情况下，在视频中注释每个人的身份可能会非常昂贵。我们应该注意到，我们的方法可以自然地扩展到使用更多的属性，例如，人脸嵌入，这对于像电影这样的特写镜头很有用。PHALP 的主要假设是我们有一个良好的目标检测器用于初始包围盒/掩模检测，以及一个强大的 HMAR 网络用于提升人的单帧到三维。如果这些组件的性能不令人满意，也会影响 PHALP。关于社会影响，跟踪系统经常被用于人类监视。我们不宽恕这种使用。相反，我们相信追踪系统对于研究社会-人类互动是有价值的。

Acknowledgements: This work was supported by ONR MURI (N00014-14-1-0671), the DARPA Machine Com-mon Sense program, as well as BAIR and BDD sponsors.

鸣谢: 这项工作得到了 ONR MURI (N00014-14-1-0671)、DARPA Machine common Sense 程序以及 BAIR 和 BDD 赞助商的支持。

Method 方法	Posetrack Posetrack IDs# Id	MOTA" # MOTA“	IDF1" IDF1“	HOTA" HOTA”	MuPoTS MuPoTS (木偶) IDs# Id	MOTA" # MOTA“	IDF1" IDF1“	HOTA" HOTA”	AVA AVA IDs# Id	IDF1" # IDF1”
Trackformer [29]										
Trackformer [29]	1263	33.7	64.0	46.7	43	24.9	62.7	53.2	716	40.9
Tracktor [4]										
Tracktor [4]	702	42.4	65.2	38.5	53	51.5	70.9	50.3	289	46.8
AlphaPose [8]										
AlphaPose [8]	2220	36.9	66.9	37.6	117	37.8	67.6	41.8	939	41.9
FlowPose [39]										
流式[39]	1047	15.4	64.2	38.0	49	21.4	67.1	43.0	452	52.9
T3DP [32]										
T3DP [32]	655	55.8	73.4	50.6	38	62.1	79.1	59.2	240	61.3
PHALP										
PHALP	541	58.9	76.4	52.9	22	66.2	81.4	59.4	227	62.7

Table 2. Comparison with state-of-the-art tracking methods. We compare our method PHALP with various tracking methods in three different datasets. Our approach outperforms the other baselines across all datasets and metrics.
表 2。与最先进的跟踪方法的比较。我们将我们的方法 PHALP 与三个不同数据集中的各种跟踪方法进行比较。我们的方法在所有数据集和度量中都优于其他基线。

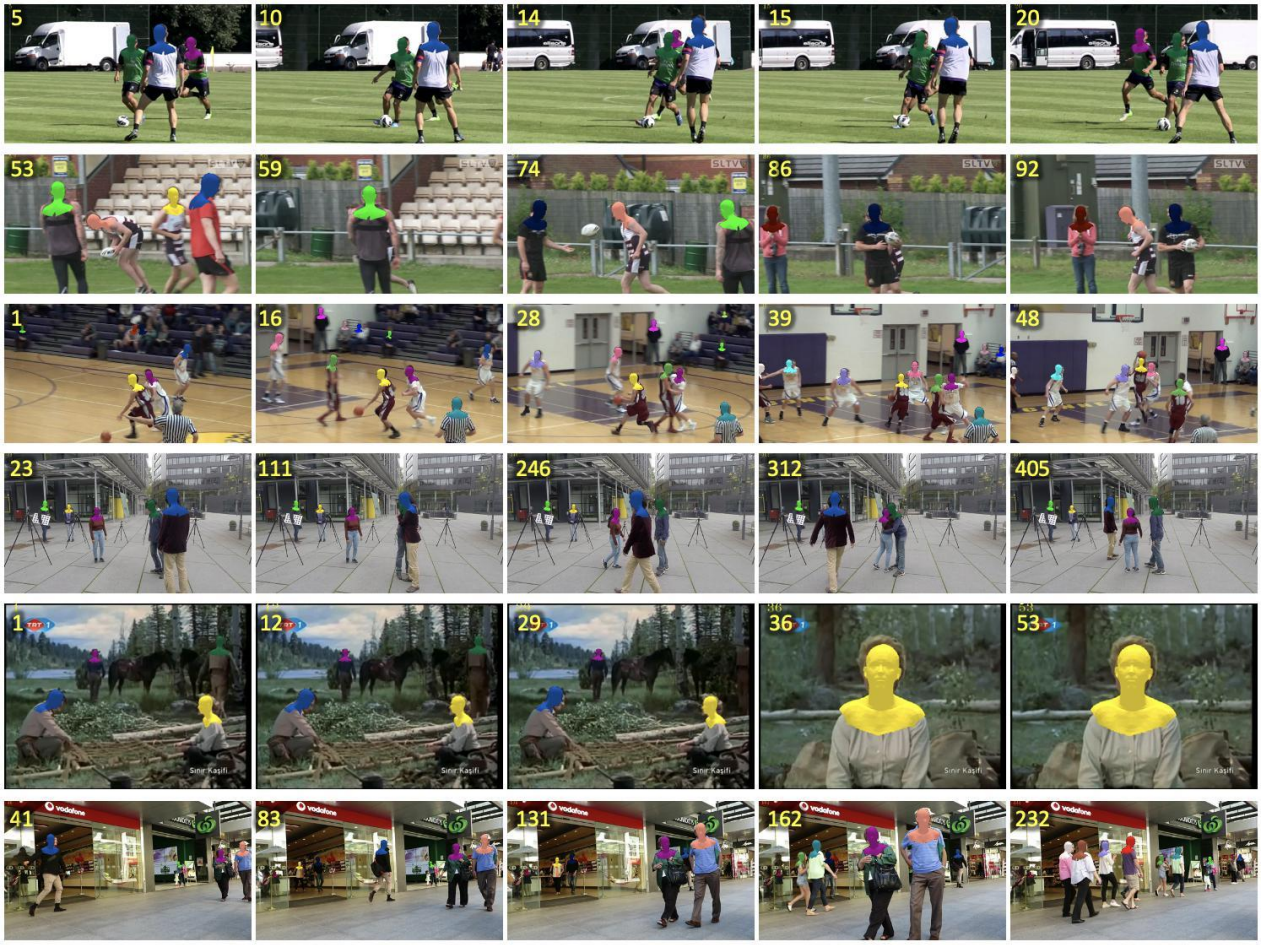


Figure 6. Qualitative Results: We show the tracking performance of PHALP in various datasets (frame number is shown at the top left corner). The first three rows are from the PoseTrack dataset [3]. These results show that even during successive occlusions our method is able to track the identity of the correct person. Note that, in the first row, although the green head-mask person and the purple head-mask person have similar appearance, our method can track each one of them successfully. In the second row, the player is going through multiple occlusions, yet recovered correctly. The third row shows the robustness of our linearization approximation for 3D location prediction, even when the motions of the players are very complex. In the MuPoTS dataset [28] (4th row), our method can handle very close interactions between people. This is due to the fact that, our modification of HMAR recovers meshes conditioned on the detected mask. We also show results (5th row) on the AVA dataset [10]. After the 3rd frame, there is a shot change in the video, and the woman is tracked successfully across the shots. Finally, we show qualitative results on a MOT17 sequence. The blue person is tracked for the whole sequence while he is going through multiple occlusions for a long time. More results at the PHALP website.

图 6。定性结果: 我们展示了 PHALP 在各种数据集中的跟踪性能(帧数显示在左上角)。前三行来自 PoseTrack 数据集[3]。这些结果表明，即使在连续闭塞期间，我们的方法也能够跟踪正确的人的身份。注意，在第一行中，尽管绿色头戴面具的人和紫色头戴面具的人具有相似的外观，但是我们的方法可以成功地跟踪每一个人。在第二排，玩家正在经历多次闭塞，但是恢复正常。第三行显示了我们的线性化近似对三维位置预测的稳健性，即使玩家的运动非常复杂。在 MuPoTS 数据集[28](第 4 行)中，我们的方法可以处理人与人之间非常密切的相互作用。这是由于事实上，我们对 HMAR 的修改恢复了对检测到的掩模条件的网格。我们还显示了 AVA 数据集的结果(第 5 行)[10]。在第三帧之后，视频中有一个镜头变化，女性在镜头之间成功跟踪。最后，我们展示了 mot17 序列的定性结果。当蓝色的人长时间处于多重闭塞状态时，他会被追踪到整个序列。更多结果请访问 PHALP 网站。

References
参考文献

[1] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. A spatio-temporal transformer for 3D human mo-tion prediction. In 3DV, 2020. 3

Emre Aksan，Peng Cao，Manuel Kaufmann，and Otmar hillges.用于三维人体运动预测的时空转换器。3DV，2020 年。3

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. So-cial LSTM: Human trajectory prediction in crowded spaces. In CVPR, 2016. 3

亚历山大·阿拉希，克拉塔斯·戈埃尔，维格内什·拉马纳坦，亚历山大·罗比凯特，李菲菲，西尔维奥·萨瓦雷斯。So-cial LSTM: 拥挤空间中的人类轨迹预测。在 CVPR，2016 年。3

[3] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A benchmark for human pose estima-tion and tracking. In CVPR, 2018. 5, 7, 8

Mykhaylo Andriluka，Umar Iqbal，Eldar Insafutdinov，Leonid Pishchulin，Anton Milan，Juergen Gall，and Bernt Schiele.PoseTrack: 人体姿态估计和跟踪的基准。在 CVPR，2018 年。5,7,8

[4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In ICCV, 2019. 3, 8

菲利普·伯格曼，蒂姆·迈因哈特，劳拉·利尔-泰克斯。追踪没有铃铛和哨子。在 ICCV，2019.3,8

[5] Christoph Bregler and Jitendra Malik. Tracking people with twists and exponential maps. In CVPR, 1998. 3

克里斯托弗·布雷格勒和吉特德拉·马利克。用曲折和指数地图追踪人。在 CVPR，1998 年

[6] Gioele Ciaparrone, Francisco Luque Sanchez,´ Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. Neu-rocomputing, 381:61–88, 2020. 3

Gioele Ciaparrone Francisco Luque Sanchez Siham Tabik Luigi Troiano Roberto Tagliaferri 还有 Francisco Herrera。视频多目标跟踪中的深度学习：一项调查。Neu-rocomputing，381:61-88,2020.3

[7] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixe´. MOTChallenge: A benchmark for single-camera multiple target tracking. IJCV, 129(4):845–881, 2021. 3

Patrick Dendorfer，Aljosa Osep，Anton Milan，Konrad Schindler，Daniel Cremers，Ian Reid，Stefan Roth，and Laura Leal-Taixe.MOTChallenge: 单摄像头多目标跟踪的基准。IJCV，129(4) : 845-881,2021.3

[8] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In ICCV, 2017. 8

郝舒芳，谢淑琴，余永泰，吕策武。 RMPE: 区域多人姿态估计。 ICCV，2017.8

[9] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-track: Efficient pose estimation in videos. In CVPR, 2018. 3

Rohit Girdhar，Georgia Gkioxari，Lorenzo Torresani，Manohar Paluri，and Du Tran.检测和跟踪: 视频中有效的姿态估计。在 CVPR，2018 年。3

[10] Chunhui Gu, Chen Sun, David A Ross, Carl Von-drick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijaya-narasimhan, George Toderici, Susanna Ricco, Rahul Suk-thankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In CVPR, 2018. 2, 7, 8

顾春晖，陈孙，大卫·罗斯，卡尔·冯·德里克，卡洛琳·潘托法鲁，李叶青，苏珊德拉·维贾亚·纳拉西姆汉，乔治·托德里奇，苏珊娜·里科，拉胡尔·苏克·谢卡尔，科迪莉娅·施密德，吉特德拉·马利克。AVA: 时空定位原子视觉行为的视频数据集。在 CVPR，2018 年。2,7,8

[11] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In ICCV, 2009. 3

Peng Guan，Alexander Weiss，Alexandru o Balan，and Michael j Black 彭观，亚历山大·维斯，亚历山大·奥巴兰，迈克尔·j·布莱克。从一张图像估算人体形状和姿势。在 ICCV，2009 年。3

[12] Kaiming He, Georgia Gkioxari, Piotr Dollar,´ and Ross Gir-shick. Mask R-CNN. In ICCV, 2017. 3, 7

卡明赫，乔治亚·格基奥萨里，皮奥特尔·美元，罗斯·吉尔希克，面具 R-CNN，ICCV，2017.3,7

[13] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. MovieNet: A holistic dataset for movie under-standing. In ECCV, 2020. 7

黄庆秋、余雄、饶安、王家泽、林大华。MovieNet: 理解电影的整体数据集。在 ECCV，2020 年。7

[14] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In CVPR, 2018. 3

Angjoo Kanazawa Michael j Black David w Jacobs 还有 Jitendra Malik。人体形状和姿势的端到端恢复。在 CVPR，2018 年。3

[15] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jiten-dra Malik. Learning 3D human dynamics from video. In CVPR, 2019. 2, 3, 5

Angjoo Kanazawa，Jason y Zhang，Panna Felsen，和 Jiten-dra Malik。从视频中学习 3d 人体动力学。在 CVPR，2019 年。2,3,5

[16] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina Korzhova, and Jing Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. PAMI, 2008. 7

Rangachar Kasturi Dmitry Goldgof Padmanabhan soundarajan Vasant Manohar John Garofolo Rachel Bowers Matthew Boonstra Valentina Korzhova 和 Jing

Zhang。视频中人脸、文本、车辆检测和跟踪的性能评估框架: 数据、度量和协议。PAMI，2008.7

[17] Tarasha Khurana, Achal Dave, and Deva Ramanan. Detect-ing invisible people. In ICCV, 2021. 3

Tarasha Khurana，Achal Dave，and Deva Ramanan。《探测隐形人》，ICCV，2021.3

[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In CVPR, 2020. 3

Muhammed Kocabas Nikos Athanasiou 和 Michael j Black。VIBE: 人体姿势和形状估计的视频推断。在 CVPR，2020 年。3

[19] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In ICCV, 2021. 3

Muhammed Kocabas，Chun-Hao p Huang，Otmar Hilliges，and Michael j Black.PARE: 用于三维人体估计的部分注意力回归器。在 ICCV，2021 年。3

[20] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Muller, Otmar Hilliges, and Michael J Black. SPEC: Seeing people in the wild with an estimated camera. In ICCV, 2021. 3

Muhammed Kocabas，Chun-Hao p Huang，Joachim Tesch，Lea Muller，Otmar hillges 和 Michael j Black。规格: 用估计摄像机观察野外的人们。在 ICCV，2021 年。3

[21] Jan J Koenderink. Optic flow. Vision research, 26(1):161– 179, 1986. 4

科恩德林克。光流。视觉研究，26(1) : 161-179,1986.4

[22] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In ICCV, 2019. 3

Nikos Kolotouros Georgios Pavlakos Michael j Black 还有 Kostas Daniilidis。通过在循环中进行模型拟合，学习重建 3d 人体姿势和形状。在 ICCV，2019 年。3

[23] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In ICCV, 2021. 3

Nikos Kolotouros，Georgios Pavlakos，Dinesh Jayaraman，and Kostas Daniilidis。人体网格恢复的概率建模

[24] Oh-Hun Kwon, Julian Tanke, and Juergen Gall. Recursive bayesian filtering for multiple human pose tracking from multiple cameras. In ACCV, 2020. 3

Oh-Hun Kwon Julian Tanke 和 Juergen Gall。基于递归贝叶斯滤波的多人姿态跟踪算法。在 ACCV，2020 年。3

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. ACM Transactions on Graphics (TOG), 34(6):1–16, 2015. 3

Matthew Loper，Naureen Mahmood，Javier Romero，Gerard Pons-Moll，and Michael j Black.SMPL: 一个带皮肤的多人线性模型。ACM 图形交易(TOG)，34(6) : 1-16,2015。3

[26] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixe,´ and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. IJCV, 2021. 7

Jonathon Luiten Aljosa Osep Patrick Dendorfer Philip Torr Andreas Geiger Laura Leal-Taixe 和 Bastian Leibe。HOTA：用于评估多目标跟踪的高阶度量。IJCV，2021.7

[27] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3D human motion estimation via motion compression and re-finement. In ACCV, 2020. 3

罗正义，s Alireza Golestaneh，和 Kris m Kitani。通过运动压缩和精化的三维人体运动估计。在 ACCV，2020 年。3

[28] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In 3DV, 2018. 7, 8

Dushyant Mehta，Oleksandr Sotnychenko，Franziska Mueller，Weipeng Xu，Srinath Sridhar，Gerard Pons-Moll，and Christian Theobalt.单眼 RGB 单镜头多人 3d 姿态估计。在 3dv，2018 年。7,8

[29] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object track-ing with transformers. arXiv preprint arXiv:2101.02702, 2021. 3, 8

提姆·迈因哈特，亚历山大·基里洛夫，劳拉·利尔-泰克斯和克里斯托夫·费希滕霍夫

[30] John A Nelder and Roger Mead. A simplex method for function minimization. The computer journal, 7(4):308–313, 1965. 7

约翰·尼尔德和罗杰·米德。函数最小化的单纯形法。计算机杂志，7(4) : 308-313,1965.7

[31] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. arXiv preprint arXiv:2012.09843, 2020. 3

Georgios Pavlakos，Jitendra Malik，and Angjoo Kanazawa。从多个镜头中恢复人体网格。arXiv preprint arXiv: 2012.09843,2020.3

[32] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3D representations. In NeurIPS, 2021. 1, 2, 3, 7, 8

Jathushan Rajasegaran Georgios Pavlakos Angjoo Kanazawa 和 Jitendra Malik。用 3d 图像追踪人们。在 NeurIPS，2021 年。1,2,3,7,8

[33] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In ECCV, 2016. 7

Ergys Ristani，Francesco Solera，Roger Zou，Rita Cucchiara，and Carlo Tomasi.多目标、多摄像机跟踪的性能测量和数据集。在 ECCV，2016 年。7

[34] Michael Snower, Asim Kadav, Farley Lai, and Hans Peter Graf. 15 keypoints is all you need. In CVPR, 2020. 3

Michael Snower，Asim Kadav，Farley Lai，and Hans Peter Graf。15 个关键点就是你所需要的一切

[35] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. In ICCV, 2021. 3

Pavel Tokmakov，Jie Li，Wolfram Burgard，and Adrien Gaidon。学习跟踪对象的永久性。在 ICCV，2021.3

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, 2017. 3, 5

Ashish Vaswani Noam Shazeer Niki Parmar Jakob Uszko-reit Llion Jones Aidan n Gomez ukasz Kaiser 还有 ilia Polosukhin。你需要的只是关注。2017 年，在 NIPS。3,5

[37] Xinshuo Weng, Yongxin Wang, Yunze Man, and Kris M Kitani. GNN3DMOT: Graph neural network for 3D multi-object tracking with 2D-3D multi-feature learning. In CVPR, 2020. 3

王永新，王云泽，Kris m Kitani。GNN3DMOT: 用于三维多目标跟踪和二维三维多特征学习的图形神经网络。在 CVPR，2020 年。3

[38] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In ECCV, 2018. 3

Bin Xiao，Haiping Wu，and ychen Wei. 人体姿态估计和跟踪的简单基线. In ECCV，2018.3

[39] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. arXiv preprint arXiv:1802.00977, 2018. 8

余亮秀、李杰峰、王浩宇、英鸿方和吕策武。Pose Flow: 有效的在线姿势跟踪。arXiv 预印 arXiv: 1802.00977,2018。8

[40] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In ICCV, 2019. 3

徐佳瑞，曹岳，郑章，韩虎。用于多目标跟踪的时空关系网络。ICCV，2019.3

[41] Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Leal-Taixe,´ and Xavier Alameda-Pineda. How to train your deep multi-object tracker. In CVPR, 2020. 3

许一宏，Aljosa Osep，Ban Yutong，Radu Horaud，Laura Leal-Taixe，and Xavier Alameda-Pineda。如何训练你的多目标跟踪器。在 CVPR，2020 年。3

[42] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object track-ing: A survey. Acm computing surveys (CSUR), 38(4):13–es, 2006. 3

目标跟踪: acm 计算概观调查(CSUR) ， 38(4) ： 13-es，2006.3

[43] Jason Y Zhang, Panna Felsen, Angjoo Kanazawa, and Jiten-dra Malik. Predicting 3D human dynamics from video. In ICCV, 2019. 3

Jason y Zhang，Panna Felsen，Angjoo Kanazawa，and Jiten-dra Malik。根据视频预测 3d 人体动力学。发表于 ICCV，2019.3

[44] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4D association graph for realtime multi-

person motion capture using multiple video cameras. In CVPR, 2020. 3

张玉香，梁安，陶玉，李秀丽，李坤，刘叶斌。使用多个摄像机实时捕捉多人运动的 4d 关联图。在 CVPR，2020 年。3

[45] Yang Zhang, Hao Sheng, Yubin Wu, Shuai Wang, Weifeng Lyu, Wei Ke, and Zhang Xiong. Long-term tracking with deep tracklet association. IEEE Transactions on Image Processing, 29:6694–6706, 2020. 3

杨章、郝胜、吴宇斌、王帅、吕伟峰、魏可和张雄。长期跟踪，深度跟踪协会。IEEE Transactions on Image pro-processing，29:6694-6706,2020.3