# RoboPEPP: Vision-Based Robot Pose and Joint Angle Estimation through mbedding redictive re-Training

# RoboPEPP:基于视觉的机器人姿态和关节角度估计通过 嵌入 预测 再训练

Raktim Gautam Goswami ami Prashanth Krishnamurthy Yann LeCun Farshad Khorrami1

Raktim Gautam Goswami ami Prashanth Krishnamurthy Yann LeCun Farshad Khorrami1

New York University Meta-FAIR

纽约大学 Meta-FAIR

# Abstract

# 摘要

Vision-based pose estimation of articulated robots with unknown joint angles has applications in collaborative robotics and human-robot interaction tasks. Current frameworks use neural network encoders to extract image features and downstream layers to predict joint angles and robot pose. While images of robots inherently contain rich information about the robot’s physical structures, existing methods often fail to leverage it fully; therefore, limiting performance under occlusions and truncations. To address this, we introduce RoboPEPP, a method that fuses information about the robot’s physical model into the encoder using a masking-based self-supervised embedding-predictive architecture. Specifically, we mask the robot’s joints and pre-train an encoder-predictor model to infer the joints’ em-beddings from surrounding unmasked regions, enhancing the encoder’s understanding of the robot’s physical model. The pre-trained encoder-predictor pair, along with joint angle and keypoint prediction networks, is then fine-tuned for pose and joint angle estimation. Random masking of input during fine-tuning and keypoint filtering during evaluation further improves robustness. Our method, evaluated on several datasets, achieves the best results in robot pose and joint angle estimation while being the least sensitive to occlusions and requiring the lowest execution time.

基于视觉的关节机器人姿态估计在未知关节角度的情况下，在协作机器人和人机交互任务中具有应用价值。当前的框架使用神经网络编码器提取图像特征，并通过下游层预测关节角度和机器人姿态。虽然机器人图像本身包含丰富的机器人物理结构信息，但现有方法往往未能充分利用这些信息，因此在遮挡和截断情况下的性能受限。为了解决这个问题，我们提出了RoboPEPP，一种通过基于掩码的自监督嵌入预测架构将机器人物理模型信息融合到编码器中的方法。具体来说，我们掩码机器人的关节，并预训练一个编码器-预测器模型，从周围未掩码区域推断关节的嵌入，从而增强编码器对机器人物理模型的理解。预训练的编码器-预测器对，连同关节角度和关键点预测网络，随后被微调以进行姿态和关节角度估计。在微调期间对输入进行随机掩码，并在评估期间进行关键点过滤，进一步提高了鲁棒性。我们的方法在多个数据集上进行了评估，在机器人姿态和关节角度估计方面取得了最佳结果，同时对遮挡最不敏感，并且执行时间最短。

# 1. Introduction

# 1. 引言

Estimating the pose and joint angles of an articulated robot in the coordinate frame of an external camera is valuable for facilitating collaborative applications wherein an agent (e.g., a human or another robot) operates in a shared space with the articulated robot . Traditional robot pose estimation methods assume known joint angles and capture multiple images with fiducial markers attached to the robot’s end-effector to establish 2D-3D correspondences between the image pixels and the robot’s frame. Recent advancements in deep learning enable the prediction of keypoints on robot joints from a single image . However, the assumption of known joint angles is not valid in many practical settings such as in collaborative robotics and human-robot interaction, where the joint angles may be unreliable or completely unknown. This challenge of simultaneously estimating joint angles and robot poses is particularly complex due to the high degrees of freedom in robotic systems and the infinite space of potential robot poses and joint angle configurations. RoboPose [16] pioneered the field of robot pose estimation with unknown joints by using an iterative render-and-compare strategy. Later works enhanced the efficiency by employing neural networks that predict joint angles and robot poses in a single feed-forward pass. While input images provide rich information about the robot’s physical structures and constraints, existing methods fail to leverage this fully, resulting in low performance in challenging scenarios like occlusions and truncations (i.e., instances where only part of the robot is visible).

在外部摄像机的坐标系中估计关节机器人的姿态和关节角度对于促进协作应用非常有价值，其中代理(例如，人类或另一个机器人)与关节机器人在共享空间中操作 。传统的机器人姿态估计方法假设已知关节角度，并通过在机器人末端执行器上附加基准标记 来捕获多张图像，以建立图像像素与机器人坐标系之间的2D-3D对应关系。深度学习的最新进展使得能够从单张图像中预测机器人关节上的 关键点 。然而，在许多实际场景中，如协作机器人和人机交互，已知关节角度的假设并不成立，因为关节角度可能不可靠或完全未知。由于机器人系统的高自由度以及潜在机器人姿态和关节角度配置的无限空间，同时估计关节角度和机器人姿态的挑战尤为复杂。RoboPose [16] 通过使用迭代渲染和比较策略开创了未知关节的机器人姿态估计领域。后续工作 通过采用神经网络在一次前向传递中预测关节角度和机器人姿态，提高了效率。虽然输入图像提供了关于机器人物理结构和约束的丰富信息，但现有方法未能充分利用这些信息，导致在遮挡和截断(即仅部分机器人可见的情况)等挑战性场景中性能低下。

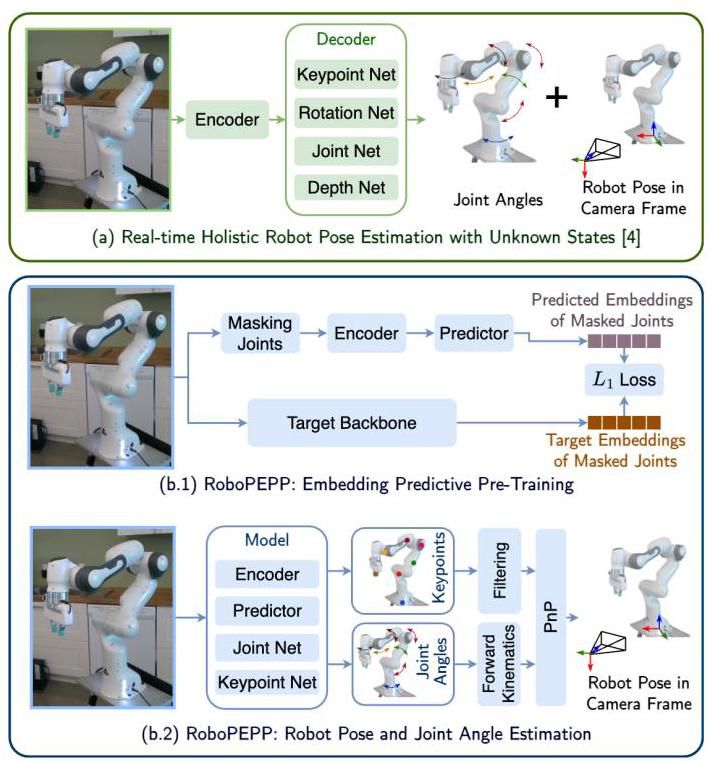


Figure 1. Comparison of an existing robot pose estimation method [5] with our RoboPEPP framework. RoboPEPP integrates joint masking-based pre-training (b.1) to enhance the encoder’s grasp of the robot’s physical model, combined with downstream networks, and keypoint filtering (b.2) to achieve high accuracy.

图1. 现有机器人姿态估计方法[5]与我们的RoboPEPP框架的比较。RoboPEPP集成了基于关节掩码的预训练(b.1)以增强编码器对机器人物理模型的理解，结合下游网络和关键点过滤(b.2)以实现高精度。

[[1]](#footnote-29)

Recently, self-supervised learning [3, 7] has shown that embedding predictive pre-training helps encoders develop a deeper semantic understanding of images. Inspired by such works, we propose RoboPEPP (Fig. 1), a robot pose estimation framework that integrates a joint-masking-based pretraining strategy to help the encoder better understand the robot’s physical model. In this approach, the encoder extracts embeddings from the unmasked regions, which a predictor uses to estimate embeddings of the masked joints. In other words, the encoder-predictor network is trained to predict the embeddings of the joints using the context around them, thus improving the network’s understanding of the robot’s structure. While this pre-trained encoder supports various robotics tasks, we focus on robot pose estimation.

最近，自监督学习(self-supervised learning)[3, 7]表明，嵌入预测预训练有助于编码器对图像进行更深层次的语义理解。受此类工作的启发，我们提出了RoboPEPP(图1)，这是一个机器人姿态估计框架，集成了基于关节掩码的预训练策略，以帮助编码器更好地理解机器人的物理模型。在该方法中，编码器从非掩码区域提取嵌入，预测器则使用这些嵌入来估计掩码关节的嵌入。换句话说，编码器-预测器网络被训练为使用关节周围的上下文来预测关节的嵌入，从而提升网络对机器人结构的理解。虽然这个预训练的编码器支持多种机器人任务，但我们专注于机器人姿态估计。

Following pre-training, the encoder and predictor are fine-tuned using downstream layers for joint angle prediction and 2D keypoint heatmap generation, allowing for end-to-end training. We further enhance the model’s occlusion robustness by randomly masking the input while fine-tuning. During inference, the pixels with the highest values in the heatmaps are identified as 2D keypoints, while corresponding 3D keypoints in the robot’s frame are computed using the forward kinematics and predicted joint angles. For cases where only part of the robot is visible in the image, we apply confidence-based keypoint filtering. Finally, we use the perspective- -point (PnP) algorithm [19] on the filtered 2D-3D correspondences to estimate the robot’s pose. In summary, our contributions are:

在预训练之后，编码器和预测器通过下游层进行微调，用于关节角度预测和2D关键点热图生成，从而实现端到端训练。我们通过在微调过程中随机掩码输入来进一步增强模型的遮挡鲁棒性。在推理过程中，热图中值最高的像素被识别为2D关键点，而机器人坐标系中的对应3D关键点则通过前向运动学和预测的关节角度计算得出。对于图像中仅部分机器人可见的情况，我们应用基于置信度的关键点过滤。最后，我们在过滤后的2D-3D对应关系上使用透视-n-点(PnP)算法[19]来估计机器人的姿态。总结来说，我们的贡献包括:

* A robot pose and joint angle estimation framework with embedding-predictive pre-training to enhance the network’s understanding of the robot’s physical model.
* 一个机器人姿态和关节角度估计框架，通过嵌入预测预训练增强网络对机器人物理模型的理解。
* An efficient network for robot pose and joint angle estimation using the pre-trained encoder-predictor alongside joint angle and keypoint estimators, trained using randomly masked inputs to enhance occlusion robustness.
* 一个高效的网络，使用预训练的编码器-预测器以及关节角度和关键点估计器进行机器人姿态和关节角度估计，并通过随机掩码输入训练以增强遮挡鲁棒性。
* A confidence-based keypoint filtering method to handle cases where only part of the robot is visible in the image.
* 一种基于置信度的关键点过滤方法，用于处理图像中仅部分机器人可见的情况。
* Extensive experiments showing RoboPEPP’s superior pose estimation, joint angle prediction, occlusion robustness, and computational efficiency.
* 大量实验表明，RoboPEPP在姿态估计、关节角度预测、遮挡鲁棒性和计算效率方面表现优异。

# 2. Related Work

# 2. 相关工作

Classical methods for robot pose estimation typically involve attaching fiducial markers, such as ArUco [13], April-Tag [25], or ARTag [12], to the robot’s end-effector to obtain easily detectable pixels in images. The corresponding 3D points in the robot’s base coordinate frame are calculated using the robot’s joint angles and forward kinematics. Using these correspondences and the camera intrinsics, an optimization problem is solved to find the robot-to-camera transformation [15, 27, 37], referred to here as the robot pose. This, however, requires multiple sets of correspondences from images taken at different robot configurations.

传统的机器人姿态估计方法通常涉及在机器人的末端执行器上附加基准标记，例如ArUco [13]、April-Tag [25]或ARTag [12]，以在图像中获得易于检测的像素。机器人基坐标系中的对应3D点通过机器人的关节角度和前向运动学计算得出。使用这些对应关系和相机内参，通过求解优化问题来找到机器人到相机的变换[15, 27, 37]，这里称为机器人姿态。然而，这需要从不同机器人配置下拍摄的图像中获取多组对应关系。

To streamline this process, DREAM [18] introduced a learning-based approach to detect multiple keypoint correspondences from a single image, estimating the pose using the PnP [19] algorithm. This method achieved performance comparable to classical approaches while requiring a single image. Building on this, PoseFusion [14] used multi-scale feature fusion to improve keypoint prediction accuracy. G-SAM [38] further improved robustness by adding a grouping and soft-argmax module, particularly useful when only part of the robot is visible. CTRNet [24] introduced a self-supervised sim-to-real approach, using differentiable PnP solver [9] and mesh renderer [28] to predict robot masks, which are compared against foreground segmentation for training. To incorporate information from prior frames, SG-TAPose [33] employed a temporal attention framework.

为了简化这一过程，DREAM [18]引入了一种基于学习的方法，从单张图像中检测多个关键点对应关系，并使用PnP [19]算法估计姿态。该方法在仅需单张图像的情况下，实现了与传统方法相当的性能。在此基础上，PoseFusion [14]使用多尺度特征融合来提高关键点预测的准确性。G-SAM [38]通过添加分组和软-argmax模块进一步提高了鲁棒性，特别是在仅部分机器人可见时尤为有用。CTRNet [24]引入了一种自监督的从仿真到真实的方法，使用可微PnP求解器[9]和网格渲染器[28]来预测机器人掩码，并将其与前景分割进行比较以进行训练。为了结合先前帧的信息，SG-TAPose [33]采用了时间注意力框架。

These approaches, however, assume known joint angles, which is often impractical in real-world settings like collaborative robotics and human-robot interaction. To address this, RoboPose [16] estimated pose with unknown joint angles using an iterative render-and-compare approach, yielding strong results but at the cost of computational efficiency. An efficient framework [5] was later developed to predict joint angles and pose in a single, real-time feed-forward pass. RoboKeyGen [35] proposed another efficient method by lifting 2D keypoints to 3D using a stable diffusion model. While some frameworks used depth cameras for pose prediction, we restrict the discussion to methods using monocular RGB images, which are more widely accessible and avoid the need for specialized sensors.

然而，这些方法假设已知关节角度，这在实际场景中(如协作机器人和人机交互)往往不切实际。为了解决这个问题，RoboPose [16]使用迭代渲染和比较的方法在未知关节角度的情况下估计姿态，取得了良好的结果，但牺牲了计算效率。随后开发了一种高效框架[5]，能够在单次实时前馈过程中预测关节角度和姿态。RoboKeyGen [35]提出了另一种高效方法，通过使用稳定扩散模型将2D关键点提升到3D。虽然一些框架 使用深度相机进行姿态预测，但我们将讨论限制在使用单目RGB图像的方法上，这些方法更广泛可用且无需专用传感器。

Current robot pose estimation methods do not fully utilize the rich features of the robot’s physical model available in images. Meanwhile, self-supervised learning frameworks have advanced encoder training to extract robust image features with Joint-Embedding Predictive Architecture (JEPA) [3] using a masking-based pretraining strategy to enhance the encoder’s semantic understanding of the image. Inspired by JEPA, we pre-train an encoder-predictor pair by masking regions around the robot’s joints and predicting embeddings of the masked regions based on the surrounding context, thus enhancing the encoder’s understanding of the robot’s physical model. The pre-trained encoder-predictor pair is then fine-tuned along with joint and keypoint prediction networks, applying random masking during fine-tuning and confidence-based key-point filtering at evaluation for improved robustness.

当前的机器人姿态估计方法 并未充分利用图像中机器人物理模型的丰富特征。与此同时，自监督学习框架 通过使用基于掩码的预训练策略，利用联合嵌入预测架构(JEPA)[3]推进了编码器训练，以提取鲁棒的图像特征，从而增强编码器对图像的语义理解。受JEPA启发，我们通过掩码机器人关节周围的区域，并基于周围上下文预测掩码区域的嵌入，预训练了一个编码器-预测器对，从而增强编码器对机器人物理模型的理解。预训练的编码器-预测器对随后与关节和关键点预测网络一起进行微调，在微调过程中应用随机掩码，并在评估时基于置信度的关键点过滤以提高鲁棒性。

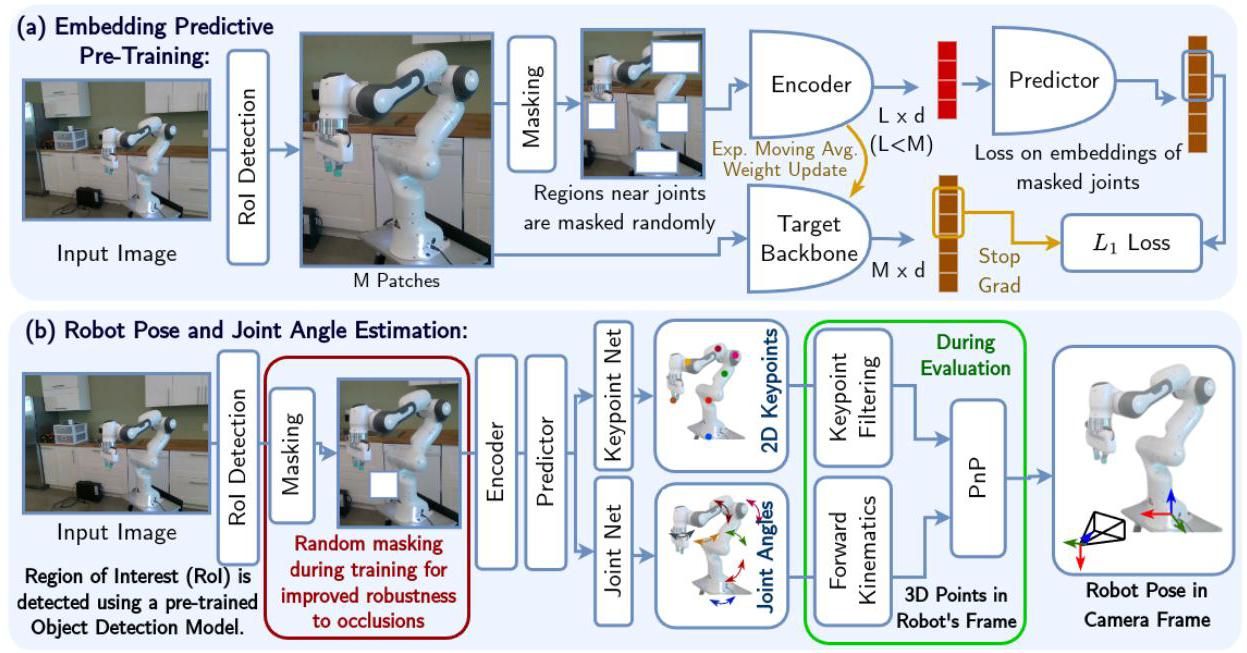


Figure 2. Overview of the RoboPEPP framework for robot pose and joint angle estimation. (a) Joint regions are masked to pre-train an encoder-predictor pair using an embedding predictive architecture. (b) The pre-trained encoder-predictor network is fine-tuned for robot pose estimation with Joint and Keypoint Prediction networks, using random masking during training to enhance occlusion robustness. During evaluation, keypoints are filtered, and a PnP algorithm estimates the robot’s pose from the filtered 2D-3D correspondences.

图2. 用于机器人姿态和关节角度估计的RoboPEPP框架概述。(a) 关节区域被掩码，以使用嵌入预测架构预训练编码器-预测器对。(b) 预训练的编码器-预测器网络通过联合和关键点预测网络进行微调，用于机器人姿态估计，训练过程中使用随机掩码以增强遮挡鲁棒性。在评估过程中，关键点被过滤，并通过PnP算法从过滤后的2D-3D对应关系中估计机器人的姿态。

# 3. Methodology

# 3. 方法论

Problem Description: Given a color image capturing an articulated robot with joints, our objective is to estimate the joint angles and the robot-to-camera rigid transformation matrix , with the robot frame being defined at its base. The robot’s forward kinematics and the camera’s intrinsic parameters are assumed to be known.

问题描述:给定一张捕捉到具有 关节的铰接机器人的彩色图像，我们的目标是估计关节角度 和机器人到相机的刚性变换矩阵 ，其中机器人框架定义在其基座上。假设机器人的正向运动学和相机的内参已知。

Method Overview: Our proposed framework (Fig. 2) comprises two stages: self-supervised pre-training of an encoder-predictor network (Sec. 3.1); and fine-tuning of the pre-trained encoder-predictor alongside 2D keypoint detection and joint angle estimation networks (Sec. 3.2). Predicted joint angles and forward kinematics yield 3D joint coordinates, which, combined with detected 2D keypoints, are used in a PnP solver to estimate pose (Sec. 3.3). During evaluation, confidence-based keypoint filtering and self-supervised fine-tuning on real-world data enhance accuracy.

方法概述:我们提出的框架(图2)包括两个阶段:编码器-预测器网络的自监督预训练(第3.1节)；以及预训练的编码器-预测器与2D关键点检测和关节角度估计网络的微调(第3.2节)。预测的关节角度和正向运动学生成3D关节坐标，与检测到的2D关键点结合，用于PnP求解器以估计姿态(第3.3节)。在评估过程中，基于置信度的关键点过滤和真实数据上的自监督微调提高了准确性。

# 3.1. Embedding Predictive Pre-Training

# 3.1. 嵌入预测预训练

Building on embedding predictive architectures [3, 7], we employ a masking-based pre-training strategy tailored for robotic applications like pose and joint estimation. Masks are selected to occlude the regions around four randomly selected robot joints, or a random area if a joint is outside the camera’s field of view. Each mask covers of the image with an aspect ratio between 0.75 and 1.5 .

基于嵌入预测架构[3, 7]，我们采用了一种针对机器人应用(如姿态和关节估计)的基于掩码的预训练策略。掩码被选择以遮挡四个随机选择的机器人关节周围的区域，或者如果关节在相机视野外，则遮挡随机区域。每个掩码覆盖图像的 ，长宽比在0.75到1.5之间。

The original image consists of patches, each sized pixels. Let represent the -th patch, where , and let denote the set of indices for the unmasked patches, with . With patches , for , as the context, a Vision Transformer (VIT) [11] encoder produces context embeddings for . These context embeddings are then passed to a VIT-based predictor, which infers embeddings for all patches of the original image, denoted for .

原始图像由 个补丁组成，每个补丁大小为 像素。设 表示第 个补丁，其中 ，并设 表示未掩码补丁的索引集，其中 。以补丁 (对于 )为上下文，视觉变换器(VIT)[11]编码器生成上下文嵌入 (对于 )。这些上下文嵌入随后传递给基于VIT的预测器，预测器推断出原始图像所有 个补丁的嵌入，记为 (对于 )。

Meanwhile, a target backbone with the same architecture as the encoder extracts embeddings for directly from the original image. The em-beddings for the masked patches, corresponding to indices (where denotes the set of masked patch indices), are used to compute the loss during training, given by:

同时，一个与编码器架构相同的目标骨干网络直接从原始图像中提取 的嵌入 。对应于索引 (其中 表示被掩码的补丁索引集)的被掩码补丁的嵌入用于在训练期间计算 损失，公式如下:

Backpropagating through the encoder-predictor network and target backbone simultaneously risks trivial solutions, like constant predictions across all networks. To avoid this, we follow [3], backpropagating only through the encoder-predictor branch and updating the target backbone with an exponential moving average of the encoder’s weights.

通过编码器-预测器网络和目标骨干网络同时进行反向传播可能会导致平凡解，例如所有网络的恒定预测。为了避免这种情况，我们遵循[3]，仅通过编码器-预测器分支进行反向传播，并使用编码器权重的指数移动平均值更新目标骨干网络。

Our approach differs from JEPA [3] by using context-informed masking at joint locations. While JEPA learns deeper semantic representations by randomly masking the input for tasks like object detection, we focus on encoding the robot’s physical properties by specifically masking joint regions. This trains the encoder to infer the robot’s joint-related information based on the surroundings, emulating a predictive understanding similar to how humans or animals deduce missing information about physical structures.

我们的方法与JEPA [3]不同之处在于在关节位置使用上下文感知的掩码。虽然JEPA通过随机掩码输入以学习更深层次的语义表示，用于目标检测等任务，我们则通过专门掩码关节区域来编码机器人的物理属性。这训练编码器根据周围环境推断机器人的关节相关信息，模拟类似于人类或动物推断物理结构缺失信息的预测理解。

# 3.2. Keypoint Detection and Joint Angle Estimation

# 3.2. 关键点检测与关节角度估计

The pre-trained encoder and predictor are then fine-tuned, where they extract embeddings for from images, which are used by the Joint Net and Keypoint Net to predict joint angles and 2D keypoints, respectively. To further increase occlusion robustness, random masks covering up to of the image are applied during training. Consistent with Sec. 3.1, the predictor outputs all patch embeddings, including masked ones. This framework is trained using the loss functions in Sec. 3.2.3.

然后对预训练的编码器和预测器进行微调，它们从图像中提取 的嵌入 ，这些嵌入分别由关节网络和关键点网络用于预测关节角度和2D关键点。为了进一步提高遮挡鲁棒性，在训练期间应用了覆盖图像高达 的随机掩码。与第3.1节一致，预测器输出所有补丁的嵌入，包括被掩码的补丁。该框架使用第3.2.3节中的损失函数进行训练。

# 3.2.1. Joint Net

# 3.2.1. 关节网络

Using the patch embeddings, , as input, the Joint Net predicts the angles for each of the robot’s joints. A global average pooling layer aggregates the patch embeddings (for ) into a single embedding to generate a global representation of the image. An iterative MLP-based approach [5] is then used to refine the joint angle predictions. Starting with a zero vector as the initial estimate , the joint angles are iteratively updated through the MLP over refinement steps (Fig. 3). The same MLP layer is used across all iterations, progressively refining the predicted joint angles for improved accuracy.

使用补丁嵌入 作为输入，关节网络预测机器人 个关节的角度。全局平均池化层将补丁嵌入 (对于 )聚合成单个嵌入 ，以生成图像的全局表示。然后使用基于MLP的迭代方法[5]来细化关节角度预测。从零向量作为初始估计 开始，通过MLP在 次细化步骤中迭代更新关节角度(图3)。在所有迭代中使用相同的MLP层，逐步细化预测的关节角度 以提高准确性。

# 3.2.2. Keypoint Net

# 3.2.2. 关键点网络

The Keypoint Net uses the patch embeddings to predict heatmaps for each of the keypoints. The matrix , contianing the patch embed-dings, is reshaped into , where . With input image of pixels and a patch size of pixels, . The Keypoint Net takes as input and applies four upsampling layers with output dimensions shown in Table 1. Each upsampling layer includes a transpose convolutional layer with a kernel size of 4, stride of 2, and one-pixel wide zero padding, followed by batch normalization, ReLU activation, and dropout. The channel dimension is gradually reduced from to 256 across these layers. The output is then passed through a linear layer that reduces the channel dimension to , followed by a sigmoid activation to produce heatmaps . Typically, each keypoint is defined at a joint of the robot, with an additional keypoint at the base, making .

关键点网络使用补丁嵌入来预测每个 关键点的热图。包含补丁嵌入的矩阵 被重塑为 ，其中 。对于 像素的输入图像和 像素的补丁大小， 。关键点网络以 作为输入，并应用四个上采样层，输出维度如表1所示。每个上采样层包括一个核大小为4、步幅为2、一像素宽零填充的转置卷积层，随后是批量归一化、ReLU激活和丢弃。在这些层中，通道维度逐渐从 减少到256。然后，输出通过一个线性层，将通道维度减少到 ，随后通过sigmoid激活生成热图 。通常，每个关键点定义在机器人的关节处，基座处还有一个额外的关键点，使得 。

# 3.2.3. Loss Functions

# 3.2.3. 损失函数

For joint angles, we employ a mean squared error loss:

对于关节角度，我们采用均方误差损失:

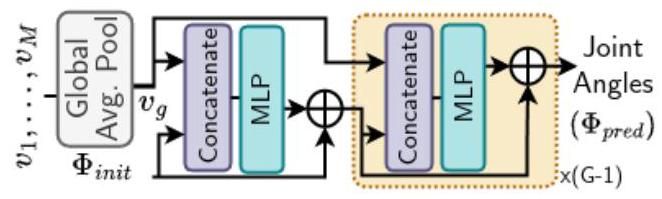


Figure 3. Joint Net: A global average pooling layer aggregates the patch embeddings, , into , which is then iteratively refined using an MLP to estimate the joint angles.

图3. 关节网络:全局平均池化层将补丁嵌入 聚合为 ，然后使用MLP迭代优化以估计关节角度。

|  |  |  |
| --- | --- | --- |
| Layer |  | Channels |
| Input Size |  | 768 |
| Upsample 1 |  | 256 |
| Upsample 2 |  | 256 |
| Upsample 3 |  | 256 |
| Upsample 4 |  | 256 |
| Linear (Heatmaps) |  | k |

|  |  |  |
| --- | --- | --- |
| 层 |  | 通道 |
| 输入尺寸 |  | 768 |
| 上采样 1 |  | 256 |
| 上采样 2 |  | 256 |
| 上采样 3 |  | 256 |
| 上采样 4 |  | 256 |
| 线性(热图) |  | k |

Table 1. Layer Output Sizes in Keypoint Net: Patch embeddings are progressively upsampled through four layers and the channel dimension is reduced to (the number of keypoint).

表1. 关键点网络中的层输出尺寸:通过四个层逐步上采样补丁嵌入，并将通道维度减少到 (关键点数量)。

where and represent the predicted and ground truth joint angles, respectively. To enhance training convergence, mean-variance normalization is applied to . For keypoint detection, we utilize the focal loss [21]:

其中 和 分别表示预测的关节角度和真实关节角度。为了增强训练收敛性，对 应用均值-方差归一化。对于关键点检测，我们使用焦点损失[21]:

where and denote the predicted and ground truth heatmaps, respectively. is generated using unnormalized Gaussian probability density functions centered at each keypoint with a 2-pixel standard deviation.

其中 和 分别表示预测的热图和真实热图。 是使用以每个关键点为中心、标准差为2像素的未归一化高斯概率密度函数生成的。

The overall training loss is a weighted combination of the two losses: , where , dependent on epoch , balances their relative importance. Since the joint angles are predicted in radians, tends to be much smaller than , especially in early training. To address this, is initialized at 0.0001, increased to 0.01 after 5 epochs, 0.1 after 10 epochs, and finally to 1 after 40 epochs, ensuring a balanced curriculum for training.

总体训练损失是两个损失的加权组合: ，其中 依赖于时期 ，平衡它们的相对重要性。由于关节角度以弧度预测， 通常比 小得多，尤其是在训练早期。为了解决这个问题， 初始化为0.0001，5个时期后增加到0.01，10个时期后增加到0.1，40个时期后最终增加到1，确保训练的平衡课程。

# 3.3. Robot Pose Estimation

# 3.3. 机器人姿态估计

Keypoint Filtering: The final layer of the Keypoint Net contains a sigmoid nonlinearity, that produces heatmaps with pixel values between 0 and 1, representing keypoint confidence at each pixel. The pixel with the highest confidence indicates the keypoint location. However, when only a portion of the robot is visible, some keypoints may lie outside the image, leading to low confidence scores across the heatmap for these keypoints (Fig. 4). Selecting the pixel with the highest confidence in such cases can be misleading, as no pixel accurately represent the true keypoint. To address this, during evaluation, we apply a threshold, only considering keypoints with confidence above it. For use

关键点过滤:关键点网络的最后一层包含一个Sigmoid非线性函数，生成像素值在0到1之间的热图，表示每个像素的关键点置信度。置信度最高的像素表示关键点位置。然而，当机器人仅部分可见时，一些关键点可能位于图像之外，导致这些关键点的热图置信度较低(图4)。在这种情况下选择置信度最高的像素可能会产生误导，因为没有像素能准确表示真实关键点。为了解决这个问题，在评估过程中，我们应用一个阈值，仅考虑置信度高于该阈值的关键点。对于使用

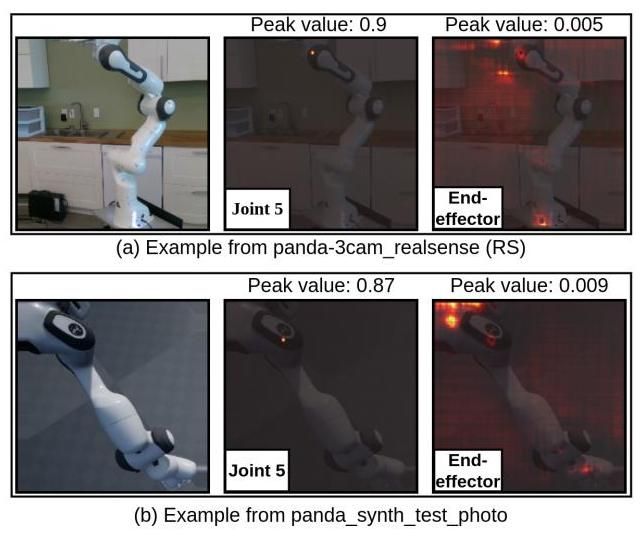


Figure 4. The examples show predicted heatmaps for Joint 5 and the End-Effector overlaid on the original image. The End-Effector, being positioned outside the field of view, produces noisy heatmaps with lower confidence (measured by peak values). Heatmap pixel values are normalized here for visual clarity.

图4. 示例显示了关节5和末端执行器的预测热图叠加在原始图像上。末端执行器位于视野之外，产生噪声较大且置信度较低的热图(通过峰值测量)。热图像素值在此处进行了归一化处理，以便视觉清晰。

with a algorithm [19] for pose estimation, we require a minimum of four 2D-3D correspondences. If fewer than four keypoints remain after filtering, we iteratively reduce by 0.025 until at least four keypoints are retained.

使用 算法[19]进行姿态估计时，我们至少需要四个2D-3D对应点。如果过滤后剩余的关键点少于四个，我们将 每次减少0.025，直到至少保留四个关键点。

Pose Estimation: The robot’s pose is estimated using the EPnP algorithm [19] with the filtered 2D-3D correspondences and known camera intrinsics. As keypoints are defined on joints, we obtain the points corresponding to the 2D keypoints using the robot’s forward kinematics and predicted joint angles.

姿态估计:使用EPnP算法[19]和过滤后的2D-3D对应点以及已知的相机内参来估计机器人的姿态。由于关键点定义在关节上，我们使用机器人的正向运动学和预测的关节角度来获得与2D关键点对应的 点。

Sim-to-Real Self-Supervised Training: In addition to supervised training for pose estimation, our method supports self-supervised fine-tuning of the trained models on real-world data to bridge the sim-to-real gap. Specifically, we use a differentiable PnP algorithm and estimate the robot’s pose and transform 3D joint locations from the robot to the camera frame. These transformed points are projected onto the image plane, yielding the projected key-points, . We then minimize the mean squared error between and the predicted keypoints

从模拟到真实的自监督训练:除了用于姿态估计的监督训练外，我们的方法还支持在真实数据上对训练模型进行自监督微调，以弥合模拟与真实之间的差距。具体来说，我们使用可微分的PnP算法 来估计机器人的姿态，并将3D关节位置从机器人转换到相机坐标系。这些转换后的点投影到图像平面上，得到投影关键点 。然后我们最小化 与预测关键点 之间的均方误差

where and are the keypoint in and , respectively. During sim-to-real finetuning, we use a lower learning rate than in the prior supervised training. Further, to prevent model collapse, the Keypoint Net’s learning rate is set close to zero.

其中 和 分别是 关键点在 和 中的位置。在从模拟到真实的微调过程中，我们使用比之前监督训练更低的学习率。此外，为了防止模型崩溃，关键点网络的学习率设置为接近零。

Region of Interest (RoI) Detection: During evaluation, we utilize the GroundingDINO [22] object detection model to automatically locate the region of interest around the robot. The detected region is cropped and resized to pixels. During training, we use ground truth bounding box information. However, to ensure our model’s robustness to region-of-interest detection, we employ a training curriculum: we use the ground truth bounding boxes and expand them by adding random offsets sampled from the uniform distribution to their edges, with progressing from 0 (first 30 epochs) to 30 pixels at epoch 30,50 pixels at epoch 50,80 pixels at epoch 70,100 pixels at epoch 90, and 120 pixels at epoch 110. This approach encourages the model to generalize effectively, even with noisy region-of-interest detection during inference.

感兴趣区域(RoI)检测:在评估过程中，我们利用GroundingDINO [22]目标检测模型自动定位机器人周围的感兴趣区域。检测到的区域被裁剪并调整为 像素。在训练过程中，我们使用真实边界框信息。然而，为了确保我们的模型对感兴趣区域检测的鲁棒性，我们采用了一种训练策略:我们使用真实边界框，并通过从均匀分布 中采样的随机偏移量扩展其边缘，其中 从0(前30个epoch)逐步增加到第30个epoch的30像素，第50个epoch的50像素，第70个epoch的80像素，第90个epoch的100像素，以及第110个epoch的120像素。这种方法鼓励模型在推理过程中即使存在噪声的感兴趣区域检测也能有效地泛化。

# 4. Experiments

# 4. 实验

# 4.1. Dataset and Implementation Details

# 4.1. 数据集与实现细节

We evaluate our framework on the DREAM dataset [18] that includes three robots (Franka Emika Panda, Kuka iiwa7, Rethink Baxter) and contains the following for each robot: Panda - synthetic domain-randomized (DR) training, DR test (Panda DR), photo-realistic test (Panda Photo), four real-world test (Panda AK, XK, RS, ORB) sequences; Kuka - synthetic DR training, DR test (Kuka DR), photo-realistic test (Kuka Photo) sequences; Baxter - synthetic DR training and test (Baxter DR) sequences.

我们在DREAM数据集[18]上评估我们的框架，该数据集包括三个机器人(Franka Emika Panda、Kuka iiwa7、Rethink Baxter)，并为每个机器人包含以下内容:Panda - 合成域随机化(DR)训练、DR测试(Panda DR)、照片级真实测试(Panda Photo)、四个真实世界测试(Panda AK、XK、RS、ORB)序列；Kuka - 合成DR训练、DR测试(Kuka DR)、照片级真实测试(Kuka Photo)序列；Baxter - 合成DR训练和测试(Baxter DR)序列。

The encoder is pre-trained using our self-supervised embedding predictive strategy (Sec. 3.1) for 200 epochs on the DR training sequences of all robots, using AdamW [23] optimizer with an initial learning rate of . For end-to-end fine-tuning (Sec. 3.2), models are trained separately for each robot for 200 epochs with AdamW optimizer (learning rate ). Sim-to-real fine-tuning is performed for 10 epochs. More details are in the supplementary material.

编码器使用我们的自监督嵌入预测策略(第3.1节)在所有机器人的DR训练序列上进行200个epoch的预训练，使用AdamW [23]优化器，初始学习率为 。对于端到端微调(第3.2节)，每个机器人分别使用AdamW优化器(学习率 )训练200个epoch。模拟到真实的微调进行10个epoch。更多细节见补充材料。

# 4.2. Results

# 4.2. 结果

# 4.2.1. Robot Pose Prediction

# 4.2.1. 机器人姿态预测

We evaluate RoboPEPP by computing the average distance between predicted and ground truth joint positions in the camera frame for each image, where is the predicted robot pose in the camera frame, is the estimated position of joint in the robot’s frame (computed from predicted joints and forward kinematics), and is the ground truth joint position in the camera frame ( signifies the robot base).

我们通过计算每个图像中预测的关节位置与真实关节位置在相机坐标系中的平均距离 来评估RoboPEPP，其中 是预测的机器人姿态在相机坐标系中的位置， 是关节 在机器人坐标系中的估计位置(通过预测的关节和正向运动学计算)， 是真实关节位置在相机坐标系中的位置( 表示机器人基座)。

In Table 2, we report the area-under-the-curve (AUC) of the average distance (ADD) across various thresholds, where higher AUC values indicate greater accuracy. We compare RoboPEPP with Real-Time Holistic Robot Pose Estimation with Unknown States (referred to as HPE in this manuscript) [5], RoboPose [16], and three variants of DREAM [18], though the latter assume known joint angles. To the best of our knowledge, RoboPose, HPE, and RoboKeyGen [35] are the only approaches besides RoboPEPP that predict robot pose with unknown joint angles. However, RoboKeyGen evaluates on a different dataset, and its code is unavailable, preventing direct comparison. Nonetheless, its reported AUC on similar datasets is lower than ours. Moreover, HPE [5] assumes known bounding boxes during evaluation, a condition often unrealistic in practice. Therefore, we also evaluate HPE with our bounding box detection strategy (denoted ) in Table 2.

在表2中，我们报告了不同阈值下平均距离(ADD)的曲线下面积(AUC)，其中较高的AUC值表示更高的准确性。我们将RoboPEPP与实时整体机器人姿态估计(本文中称为HPE)[5]、RoboPose [16]以及DREAM [18]的三个变体进行比较，尽管后者假设已知关节角度。据我们所知，除了RoboPEPP之外，RoboPose、HPE和RoboKeyGen [35]是唯一预测未知关节角度的机器人姿态的方法。然而，RoboKeyGen在不同的数据集上进行评估，且其代码不可用，无法直接比较。尽管如此，其在类似数据集上报告的AUC低于我们的结果。此外，HPE [5]在评估过程中假设已知边界框，这一条件在实际中往往不现实。因此，我们还在表2中使用我们的边界框检测策略(记为 )评估了HPE。

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Method | Known Joint Angles | Known Bounding Box | Panda | | | | | | Kuka | | Baxter |
| Synthetic | | Real | | | | Synthetic | | Synthetic |
| Photo | DR | AK | XK | RS | ORB | Photo | DR | DR |
| DREAM-F | Yes | No | 79.5 | 81.3 | 68.9 | 24.4 | 76.1 | 61.9 | - | - | - |
| DREAM-Q | Yes | No | 74.3 | 77.8 | 52.4 | 37.5 | 78.0 | 57.1 | - | - | 75.5 |
| DREAM-H | Yes | No | 81.1 | 82.9 | 60.5 | 64.0 | 78.8 | 69.1 | 72.1 | 73.3 | - |
| HPE | No | Yes | 82.0 | 82.7 | 82.2 | 76.0 | 75.2 | 75.2 | 73.9 | 75.1 | 58.8 |
| RoboPose | No | No | 79.7 | 82.9 | 70.4 | 77.6 | 74.3 | 70.4 | 73.2 | 80.2 | 32.7 |
| HPE\* | No | No | 40.7 | 41.4 | 66.7 | - | 49.1 | 51.6 | 56.7 | 56.2 | 9.8 |
| RoboPEPP (Ours) | No | No | 84.1 | 83.0 | 75.3 | 78.5 | 80.5 | 77.5 | 76.1 | 76.2 | 34.4 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 方法 | 已知关节角度 | 已知边界框 | 熊猫 | | | | | | 库卡 | | 巴克斯特 |
| 合成 | | 真实 | | | | 合成 | | 合成 |
| 照片 | DR | AK | XK | RS | ORB | 照片 | DR | DR |
| DREAM-F | 是 | 否 | 79.5 | 81.3 | 68.9 | 24.4 | 76.1 | 61.9 | - | - | - |
| DREAM-Q | 是 | 否 | 74.3 | 77.8 | 52.4 | 37.5 | 78.0 | 57.1 | - | - | 75.5 |
| DREAM-H | 是 | 否 | 81.1 | 82.9 | 60.5 | 64.0 | 78.8 | 69.1 | 72.1 | 73.3 | - |
| HPE | 否 | 是 | 82.0 | 82.7 | 82.2 | 76.0 | 75.2 | 75.2 | 73.9 | 75.1 | 58.8 |
| RoboPose | 否 | 否 | 79.7 | 82.9 | 70.4 | 77.6 | 74.3 | 70.4 | 73.2 | 80.2 | 32.7 |
| HPE\* | 否 | 否 | 40.7 | 41.4 | 66.7 | - | 49.1 | 51.6 | 56.7 | 56.2 | 9.8 |
| RoboPEPP(我们的) | 否 | 否 | 84.1 | 83.0 | 75.3 | 78.5 | 80.5 | 77.5 | 76.1 | 76.2 | 34.4 |

Table 2. Comparison of robot pose estimation using AUC on the ADD metric. Best values among methods using unknown joint angles and bounding boxes during evaluation are bolded. HPE denotes HPE [5] evaluated with the same off-the-shelf bounding box detector as RoboPEPP. HPE\* was not evaluated on Panda XK since corresponding model weights were unavailable.

表2. 使用AUC在ADD指标上对机器人姿态估计的比较。在评估期间使用未知关节角度和边界框的方法中，最佳值以粗体显示。HPE 表示HPE [5] 使用与RoboPEPP相同的现成边界框检测器进行评估。HPE\* 未在Panda XK上进行评估，因为相应的模型权重不可用。

RoboPEPP yields the highest scores across all sequences (except for Kuka DR where it remains competitive) among methods with unknown joint angles and bounding boxes. HPE [5], on the other hand, shows sensitivity to bounding box selection, with its performance dropping when ground truth bounding boxes are unavailable. Our analysis has shown that using bounding boxes just 5 pixels wider than ground truth reduces HPE’s accuracy by up to on the Panda Photo test set and by around with 10 -pixel wider boxes. While both RoboPEPP and HPE are trained using ground truth bounding boxes, RoboPEPP’s training strategy (Sec. 3.3) reduces its dependency on them. Notably, RoboPEPP outperforms HPE on most sequences even when HPE has acces to known bounding boxes during evaluation.

RoboPEPP在未知关节角度和边界框的方法中，在所有序列中得分最高(除了Kuka DR，它仍然具有竞争力)。另一方面，HPE [5] 对边界框选择表现出敏感性，当真实边界框不可用时，其性能下降。我们的分析表明，使用比真实边界框宽5像素的边界框会使HPE在Panda Photo测试集上的准确性降低多达 ，而使用宽10像素的边界框则降低约 。虽然RoboPEPP和HPE都使用真实边界框进行训练，但RoboPEPP的训练策略(第3.3节)减少了对它们的依赖。值得注意的是，即使在评估期间HPE可以访问已知的边界框，RoboPEPP在大多数序列上仍优于HPE。

A qualitative comparison of RoboPEPP with Robo-Pose [16] and HPE [5] on Panda Photo test dataset (example 1) and the occlusion dataset of Sec. 4.2.3 (examples 2 and 3) is presented in Fig. 5. Each method uses the input image to predict pose and joint angles, rendering a robot mesh that is projected and overlaid onto the original image, where closer alignment indicates higher prediction accuracy. Example 1 depicts a case where only part of the robot is visible and examples 2 and 3 show cases of occlusions (detailed in Sec. 4.2.3). In these challenging scenarios, RoboPEPP achieves highly accurate overlays while other methods are less precise, as highlighted by the red rectangles. In Fig. 5, HPE is used with ground truth bounding boxes.

图5展示了RoboPEPP与Robo-Pose [16] 和HPE [5] 在Panda Photo测试数据集(示例1)和第4.2.3节的遮挡数据集(示例2和3)上的定性比较。每种方法使用输入图像预测姿态和关节角度，渲染一个机器人网格，并将其投影并叠加到原始图像上，其中更接近的对齐表示更高的预测准确性。示例1描绘了机器人仅部分可见的情况，示例2和3展示了遮挡的情况(详见第4.2.3节)。在这些具有挑战性的场景中，RoboPEPP实现了高度准确的重叠，而其他方法则不够精确，如红色矩形所突出显示。在图5中，HPE使用了真实边界框。

# 4.2.2. Joint Prediction

# 4.2.2. 关节预测

In Table 3, we report the mean absolute error (in degrees) for joint angle prediction on the Panda Photo, Panda DR, Kuka Photo, and Kuka DR test datasets. The keypoints corresponding to the end-effector and the final joint (i.e., the joint nearest to the end-effector) lie along the axis of rotation of this joint, making their locations independent of this joint’s angle. Consequently, we predict the angles of all joints except the last one, assigning a random angle to it during evaluation. Thus, Table 3 presents the mean absolute error for the first six joint angles. We compare RoboPEPP with HPE [5] and RoboPose [16]. RoboPEPP demonstrates the lowest average joint prediction error on all datasets. Although HPE utilizes known bounding boxes for region-of-interest detection, RoboPEPP still outperforms HPE by over 15% on average across all datasets in Table 3. When HPE is tested without ground truth bounding boxes, its performance drops to an average error of 7.2 degrees.

在表3中，我们报告了在Panda Photo、Panda DR、Kuka Photo和Kuka DR测试数据集上关节角度预测的平均绝对误差(以度为单位)。与末端执行器和最终关节(即最接近末端执行器的关节)对应的关键点位于该关节的旋转轴上，使得它们的位置与该关节的角度无关。因此，我们预测除最后一个关节外的所有关节的角度，在评估期间为其分配一个随机角度。因此，表3展示了前六个关节角度的平均绝对误差。我们将RoboPEPP与HPE [5] 和RoboPose [16] 进行比较。RoboPEPP在所有数据集上展示了最低的平均关节预测误差。尽管HPE使用已知的边界框进行感兴趣区域检测，但RoboPEPP在表3中的所有数据集上平均仍优于HPE超过15%。当HPE在没有真实边界框的情况下进行测试时，其性能下降到平均误差为7.2度。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Method | J1 | J2 | J3 | J4 | J5 | J6 | Avg. |
| Panda | Photo | RoboPose | 7.7 | 3.5 | 4.3 | 3.4 | 7.3 | 8.1 | 5.7 |
| HPE (Known BBox) | 6.1 | 2.2 | 3.6 | 2.0 | 6.2 | 6.6 | 4.5 |
| RoboPEPP | 4.4 | 1.8 | 2.2 | 1.8 | 4.4 | 4.8 | 3.2 |
|  | RoboPose | 6.1 | 2.7 | 3.6 | 2.5 | 6.3 | 8.1 | 4.9 |
| DR | HPE (Known BBox) | 6.2 | 2.2 | 3.9 | 1.9 | 5.9 | 6.6 | 4.4 |
|  | RoboPEPP | 4.9 | 2.3 | 2.7 | 2.2 | 4.9 | 5.4 | 3.8 |
|  | Photo | RoboPose | 4.9 | 5.1 | 6.7 | 6.0 | 10.8 | 9.6 | 7.2 |
|  | HPE (Known BBox) | 4.8 | 3.8 | 5.0 | 2.8 | 4.9 | 5.9 | 4.5 |
| Kuka | RoboPEPP | 3.8 | 2.8 | 4.6 | 3.1 | 3.8 | 5.4 | 3.9 |
| DR | | RoboPose | 4.4 | 2.8 | 5.4 | 3.4 | 12.5 | 8.5 | 6.2 |
|  | | HPE (Known BBox) | 4.6 | 3.6 | 4.9 | 2.8 | 5.2 | 6.1 | 4.5 |
|  | | RoboPEPP | 3.7 | 3.5 | 5.1 | 3.5 | 4.1 | 6.2 | 4.3 |
| Avg. | | RoboPose | 5.8 | 3.5 | 5.0 | 3.8 | 9.2 | 8.6 | 6.0 |
|  | | HPE (Known BBox) | 5.4 | 3.0 | 4.4 | 2.4 | 5.6 | 6.3 | 4.5 |
|  | | RoboPEPP | 4.2 | 2.6 | 3.7 | 2.7 | 4.3 | 5.5 | 3.8 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | 方法 | J1 | J2 | J3 | J4 | J5 | J6 | 平均 |
| 熊猫 | 照片 | 机器人姿态 | 7.7 | 3.5 | 4.3 | 3.4 | 7.3 | 8.1 | 5.7 |
| 人体姿态估计(已知边界框) | 6.1 | 2.2 | 3.6 | 2.0 | 6.2 | 6.6 | 4.5 |
| 机器人姿态估计与预测 | 4.4 | 1.8 | 2.2 | 1.8 | 4.4 | 4.8 | 3.2 |
|  | 机器人姿态 | 6.1 | 2.7 | 3.6 | 2.5 | 6.3 | 8.1 | 4.9 |
| 动态范围 | 人体姿态估计(已知边界框) | 6.2 | 2.2 | 3.9 | 1.9 | 5.9 | 6.6 | 4.4 |
|  | 机器人姿态估计与预测 | 4.9 | 2.3 | 2.7 | 2.2 | 4.9 | 5.4 | 3.8 |
|  | 照片 | 机器人姿态 | 4.9 | 5.1 | 6.7 | 6.0 | 10.8 | 9.6 | 7.2 |
|  | 人体姿态估计(已知边界框) | 4.8 | 3.8 | 5.0 | 2.8 | 4.9 | 5.9 | 4.5 |
| 库卡 | 机器人姿态估计与预测 | 3.8 | 2.8 | 4.6 | 3.1 | 3.8 | 5.4 | 3.9 |
| 动态范围 | | 机器人姿态 | 4.4 | 2.8 | 5.4 | 3.4 | 12.5 | 8.5 | 6.2 |
|  | | 人体姿态估计(已知边界框) | 4.6 | 3.6 | 4.9 | 2.8 | 5.2 | 6.1 | 4.5 |
|  | | 机器人姿态估计与预测 | 3.7 | 3.5 | 5.1 | 3.5 | 4.1 | 6.2 | 4.3 |
| 平均 | | 机器人姿态 | 5.8 | 3.5 | 5.0 | 3.8 | 9.2 | 8.6 | 6.0 |
|  | | 人体姿态估计(已知边界框) | 5.4 | 3.0 | 4.4 | 2.4 | 5.6 | 6.3 | 4.5 |
|  | | 机器人姿态估计与预测 | 4.2 | 2.6 | 3.7 | 2.7 | 4.3 | 5.5 | 3.8 |

Table 3. Mean absolute error between the predicted and actual joint angles (in degrees) for the Panda and Kuka synthetic test sets.

表3. Panda和Kuka合成测试集的预测与实际关节角度之间的平均绝对误差(以度为单位)。

# 4.2.3. Robustness to Occlusions

# 4.2.3. 对遮挡的鲁棒性

In addition to achieving high performance across various metrics and datasets, RoboPEPP demonstrates robustness to occlusions. To evaluate this, we compared the performance of RoboPEPP with other methods [5, 16] on a custom dataset, created by adding synthetic occlusions to Panda Photo. Specifically, we overlaid black rectangular or circular masks at random positions on the robot, ensuring that the masks covered at least some part of the robot (and not just the background). We generated four test sequences with occlusion ratios of0.1,0.2,0.3, and 0.4 on the RoI area in each image, respectively. This approach differs from the masking used during our training, where the model is informed about the number of masked patches. Here, the model processes occluded images as it would any other input image, without any knowledge of the occlusion.

除了在各种指标和数据集上实现高性能外，RoboPEPP还展示了对遮挡的鲁棒性。为了评估这一点，我们在一个自定义数据集上比较了RoboPEPP与其他方法[5, 16]的性能，该数据集通过在Panda Photo上添加合成遮挡创建。具体来说，我们在机器人上随机位置叠加黑色矩形或圆形掩码，确保掩码至少覆盖机器人的一部分(而不仅仅是背景)。我们在每张图像的感兴趣区域(RoI)上生成了四个测试序列，遮挡比例分别为0.1、0.2、0.3和0.4。这种方法与我们训练期间使用的掩码不同，训练时模型知道被掩码的补丁数量。在这里，模型处理遮挡图像时与处理其他输入图像一样，没有任何关于遮挡的先验知识。

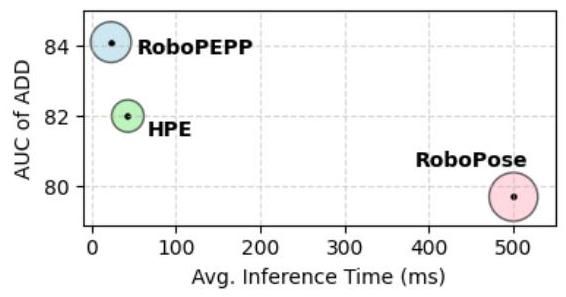


Figure 7. Execution time and computation analysis on the Panda Photo test dataset with RoboPEPP showing best performance and accuracy. The circle sizes in the plot represent model FLOPs.

图7. 在Panda Photo测试数据集上，RoboPEPP的执行时间和计算分析展示了最佳性能和准确性。图中的圆圈大小代表模型的浮点运算次数(FLOPs)。

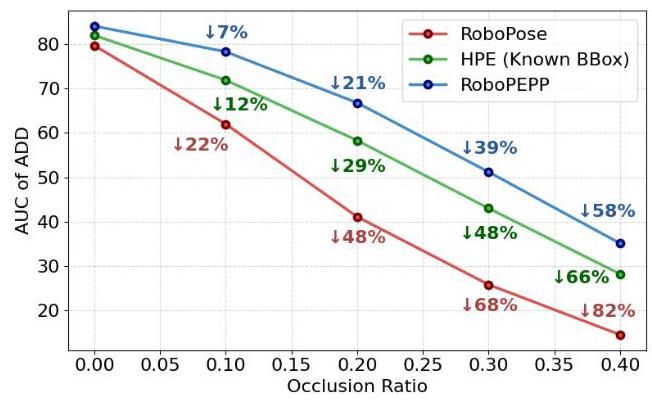


Figure 6. AUC comparison of the distance metric under varying occlusion levels, evaluated on the dataset in Sec. 4.2.3. Percentages next to the lines indicate the relative drop in each method’s performance compared to their performance without occlusions.

图6. 在不同遮挡水平下，距离度量的AUC比较，基于第4.2.3节中的数据集进行评估。线条旁边的百分比表示每种方法在遮挡情况下的性能相对于无遮挡时的性能下降比例。

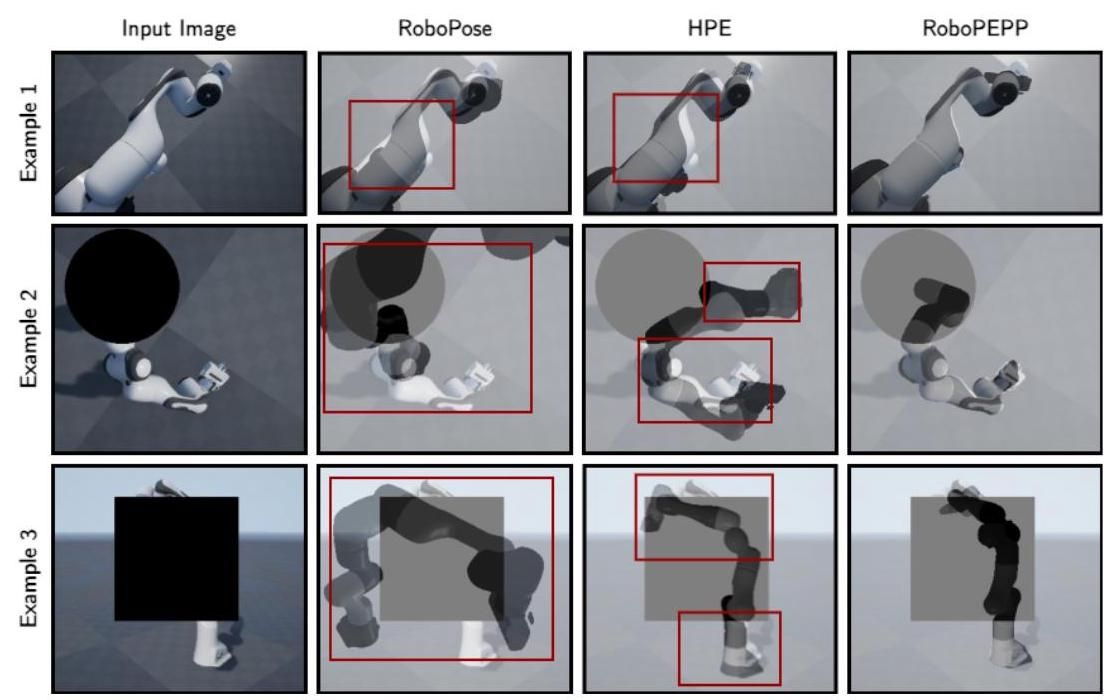


Figure 5. Qualitative Comparison on Panda Photo (Example 1) and Occlusion (Example 2 and 3) datasets: Predicted poses and joint angles are used to generate a mesh overlaid on the original image, where closer alignment indicates greater accuracy. Highlighted rectangles indicate regions where other methods’ meshes misalign, while RoboPEPP achieves high precision.

图5. Panda Photo(示例1)和遮挡(示例2和3)数据集上的定性比较:使用预测的姿势和关节角度生成覆盖在原始图像上的网格，对齐越接近表示准确性越高。高亮矩形表示其他方法的网格未对齐的区域，而RoboPEPP实现了高精度。

In Fig. 6, we plot the AUC of the ADD against the occlusion ratio. The plot also includes the percentage decrease in AUC relative to the respective model’s performance without occlusion. RoboPEPP demonstrates superior robustness to occlusion, achieving an AUC score of 35.1 even when of the RoI is occluded, compared to 28.2 for HPE and 14.5 for RoboPose. Examples 2 and 3 in Fig. 5 provide qualitative comparisons of RoboPEPP, HPE [5], and Robo-Pose [16] on the occlusion dataset. RoboPEPP demonstrates superior performance, even in challenging cases like example 3, where most of the robot is occluded. In both examples 2 and 3, RoboPose produces inaccurate results, while HPE shows partial overlap but still exhibits notable inaccuracies, highlighted by the red rectangles.

在图6中，我们绘制了ADD的AUC与遮挡比例的关系图。图中还包括了AUC相对于各自模型在无遮挡情况下的性能下降百分比。RoboPEPP展示了对遮挡的卓越鲁棒性，即使在RoI的 被遮挡时，AUC得分仍为35.1，而HPE为28.2，RoboPose为14.5。图5中的示例2和3提供了RoboPEPP、HPE [5]和Robo-Pose [16]在遮挡数据集上的定性比较。RoboPEPP展示了卓越的性能，即使在像示例3这样具有挑战性的情况下，机器人大部分被遮挡时也是如此。在示例2和3中，RoboPose产生了不准确的结果，而HPE显示了部分重叠，但仍然存在显著的不准确性，如红色矩形所示。

# 4.2.4. Percentage of Correct Keypoints

# 4.2.4. 关键点正确率

The accuracy of keypoint detection affects the overall performance of RoboPEPP. Therefore, in Table 4, we report the percentage of correct keypoints (PCK) within thresholds of 2.5,5, and 10 pixels on the Panda Photo dataset. Since HPE [5] and RoboPose [16] do not rely on 2D keypoint detection for pose estimation, we include only DREAM [18] as a baseline for comparison. RoboPEPP achieves high average PCK scores, with 0.43 @ 2.5 pixels, 0.73 @ 5 pixels, and 0.95 @ 10 pixels. While DREAM outperforms RoboPEPP on certain metrics, such as PCK@2.5 pixels on the Panda AK and Panda XK, RoboPEPP demonstrates superior accuracy across most other metrics and on average, highlighting its robust keypoint detection performance.

关键点检测的准确性影响RoboPEPP的整体性能。因此，在表4中，我们报告了在Panda Photo数据集上，关键点正确率(PCK)在2.5、5和10像素阈值内的百分比。由于HPE [5]和RoboPose [16]不依赖2D关键点检测进行姿态估计，我们仅将DREAM [18]作为基线进行比较。RoboPEPP实现了较高的平均PCK得分，分别为0.43 @ 2.5像素、0.73 @ 5像素和0.95 @ 10像素。虽然DREAM在某些指标上优于RoboPEPP，例如在Panda AK和Panda XK上的PCK@2.5像素，但RoboPEPP在大多数其他指标和平均表现上展示了更高的准确性，突显了其强大的关键点检测性能。

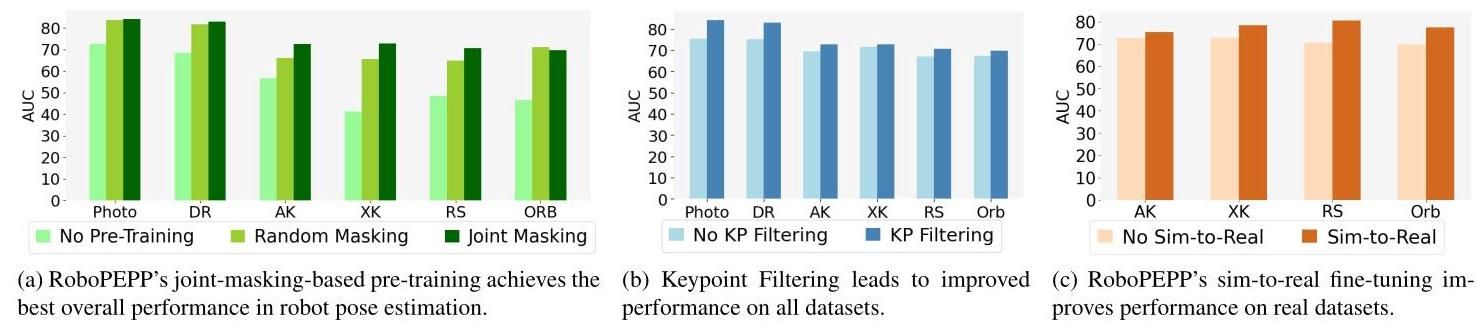


Figure 8. Ablation studies on (a) Pre-Training, (b) Keypoint Filtering, and (c) Sim-to-Real fine-tuning on the Panda test datasets.

图8. 在Panda测试数据集上进行的消融研究，包括(a)预训练，(b)关键点过滤，和(c)从模拟到现实的微调。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Method | PCK @ (pixel) | | |
| 2.5 | 5 | 10 |
| DR | DREAM | 0.79 | 0.88 | 0.9 |
| RoboPEPP | 0.84 | 0.91 | 0.93 |
| Photo | DREAM | 0.77 | 0.87 | 0.9 |
| RoboPEPP | 0.87 | 0.92 | 0.94 |
| AK | DREAM | 0.36 | 0.65 | 0.9 |
| RoboPEPP | 0.16 | 0.62 | 0.93 |
| XK | DREAM | 0.15 | 0.37 | 0.59 |
| RoboPEPP | 0.09 | 0.37 | 0.96 |
| RS | DREAM | 0.24 | 0.83 | 0.96 |
| RoboPEPP | 0.31 | 0.82 | 0.97 |
| ORB | DREAM | 0.28 | 0.67 | 0.83 |
| RoboPEPP | 0.28 | 0.73 | 0.96 |
|  | DREAM | 0.43 | 0.71 | 0.85 |
| RoboPEPP | 0.43 | 0.73 | 0.95 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | 方法 | PCK @ (像素) | | |
| 2.5 | 5 | 10 |
| DR | DREAM | 0.79 | 0.88 | 0.9 |
| RoboPEPP | 0.84 | 0.91 | 0.93 |
| 照片 | DREAM | 0.77 | 0.87 | 0.9 |
| RoboPEPP | 0.87 | 0.92 | 0.94 |
| AK | DREAM | 0.36 | 0.65 | 0.9 |
| RoboPEPP | 0.16 | 0.62 | 0.93 |
| XK | DREAM | 0.15 | 0.37 | 0.59 |
| RoboPEPP | 0.09 | 0.37 | 0.96 |
| RS | DREAM | 0.24 | 0.83 | 0.96 |
| RoboPEPP | 0.31 | 0.82 | 0.97 |
| ORB | DREAM | 0.28 | 0.67 | 0.83 |
| RoboPEPP | 0.28 | 0.73 | 0.96 |
|  | DREAM | 0.43 | 0.71 | 0.85 |
| RoboPEPP | 0.43 | 0.73 | 0.95 |

Table 4. Comparison of Percentage of Correct Keypoints (PCK) at different pixel thresholds across the Panda test datasets.

表4. 熊猫测试数据集在不同像素阈值下的正确关键点百分比(PCK)比较。

# 4.2.5. Execution Time

# 4.2.5. 执行时间

To demonstrate the practical effectiveness of the proposed RoboPEPP method, we compare execution times (in milliseconds) in Fig. 7. The circle sizes in the figure correspond to the relative number of floating-point operations (FLOPs) required by each model. All evaluations were conducted on a system equipped with an Nvidia RTX A4000 GPU, an Intel(R) i9 CPU, and 128 GB RAM, using the Panda Photo test dataset. Consistent with previous work [5], we report only model execution time, excluding pre-processing steps such as data loading and RoI detection. Despite having a slightly higher FLOPs count than HPE [5], RoboPEPP achieves the highest AUC ADD score and the fastest execution time, completing inference in just 23 milliseconds.

为了展示所提出的RoboPEPP方法的实际效果，我们在图7中比较了执行时间(以毫秒为单位)。图中的圆圈大小对应于每个模型所需的浮点运算(FLOPs)的相对数量。所有评估均在配备Nvidia RTX A4000 GPU、Intel(R) i9 CPU和128 GB RAM的系统上进行，使用熊猫照片测试数据集。与之前的工作[5]一致，我们仅报告模型执行时间，不包括数据加载和感兴趣区域(RoI)检测等预处理步骤。尽管RoboPEPP的FLOPs计数略高于HPE[5]，但它实现了最高的AUC ADD分数和最快的执行时间，仅需23毫秒完成推理。

# 4.3. Ablation Studies

# 4.3. 消融研究

Embedding Predictive Pre-Training: To assess the impact of embedding predictive pre-training in RoboPEPP, we conducted an ablation study comparing three versions of the model: a version of RoboPEPP without pre-training, and a version pre-trained with random masking instead of joint-specific masking, and standard RoboPEPP (i.e., pre-trained with joint masking). For all experiments, we utilized the same model architecture and training settings. The bar graphs in Fig. 8a illustrate that pre-training significantly improves performance. While the model trained with the default masking strategy demonstrated competitive results on synthetic test datasets, the model trained with joint-specific masking achieved better performance on real-world datasets in general. Note that the real-world results shown here do not include the sim-to-real fine-tuning of Sec. 3.3. Further, on the occlusion dataset (Sec. 4.2.3) with a 0.4 occlusion ratio, the model without pre-training achieves AUC of 22.6, the one with random masking reaches 30, and RoboPEPP achieves 35.1, highlighting the latter’s occlusion robustness.

嵌入预测预训练:为了评估嵌入预测预训练在RoboPEPP中的影响，我们进行了一项消融研究，比较了三个版本的模型:一个没有预训练的RoboPEPP版本，一个使用随机掩码而不是关节特定掩码进行预训练的版本，以及标准的RoboPEPP(即使用关节掩码进行预训练)。所有实验均使用相同的模型架构和训练设置。图8a中的柱状图显示，预训练显著提高了性能。虽然使用默认掩码策略训练的模型在合成测试数据集上表现出竞争力，但使用关节特定掩码训练的模型在真实世界数据集上通常表现更好。请注意，此处展示的真实世界结果不包括第3.3节中的模拟到现实的微调。此外，在遮挡率为0.4的遮挡数据集(第4.2.3节)上，没有预训练的模型的AUC为22.6，使用随机掩码的模型达到30，而RoboPEPP达到35.1，突显了后者在遮挡情况下的鲁棒性。

Keypoint Filtering: In Fig. 8b, we demonstrate that the integration of keypoint filtering (KP filtering) enhances performance across all datasets by helping in filtering out key-points that fall outside the camera’s field of view. Similar to Fig. 8a, the real-world results presented in Fig. 8b do not include any sim-to-real fine-tuning.

关键点过滤:在图8b中，我们展示了关键点过滤(KP过滤)的集成通过帮助过滤掉落在相机视野之外的关键点，提高了所有数据集的性能。与图8a类似，图8b中展示的真实世界结果不包括任何模拟到现实的微调。

Sim-to-Real Fine-Tuning: In Fig. 8c, we show performance gains from sim-to-real self-supervised training, with the model’s accuracy improving by an average of points after fine-tuning. Our sim-to-real training requires only 10 epochs, with each epoch lasting around 2 minutes.

模拟到现实微调:在图8c中，我们展示了模拟到现实自监督训练带来的性能提升，微调后模型的准确性平均提高了 点。我们的模拟到现实训练仅需10个epoch，每个epoch持续约2分钟。

# 5. Conclusion

# 5. 结论

We introduced a novel framework RoboPEPP enhancing robot pose and joint angle estimation using an embedding predictive pre-training strategy. RoboPEPP uses a joint-masking-based method to pre-train an encoder-predictor pair to be integrated into downstream networks for joint and pose predictions. Experimental results show RoboPEPP’s superior performance, particularly in handling occlusions due to the combined effects of pre-training and random masking during fine-tuning. RoboPEPP’s training helps fuse knowledge of the robot’s physical model within the encoder, making RoboPEPP effective for pose estimation and versatile for broader applications such as system dynamic prediction and imitation learning.

我们引入了一种新颖的框架RoboPEPP，通过嵌入预测预训练策略增强机器人姿态和关节角度估计。RoboPEPP使用基于关节掩码的方法预训练编码器-预测器对，并将其集成到下游网络中以进行关节和姿态预测。实验结果表明，RoboPEPP表现出卓越的性能，特别是在处理遮挡方面，这得益于预训练和微调期间随机掩码的综合效果。RoboPEPP的训练有助于将机器人物理模型的知识融合到编码器中，使RoboPEPP在姿态估计方面有效，并在系统动态预测和模仿学习等更广泛的应用中具有多功能性。

References

参考文献

[1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. arXiv preprint arXiv:1911.05371, 2019. 2

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. 通过同时聚类和表示学习进行自标记. arXiv预印本 arXiv:1911.05371, 2019. 2

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bo-janowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In Proceedings of the European Conference on Computer Vision, pages 456-473. Springer, 2022.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bo-janowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. 用于标签高效学习的掩码孪生网络. 欧洲计算机视觉会议论文集, 第456-473页. Springer, 2022.

[3] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bo-janowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15619-15629, 2023. 2, 3, 11

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bo-janowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 使用联合嵌入预测架构从图像中进行自监督学习. IEEE/CVF计算机视觉与模式识别会议论文集, 第15619-15629页, 2023. 2, 3, 11

[4] Jimmy Lei Ba. Layer normalization. arXiv preprint arXiv:1607.06450, 2016. 11

Jimmy Lei Ba. 层归一化. arXiv预印本 arXiv:1607.06450, 2016. 11

[5] Shikun Ban, Juling Fan, Xiaoxuan Ma, Wentao Zhu, Yu Qiao, and Yizhou Wang. Real-time holistic robot pose estimation with unknown states. arXiv preprint arXiv:2402.05655, 2024. 1, 2, 4, 5, 6, 7, 8, 12, 13

班世坤, 范居玲, 马晓轩, 朱文涛, 乔宇, 王亦洲. 实时整体机器人姿态估计与未知状态. arXiv预印本 arXiv:2402.05655, 2024. 1, 2, 4, 5, 6, 7, 8, 12, 13

[6] Adrien Bardes, Jean Ponce, and Yann LeCun. Vi-creg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906, 2021. 2

阿德里安·巴德斯(Adrien Bardes)、让·庞斯(Jean Ponce)和扬·勒昆(Yann LeCun)。Vi-creg:用于自监督学习的方差-不变性-协方差正则化。arXiv预印本 arXiv:2105.04906, 2021. 2

[7] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-JEPA: Latent video prediction for visual representation learning, 2024. 2, 3

阿德里安·巴德斯(Adrien Bardes)、昆汀·加里多(Quentin Garrido)、让·庞斯(Jean Ponce)、陈新雷(Xinlei Chen)、迈克尔·拉巴特(Michael Rabbat)、杨立昆(Yann LeCun)、米多·阿斯拉恩(Mido Assran)和尼古拉斯·巴拉斯(Nicolas Ballas)。V-JEPA:用于视觉表示学习的潜在视频预测，2024年。2, 3

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems, 33:9912- 9924, 2020. 2

马蒂尔德·卡隆(Mathilde Caron)、伊山·米斯拉(Ishan Misra)、朱利安·梅拉尔(Julien Mairal)、普里亚·戈亚尔(Priya Goyal)、彼得·博亚诺夫斯基(Piotr Bojanowski)和阿曼德·朱林(Armand Joulin)。通过对比聚类分配进行视觉特征的无监督学习。《神经信息处理系统进展》，33卷:9912-9924页，2020年。2

[9] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, and Tat-Jun Chin. End-to-end learnable geometric vision by backprop-agating PnP optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages

陈博(Bo Chen)、阿尔瓦罗·帕拉(Alvaro Parra)、曹杰伟(Jiewei Cao)、李楠(Nan Li)和秦达君(Tat-Jun Chin)。通过反向传播PnP优化实现端到端可学习的几何视觉。发表于IEEE/CVF计算机视觉与模式识别会议论文集，页码

[10] Sammy Christen, Wei Yang, Claudia Pérez-D’Arpino, Ot-mar Hilliges, Dieter Fox, and Yu-Wei Chao. Learning human-to-robot handovers from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9654-9664, 2023. 1

萨米·克里斯滕(Sammy Christen)、杨伟(Wei Yang)、克劳迪娅·佩雷斯-达尔皮诺(Claudia Pérez-D’Arpino)、奥特马尔·希利格斯(Otmar Hilliges)、迪特尔·福克斯(Dieter Fox)和赵宇伟(Yu-Wei Chao)。从点云中学习人机交接。在IEEE/CVF计算机视觉与模式识别会议论文集(Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition)中，第9654-9664页，2023年。1

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-vain Gelly, et al. An image is worth words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, 2020. 3, 11

阿列克谢·多索维茨基(Alexey Dosovitskiy)、卢卡斯·拜尔(Lucas Beyer)、亚历山大·科列斯尼科夫(Alexander Kolesnikov)、德克·魏森博恩(Dirk Weissenborn)、翟小华(Xiaohua Zhai)、托马斯·昂特希纳(Thomas Unterthiner)、莫斯塔法·德赫加尼(Mostafa Dehghani)、马蒂亚斯·明德雷尔(Matthias Minderer)、格奥尔格·海戈尔德(Georg Heigold)、西尔万·盖利(Sylvain Gelly)等。一张图片胜过千言万语(An image is worth words):大规模图像识别中的Transformer。发表于国际学习表征会议(International Conference on Learning Representations)论文集，2020年。3, 11

[12] Mark Fiala. Artag, a fiducial marker system using digital techniques. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 590-596. IEEE, 2005. 1, 2

Mark Fiala. Artag, a fiducial marker system using digital techniques. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 590-596. IEEE, 2005. 1, 2

[13] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Fran-

塞尔吉奥·加里多-华拉多(Sergio Garrido-Jurado)，拉斐尔·穆尼奥斯-萨利纳斯(Rafael Muñoz-Salinas)，弗兰-

cisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognition, 47(6):2280-2292, 2014. 1, 2

cisco José Madrid-Cuevas 和 Manuel Jesús Marín-Jiménez。遮挡条件下高可靠性基准标记的自动生成与检测。模式识别(Pattern Recognition)，47(6):2280-2292, 2014. 1, 2

[14] Xujun Han, Shaochen Wang, Xiucai Huang, and Zhen Kan. Posefusion: Multi-scale keypoint correspondence for monocular camera-to-robot pose estimation in robotic manipulation. In Proceedings of the IEEE International Conference on Robotics and Automation, pages 795-801, 2024. 1,2

韩旭俊, 王绍晨, 黄秀才, 阚振. Posefusion: 用于机器人操作中单目相机到机器人姿态估计的多尺度关键点对应方法. 在IEEE国际机器人与自动化会议论文集, 第795-801页, 2024. 1,2

[15] Jarmo Ilonen and Ville Kyrki. Robust robot-camera calibration. In Proceedings of the International Conference on Advanced Robotics, pages 67-74. IEEE, 2011. 2

亚尔莫·伊洛宁(Jarmo Ilonen)和维尔·基尔基(Ville Kyrki)。鲁棒机器人-相机标定。在国际先进机器人会议(International Conference on Advanced Robotics)论文集，第67-74页。IEEE，2011年。2

[16] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Single-view robot pose and joint angle estimation via render & compare. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1654-1663, 2021. 2, 5, 6, 7, 12, 13

Yann Labbé, Justin Carpentier, Mathieu Aubry, 和 Josef Sivic。通过渲染与比较进行单视图机器人姿态和关节角度估计。在IEEE/CVF计算机视觉与模式识别会议论文集(Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition)中，第1654-1663页，2021年。2, 5, 6, 7, 12, 13

[17] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. In Proceedings of the International Conference on Learning Representations, 2022. 11

古斯塔夫·拉尔森(Gustav Larsson)、迈克尔·梅尔(Michael Maire)和格雷戈里·沙赫纳罗维奇(Gregory Shakhnarovich)。Fractalnet:无残差的超深度神经网络。发表于《国际学习表征会议论文集》(Proceedings of the International Conference on Learning Representations)，2022年。11

[18] Timothy E Lee, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Oliver Kroemer, Dieter Fox, and Stan Birch-field. Camera-to-robot pose estimation from a single image. In Proceedings of the IEEE International Conference on Robotics and Automation, pages 9426-9432. IEEE, 2020. 1,2,5,7,11,12

蒂莫西·E·李(Timothy E Lee)、乔纳森·特伦布莱(Jonathan Tremblay)、唐·托(Thang To)、贾成(Jia Cheng)、特里·莫西尔(Terry Mosier)、奥利弗·克雷默(Oliver Kroemer)、迪特尔·福克斯(Dieter Fox)和斯坦·伯奇菲尔德(Stan Birch-field)。从单张图像中估计相机到机器人的姿态(Camera-to-robot pose estimation from a single image)。在IEEE国际机器人与自动化会议(Proceedings of the IEEE International Conference on Robotics and Automation)中，页码9426-9432。IEEE，2020年。1,2,5,7,11,12

[19] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate solution to the problem. International journal of computer vision, 81:155-166, 2009. 2, 5

Vincent Lepetit, Francesc Moreno-Noguer, 和 Pascal Fua。EPnP:一种精确的 解决方案，用于 问题。国际计算机视觉杂志，81:155-166, 2009。2, 5

[20] Shushuai Li, Christophe De Wagter, and Guido CHE De Croon. Self-supervised monocular multi-robot relative localization with efficient deep neural networks. In Proceedings of the International Conference on Robotics and Automation, pages 9689-9695. IEEE, 2022. 1

Shushuai Li, Christophe De Wagter, 和 Guido CHE De Croon。自监督单目多机器人相对定位与高效深度神经网络。在国际机器人与自动化会议论文集，页码9689-9695。IEEE, 2022。1

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 2980-2988, 2017. 4

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, 和 Piotr Dollár。密集目标检测的焦点损失。在IEEE国际计算机视觉会议论文集，页码2980-2988, 2017。4

[22] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding Dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499, 2023. 5, 11

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, 等。Grounding Dino:将Dino与接地预训练结合用于开放集目标检测。arXiv预印本arXiv:2303.05499, 2023。5, 11

[23] I Loshchilov. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017. 5, 11

I Loshchilov。解耦权重衰减正则化。arXiv预印本arXiv:1711.05101, 2017。5, 11

[24] Jingpei Lu, Florian Richter, and Michael C. Yip. Markerless camera-to-robot pose estimation via self-supervised sim-to-real transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21296- 21306, 2023. 1, 2, 5

Jingpei Lu, Florian Richter, 和 Michael C. Yip。无标记相机到机器人姿态估计通过自监督模拟到真实迁移。在IEEE/CVF计算机视觉与模式识别会议论文集，页码21296-21306, 2023。1, 2, 5

[25] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In Proceedings of the IEEE International Conference on Robotics and Automation, pages 3400-3407. IEEE, 2011.1,2

Edwin Olson。Apriltag:一种鲁棒且灵活的视觉基准系统。在IEEE国际机器人与自动化会议论文集，页码3400-3407。IEEE, 2011。1,2

[26] Andreas Papadimitriou, Sina Sharif Mansouri, and George Nikolakopoulos. Range-aided ego-centric collaborative pose estimation for multiple robots. Expert Systems with Applications, 202:117052, 2022. 1

Andreas Papadimitriou, Sina Sharif Mansouri, 和 George Nikolakopoulos。范围辅助自我中心协作多机器人姿态估计。专家系统与应用，202:117052, 2022。1

[27] Frank C Park and Bryan J Martin. Robot sensor calibration: solving on the euclidean group. IEEE Transactions on Robotics and Automation, 10(5):717-721, 1994. 2

Frank C Park 和 Bryan J Martin。机器人传感器校准:在欧几里得群上求解 。IEEE机器人与自动化汇刊，10(5):717-721, 1994。2

[28] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. arXiv preprint arXiv:2007.08501, 2020. 2

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, 和 Georgia Gkioxari。使用pytorch3d加速3D深度学习。arXiv预印本arXiv:2007.08501, 2020。2

[29] Yara Rizk, Mariette Awad, and Edward W Tunstel. Cooperative heterogeneous multi-robot systems: A survey. ACM Computing Surveys (CSUR), 52(2):1-31, 2019. 1

Yara Rizk, Mariette Awad, 和 Edward W Tunstel。协作异构多机器人系统:综述。ACM计算调查(CSUR)，52(2):1-31, 2019。1

[30] Alessandro Simoni, Stefano Pini, Guido Borghi, and Roberto Vezzani. Semi-perspective decoupled heatmaps for 3D robot pose estimation from depth maps. IEEE Robotics and Automation Letters, 7(4):11569-11576, 2022. 2

Alessandro Simoni, Stefano Pini, Guido Borghi, 和 Roberto Vezzani。半透视解耦热图用于从深度图进行3D机器人姿态估计。IEEE机器人与自动化快报，7(4):11569-11576, 2022。2

[31] Alessandro Simoni, Francesco Marchetti, Guido Borghi, Federico Becattini, Lorenzo Seidenari, Roberto Vezzani, and Alberto Del Bimbo. Robot pose nowcasting: Forecast the future to improve the present. arXiv preprint arXiv:2308.12914, 2023. 2

Alessandro Simoni, Francesco Marchetti, Guido Borghi, Federico Becattini, Lorenzo Seidenari, Roberto Vezzani, 和 Alberto Del Bimbo。机器人姿态即时预测:预测未来以改善现在。arXiv预印本arXiv:2308.12914, 2023。2

[32] Mikael Svenstrup, Soren Tranberg, Hans Jorgen Andersen, and Thomas Bak. Pose estimation and adaptive robot behaviour for human-robot interaction. In Proceedings of the IEEE International Conference on Robotics and Automation, pages 3571-3576. IEEE, 2009. 1

Mikael Svenstrup, Soren Tranberg, Hans Jorgen Andersen, 和 Thomas Bak。姿态估计与自适应机器人行为用于人机交互。在IEEE国际机器人与自动化会议论文集，页码3571-3576。IEEE, 2009。1

[33] Yang Tian, Jiyao Zhang, Zekai Yin, and Hao Dong. Robot structure prior guided temporal attention for camera-to-robot pose estimation from image sequence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8917-8926, 2023. 2

Yang Tian, Jiyao Zhang, Zekai Yin, 和 Hao Dong。机器人结构先验引导的时间注意力用于从图像序列进行相机到机器人姿态估计。在IEEE/CVF计算机视觉与模式识别会议论文集，页码8917-8926, 2023。2

[34] Yang Tian, Jiyao Zhang, Zekai Yin, and Hao Dong. Robot structure prior guided temporal attention for camera-to-robot pose estimation from image sequence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8917-8926, 2023. 1

杨天、张继尧、尹泽凯、董浩。机器人结构先验引导的时间注意力用于图像序列中的相机到机器人姿态估计。在IEEE/CVF计算机视觉与模式识别会议论文集，第8917-8926页，2023年。1

[35] Yang Tian, Jiyao Zhang, Guowei Huang, Bin Wang, Ping Wang, Jiangmiao Pang, and Hao Dong. Robokeygen: Robot pose and joint angles estimation via diffusion-based 3D key-point generation. arXiv preprint arXiv:2403.18259, 2024. 2, 5

杨天、张继尧、黄国伟、王斌、王平、庞江淼、董浩。Robokeygen:通过基于扩散的3D关键点生成进行机器人姿态和关节角度估计。arXiv预印本arXiv:2403.18259，2024年。2, 5

[36] Chengjun Xu, Xinyi Yu, Zhengan Wang, and Linlin Ou. Multi-view human pose estimation in human-robot interaction. In Proceedings of the Annual Conference of the IEEE Industrial Electronics Society, pages 4769-4775. IEEE, 2020. 1

徐成俊、余欣怡、王振干、欧林林。人机交互中的多视角人体姿态估计。在IEEE工业电子学会年会议论文集，第4769-4775页。IEEE，2020年。1

[37] Dekun Yang and John Illingworth. Calibrating a robot camera. In Proceedings of the British Machine Vision Conference, pages 1-10, 1994. 2

杨德坤、John Illingworth。校准机器人相机。在英国机器视觉会议论文集，第1-10页，1994年。2

[38] Xiaopin Zhong, Wenxuan Zhu, Weixiang Liu, Jianye Yi, Chengxiang Liu, and Zongze Wu. G-SAM: A robust one-shot keypoint detection framework for based robot pose estimation. Journal of Intelligent & Robotic Systems, 109(2): 28, 2023. 2

钟晓品、朱文轩、刘伟祥、易建业、刘成祥、吴宗泽。G-SAM:一个基于 的机器人姿态估计的鲁棒一次性关键点检测框架。智能与机器人系统杂志，109(2): 28，2023年。2

# Supplementary Material

# 补充材料

# A1. Encoder and Predictor Architectures

# A1. 编码器和预测器架构

As described in Sec. 3.1, we use Vision Transformer (ViT) [11] architectures for both the encoder and predictor, similar to [3]. The input image, originally sized at pixels, is cropped based on the region of interest, resized to obtain 224 pixels along its longer side, and padded to yield a resolution. A convolutional layer with a kernel size of 16 and a stride of 16 serves as the patch embedding layer, converting the image into patches of size each with a channel dimension of . These patches are flattened, and learnable positional embed-dings, initialized as 2D sinusoidal functions, are added to the patches. The combined representations are then passed through 12 transformer blocks. Each block contains multiheaded self-attention with 12 heads, drop-path regularization [17], layer normalization [4], and a multi-layer per-ceptron (MLP). The output of the final transformer block undergoes another layer normalization step, resulting in the encoder output for .

如第3.1节所述，我们使用Vision Transformer (ViT) [11]架构作为编码器和预测器，类似于[3]。输入图像最初大小为 像素，根据感兴趣区域裁剪，调整大小使其较长边为224像素，并填充以获得 分辨率。一个核大小为16、步幅为16的卷积层作为补丁嵌入层，将图像转换为 个大小为 的补丁，每个补丁的通道维度为 。这些补丁被展平，并添加可学习的位置嵌入，初始化为2D正弦函数。组合后的表示通过12个Transformer块。每个块包含12个头的多头自注意力、drop-path正则化[17]、层归一化[4]和多层感知器(MLP)。最后一个Transformer块的输出经过另一层层归一化步骤，生成编码器输出 用于 。

During evaluation, for the image of size and a patch size of , the number of patches is computed as

在评估期间，对于大小为 的图像和补丁大小为 ，补丁数量计算为

However, during training, only the unmasked patches are considered, so , i.e., .

然而，在训练期间，仅考虑未掩码的补丁，因此 ，即 。

The predictor takes the encoder output and reduces the embedding dimension of the patches from 768 to 384 using a linear layer. It also adds positional embeddings, similar to the encoder. During training, the embeddings corresponding to the unmasked patches and(M - L)learnable mask tokens are concatenated to represent all patches of the original image, including the masked ones. These em-beddings are then processed through 12 transformer blocks. The final output’s dimension is increased to 768 to match the encoder’s output dimension, resulting in the predictor output for .

预测器接收编码器输出，并使用线性层将补丁的嵌入维度从768减少到384。它还添加了位置嵌入，类似于编码器。在训练期间，对应于未掩码补丁的 嵌入和(M - L)个可学习的掩码标记被连接起来，以表示原始图像的所有补丁，包括被掩码的补丁。这些嵌入然后通过12个Transformer块处理。最终输出的维度增加到768，以匹配编码器的输出维度，生成预测器输出 用于 。

The target backbone uses the same architecture as the encoder but directly operates on all patches during training. It produces outputs for . As outlined in the manuscript, during embedding predictive pre-training, an loss between and is used to update the weights of the encoder and predictor. Following [3], the target backbone is updated using an exponential moving average of the encoder’s weights.

目标骨干网络使用与编码器相同的架构，但在训练期间直接操作所有 个补丁。它生成输出 用于 。如手稿所述，在嵌入预测预训练期间，使用 损失在 和 之间更新编码器和预测器的权重。根据[3]，目标骨干网络使用编码器权重的指数移动平均值进行更新。

# A2. Training Settings

# A2. 训练设置

Embedding Predictive Pre-Training: The AdamW optimizer [23] with an initial learning rate of is used for embedding predictive pre-training. The learning rate is linearly increased to over the first 10 epochs and subsequently decreased to using a cosine annealing scheduler. The network is pre-trained for a total of 200 epochs with a batch size of 320 . Weight decay is linearly increased from 0.04 to 0.4 during pre-training. For the exponential moving average (EMA) update of the target backbone’s weights, a momentum value of 0.996 is used, which is linearly increased to 1.0 over the training process.

嵌入预测预训练:使用初始学习率为 的AdamW优化器[23]进行嵌入预测预训练。学习率在前10个周期内线性增加到 ，随后使用余弦退火调度器将其降低到 。网络预训练总共进行200个周期，批量大小为320。在预训练期间，权重衰减从0.04线性增加到0.4。对于目标骨干网络权重的指数移动平均(EMA)更新，使用动量值为0.996，并在训练过程中线性增加到1.0。

Keypoint Detection and Joint Angle Estimation: As detailed in the manuscript, the pre-trained encoder-predictor pair is fine-tuned along with the Keypoint Net and Joint Net. An AdamW optimizer [23] is used with an initial learning rate of , which is decreased to using a cosine annealing scheduler. The network is trained for a total of 200 epochs with a batch size of 140 .

关键点检测和关节角度估计:如手稿中所述，预训练的编码器-预测器对与关键点网络和关节网络一起进行微调。使用初始学习率为 的AdamW优化器[23]，并使用余弦退火调度器将其降低到 。网络训练总共进行200个周期，批量大小为140。

Sim-to-Real Self-Supervised Training: To bridge the sim-to-real gap, the trained networks are fine-tuned on real datasets with self-supervised training, as described in Sec. 3.3. An AdamW optimizer [23] is used with learning rates of for the encoder and predictor, for the joint network. The learning rate for the keypoint network is set close to zero to prevent model collapse. These learning rates are decreased by a factor of over the training process. Models are fine-tuned separately for each real-world dataset for 10 epochs with a batch size of 64 .

从仿真到现实的自监督训练:为了弥合仿真与现实的差距，如第3.3节所述，训练好的网络在真实数据集上通过自监督训练进行微调。使用AdamW优化器[23]，编码器和预测器的学习率为 ，关节网络的学习率为 。关键点网络的学习率设置为接近零，以防止模型崩溃。这些学习率在训练过程中按 的因子降低。每个现实世界数据集分别进行10个周期的微调，批量大小为64。

# A3. Region of Interest Detection

# A3. 感兴趣区域检测

We utilize the pre-trained GroundingDINO [22] object detection model to identify the region of interest, as described in Sec. 3.3. GroundingDINO is a highly accurate open-set object detector that accepts an (image, text) pair as input and outputs bounding boxes corresponding to regions of the image described by the text query. For all real and photo-realistic test datasets, we use the text query "robotic arm." However, for the Panda, Kuka, and Baxter domain-randomized datasets, we use the query "robot" because these images often contain multiple objects, some of which resemble arms and can confuse the detection model. All other parameters of GroundingDINO are left at their default values. To address scenarios where only a portion of the robot is detected, we expand all the detected bounding boxes, especially for real datasets. Increasing all the bounding box sizes by 100 pixels on all sides generally yields robust robot pose estimation results. However, some fine-tuning of this parameter may be necessary for optimal performance depending on the specific dataset. Nonetheless, high performance is obtained even without fine-tuning.

我们使用预训练的GroundingDINO [22]目标检测模型来识别感兴趣区域，如第3.3节所述。GroundingDINO是一个高度准确的开集目标检测器，它接受(图像，文本)对作为输入，并输出与文本查询描述的图像区域相对应的边界框。对于所有真实和照片级真实感测试数据集，我们使用文本查询“机械臂”。然而，对于Panda、Kuka和Baxter的域随机化数据集，我们使用查询“机器人”，因为这些图像通常包含多个对象，其中一些对象类似于手臂，可能会混淆检测模型。GroundingDINO的所有其他参数均保持默认值。为了解决仅检测到机器人一部分的情况，我们扩展了所有检测到的边界框，特别是对于真实数据集。将所有边界框的尺寸在所有方向上增加100像素通常会产生稳健的机器人姿态估计结果。然而，根据具体数据集的不同，可能需要对这一参数进行一些微调以获得最佳性能。尽管如此，即使不进行微调，也能获得高性能。

# A4. Dataset Details

# A4. 数据集详情

We evaluate our method on the DREAM dataset [18], which includes sequences from three robots: Franka Emika Panda (Panda), Kuka iiwa7 (Kuka), and Rethink Robotics Baxter (Baxter). The dataset provides training and testing sequences in both synthetic and real-world settings, as detailed in Table A2. The synthetic data comprises domain-randomized (DR) and photo-realistic (Photo) sequences. For real-world data, sequences of the Panda robot were captured using Microsoft Azure Kinect (AK), Xbox 360 Kinect (XK), and Intel RealSense D415 (RS) cameras, with the cameras positioned at fixed locations. Additionally, the Panda ORB dataset was collected using a RealSense camera but with varying camera placements. Example images from each dataset sequence are illustrated in Fig. A1.

我们在DREAM数据集[18]上评估了我们的方法，该数据集包括来自三个机器人的序列:Franka Emika Panda(Panda)、Kuka iiwa7(Kuka)和Rethink Robotics Baxter(Baxter)。该数据集提供了合成和现实环境中的训练和测试序列，如表A2中所述。合成数据包括域随机化(DR)和照片级真实感(Photo)序列。对于现实世界数据，Panda机器人的序列使用Microsoft Azure Kinect(AK)、Xbox 360 Kinect(XK)和Intel RealSense D415(RS)相机捕获，相机位于固定位置。此外，Panda ORB数据集是使用RealSense相机收集的，但相机位置不同。每个数据集序列的示例图像如图A1所示。

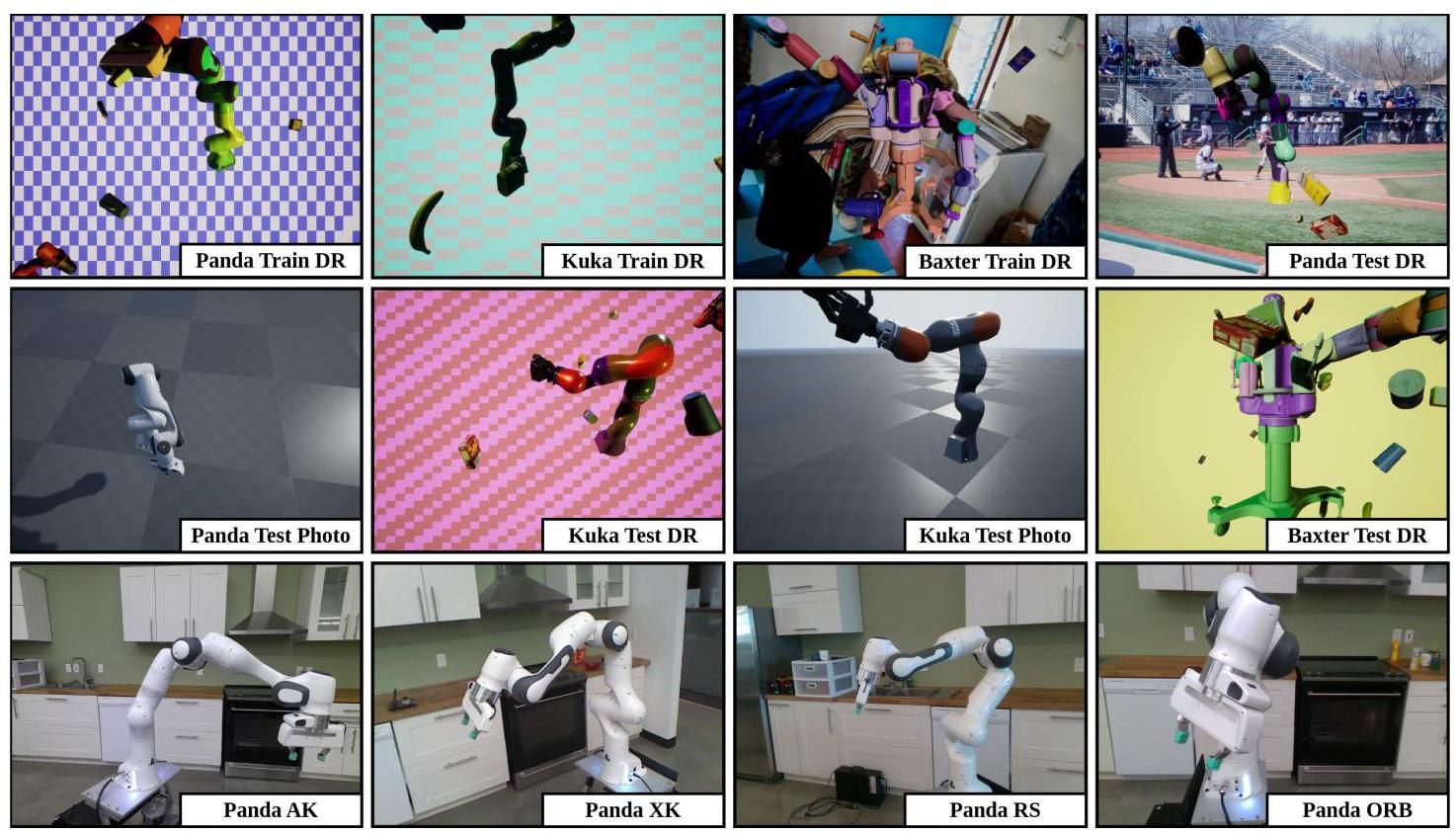


Figure A1. Example images from each of the training and test sequences from the DREAM dataset [18].

图A1. DREAM数据集[18]中每个训练和测试序列的示例图像。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Known Joint Angles | Known Bounding Box |  | | | | Average |
| Panda AK | Panda XK | Panda RS | Panda ORB |
| DREAM-F |  |  | 11413 | 491911 | 2077 | 95319 | 150180 |
| DREAM-Q |  |  | 78089 | 54178 | 27 | 64248 | 49136 |
| DREAM-H | Yes |  | 57 | 7382 | 24 | 25685 | 8287 |
| HPE |  |  | 19 | 24 | 25 | 25 | 23 |
| RoboPose |  |  | 34 | 22 | 26 | 30 | 28 |
| HPE\* |  |  | 46 | - | 61 | 52 | 53 |
| RoboPEPP (Ours) |  |  | 29 | 22 | 23 | 27 | 26 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 已知关节角度 | 已知边界框 |  | | | | 平均值 |
| 熊猫AK | 熊猫XK | 熊猫RS | 熊猫ORB |
| DREAM-F |  |  | 11413 | 491911 | 2077 | 95319 | 150180 |
| DREAM-Q |  |  | 78089 | 54178 | 27 | 64248 | 49136 |
| DREAM-H | 是 |  | 57 | 7382 | 24 | 25685 | 8287 |
| HPE |  |  | 19 | 24 | 25 | 25 | 23 |
| RoboPose |  |  | 34 | 22 | 26 | 30 | 28 |
| HPE\* |  |  | 46 | - | 61 | 52 | 53 |
| RoboPEPP(我们的) |  |  | 29 | 22 | 23 | 27 | 26 |

Table A1. Comparison of robot pose estimation using mean ADD (in millimeters), with lower a value signifying better performance. The best values among methods that use unknown joint angles and unknown bounding boxes during evaluation are bolded. HPE denotes HPE [5] evaluated with the same off-the-shelf bounding box detector as RoboPEPP. HPE\* was not evaluated on Panda XK since corresponding model weights were unavailable.

表A1. 使用平均ADD(以毫米为单位)进行机器人姿态估计的比较，较低的值表示更好的性能。在评估期间使用未知关节角度和未知边界框的方法中，最佳值以粗体显示。HPE 表示使用与RoboPEPP相同的现成边界框检测器评估的HPE [5]。由于相应的模型权重不可用，HPE\*未在Panda XK上进行评估。

# A5. Additional Results

# A5. 附加结果

# A5.1. Mean ADD

# A5.1. 平均ADD

In Table A1, we present the mean ADD (Average Distance) values (ADD defined in Sec. 4.2.1) on the Panda real-world datasets. Consistent with Table 2, we compare our method, RoboPEPP, against DREAM [18], RoboPose [16], HPE [5], and (HPE using our bounding box detection strategy). RoboPEPP achieves the lowest mean ADD across all real-world data sequences among methods that operate with unknown joint angles and bounding boxes. DREAM [18], which detects keypoints and employs them in a solver to estimate the robot pose, is highly sensitive to key-point detection errors. Even a single incorrectly detected keypoint can cause DREAM to fail in pose estimation, leading to high ADD.

在表A1中，我们展示了Panda真实世界数据集上的平均ADD(平均距离)值(ADD定义见第4.2.1节)。与表2一致，我们将我们的方法RoboPEPP与DREAM [18]、RoboPose [16]、HPE [5]和 (使用我们的边界框检测策略的HPE)进行了比较。在未知关节角度和边界框的情况下，RoboPEPP在所有真实世界数据序列中实现了最低的平均ADD。DREAM [18]检测 关键点并在 求解器中使用它们来估计机器人姿态，对关键点检测错误非常敏感。即使是一个错误检测的关键点也可能导致DREAM在姿态估计中失败，从而导致高ADD。

|  |  |  |  |
| --- | --- | --- | --- |
|  | Dataset | Real | #Images |
| Training | Panda Train DR | ✘ | 104972 |
| Kuka Train DR | ✘ | 104977 |
| Baxter Train DR | ✘ | 104982 |
| Testing | Panda Photo | ✘ | 5997 |
| Panda DR | ✘ | 5998 |
| Panda AK | ✓ | 6369 |
| Panda XK | ✓ | 4966 |
| Panda RS | ✓ | 5944 |
| Panda ORB | ✓ | 32315 |
| Kuka Photo | ✘ | 5999 |
| Kuka DR | ✘ | 5997 |
| Baxter DR | ✘ | 5982 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 数据集 | 真实 | #图像 |
| 训练 | 熊猫训练DR | ✘ | 104972 |
| 库卡训练DR | ✘ | 104977 |
| 巴克斯特训练DR | ✘ | 104982 |
| 测试 | 熊猫照片 | ✘ | 5997 |
| 熊猫DR | ✘ | 5998 |
| 熊猫AK | ✓ | 6369 |
| 熊猫XK | ✓ | 4966 |
| 熊猫RS | ✓ | 5944 |
| 熊猫ORB | ✓ | 32315 |
| 库卡照片 | ✘ | 5999 |
| 库卡DR | ✘ | 5997 |
| 巴克斯特DR | ✘ | 5982 |

Table A2. Number of images in each sequence of the dataset.

表A2. 数据集中每个序列的图像数量。

# A5.2. Ablation: Occlusion Robustness

# A5.2. 消融实验:遮挡鲁棒性

In this section, we evaluate the methods from the - bedding Predictive Pre-Training ablation studies (Sec. 4.3) on the occlusion dataset described in Sec. 4.2.3. Specifically, we compare the following models: (1) a version of RoboPEPP without pre-training, (2) a version pre-trained with random masking instead of joint-specific masking, (3) the standard RoboPEPP (pre-trained with joint masking), and (4) a model pre-trained with joint masking but fine-tuned without masking during the encoder-predictor fine-tuning phase. As shown in Fig. A2, and similar to Fig. 6, we plot the AUC of the ADD metric against the occlusion ratio. Additionally, the percentage decrease in AUC relative to the performance without occlusion is annotated on the plot. Among the methods, RoboPEPP achieves the best performance across all occlusion ratios. While the framework with random-masking-based pre-training and the one fine-tuned without masking achieve performance comparable to RoboPEPP under zero occlusion, their performances degrade more rapidly as the occlusion ratio increases.

在本节中，我们评估了 - bedding预测预训练消融研究(第4.3节)中描述的方法在遮挡数据集上的表现(第4.2.3节)。具体来说，我们比较了以下模型:(1)未进行预训练的RoboPEPP版本，(2)使用随机掩码而非关节特定掩码进行预训练的版本，(3)标准的RoboPEPP(使用关节掩码进行预训练)，以及(4)使用关节掩码进行预训练但在编码器-预测器微调阶段未使用掩码进行微调的模型。如图A2所示，与图6类似，我们绘制了ADD指标的AUC随遮挡比例的变化。此外，图中还标注了相对于无遮挡情况下AUC的百分比下降。在这些方法中，RoboPEPP在所有遮挡比例下均表现最佳。虽然基于随机掩码预训练的框架和未使用掩码微调的框架在零遮挡情况下表现与RoboPEPP相当，但随着遮挡比例的增加，它们的性能下降得更快。

# A6. Additional Qualitative Comparison

# A6. 额外的定性比较

In this section, we provide additional examples of qualitative comparisons. Fig. A3 presents examples from the occlusion dataset discussed in Sec. 4.2.3. Fig. A4 shows comparisons on the Franka Photo dataset, while Fig. A5 highlights results on the real-world datasets Franka RS and AK. Lastly, Fig. A6 focuses on real-world images of the Franka robot collected in the lab under highly cluttered and occluded conditions. For all examples, comparisons are made against RoboPose [16] and HPE [5]. Rectangles are used to emphasize areas where these methods perform poorly, while RoboPEPP shows higher accuracy.

在本节中，我们提供了额外的定性比较示例。图A3展示了第4.2.3节中讨论的遮挡数据集的示例。图A4展示了在Franka Photo数据集上的比较，而图A5则突出了在真实世界数据集Franka RS和AK上的结果。最后，图A6聚焦于在实验室中收集的高度杂乱和遮挡条件下的Franka机器人的真实世界图像。对于所有示例，我们与RoboPose [16]和HPE [5]进行了比较。矩形用于强调这些方法表现不佳的区域，而RoboPEPP显示出更高的准确性。

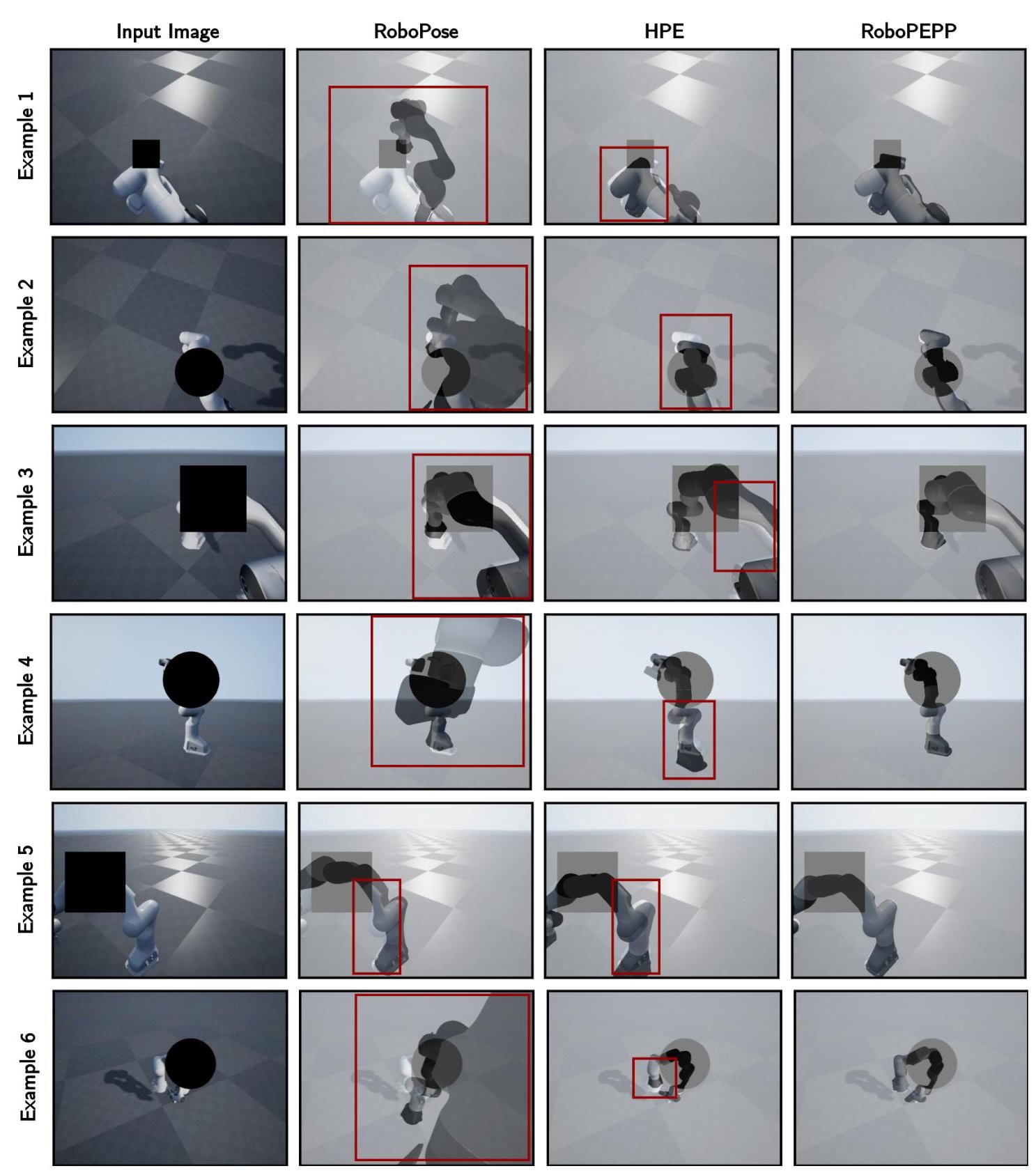


Figure A3. Qualitative Comparison on Occlusion dataset: Predicted poses and joint angles are used to generate a mesh overlaid on the original image, where closer alignment indicates greater accuracy. Highlighted rectangles indicate regions where other methods’ meshes misalign, while RoboPEPP achieves high precision.

图A3. 遮挡数据集上的定性比较:预测的姿势和关节角度用于生成覆盖在原始图像上的网格，其中更接近的对齐表示更高的准确性。高亮矩形表示其他方法的网格未对齐的区域，而RoboPEPP实现了高精度。

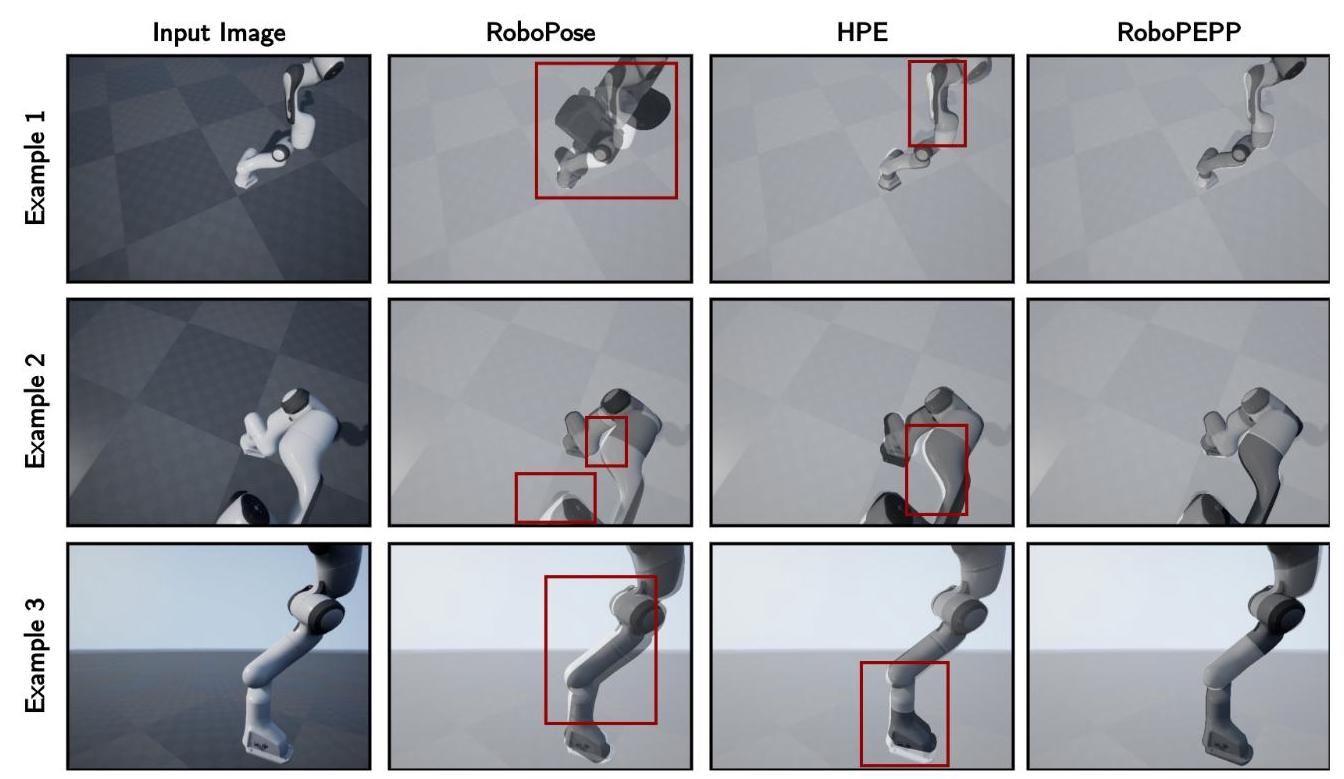


Figure A4. Qualitative Comparison on Panda Photo dataset: Predicted poses and joint angles are used to generate a mesh overlaid on the original image, where closer alignment indicates greater accuracy. Highlighted rectangles indicate regions where other methods’ meshes misalign, while RoboPEPP achieves high precision.

图A4. Panda Photo数据集上的定性比较:预测的姿势和关节角度用于生成覆盖在原始图像上的网格，其中更接近的对齐表示更高的准确性。高亮矩形表示其他方法的网格未对齐的区域，而RoboPEPP实现了高精度。

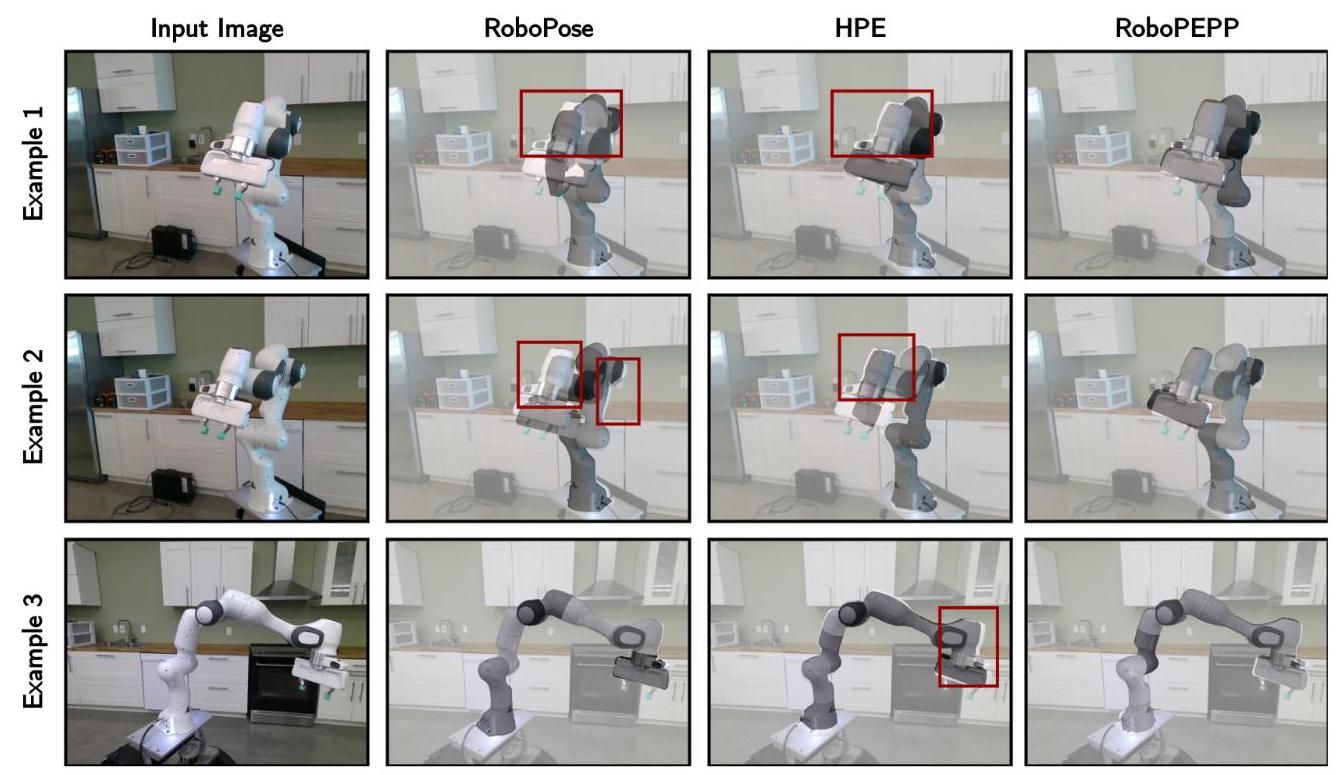


Figure A5. Qualitative Comparison on Panda RS (Example 1 and 2) and Panda AK (Example 3) datasets: Predicted poses and joint angles are used to generate a mesh overlaid on the original image, where closer alignment indicates greater accuracy. Highlighted rectangles indicate regions where other methods’ meshes misalign, while RoboPEPP achieves high precision.

图A5. Panda RS(示例1和2)和Panda AK(示例3)数据集上的定性比较:预测的姿势和关节角度用于生成覆盖在原始图像上的网格，其中更接近的对齐表示更高的准确性。高亮矩形表示其他方法的网格未对齐的区域，而RoboPEPP实现了高精度。

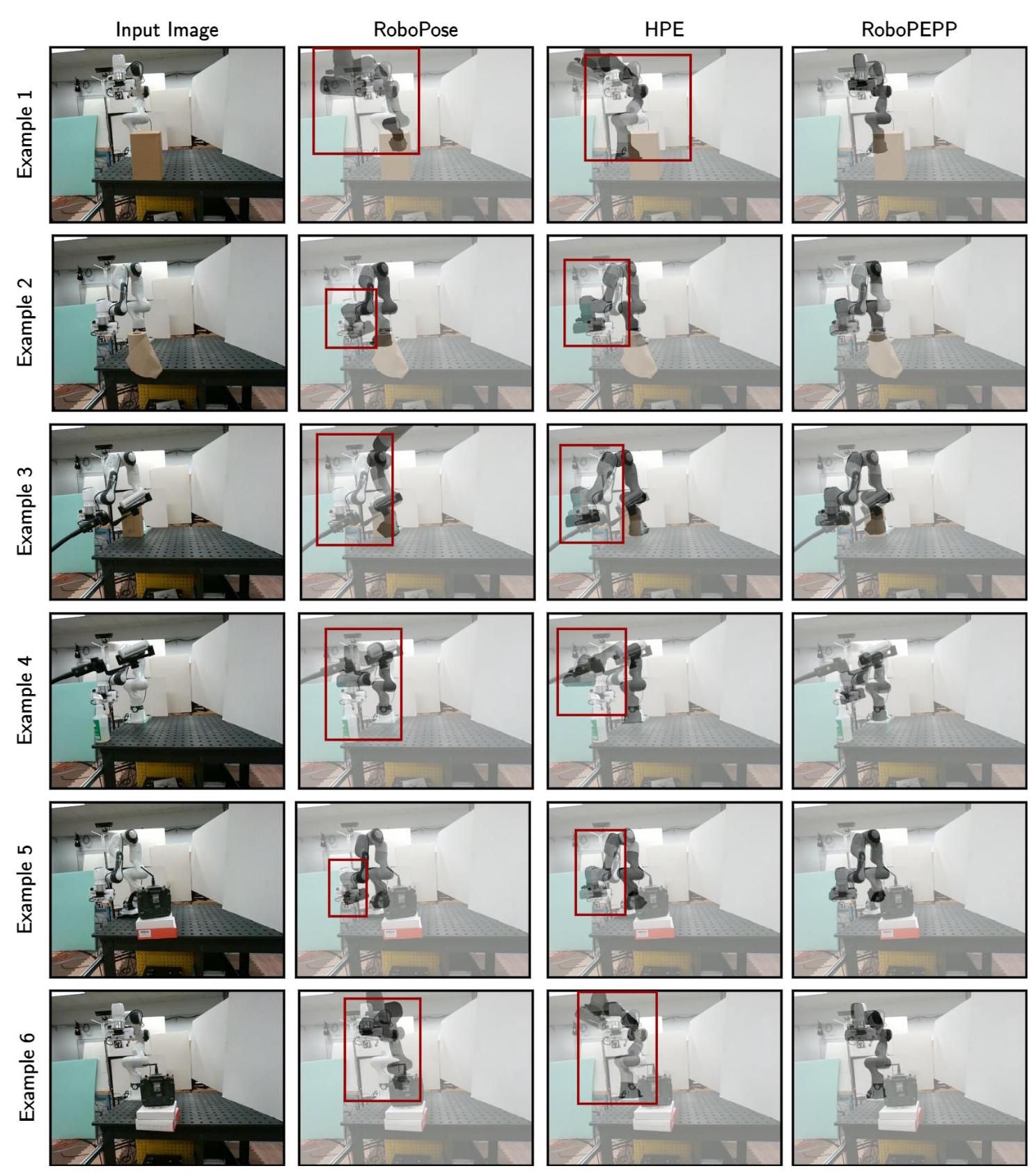


Figure A6. Qualitative Comparison on Additional Real-World Images: These images are collected in highly cluttered environments with robot occlusions. Predicted poses and joint angles generate a mesh overlaid on the original image, where closer alignment indicates greater accuracy. Highlighted rectangles indicate regions where other methods’ meshes misalign, while RoboPEPP achieves high precision.

图A6. 额外真实世界图像上的定性比较:这些图像是在高度杂乱的环境中收集的，存在机器人遮挡。预测的姿势和关节角度用于生成覆盖在原始图像上的网格，其中更接近的对齐表示更高的准确性。高亮矩形表示其他方法的网格未对齐的区域，而RoboPEPP实现了高精度。

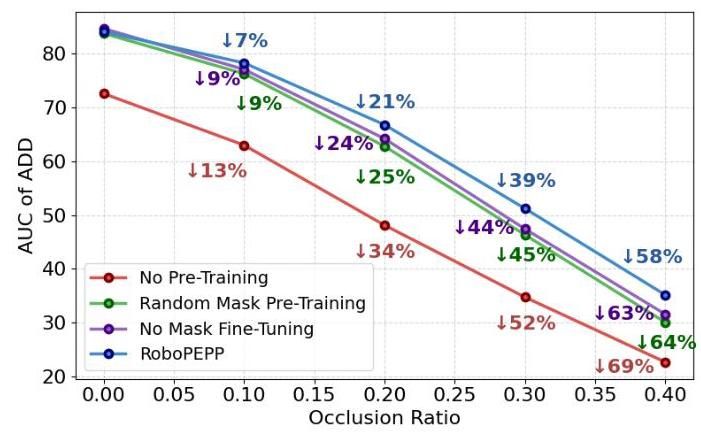


Figure A2. AUC comparison of the distance metric under varying occlusion levels, evaluated on the dataset in Sec. 4.2.3. Percentages next to the lines indicate the relative drop in each method’s performance compared to their performance without occlusions.

图A2. 不同遮挡水平下距离指标的AUC比较，评估基于第4.2.3节中的数据集。线条旁边的百分比表示每种方法相对于无遮挡情况下性能的相对下降。

1. \*Corresponding author: rgg9769@nyu.edu.This paper is supported in part by the Army Research Office under grant number W911NF- 21-1-0155 and by the New York University Abu Dhabi (NYUAD) Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010.

   \*通讯作者:rgg9769@nyu.edu。本文部分由陆军研究办公室资助，资助号为W911NF-21-1-0155，并由纽约大学阿布扎比(NYUAD)人工智能与机器人中心资助，该中心由Tamkeen通过NYUAD研究院奖CG010资助。 [↑](#footnote-ref-29)