# Abstract

**Given 2d joint locations, predict 3d postitions, includes an array of off-the-self 2d detector that have ben trained end-to-end for 2d.**
**results**:

* a large portion of the error of 3d stems from visual analysis
* suggests direction to further advance the 3d human pose estimation

# 1 Introduction

之前为了保证在提供一张2d表示预测3d时，算法不随光线、形状、颜色等一些背景因素改变，一般使用特征如silhouettes(剪影）、shape context、SIFT descriptors等，但现在基于深度学习的方法表现更好。
我们将3d预测问题分解为：

* 2d pose estimation
* 3d pose estimation from 3d joint detections

**疑问**：标签怎么办？是用2d的还是3d的，如果3d的，哪来这么多数据?
**回答**： we can train data-hungry algorithms for the 2d-to-3d problem with large amounts of 3d mocap data captured in controlled environments.
**模型的效率**： 对64的一个batch, 3ms（即300fps)
论文说这种效率的来源是一些简单的想法： estimating 3d joints in camera coordinate frame, adding residual connections and using batch normallization.
如果直接在2d的真值上预测能减少30%的错误率。 所以在2d预测方面还有较大的提升空间去优化最终3d预测的结果。

# 2 Previous work

## 2.1 Depth from images

## 2.2 Top-down 3d reasoning

reasoning about 3d human posture from a minimal representation such as 2d projections.

## 2.3 2d to 3d joints

many methods such as using binary decision tree、nearest neighbor queries. and **Deep-net-based 2d to 3d joints**(such as hourglass architecture that instead of regressing 2d joint probability heatmaps, maps to probability distributions in 3d space).

## 2.4 2d to 3d angular pose

estimate the body configuration in terms of angles.

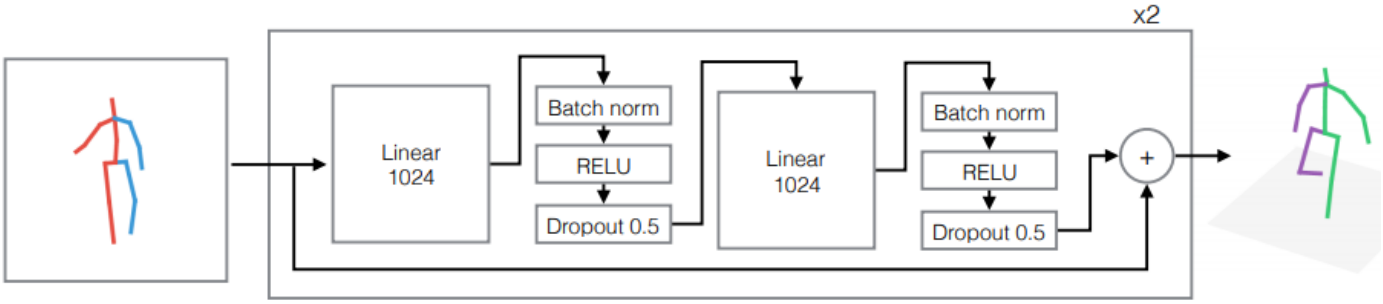
# 3 Solution methodology

输入时一系列2d点$x \in R^{2n}$, 输出是一系列3d点$y \in R^{3d}$, 我们要学习函数$f^*: R^{2n} \rightarrow R^{3n}$,从而尽可能减小数据集额误差:

$$f^* = \min_f \frac{1}{N} \sum_{i=1}^{N} \zeta(f(x_i) - y_i)$$

## 3.1 network design

based on a simple, deep, multilayer neural network with batch normalization , dropout and RELUs as well as residual connections. 如下图所示:




上图中没有展现的，网络在会将输入扩展到1024维，然后会将输出变成3n，此外，该网络一般会使用两个残差网络，意味着总共由6个线性层。

## 3.2 Data preprocessing

- **Camera coordinates**: use the camera frame as a natural choice of global coordinate frame, sinace this makes the 2d to 3d problem similar across different cameras.为防止过拟合，避免因图片的旋转造成预测结果的任意性，it rotating and translatingthe 3d ground-truth according to the inverse transform of the camera.
- **2d detections**: use the stacked hourglass network. cause it is about 10 times faster than CPM to evaluate. when fine-tuned the stacked hourglass model, we get more accurate results
- **Training details**: train our network for 200 epochs usig Adam, starting learning rate of 0.001 and exponential decay. mini-batches of size 64, 5ms for a forward+backward pass on a Titan Xp GPU,which can be **run in real time**. and 2 min for one epoch, 400 min for the training.