

# Semi-supervised Learning for Reconstruction of Multiple Humans from a Single Image

謝宇星

Dept. of CSIE

National Taiwan University  
Taiwan  
r10922024@ntu.edu.tw

陳冠盛

Dept. of CSIE

National Taiwan University.  
Taiwan  
r10922052@ntu.edu.tw

黃奕誠

Dept. of CSIE

National Taiwan University  
Taiwan  
r10922136@ntu.edu.tw

**Abstract**—從單張影像來還原場景中出現的人的位置、姿態等資訊，是一件有挑戰性的事情，除了需要從單張 2D 平面影像來判斷深度，決定其在 3D 空間中的位置，也需要考慮到還原出來的人體姿態是否符合正常的情況。我們參考的論文 [1]，使用了多個資料集（主要是合成資料集 Human3.6m [2]），透過一些 loss 的設定以及 SMPL [3]，使模型可以還原影像中的人物。考量到合成影像以及真實影像，還是有一定的落差，我們期望透過增加 unlabeled 的真實影像，利用 Semi-supervised 的方法，彌補其落差，使模型在真實影像上表現的更加出色。程式碼位於 [https://github.com/cocrs/3DCV\\_final\\_project](https://github.com/cocrs/3DCV_final_project)。

## I. INTRODUCTION

### A. 原始 *Proposal* 調整

我們原先參考的論文 [4] 是利用單張影像來重建場景中的物體，如 Fig.1，當中主要利合成資料集進行訓練，並測試在真實資料集上，因此我們原本打算利用 semi-supervised 的方式加入更多 unlabeled 的真實影像協助訓練，提升模型的效能，但部分因為程式碼及資料集的問題，後續只好繼續尋找其他的論文。



Fig. 1. Corenet [4]

### B. 最後使用之方法

最後選用的論文 [1] 是以重建 multiple human 的 3D pose shape 為目的，不過在過程中，我們有發現，儘管他們主要仍是以合成資料集進行訓練，但其實也利用了不少真實資料集協助訓練，所以我們有意識到可能 semi-supervised 在這邊不會發揮那麼大的功用，不過因為先前嘗試其他論文的程式碼已經花上不少時間，所以還是決定先保留原主題繼續嘗試。

## II. METHODOLOGY (NOVELTY)

### A. Paper

我們所參考的 paper [1] 以 R-CNN [5] 架構為基礎，偵測輸入影像中的人體並以 [3] 提出之 SMPL model 的形式輸出，如 Fig.2 所示。使用 SMPL model 使他們得以將與人體相關的限制加入 loss function，進而可以避免模型產生不合理的肢體動作或發生人體在空間中重疊、複數人體在空間中的深度排序錯

誤等。本次專題保留 [1] 中的主要架構，並針對 [1] 在訓練模型的過程使用的資料集嘗試進行修改。

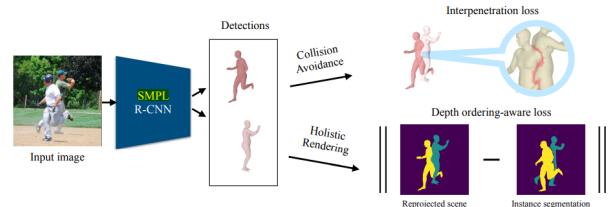


Fig. 2. multiperson framework [1]

### B. Semi-supervised method

我們在本次專題中主要選擇利用 pseudo label 的方法來進行測試，同時我們也參考 [6] 方法，嘗試藉由保留更 confident 的 pseudo label 來提升最終結果。

1) *pseudo label*: 基礎的 pseudo label 流程就如 Fig.3 所示，首先我們會利用 labeled data 訓練出一個 pretrained model，並以此 model 為 unlabeled data 進行 pseudo label 標註，接著即可用 labeled data 加上生成的 pseudo labeled data 進行 semi-supervised 的訓練，其中只利用 labeled data 的 pretrained model 即可當作 baseline 來進行比較。

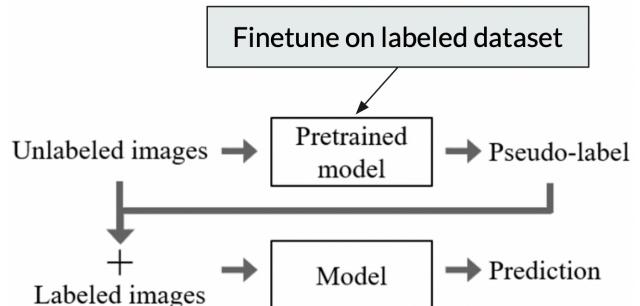


Fig. 3. Pseudo Label 流程圖

2) *confident pseudo label*: 考慮到 pseudo label 中常會包含許多 noisy label，可能會造成 model 訓練上的困難，因此我們希望利用一些方法來增加 pseudo label 的可信度，最後我們從 [6] 得到靈感並進行一些改動，原 paper 的方法主要是應用於 classification 上，他們首先將一張輸入圖片經過 K 個不同的 augmentation 產生 K 張對應圖片，並將所有圖通過 classifier 得到的 prediction 進行 average，來取得較為 confident 的 class distribution。而考慮到我們這篇文章的輸出會包含 bounding box(bbox) 及 keypoints 等資訊，一方面不

適合採取會使圖片大小更改的 augmentation，一方面在 bbox 上取 average 容易受到 outlier 影響，反而造成座標偏差，因此我們設定若以下條件達成，就直接將原圖的 prediction 當作 confident pseudo label：

- (1) 先在原輸入圖上進行一次 prediction，計算 bbox 的數量
- (2) 利用 color jitter 的方式產生 8 張與原圖大小相同的圖片，同樣透過 model 取得 prediction(我們會認為 confident 的圖片，在經過 color jitter 之後 predict 的 bbox 的數量還是會保持一致)
- (3) 為了避免條件太過嚴苛導致 pseudo label 不足，最後設的條件為原圖 bbox 數量和經過 color jitter 產生的 8 個 bbox 數量之標準差不得超過 0.42，且 8 個 bbox 數量中至少有 6 個與原圖的數量相同。

### III. EXPERIMENT (IMPLEMENTATION)

在原 paper 中他們使用了多個 dataset 進行訓練，包含 Human3.6m [2]、COCO [7]、PoseTrack [8]、MPI-INF-3DHP [9]、MPIII [10]，其中以 synthetic 的 Human3.6m 為主，其餘則是 real 的 dataset，可以注意到並不是每個 dataset 都有提供完整的 label，例如 pose、shape 只有 human36m 及 MPI-INF-3DHP 有。因為有些 dataset 無法取得、有些大小過大，考慮到時間因素，最後我們決定從 train、test 各選一個 dataset 來進行測試，train 選擇了包含 pose、shape 的 MPI-INF-3DHP，是一個 single person 的 dataset，test 則是使用 panoptic [11]，當然我們有注意到在無法還原同樣 training 的情況下，我們將很難超越原 paper 的表現。

同樣因為前面提到的 dataset 等因素，我們設計了一個情境來進行 semi-supervised 的測試。由於我們選用的 labeled 的 dataset 數量不算小(約 90000)，在短時間無法蒐集到如此大量的 unlabeled image，所以我們決定從 MPI-INF-3DHP 中隨機 sample 出一個 500 的 subset 當作 training set(MPI 500)，並以 cityscapes [12] 的 5000 張當作 unlabeled dataset 進行測試。為了公平比較，我們兩個 model 都以 multiperson github 提供的 checkpoint 為基礎，一個只在 MPI 500 上 fintune 作為 baseline，並利用先前提到的方法生成 pseudo label 及 confident pseudo label，來訓練我們的 semi-supervised model。測試於 Panoptic 各 sequence 的結果如 Tab.I 和 Fig.4 所示，其中使用 MPJPE (Mean Per Joint Postion Error ) 進行比較。

Method	Haggling	Mafia	Ultim.	Pizza	Mean
github ckpt	129.1	132.8	153.0	153.6	142.1
our baseline	132.5	137.4	157.9	158.9	146.7
pseudo label	132.1	134.4	153.4	157.6	144.4
conf. pseudo label	130.4	135.7	153.6	156.3	144.0

TABLE I  
PANOPTIC 實驗結果 (MPJPE)

可以發現到在 MPI 500 進行 fintune 後結果會下降，但因為我們無法取得所有 dataset 來訓練，所以也算是預期中的結果，而底下的兩種方法相較於 baseline 都表現得比較好，可能表示在我們的設置情境下，semi-supervised 還是發揮了效果。另外值得注意的是選出的 confident pseudo label(480) 實比 pseudo label(4824) 少掉許多，但表現卻能略好於 psuedo label，表示選出 confident data 來訓練是有助於提升結果的。

### IV. DISCUSSION (WHAT WE HAVE LEARNED)

這次的實驗結果不如我們 proposal 的目標，我們認為如同前面提過的幾點，包含 multiperson [1] 已使用滿多的 real dataset 進行訓練，另外 dataset 以及時間的限制等問題，使

得沒有在最適當的情境下測試 semi supervised 的方法，但我們同時也從測試中得到了一些收穫，也就是相較於大量的 noisy data，可能較少 confident data 能帶來更好效果。

### V. DIVISION OF WORK

#### A. 謝宇星

- 1) Survey paper
- 2) Find appropriate unlabeled dataset
- 3) Test github code

#### B. 陳冠盛

- 1) Survey paper
- 2) Test github code
- 3) Record demo video

#### C. 黃奕誠

- 1) Survey paper
- 2) Test github code

### REFERENCES

- [1] Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., & Daniilidis, K. (2020). Coherent reconstruction of multiple humans from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5579-5588).
- [2] Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2013). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence, 36(7), 1325-1339.
- [3] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A skinned multi-person linear model. ACM transactions on graphics (TOG), 34(6), 1-16.
- [4] Popov, S., Bauszat, P., & Ferrari, V. (2020, August). Corenet: Coherent 3d scene reconstruction from a single rgb image. In European Conference on Computer Vision (pp. 366-383). Springer, Cham.
- [5] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.
- [6] Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. Advances in Neural Information Processing Systems, 32.
- [7] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.
- [8] Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., & Schiele, B. (2018). Posetrack: A benchmark for human pose estimation and tracking. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5167-5176).
- [9] Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., & Theobalt, C. (2017, October). Monocular 3d human pose estimation in the wild using improved cnn supervision. In 2017 international conference on 3D vision (3DV) (pp. 506-516). IEEE.
- [10] Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (pp. 3686-3693).
- [11] Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., ... & Sheikh, Y. (2015). Panoptic studio: A massively multiview system for social motion capture. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3334-3342).
- [12] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3213-3223).

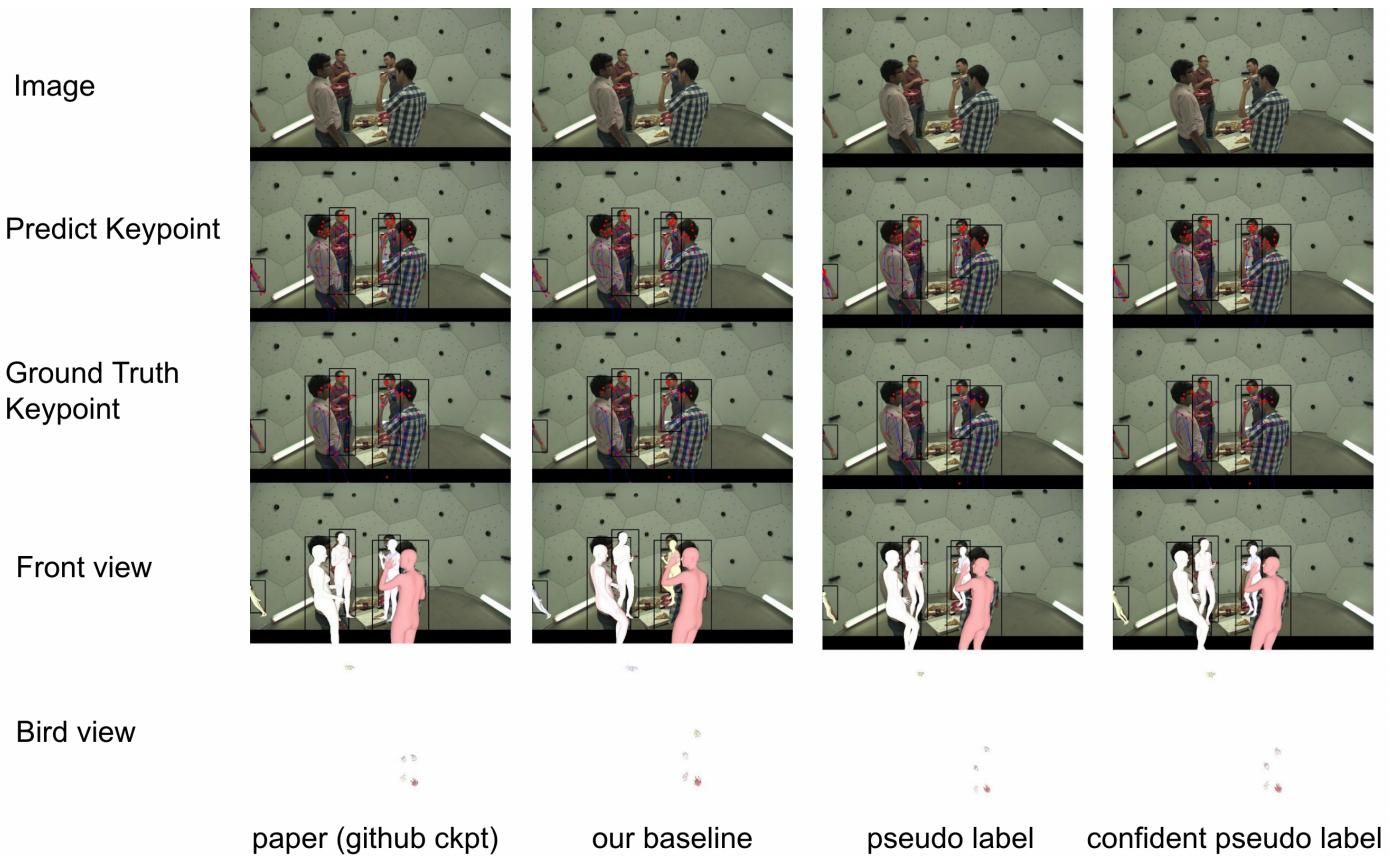


Fig. 4. 視覺化結果