Body-Face Joint Detection via Embedding and Head Hook (Supplementary Materials)

Junfeng Wan*† Jiangfan Deng* Xiaosong Qiu Feng Zhou
Algorithm Research, Aibee Inc.
{jfwan, jfdeng, xsqiu, fzhou}@aibee.com

Method	greedy	assignment	mMR^{-2}/%
FPN + BFJ	✓		52.5
		✓	53.7
RetinaNet + BFJ	✓		63.7
		✓	63.2

Table 1. Comparison of the original *greedy* approach and *assignment* approach in the association process of the BFJ detectors.

1. Further Exploration

In Sec.3 of our paper, after acquisition of the fused similarity matrix $\mathbf{S}$, we simply use a greedy "maximum" operation to find the associated face box for each body (denoted as *greedy* approach). However, there is an alternative solution following the position mode (POS) baseline: solving a linear assignment problem with Hungarian algorithm using $\mathbf{S}$ as the cost matrix (denoted as *assignment* approach). Table. 1 shows the comparison results on CrowdHuman. In FPN, the greedy method outperforms its assignment counterpart by **1.2%** (52.5% *vs.* 53.7%) in mMR^{-2}, while in RetinaNet, the assignment strategy has a marginal superiority of **0.5%** (63.2% *vs.* 63.7%). The results suggest that the similarity values in $\mathbf{S}$ are with enough discriminability so that a simple maximum operation can works well. Moreover, by taking efficiency into consideration, we think our greedy approach is a better choice.

2. Ablations of EML

In the Embedding Matching Loss of our method in Sec.3, we build three types of pairs: body-to-body (*bb*), face-to-face (*ff*) and body-to-face (*bf*), in which only the distances between bodies and faces (pairs of the case *bf*) are actually used during inference. Therefore, we conduct ablation studies on CrowdHuman to demonstrate the effect of the other two types of pairs during training (shown in Table. 2). Furthermore, we also make another trial: constructing embedding pairs between proposals and gt-boxes (the 3rd line of Table. 2, depicted as GT-mode). These ablation experiments indicate that our approach is the most effective one.

Method	mMR^{-2}
FPN + BFJ (ours)	52.5
FPN + BFJ (w/o bb & ff)	54.1
FPN + BFJ GT-mode	53.7

Table 2. Ablation experiments of EML on CrowdHuman.

Dataset	*body*	*head*	*face*
CrowdHuman	339565	339565	190083
CityPersons	14762	14554	6487

Table 3. Stastics of our benchmark datasets (train set).

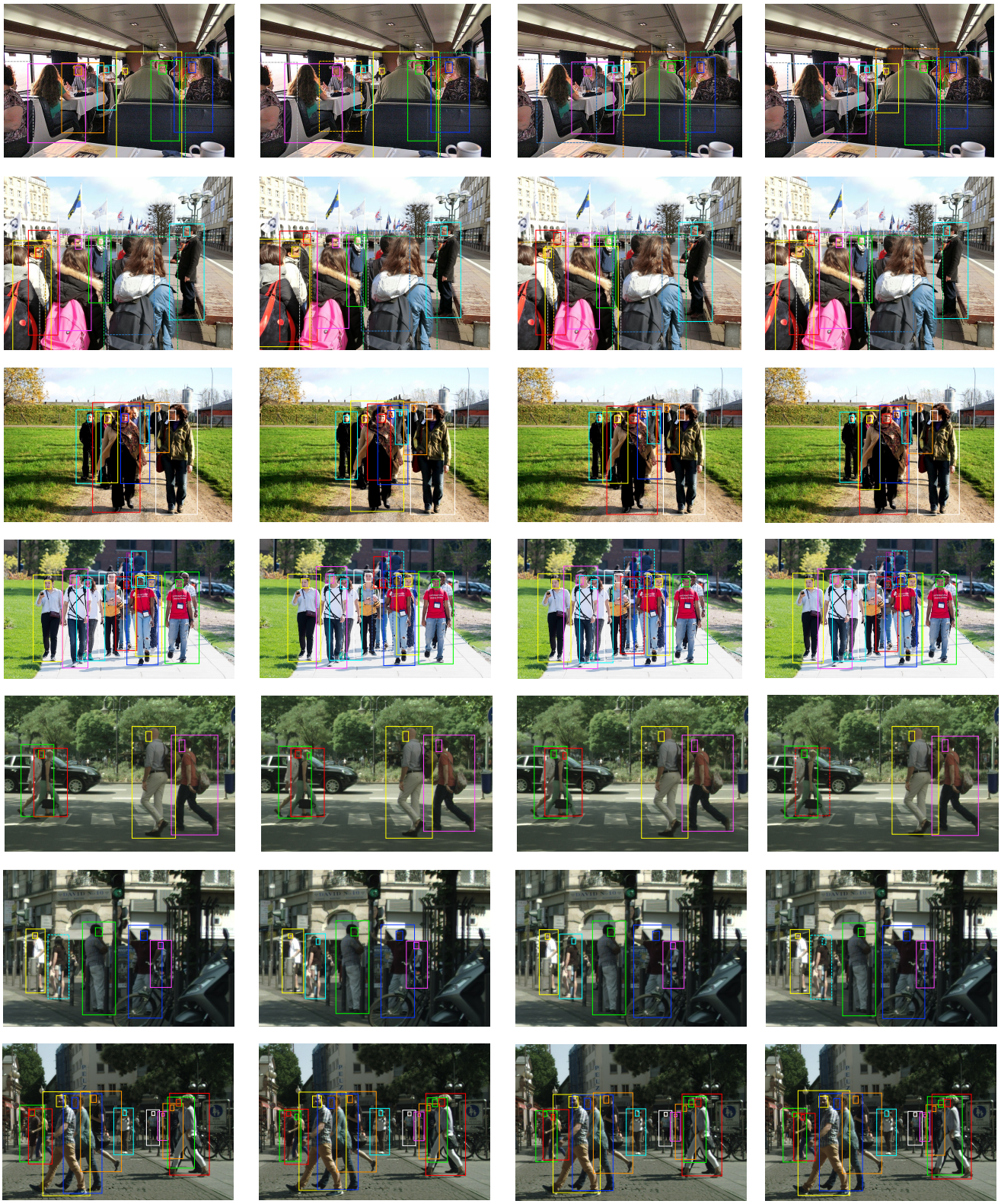
3. Evidence for the existance of heads

As shown in Table. 3. In CrowdHuman, every person has a body box and a head box. In CityPersons, 98.7% of people have both body and head box. So we have the conclusion that "head almost always virtually exists for each pedestrian".

4. Supplementary Visual Comparisons

We post supplementary visual comparisons in Fig. 1. Qualitively, our method mainly solves the miss-matching issues when people are very close with each other. For example, in the first three columns (baselines) of the first row in Fig. 1, bodies of people who are back to the camera (their faces are invisible) are likely to be matched with faces of people in the distance who are face to the camera. In the contrast, our BFJ detector can make correct associations. In the first (CAS) and second (POS) column of the fourth row, bodies of two people in the distance (upper middle of the image) are matched with faces of people in the nearby place only because the bottom region of their full body boxes involve face of another people. In BFJ, these bad cases can be easily avoided thanks to the powerful embedding and head hook guidance.

*Equal contribution.
†Work done during Junfeng's internship at Aibee.



CAS POS POSH BFJ

Figure 1. Supplementary visual comparisons on the FPN detector of the CAS, POS, POSH baselines and our BFJ respectively. Images in the first four lines are from CrowdHuman and those in the last three lines come from CityPersons. Solid boxes with the same color denote one pair of body and face associated. Dashed box denotes the detected body or face that is not associated successfully.