6. Appendix

This document presents visualizations of the model's performance when trained with various relative point strategies and matching approximations, as well as a comparison with the baseline in different scenarios.

6.1. Relative points effect

Deformable DETR is a method that predicts relative points for sampling image regions, rather than using full-image features. This approach has been found to be more efficient in terms of training time required to achieve good convergence. In the case of pair prediction, the relative point is used to represent two bounding boxes instead of one. However, this can lead to convergence issues during training, as the relative point will attempt to be positioned between the centers of the two bounding boxes.

Our results indicate that the most effective approach was to use adaptive relative points. It is important to note that in all cases, the predicted pair coordinates represent an offset from the predicted relative point. The main difference between the experiments was the adaptive use of sampling regions for deformable cross-attention. As shown in Fig. 5, the adaptive relative points model predicts the relative points around the face in face-body pairs and around the body box center when the class consists only of bodies, unlike other experiments that attempted to predict the relative point in an intermediate position between the body and the face.

Whether we use face-relative points or body-relative points, the outcome is similar, as the relative point still aims to stabilize in the middle of the line connecting the pair's bounding boxes. This results in a decrease in the prediction of one class and an improvement in the other.

6.2. Matching strategy effect

We provide a visual comparison of different algorithms for approximating the matching process. However, this visualization does not fully demonstrate the strength of our method, as the association between predictions and ground-truth is not affected as much as the AP results, regardless of the matching strategy used.

In Figure 6, we show the predictions of four different matching approximations for training:

1. Our own matching strategy, which uses head annotations or relatively expected classes for invisible parts.
2. The use of MinCost MaxFlow to match predictions and ground-truth.
3. A simple matching approach that uses Hungarian directly without approximating the cost, so that the cost of one object matching is treated as matching pairs.
4. A body matching approximation, where the body cost is multiplied by two to prioritize pairs matching during training. This means that the body cost is calculated twice when the face is invisible.

6.3. Visualization comparison against BFJ

In order to compare our results with the BFJ method, we have visualized more images from the validation set. Figure 7 displays the performance of each method in various scenarios, including very crowded scenes. The visualization primarily focuses on the association part, hence the bounding boxes are only visualized when there is an association.

6.4. Computational Effort

In table 8, all models use the same ResNet-50 backbone and resolution to measure the computation efficiency of our method against the baseline. Our method requires less number of arithmetic operations and doesn't need any post-processing (nms and matching post-processing weren't included in the flops calculation). We also provide information about the CO2 Emissions (CE) in Kg and Power consumption (PC) in KWh, for training BFJ and PairDETR on 4K samples.

Table 8. Comparison between Our model and the state of the art methods In terms of the number of GFlops, power consumption (PC) and CO2 Emissions (CE) with the same image size

Model	GFlops↓	Params↓	PC↓	CE↓	E2E
POS	218.9	41.3 M	-	-	✗
FPN + BFJ	238.1	55.2 M	0.28	0.096	✗
PairDETR (ours)	**182.2**	**40.7 M**	0.325	0.113	✓

6.5. Geometric approach

It is important to clarify why basic geometric approaches often fail in association tasks, particularly in crowded scenes. Assuming that we have a perfect detector that can detect all faces and bodies in a given image, the problem transforms into connecting two bounding boxes of the same person. So we need to find a cost function C to represent the cost of connecting the face and body. The objective is then to minimize these costs. Using greedy matching produces poor results. It is better to represent this as a graph problem where we have a set of nodes of type A (faces) connected to another set of nodes of type B (body). In the end, we get a bipartite graph that we can solve using Hungarian, maximum-matching, or max-flow algorithms. The definition of the cost function is crucial, and the optimal cost criteria may vary from image to image. Since the geometric approach based only on locations cannot be generalized, researchers conducting this approach added additional outputs to the detection model to extract local data embedding to help build a more general cost function, which is the BFJ baseline, which is very similar. While the geometric approach can find the optimal matching in most cases, there can still be multiple solutions with the same minimum cost value. As a result, the approach may return a set of solutions with the same matching cost. However, only one of these solutions is correct in reality.

6.6. Comparing PairDETR against human pose estimation in association

To verify if human pose estimation can be used for detection and association problem, we performed an experiment using pre-trained state-of-the-art top-down system. We find the results (table 9) promising, yet far from PairDETR performance in association. While finetuning both networks for CrowdHuman dataset should improve detection performance, we hypothesize that it would be inferior to PairDETR due to the two-step approach, while PairDETR is an end-to-end method.

Figure 5. Blue points represent the relative points for the face-body pairs, while green points represent the relative points for the body-only class. As shown, the adaptive method ensures that the points are closer to the center of the face when there is a pair and closer to the center of the body otherwise, while in other cases it tries to optimize the position of the relative point so that it lies between the centers.

Table 9. Detection and association via the human pose estimation compared to PairDETR

Model	resolution	mMR^{-2}	mAP (face&body)
ViTPose-B	1400	88.2	43.4
ViTPose-H	1024	86.6	41
PairDETR	1400	42.9	79.9

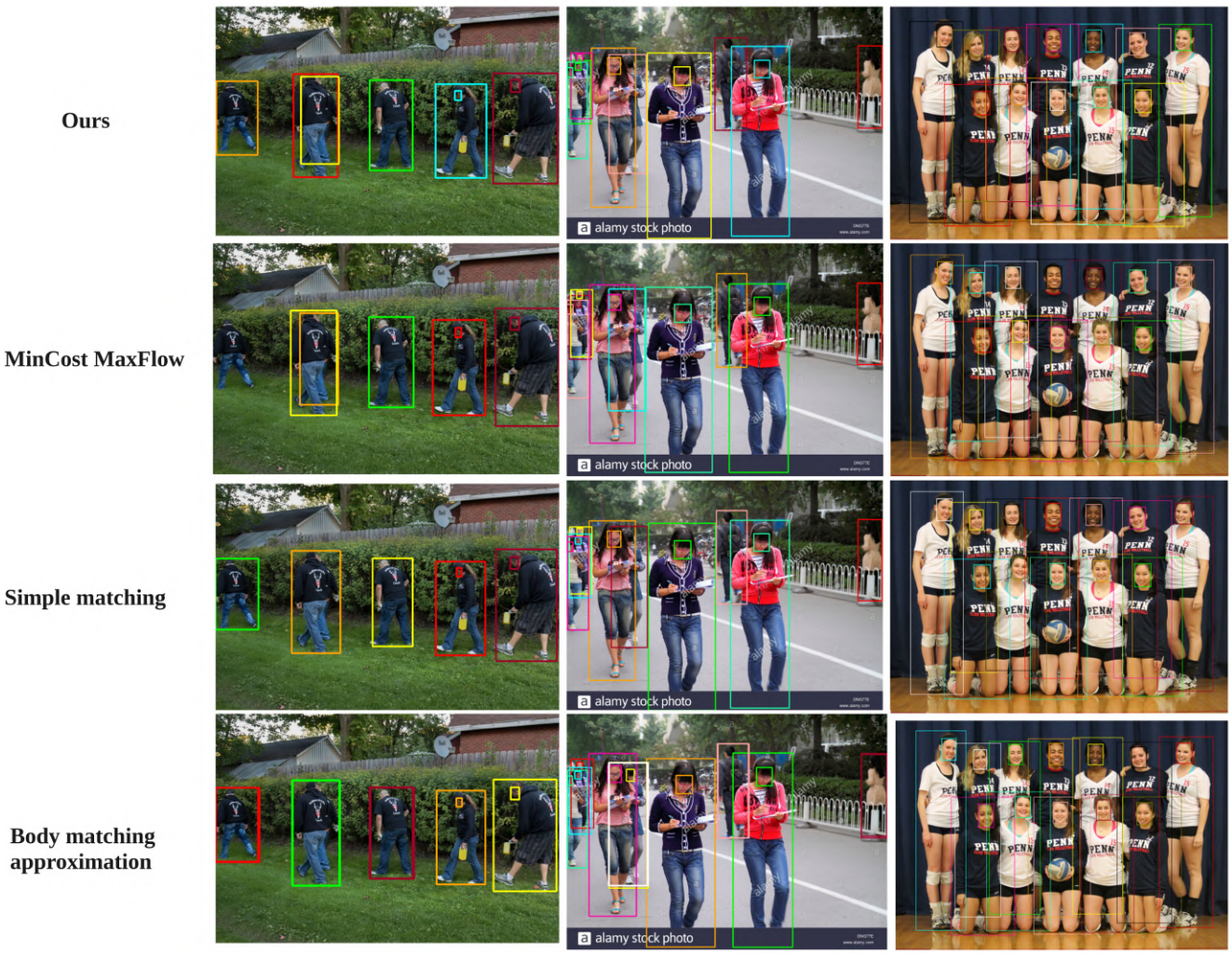
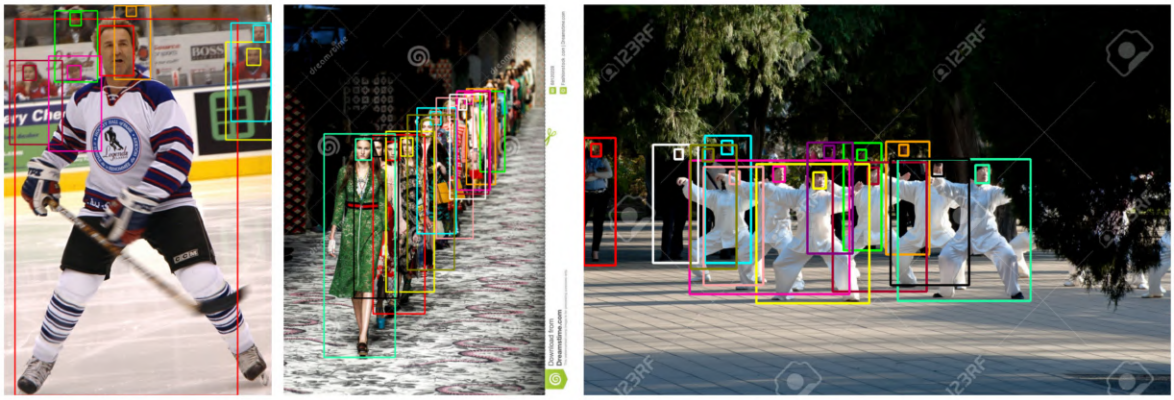
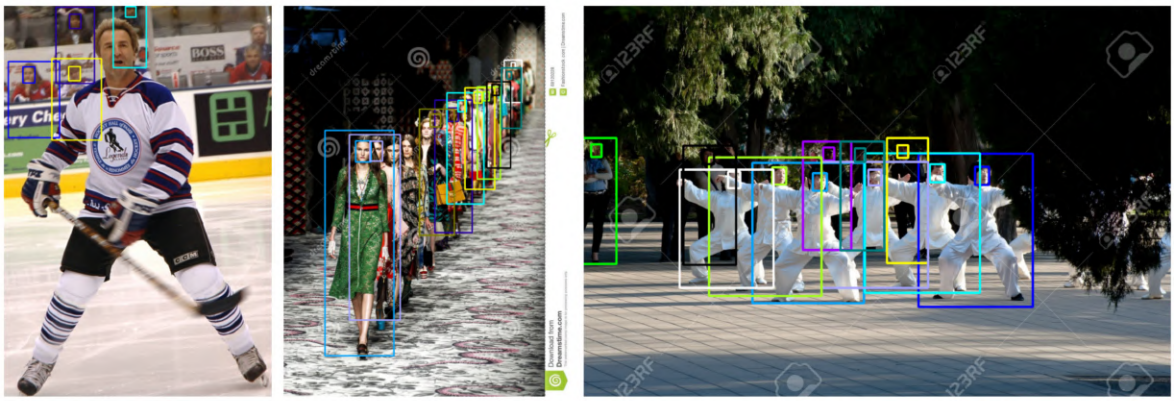


Figure 6. All models prediction use the same threshold 0.5. As shown, our proposed approximation works better than the others.

PairDETR



BFJ



PairDETR

BFJ