Differentiable Multi-Granularity Human Representation Learning for Instance-Aware Human Semantic Parsing
Supplemental Material

Tianfei Zhou[1] , Wenguan Wang[1*] , Si Liu[2] , Yi Yang[3] , Luc Van Gool[1]
[1]Computer Vision Lab, ETH Zurich [2]Institute of Artificial Intelligence, Beihang University [3]University of Technology Sydney
https://github.com/tfzhou/MG-HumanParsing

In this document, we first present more implementation details of our differentiable matching algorithm (§1). Then, we provide more qualitative instance-level parsing results on MHP$_{v2}$ [8], DensePose-COCO [1], and PASCAL-Person-Part [7] datasets (§2). Last, we offer several failure cases for more comprehensive analysis of our model (§3).

1. Differentiable Keypoint Matching Algorithm

Here, we supplement the proposed differentiable keypoint matching algorithm with more details. Recall that in the main article, we solve the following linear programming problem individually for each limb to find a real-valued assignment matrix Y:

$$\min_{Y} \text{Tr}(-AY^\top), \tag{1}$$

$$s.t. \quad Y\mathbf{1}_{N_{k'}} \leq \mathbf{1}_{N_k}, \quad Y^\top \mathbf{1}_{N_k} \leq \mathbf{1}_{N_{k'}}, \quad Y \geq 0. \tag{2}$$

Here, both the target function (Eq. (1)) and constraint functions (Eq. (2)) are convex. Although there are some standard solvers (*e.g.*, simplex method, interior-point method) [6] for such convex constrained optimization problem, they are not differentiable. Fortunately, projected gradient descent (PGD) algorithm provides us a powerful and differentiable solution to this problem. Let us denote the constraints in Eq. (2) as $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3$, where $\mathcal{C}_1 = Y\mathbf{1}_{N_{k'}} \leq \mathbf{1}_{N_k}$, $\mathcal{C}_2 = Y^\top \mathbf{1}_{N_k} \leq \mathbf{1}_{N_{k'}}$, and $\mathcal{C}_3 = Y \geq 0$. PGD estimates Y by iterating the following equation:

$$Y \leftarrow \mathcal{F}^{\mathcal{C}}(Y - \alpha \triangledown f(Y)). \tag{3}$$

Here, $f(Y) = \text{Tr}(-AY^\top)$, $\triangledown f(Y) = -A$, and parameter α denotes the step size of gradient descent. The projection operation $\mathcal{F}^C$ is also an optimization problem:

$$\mathcal{F}^{\mathcal{C}}(Y) = \underset{Y' \in \mathcal{C}}{\text{argmin}} \frac{1}{2}||Y' - Y||_2^2. \tag{4}$$

Given Y, $\mathcal{F}^{\mathcal{C}}$ tries to find a point $Y' \in \mathcal{C}$ that is closet to Y. To project Y onto the constraint set $\mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3$, we adopt Dykstra's cyclic projection algorithm, which is proved to be convergent for projection onto the non-empty intersection of convex sets in Hilbert space [4, 3]. Specifically, let $\{\mathcal{C}_k\}_{k=1}^K$ be a family of K closed convex subsets in Hilbert space such that $\bigcap_{k=1}^K \mathcal{C}_k \neq \emptyset$. The algorithm breaks the whole constraint set into multiple individual sets, then iterates by passing sequentially over the individual sets and projecting onto each one a deflected version of the previous iteration. Dykstra's algorithm makes use of several additional auxiliary variables. Starting with an initial point $X_1^0 = Y - \alpha \triangledown f(Y)$ and $u_0^1 = \cdots = u_0^K = \mathbf{0}$, the algorithm updates the following equations at each iteration (*i.e.*, $n = 1, 2, 3, \cdots$):

$$\begin{aligned}
X_n^1 &= \mathcal{F}^{\mathcal{C}_1}(X_n^0 + u_{n-1}^1), \\
u_n^1 &= X_n^0 + u_{n-1}^1 - X_n^1, \\
&\cdots \\
X_n^k &= \mathcal{F}^{\mathcal{C}_k}(X_n^{k-1} + u_{n-1}^k), \\
u_n^k &= X_n^{k-1} + u_{n-1}^k - X_n^k, \\
&\cdots \\
X_n^K &= \mathcal{F}^{\mathcal{C}_K}(X_n^{K-1} + u_{n-1}^K), \\
u_n^K &= X_n^{K-1} + u_{n-1}^K - X_n^K.
\end{aligned} \tag{5}$$

Algorithm 1 Projected Gradient Descent for Differentiable Keypoint Matching

Input: A, α, T, N
Initialization: $Y_0 = \texttt{zeros\_like}(A)$
for $t = 1, \cdots, T$ **do**
  $X_1^0 = Y_{t-1} - \alpha \triangledown f(Y_{t-1})$
  $u_0^1 = 0$, $u_0^2 = 0$, $u_0^3 = 0$
  for $n = 1, \cdots, N$ **do**
    $X_n^1 = \mathcal{F}^{\mathcal{C}_1}(X_n^0 + u_{n-1}^1)$, $u_n^1 = X_n^0 + u_{n-1}^1 - X_n^1$
    $X_n^2 = \mathcal{F}^{\mathcal{C}_2}(X_n^1 + u_{n-1}^2)$, $u_n^2 = X_n^1 + u_{n-1}^2 - X_n^2$
    $X_n^3 = \mathcal{F}^{\mathcal{C}_3}(X_n^2 + u_{n-1}^3)$, $u_n^3 = X_n^2 + u_{n-1}^3 - X_n^3$
    $X_{n+1}^0 = X_n^3$
  end for
  $Y_t = X_N^3$
end for
return $\hat{Y} = \frac{1}{T}\sum_{t=1}^T Y_t$

*Corresponding author: *Wenguan Wang*.

The sequence $\{X_n^K\}_n$ converges to the solution of Eq. (4) [2, 5].

In our problem, we have $K=3$ constraints. The projection operator $\mathcal{F}^{\mathcal{C}_k}$ with respect to each constraint $\mathcal{C}_k$ can be easily derived as follows:

$$\mathcal{F}^{\mathcal{C}_1}(Y)=\begin{cases} Y, & \text{if } Y\mathbf{1}_{N_{k'}} \leq \mathbf{1}_{N_k}, \\ Y-\frac{1}{N_{k'}}(Y\mathbf{1}_{N_{k'}}-\mathbf{1}_{N_k})\mathbf{1}_{N_{k'}}^\top, & \text{otherwise;} \end{cases}$$
$$\mathcal{F}^{\mathcal{C}_2}(Y)=\begin{cases} Y, & \text{if } Y^\top\mathbf{1}_{N_k} \leq \mathbf{1}_{N_{k'}}, \\ Y-\frac{1}{N_k}\mathbf{1}_{N_k}(\mathbf{1}_{N_k}^\top Y-\mathbf{1}_{N_k'}^\top), & \text{otherwise;} \end{cases} \quad (6)$$
$$\mathcal{F}^{\mathcal{C}_3}(Y)=Y^+.$$

Here, $\mathcal{F}^{\mathcal{C}_3}$ is a ReLU operator. All these projection operators are differentiable, thus providing us a fully end-to-end pose estimation solution.

We summarize the implementation of the above matching algorithm in Alg. 1, where T and N indicate the number of gradient descent (outer-loop) and projection (inner-loop) steps, respectively. Note that at the t-th outer step, we first get an initially corrected point $X_1^0 = Y_{t-1} - \alpha\nabla f(Y_{t-1})$, and then run Dykstra's algorithm iteratively to project this point onto each individual constraint set (*i.e.*, $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$). In the n-th inner step, we update the correction according to the difference between pre- and post-projection. The final solution is obtained by averaging the intermediate projected assignment matrices, *i.e.*, $\hat{Y}=\frac{1}{T}\sum_{t=1}^{T} Y_t$.

Implementation Details. Benefiting from limb scoring which is able to characterize pair-wise matching between joints, our network can converge very fast with only a small number of steps. In the implementation, we empirically set $T=50$, $N=5$, $\alpha=0.01$ in all experiments. This configuration leads to consistent performance improvement across various datasets, as well as high efficiency.

2. Additional Qualitative Result

We provide additional instance-level human parsing results on the three human parsing datasets, including MHP$_{v2}$ [8] `val` in Fig. 1, DensePose-COCO [1] `test` in Fig. 2, and Pascal-Person-Part [7] `test` in Fig. 3. We observe that our approach can produce compelling parsing results under various challenging situations, *e.g.*, occlusions, small objects, extreme poses, *etc*.

3. Failure Case Analysis

To give a deeper insight into our method, we provide three representative failure cases in Fig. 4. The *first* type (*i.e.*, 1^{st} row) of common mistakes is caused by non-typical, upside-down poses, which results in the failure of human parsing in all levels. Increasing the rotation augmentation visually seems to alleviate these issues, but the overall performance on DensePose-COCO [1] greatly decreases. The *second* type (*i.e.*, 2^{nd} row) of challenges is dim scenes in which background objects are visually similar to humans. Moreover, it is extremely challenging to parse humans at very small scales in such low-light scenarios. *Third*, for some cases in the presence of overlapping parts (*i.e.*, 3^{rd} row), as highlighted in the yellow box, our model finds it hard to precisely distinguish left and right parts (*e.g.*, arms). This further leads to inaccurate instance-level part discrimination. In the future, we will therefore focus on addressing these issues.

References

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 1, 2, 4, 6

[2] Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011. 2

[3] James P Boyle and Richard L Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in order restricted statistical inference*, pages 28–47, 1986. 1

[4] Richard L Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983. 1

[5] Gaffke Norbert and Mathar Rudolf. A cyclic projection algorithm via duality. *Metrika*, 36:29–54, 1989. 2

[6] Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998. 1

[7] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, 2017. 1, 2, 5

[8] Jian Zhao, Jianshu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *ACMMM*, 2018. 1, 2, 3

Figure 1: **Instance-level human semantic parsing results** on MHP$_{v2}$ [8] `val`.

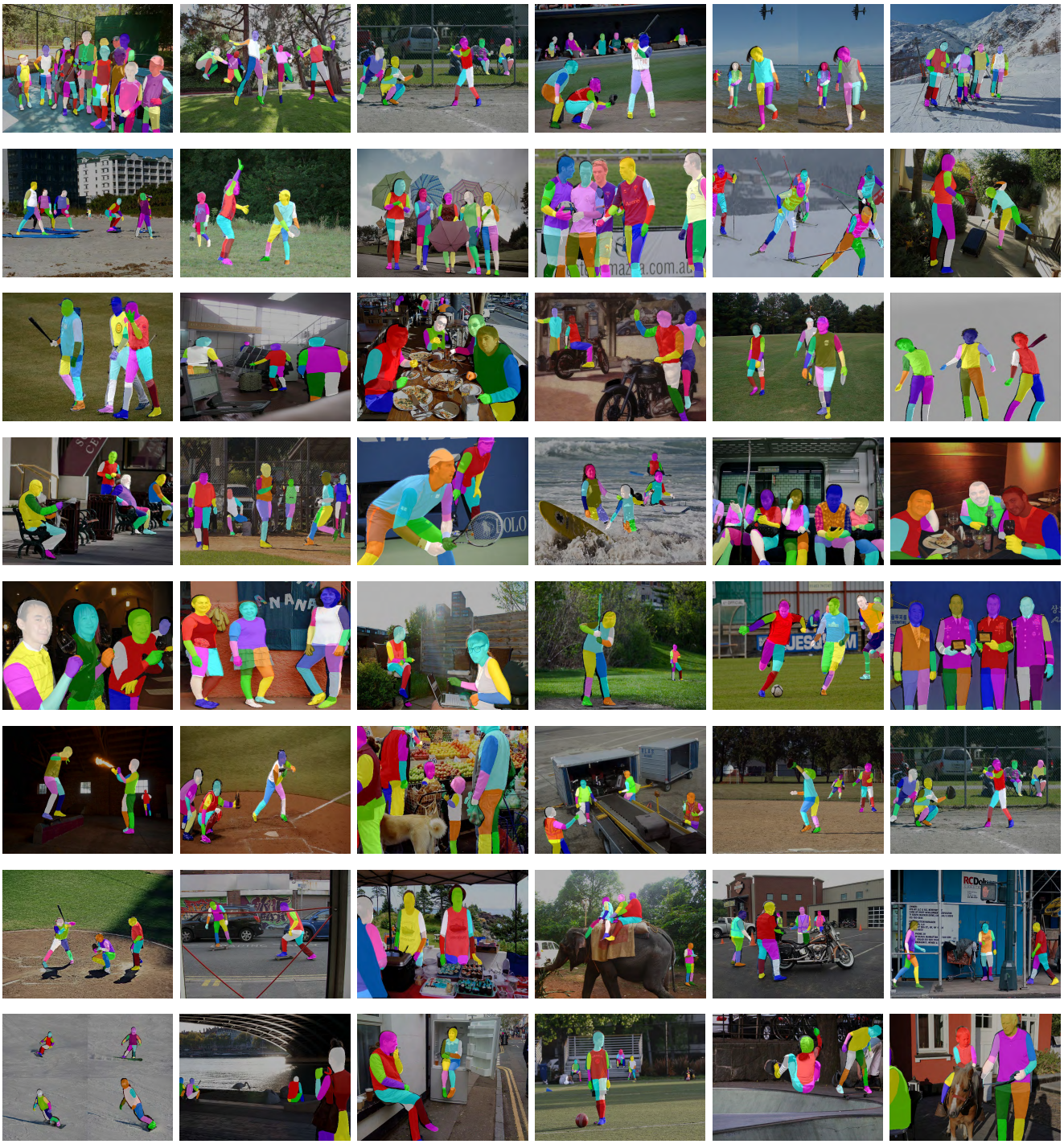

Figure 2: **Instance-level human semantic parsing results** on DensePose-COCO[1] `test`.

Figure 3: **Instance-level human semantic parsing results** on PASCAL-Person-Part[7] `test`.

Images | Semantic-Level Parsing | Pose Parsing | Instance-Level Parsing

Figure 4: Visualizations of typical failure cases on the DensePose-COCO [1] `test`.