Supplementary Materials
for
Graphonomy: Universal Human Parsing via Graph Transfer Learning

Ke Gong[1,2†], Yiming Gao[1†], Xiaodan Liang[1*],
Xiaohui Shen[3], Meng Wang[4], Liang Lin[1,2]

[1]Sun Yat-sen University [2]DarkMatter AI Research [3]ByteDance AI Lab [4]Hefei University of Technology

kegong936@gmail.com, gaoym9@mail2.sysu.edu.cn, xdliang328@gmail.com,
shenxiaohui@gmail.com, wangmeng@hfut.edu.cn, linliang@ieee.org

This supplementary material provides additional results supporting the claims of the main paper. First, we introduce the implementation details about the mentioned Intra-Graph Reasoning. Second, we conduct additional ablation studies to validate the set of graph convolution layers. Third, we report the experimental results on the MHP dataset [5]. Finally, we present more qualitative results of our proposed approach on different datasets.

1. Intra-Graph Reasoning

In this section, we provide details about the implementation for producing graph representation and re-projecting the graph to images features.

1.1. Graph Projection

According to our paper, we define an undirected graph as $G = (V, E)$ where V denotes the vertices, E denotes the edges, and $N = |V|$.

We use the feature maps $X \in \mathbb{R}^{H \times W \times C}$ as the module inputs, where H, W and C are height, width and channel number of the feature maps. And we need to project the feature maps to the high-level graph representation $Z \in \mathbb{R}^{N \times D}$ of all N vertices, where D is the desired feature dimension for each $v \in V$.

The projection process is first to learn a projection parameter $P \in \mathbb{R}^{C \times N}$, and change the features dimension of X corresponding to nodes number, formulated as:

$$X_1 = X^{HW \times C} \times P, \tag{1}$$

where $X \in \mathbb{R}^{H \times W \times C}$ is resized to $\mathbb{R}^{HW \times C}$, $\times$ is the matrix multiplication and we can get the $X_1 \in \mathbb{R}^{HW \times N}$. Next, considering the relationship between X and X_1, we calculate an intermediate feature X_2 as:

$$X_2 = X_1^T \times X^{HW \times C}, \tag{2}$$

where $X \in \mathbb{R}^{H \times W \times C}$ is resized to $\mathbb{R}^{HW \times C}$. Finally, we multiply X_2 with a trainable weight matrix $W_1 \in \mathbb{R}^{C \times D}$ to get the graph representation $Z \in \mathbb{R}^{N \times D}$:

$$Z = X_2 \times W_1. \tag{3}$$

Accordingly, the graph projection process can be formulated as the function ϕ (Equation 1 in the paper):

$$\begin{aligned} Z &= \phi(X, W) \\ &= P^T \times X^{C \times HW} \times X^{HW \times C} \times W_1. \end{aligned} \tag{4}$$

1.2. Graph Reprojection

The Graph Reprojection is the inverse process of the Graph Projection. Thus, the Graph Reprojection aims to project the graph $Z \in \mathbb{R}^{N \times D}$ to the same size as input feature maps $X \in \mathbb{R}^{H \times W \times C}$.

First, we expand the graph representation Z to $Z_1 \in \mathbb{R}^{HW \times N \times D}$, and multiply it by a trainable weight matrix $W_2 \in \mathbb{R}^{D \times 1}$ to get $Z_2 \in \mathbb{R}^{HW \times N}$:

$$Z_2 = Z_1 \times W_2. \tag{5}$$

Then we consider combining the information of the input feature maps X and graph features Z_2. We resize the feature maps X from $\mathbb{R}^{H \times W \times C}$ to $\mathbb{R}^{HW \times C}$, then multiply it by a trainable weight matrix $W_3 \in \mathbb{R}^{C \times N}$, and add it to Z_2, which can be formulated as

$$Z_3 = Z_2 + X^{HW \times C} \times W_3, \tag{6}$$

where we get the $Z_3 \in \mathbb{R}^{HW \times N}$. Second, we calculate an intermediate feature $Z_4 \in \mathbb{R}^{HW \times D}$ with the graph feature

† Equal contribution. *Corresponding Author. This work was supported by the Sun Yat-sen University Start-up Foundation Under Grant No. 76160-18841201.

Layers	Mean IoU(%)
1	67.80
3	68.34
5	68.36

Table 1. Comparison of human parsing performance with several graph convolution layers of our proposed Intra-Graph Reasoning on PASCAL-Person-Part dataset [2].

Source	Training images	Categories	Mean IoU(%)
ATR [4]	17,700	17	70.58
MHP [5]	15,403	58	70.89
CIHP [3]	28,280	19	71.14

Table 2. Comparison of human parsing performance with several source datasets on PASCAL-Person-Part dataset [2].

Method	Mean IoU(%)
DeepLab V3+ [1]	32.93
Graphonomy(PASCAL)	34.05

Table 3. Comparison of human parsing performance on MHP dataset [5].Performance on val set.

Z_3 and Z, written as:

$$\begin{aligned} Z_4 &= Z_3 \times Z \\ &= (Z_1 \times W_2 + X^{HW \times C} \times W_3) \times Z. \end{aligned} \tag{7}$$

Finally, we multiply Z_4 by a trainable weight matrix $W_4 \in \mathbb{R}^{D \times C}$, written as:

$$X^R = Z_4 \times W_4, \tag{8}$$

where we get the reprojected image features $X^R \in \mathbb{R}^{HW \times C}$ and then resize it to $\mathbb{R}^{H \times W \times C}$.

2. Ablation Study

Besides the experiments in the paper, we perform more ablation studies on PASCAL-Person-Part dataset [2].

The layers of graph convolution. To understand the effectiveness of the different layers of graph convolution, we report the different layers results in Table 1. Increasing the graph layers of the Intra-Graph leads to better performance. Using 5 layers improves the performance by about 0.56 compared with using only 1 layer, but brings about 0.02% compared with using 3 layers. However, increasing layers means more parameters, more GPU memory, and time consumption. Thus, we prefer 3 layers, which can use fewer parameters and save GPU memory and time.

Different source dataset. To understanding the effectiveness of different source dataset, we report the results of different dataset transfer to PASCAL dataset [2]. We compare the ATR dataset [4], MHP dataset [5] and CIHP dataset [3], where ATR dataset has the similarity training numbers with MHP dataset and ATR dataset has the similarity categories numbers with CIHP dataset.

We report the results in Table 2. We can see that pretrained on MHP dataset is better than pretrained on ATR dataset. We think the reason is that ATR dataset is easier than both MHP dataset and PASCAL-Person-Part dataset (the images of ATR dataset are visualized in Fig.1). Thus PASCAL-Person-Part can only benefit a little knowledge from ATR dataset. Besides, pretrained on CIHP with large training images is better than the pretrained on the MHP dataset. We think the reason is that when training on PASCAL-Person-Part dataset, the model can not benefit much prior knowledge from the more fine-grained categories, because, a lot of categories of MHP dataset, e.g. ”bikini/bra”, ”jacket/windbreaker/hoodie”, ”t-shirt”, ”polo-shirt”, ”sweater”, ”singlet”, ”torso-skin”, ”tie”,etc., only mapping the the semantic label of torso in PASCAL-Person-Part dataset. Thus, CIHP dataset with large training images number is the best choice for the PASCAL-Person-Part pretrained model.

3. Additional Experiments

Besides the three human parsing datasets used in the paper, we further evaluate our Graphonomy on another larger-scale dataset MHP [5]. We first introduce the details of the MHP dataset [5]. Then, we report the results on it.

MHP dataset [5] is a new fine-grained benchmark for human parsing task, including 25,403 images with 58 semantic categories defined and annotated except for the ”background” category (i.e.”cap/hat”, ”helmet”, ”face”, ”hair”, ”left- arm”, ”right-arm”, ”left-hand”, ”right-hand”, ”protector”, ”bikini/bra”, ”jacket/windbreaker/hoodie”, ”t-shirt”, ”polo-shirt”, ”sweater”, ”singlet”, ”torso-skin”, ”pants”, ”shorts/swim-shorts”, ”skirt”, ”stock-ings”, ”socks”, ”left-boot”, ”right-boot”, ”left-shoe”, ”right-shoe”, ”left-highheel”, ”right-highheel”, ”left-sandal”, ”right-sandal”, ”left-leg”, ”right-leg”, ”left-foot”, ”right-foot”, ”coat”, ”dress”, ”robe”, ”jumpsuits”, ”other-full-body-clothes”, ”headwear”, ”backpack”, ”ball”, ”bats”, ”belt”, ”bottle”, ”carrybag”, ”cases”, ”sunglasses”, ”eyewear”, ”gloves”, ”scarf”, ”umbrella”, ”wallet/purse”, ”watch”, ”wristband”, ”tie”, ”other-accessaries”, ”other-upper-body-clothes”, and ”other-lower-body-clothes”). Following the benchmark, we use 15,403 images for training, 5,000 images for validation and 5,000 images for testing.

Following the evaluation metric for human parsing used in our paper, we report the results in terms of the standard intersection over union(IoU) on MHP dataset in Table 3. Our results show that our Graphonomy can bring about 1% improvement.

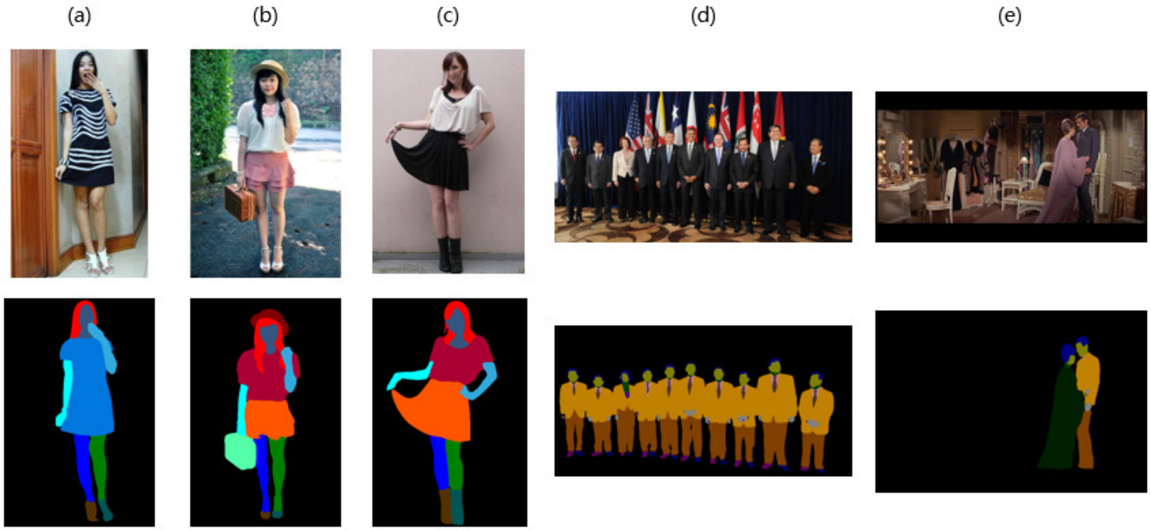

Figure 1. Visualized results predicted on ATR dataset [4] (Left) and MHP dataset [5] (Right).

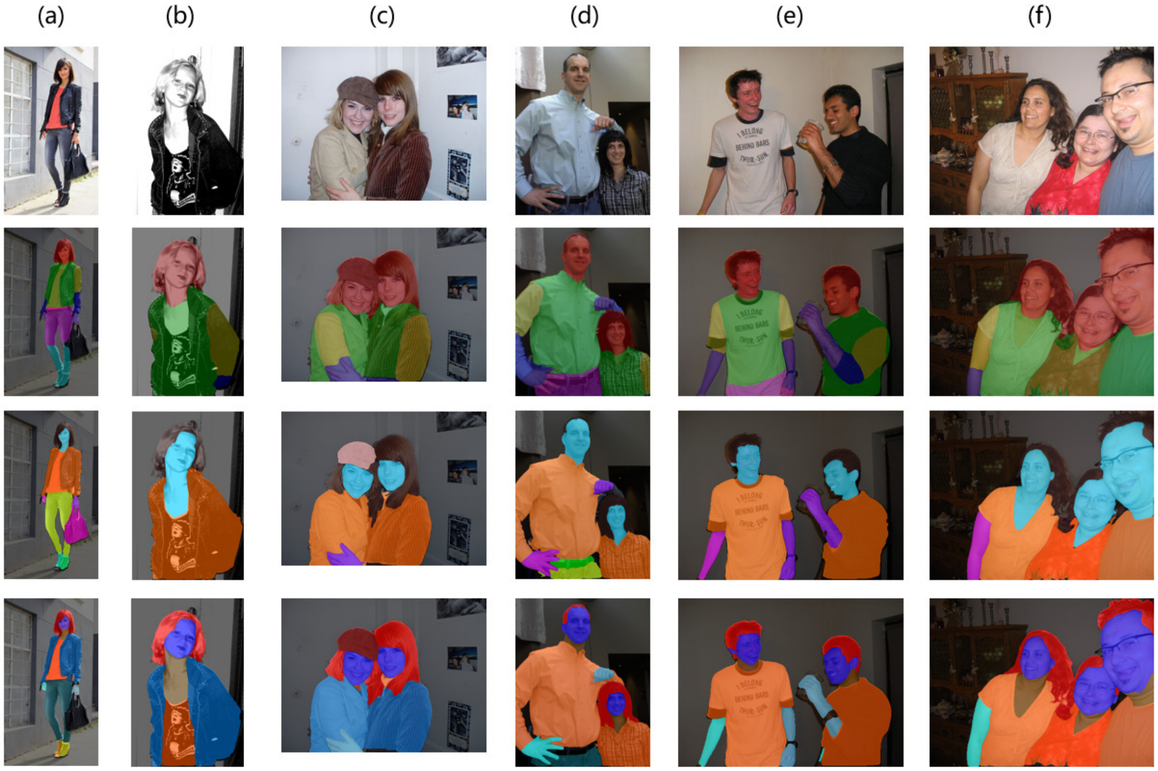

Figure 2. Visualized results predicted by our Graphonomy.

4. Qualitative Results

In this section, we first present the qualitative results generated by our Graphonomy on the ATR dataset [4] and MHP dataset [5], as visualized in Fig. 1.The above is the origin images and the below is predicted by our method.

Then, we present more qualitative universal results generated by our Graphonomy, as visualized in Fig. 2. The first row is the original images, the second row is the results of 7 human body parts labels (including background) provided by PASCAL-Person-Part dataset [2], the third row is the results of 18 semantic labels provided by the ATR dataset [4], and the last row is the results of 20 semantic part labels provided by CIHP dataset [3]. It can be observed that our Graphonomy has the capability to learn the universal representation features across datasets with different label an-

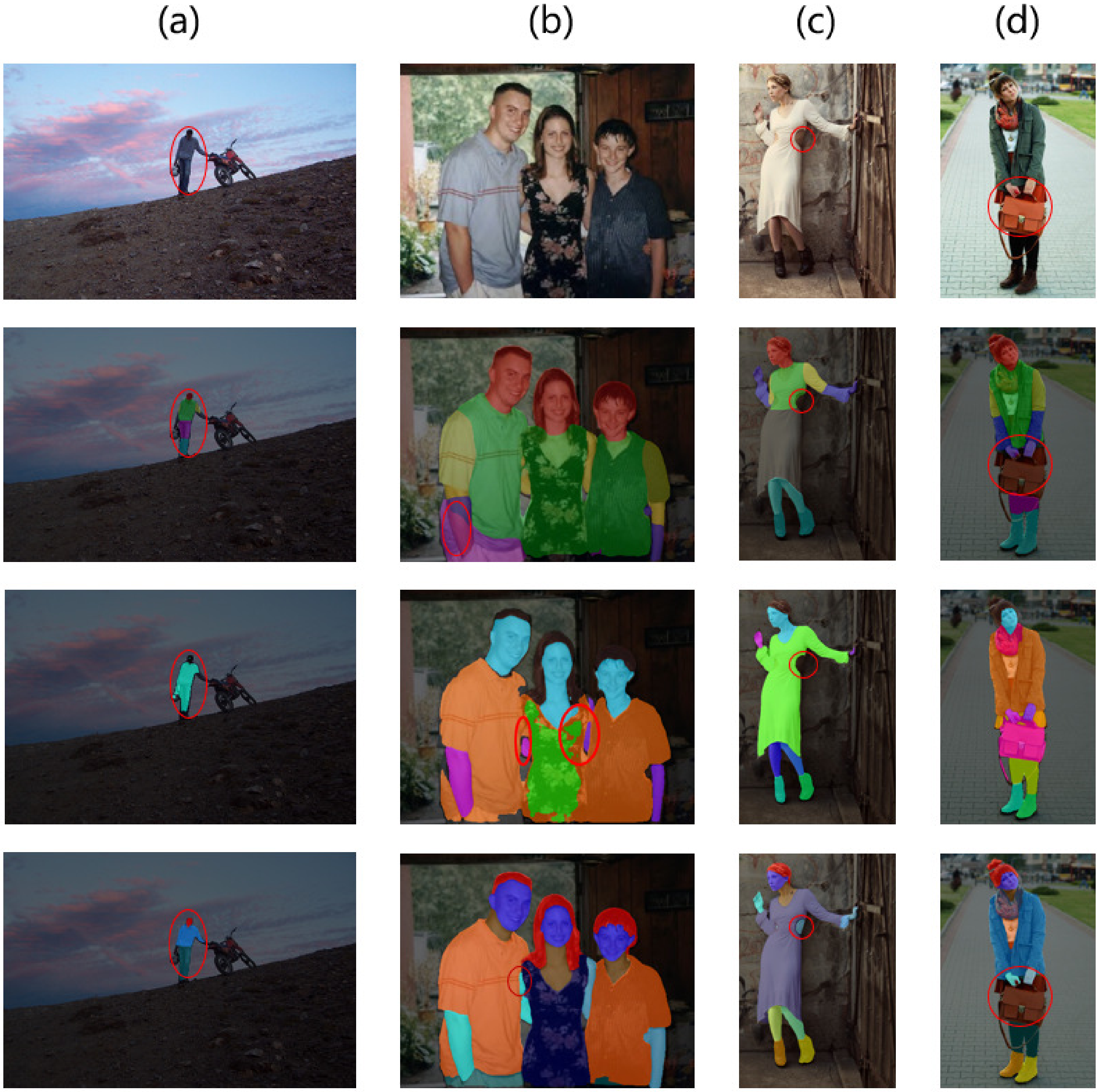

Figure 3. Failure cases of our proposed method.The incorrect regions are circled in red.

notations and efficiently predict different levels of human parsing results simultaneously.

Our method may also fail for some difficult images. We present some failure universal cases of our method in Fig. 3. We find that the person in a too small scale is difficult to predict correctly, as shown in Fig. 3 (a). Besides, when the body parts of different people touch closely (Fig. 3(b)), the arms are wrongly predicted because of the crowd and occluded persons instances.

In the future, we will investigate on incorporating more knowledge into our graph structure to tackle the challenging cases and generalize our proposed Graphonomy to more general semantic segmentation tasks and instance-level parsing tasks.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018.

[2] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, et al. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014.

[3] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *ECCV*, September 2018.

[4] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep human

parsing with active template regression. *TPAMI*, 2015.

[5] Jian Zhao, Jianshu Li, Yu Cheng, Li Zhou, Terence Sim, Shuicheng Yan, and Jiashi Feng. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. *arXiv preprint arXiv:1804.03287*, 2018.