# Hierarchical Human Semantic Parsing with Comprehensive Part-Relation Modeling *Supplemental Material*

Wenguan Wang, *Member IEEE*, Tianfei Zhou, Siyuan Qi,
Jianbing Shen, *Senior Member IEEE*, Song-Chun Zhu, *Fellow IEEE*

## 1 Overview

In this document, we present additional materials including pseudo code (§2), extra quantitative results on the PASCAL-Person-Part [1] and ATR [2] datasets (§3), ablation studies (§4), as well as qualitative results on the PASCAL-Person-Part [1], LIP [3], ATR [2], Fashion Clothing [4], and PPSS [5] datasets (§5).

## 2 Pseudo Code

Algorithms 1 and 2 give the pseudo codes of our structured human parsers, CNIF and PRHP, respectively.

## 3 Additional Diagnostic Experiment

We list the per-class performance on LIP `val` set [3] to further verify the effectiveness of our proposed two human parsers, using mIoU metric. As can be seen from Table 1, our proposed methods achieve superior performance; they outperform state-of-the-art methods across all categories. In addition, the per-class performance using F-1 score on ATR `test` set [2] is listed in Table 2. The results demonstrate again that our CNIF and PRHP outperform state-of-the-art methods across most categories.

## 4 Additional Ablation Study

### 4.1 Additional Ablation Study for CNIF based Human Parser

To analyze and quantify the effectiveness and importance of each essential components of our CNIF algorithm, Table 3 shows the detailed evaluation of our full model compared to ablated versions without some key ingredients. The experiments are performed on PASCAL-Person-Part `test` [1], reported over each part using mIoU metric. Experimental results intuitively demonstrate the superiority of our conditional and compositional information fusion approach over all the human parts on the three semantic levels.

**Algorithm 1** Our Structured Human Parser, CNIF

**Input:** A test image $I$;
**Output:** Hierarchical human parsing results $\{\hat{\mathcal{Y}}_l\}_{l=1}^3$;
*/* Human Semantic Hierarchy Initialization */*
Represent the human semantics as a directed, three-layer hierarchical graph $\mathcal{G}=(\mathcal{V}=\cup_{l=1}^3\mathcal{V}_l, \mathcal{E}, \mathcal{Y}=\cup_{l=1}^3\mathcal{Y}_l)$.
*/* Node Embedding Initialization */*
Apply a backbone network $F^{\text{B}}$ over $I$ to obtain the image representation $\boldsymbol{h}_I$, then get three level-specific embeddings $\{\boldsymbol{h}_l^{\text{LSF}}\}_{l=1}^3$, and node (part) features $\{\boldsymbol{h}_v^{(0)}\}_{v\in\mathcal{V}}$;
*/* Direct Inference */*
**for** each node $v\in\mathcal{V}$ **do**
  Get prediction from the direct inference process: $\text{logit}(\hat{y}_v|I) = F^{\uparrow}(\boldsymbol{h}_v)$;
**end for**
*/* Top-Down/Bottom-Up Inference */*
**for** each node $v\in\mathcal{V}$ **do**
  Get prediction from the top-down inference process: $\text{logit}(y_v|\hat{y}_u) = F^{\downarrow}([\hat{y}_u, \boldsymbol{h}_v])$;
  Get prediction from the bottom-up inference process: $\text{logit}(y_v|\hat{\boldsymbol{y}}_{\boldsymbol{w}}) = F^{\uparrow}([\text{PMP}([\hat{y}_w]_{w\in\boldsymbol{w}}), \boldsymbol{h}_v])$;
**end for**
*/* Conditional Neural Information Fusion */*
**for** each node $v\in\mathcal{V}$ **do**
  Estimate the confidence of direct, top-down, and bottom-up process: $\delta_v^{\uparrow} = \sigma(\boldsymbol{C}_v^{\uparrow}\cdot\text{CAP}(\boldsymbol{h}_v))$, $\delta_u^{\downarrow} = \sigma(\boldsymbol{C}_u^{\downarrow}\cdot\text{CAP}(\boldsymbol{h}_u))$, $\delta_{\boldsymbol{w}}^{\uparrow} = \sigma(\boldsymbol{C}_{\boldsymbol{w}}^{\uparrow}\cdot\text{CAP}([\boldsymbol{h}_w]_{w\in\boldsymbol{w}}))$;
  Get final prediction from the three inference process: $\text{logit}(y_v|Z) = F^{\cup}(\delta_v^{\uparrow}F_v^{\uparrow}, \delta_u^{\downarrow}F_v^{\downarrow}, \delta_{\boldsymbol{w}}^{\uparrow}F_v^{\uparrow})$;
**end for**
*/* Final Hierarchical Human Parsing Result Generation */*
**for** $l = 1\cdots 3$ **do**
  For the $l$-th layer nodes $\mathcal{V}_l$, apply *pixel-wise softmax* (PSM) for normalization: $\hat{\mathcal{Y}}_l = \{\hat{y}_v\}_{v\in\mathcal{V}_l} = \text{PSM}\big([\text{logit}(y_v|Z)]_{v\in\mathcal{V}_l}\big)$;
**end for**
**return** Hierarchical human parsing results $\{\hat{\mathcal{Y}}_l\}_{l=1}^3$.

**Algorithm 2** Our Structured Human Parser, PRHP

**Input:** A test image $I$, and total inference iteration steps $T$;
**Output:** Hierarchical human parsing results $\{\hat{\mathcal{Y}}_l^{(T)}\}_{l=1}^3$;

*/* Human Semantic Hierarchy Initialization */*

Represent the human semantics as a directed, three-layer hierarchical graph $\mathcal{G}=(\mathcal{V}=\cup_{l=1}^3\mathcal{V}_l, \mathcal{E}, \mathcal{Y}=\cup_{l=1}^3\mathcal{Y}_l)$.

*/* Node Embedding Initialization */*

Apply a backbone network $F^{\mathrm{B}}$ over $I$ to obtain the image representation $\boldsymbol{h}_I$, then get three level-specific embeddings $\{\boldsymbol{h}_l^{\mathrm{LSF}}\}_{l=1}^3$, and node (part) features $\{\boldsymbol{h}_v^{(0)}\}_{v\in\mathcal{V}}$;

*/* Message-Passing based Iterative Inference */*

**for** $t = 1\cdots T$ **do**

*/* Typed Relation Embedding Update */*

  **for** each edge $(u,v)\in\mathcal{E}$ **do**

    Update edge embedding: $\boldsymbol{h}_{u,v}^{(t-1)} = R^r([F^r(\boldsymbol{h}_u^{(t-1)}), \boldsymbol{h}_v^{(t-1)}])$, where $r\in\{\mathrm{dec}, \mathrm{com}, \mathrm{dep}\}$;

  **end for**

  **for** each node $v\in\mathcal{V}$ **do**

*/* Message Aggregation */*

    Gather information along incoming edges: $\boldsymbol{m}_v^{(t)} = \underbrace{\sum_{u\in\mathcal{P}_v}\boldsymbol{h}_{u,v}^{(t-1)}}_{\text{decomposition}} + \underbrace{\sum_{u\in\mathcal{C}_v}\boldsymbol{h}_{u,v}^{(t-1)}}_{\text{composition}} + \underbrace{\sum_{u\in\mathcal{K}_v}\boldsymbol{h}_{u,v}^{(t-1)}}_{\text{dependency}}$;

*/* Node Embedding Update */*

    Use the collected massages update node embedding: $\boldsymbol{h}_v^{(t)} = U_{\mathrm{convGRU}}(\boldsymbol{h}_v^{(t-1)}, \boldsymbol{m}_v^{(t)})$;

  **end for**

**end for**

*/* Final Hierarchical Human Parsing Result Readout */*

**for** $l = 1\cdots 3$ **do**

  For the $l$-th layer nodes $\mathcal{V}_l$, apply a convolutional readout function $O$ over the final node embeddings $\{\boldsymbol{h}_v^{(T)}\}_{v\in\mathcal{V}}$, and *pixel-wise soft-max* (PSM) for normalization: $\hat{\mathcal{Y}}_l^{(T)} = \{\hat{y}_v^{(T)}\}_{v\in\mathcal{V}_l} = \mathrm{PSM}([O(\boldsymbol{h}_v^{(T)})]_{v\in\mathcal{V}_l})$;

**end for**

**return** Hierarchical human parsing results $\{\hat{\mathcal{Y}}_l^{(T)}\}_{l=1}^3$.

### 4.2 Additional Ablation Study for PRHP Model

We herein provide more detailed experimental results about the three typed part relations (*i.e.*, decompositional, compositional and dependency relations) over different human semantic levels. The experiments are performed on PASCAL-Person-Part `test` set [1], using mIoU metric.

To analyze and quantify the effectiveness and importance of different part relations to our hierarchical parsing model, PRHP, experiments with different settings are conducted. Table 4 shows the evaluation of our full model with all three typed part relations compared to ablated versions with only one of three typed part relations. Please note that, to better show the effectiveness of different components, such experiment is carried on only one iteration while our full model performs two-iteration inference. In summary, one typed part relation can mainly enhance partial nodes in human hierarchy. To exploit more comprehensive information from the whole human hierarchy, iteratively updating node embeddings with all three typed part-relation reasoning is more effective. Experimental results of our full model with different iterations intuitively demonstrate

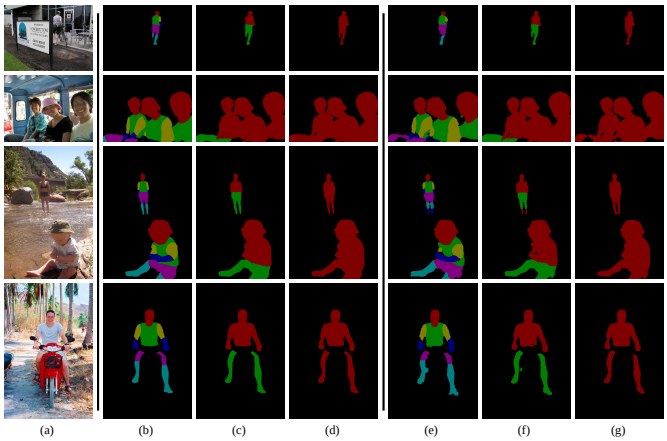


Fig. 1. Comparisons of our full CNIF model and direct inference on PASCAL-Person-Part `test` [1]. Given an input image (a), results of our full model: fine-grained human parts (b), upper/lower body (c), and full body (d) predictions; results from the direct inference: fine-grained human parts (e), upper/lower body (f), and full body (g) predictions.

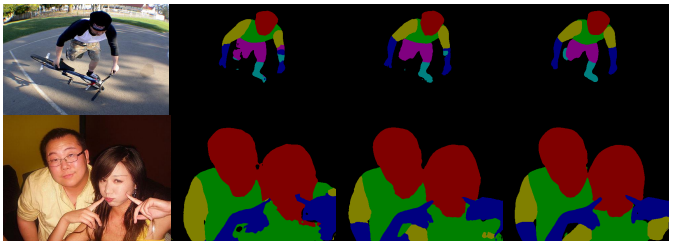


Fig. 2. Visual comparison results from our (b) backbone, (c) compositional information fusion *w/o* conditional inference, and (d) full CNIF model.

the superiority of our hierarchical human parsing approach with typed part-relation reasoning.

As shown in Table 5, different typed part relation networks with or without the corresponding attention mechanisms are evaluated. The results indicate that attention mechanisms can improve the performance of part relation networks consistently. Noted that, simply passing original part embedding as the message in dependency relation network does not bring any improvement or even be harmful.

## 5 Additional Qualitative Result

### 5.1 Additional Qualitative Result for CNIF Model

Through fusing cross-level information within human structures, our CNIF model estimates the overall part configuration more accurately. For example, as shown in Fig. 1, our full model is able to generate more accurate human parts in different levels, compared to only using the direct inference. It outputs semantically meaningful and precise predictions despite the existence of large appearance and position variations. In addition, our CNIF model is able to give clearer details of arms and legs, especially for small-scale parts or the regions with similar appearances. For example, observed from the 1st row of Fig. 1, the small regions (*e.g.*, left lower-leg) can be successfully segmented out by our method with the constraint of the top-down inference. These regions with similar appearances can be recognized

TABLE 1
Per-class comparison of mIoU with state-of-the-art methods on LIP `val` [3]. The two best scores are marked in **red** and **blue**, respectively.

| Method | Hat | Hair | Glov | Sung | Clot | Dress | Coat | Sock | Pant | Suit | Scarf | Skirt | Face | L-Arm | R-Arm | L-Leg | R-Leg | L-Sh | R-Sh | B.G. | Ave. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SegNet [6] | 26.60 | 44.01 | 0.01 | 0.00 | 34.46 | 0.00 | 15.97 | 3.59 | 33.56 | 0.01 | 0.00 | 0.00 | 52.38 | 15.30 | 24.23 | 13.82 | 13.17 | 9.26 | 6.47 | 70.62 | 18.17 |
| FCN-8s [7] | 39.79 | 58.96 | 5.32 | 3.08 | 49.08 | 12.36 | 26.82 | 15.66 | 49.41 | 6.48 | 0.00 | 2.16 | 62.65 | 29.78 | 36.63 | 28.12 | 26.05 | 17.76 | 17.70 | 78.02 | 28.29 |
| DeepLabV2 [8] | 56.48 | 65.33 | 29.98 | 19.67 | 62.44 | 30.33 | 51.03 | 40.51 | 69.00 | 22.38 | 11.29 | 20.56 | 70.11 | 49.25 | 52.88 | 42.37 | 35.78 | 33.81 | 32.89 | 84.53 | 41.64 |
| Attention [9] | 58.87 | 66.78 | 23.32 | 19.48 | 63.20 | 29.63 | 49.70 | 35.23 | 66.04 | 24.73 | 12.84 | 20.41 | 70.58 | 50.17 | 54.03 | 38.35 | 37.70 | 26.20 | 27.09 | 84.00 | 42.92 |
| Attention+SSL [3] | 59.75 | 67.25 | 28.95 | 21.57 | 65.30 | 29.49 | 51.92 | 38.52 | 68.02 | 24.48 | 14.92 | 24.32 | 71.01 | 52.64 | 55.79 | 40.23 | 38.80 | 28.08 | 29.03 | 84.56 | 44.73 |
| ASN [10] | 56.92 | 64.34 | 28.07 | 17.78 | 64.90 | 30.85 | 51.90 | 39.75 | 71.78 | 25.57 | 7.97 | 17.63 | 70.77 | 53.53 | 56.70 | 49.58 | 48.21 | 34.57 | 33.31 | 84.01 | 45.41 |
| SSL [3] | 58.21 | 67.17 | 31.20 | 23.65 | 63.66 | 28.31 | 52.35 | 39.58 | 69.40 | 28.61 | 13.70 | 22.52 | 74.84 | 52.83 | 55.67 | 48.22 | 47.49 | 31.80 | 29.97 | 84.64 | 46.19 |
| MMAN [11] | 57.66 | 65.63 | 30.07 | 20.02 | 64.15 | 28.39 | 51.98 | 41.46 | 71.03 | 23.61 | 9.65 | 23.20 | 69.54 | 55.30 | 58.13 | 51.90 | 52.17 | 38.58 | 39.05 | 84.75 | 46.81 |
| SS-NAN [12] | 63.86 | 70.12 | 30.63 | 23.92 | 70.27 | 33.51 | 56.75 | 40.18 | 72.19 | 27.68 | 16.98 | 26.41 | 75.33 | 55.24 | 58.93 | 44.01 | 41.87 | 29.15 | 32.64 | **88.67** | 47.92 |
| CE2P [13] | 65.29 | 72.54 | 39.09 | 32.73 | 69.46 | 32.52 | 56.28 | 49.67 | 74.11 | 27.23 | 14.19 | 22.51 | 75.50 | 65.14 | 66.59 | 60.10 | 58.59 | 46.63 | 46.12 | 87.67 | 53.10 |
| BraidNet [14] | 66.8 | 72.0 | 42.5 | 32.1 | 69.8 | 33.7 | 57.4 | 49.0 | 74.9 | **32.4** | 19.3 | 27.2 | 74.9 | 65.5 | 67.9 | 60.2 | 59.6 | 47.4 | 47.9 | 88.0 | 54.4 |
| CNIF (**Ours**) | **69.55** | **73.45** | **45.17** | **41.45** | **70.57** | **38.52** | **57.94** | **54.02** | **75.07** | 28.00 | **31.92** | **30.20** | **76.38** | **68.28** | **69.49** | **65.52** | **65.51** | **52.67** | **53.38** | 87.99 | **57.74** |
| PRHP (**Ours**) | **70.65** | **75.18** | **46.77** | **43.10** | **71.82** | **40.72** | **59.41** | **55.65** | **76.38** | **30.13** | **33.72** | **32.12** | **77.25** | **69.58** | **70.47** | **66.84** | **66.79** | **54.23** | **54.69** | **89.45** | **59.25** |

TABLE 2
Per-class comparison of F-1 scores with state-of-the-art methods on the ATR `test` [2]. The two best scores are marked in **red** and **blue**, respectively.

| Method | Hat | Hair | S-Gls | U-Cloth | Skirt | Pants | Dress | Belt | L-Shoe | R-Shoe | Face | L-Leg | R-Leg | L-Arm | R-Arm | Bag | Scarf |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Yamaguchi [15] | 8.44 | 59.96 | 12.09 | 56.07 | 17.57 | 55.42 | 40.94 | 14.68 | 38.24 | 38.33 | 72.10 | 58.52 | 57.03 | 45.33 | 46.65 | 24.53 | 11.43 |
| Paperdoll [16] | 1.72 | 63.58 | 0.23 | 71.87 | 40.20 | 69.35 | 59.49 | 16.94 | 45.79 | 44.47 | 61.63 | 52.19 | 55.60 | 45.23 | 46.75 | 30.52 | 2.95 |
| M-CNN [17] | 80.77 | 65.31 | 35.55 | 72.58 | 77.86 | 70.71 | 81.44 | 38.45 | 53.87 | 48.57 | 72.78 | 63.25 | 68.24 | 57.40 | 51.12 | 57.87 | 43.38 |
| ATR [2] | 77.97 | 68.18 | 29.20 | 79.39 | 80.36 | 79.77 | 82.02 | 22.88 | 53.51 | 50.26 | 74.71 | 69.07 | 71.69 | 53.79 | 58.57 | 53.66 | 57.07 |
| DeepLabv2 [8] | 72.25 | 82.58 | 44.61 | 87.12 | 80.91 | 85.80 | 79.05 | 24.96 | 65.44 | 65.70 | 85.33 | 80.21 | 80.34 | 73.04 | 74.49 | 78.33 | 46.99 |
| PSPNet [18] | 74.30 | 86.51 | 67.78 | 88.53 | 79.04 | 86.73 | 77.14 | 41.76 | 64.53 | 62.94 | 89.45 | 82.55 | 81.92 | 77.68 | 78.01 | 77.69 | 49.83 |
| Attention [9] | 76.78 | 84.62 | 56.98 | 87.56 | 83.46 | 87.95 | **82.49** | 36.20 | 68.73 | 69.36 | 87.03 | 84.29 | 83.63 | 78.43 | 78.99 | 81.60 | 48.90 |
| DeepLabv3+ [19] | 77.22 | 87.44 | 73.06 | 89.64 | **85.15** | 90.11 | 79.99 | 44.48 | 70.08 | 71.13 | 90.53 | 85.60 | 85.25 | 81.96 | 82.48 | 81.73 | 53.46 |
| Co-CNN [20] | 75.88 | 89.97 | 81.26 | 87.38 | 71.94 | 84.89 | 71.03 | 40.14 | **81.43** | **81.49** | **92.73** | 88.77 | 88.48 | 89.00 | 88.71 | 83.81 | 46.24 |
| TGPNet [4] | 80.18 | 87.13 | 70.93 | **91.01** | **88.95** | **90.72** | **87.42** | 51.73 | 75.13 | 75.36 | 89.78 | 89.06 | 88.73 | 83.91 | 83.96 | 84.72 | 52.86 |
| CNIF (**Ours**) | **86.53** | **90.25** | **82.16** | 90.25 | 81.64 | 90.30 | 80.63 | **64.72** | 79.11 | 78.97 | **92.71** | **91.60** | **91.48** | **90.47** | **90.32** | **87.82** | **68.07** |
| PRHP (**Ours**) | **87.14** | **90.58** | **83.15** | **91.04** | 82.76 | **91.09** | 81.31 | **66.32** | **81.26** | **80.15** | 92.57 | **91.76** | **91.42** | **91.45** | **90.87** | **88.79** | **70.33** |

TABLE 3
Additional ablation study for CNIF model using class mIoU on the PASCAL-Person-Part `test` [1].

| Aspects | Module | mIoU | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $\mathcal{V}_3$ | | | | | | $\mathcal{V}_2$ | | $\mathcal{V}_1$ |
| | | Head | Torso | U-Arm | L-Arm | U-Leg | L-Leg | Upper-body | Lower-body | Full-body |
| **CNIF** | **direct+bottom-up+** top-down + conditional fusion | 88.02 | 72.90 | 64.31 | 63.52 | 55.60 | 54.95 | 86.27 | 62.51 | 86.53 |
| Backbone | direct infer. *w/o* hierarchy | 85.29 | 67.61 | 54.01 | 53.65 | 48.53 | 44.52 | - | - | - |
| Variant | direct infer. | 85.46 | 68.85 | 56.38 | 55.31 | 50.06 | 46.77 | 83.79 | 57.33 | 84.22 |
| | direct+bottom-up | 85.29 | 68.93 | 57.50 | 57.87 | 51.28 | 48.54 | 84.62 | 57.96 | 86.03 |
| | direct+top-down | 86.91 | 71.43 | 61.91 | 61.57 | 54.21 | 52.33 | 85.03 | 58.47 | 84.85 |
| | direct+bottom-up+top-down | 87.63 | 71.84 | 62.80 | 62.27 | 54.80 | 52.74 | 86.02 | 62.02 | 86.23 |

and separated by the top-down guidance from their parent nodes. In general, by effectively exploiting human semantic hierarchy, our approach outputs more reasonable human parsing results.

A visual comparison between the results from our backbone network, our model only using compositional fusion and our full CNIF model can be found in Fig. 2 (b-d), which intuitively shows the improvements from our conditional and compositional information fusion.

We then add several segmentation results, compared with ground-truth, on different human parsing datasets, including the PASCAL-Person-Part `test` [1], shown in Fig. 4; the LIP `val` [3], shown in Fig. 5; the ATR `test` [2], shown in Fig. 6; the Fashion Clothing `test` [4], shown in Fig. 7, and the PPSS `test` [5], shown in Fig. 8.

### 5.2 Additional Qualitative Result for PRHP Model

By effectively and iteratively exploiting human semantic hierarchy and rich part relations (decomposition relation, composition relation and dependency relation), our PRHP model is able to generate more precise human parsing results. For example, as shown in Fig. 3, our full PRHP model is able to generate more accurate human parts in different levels, compared to only using the initial node embedding for inference.

We then provide several segmentation results, compared with ground-truth, on different human parsing datasets, including the PASCAL-Person-Part `test` [1], shown in Fig. 9; the LIP `val` [3], shown in Fig. 10; the ATR `test` [2], shown in Fig. 11; the Fashion Clothing `test` [4], shown in Fig. 12, and the PPSS `test` [5], shown in Fig. 13.

TABLE 4
Ablation study of the effectiveness of different typed part relations in our PRHP model on the PASCAL-Person-Part `test` [1].

| Module | mIoU | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{V}_3$ | | | | | | $\mathcal{V}_2$ | | $\mathcal{V}_1$ |
| | Head | Torso | U-Arm | L-Arm | U-Leg | L-Leg | Upper-body | Lower-body | Full-body |
| **Decomposition+Composition+Dependency (1 iteration)** | 89.12 | 74.35 | 65.86 | 65.03 | 57.46 | 56.87 | 86.91 | 63.87 | 86.74 |
| Initial node embedding | 85.95 | 71.55 | 62.3 | 60.26 | 53.7 | 51.93 | 85.33 | 62.22 | 85.28 |
| Decomposition relation | 87.74 | 73.72 | 65.13 | 63.90 | 56.41 | 56.39 | 86.12 | 63.05 | 85.76 |
| Composition relation | 86.52 | 71.94 | 63.17 | 61.32 | 54.48 | 53.01 | 86.03 | 62.97 | 86.32 |
| Dependency relation | 86.71 | 72.14 | 63.21 | 61.27 | 54.69 | 53.22 | 85.89 | 62.91 | 85.33 |

TABLE 5
Ablation study of the attention mechanisms in different typed part relation networks of our PRHP model on PASCAL-Person-Part `test` [1].

| Aspect | Relation Network | Relation Adaption $F^r$ | mIoU | △ |
|---|---|---|---|---|
| Reference | **Full model** | - | 73.12 | - |
| Relation | Decomposition Relation (Eq. 20) | $\boldsymbol{h}_u \odot \mathtt{att}^{\mathrm{dec}}_{u,v}(\boldsymbol{h}_u)$ | 71.38 | -1.74 |
| | | $\boldsymbol{h}_u$ | 71.04 | -2.08 |
| | Composition Relation (Eq. 23) | $\boldsymbol{h}_u \odot \mathtt{att}^{\mathrm{com}}_{v}([\boldsymbol{h}_{u'}]_{u' \in \mathcal{C}_v})$ | 69.35 | -3.77 |
| | | $\boldsymbol{h}_u$ | 69.19 | -3.93 |
| | Dependency Relation (Eq. 25) | $F^{\mathrm{cont}}(\boldsymbol{h}_u) \odot \mathtt{att}^{\mathrm{dec}}_{u,v}(F^{\mathrm{cont}}(\boldsymbol{h}_u))$ | 69.43 | -3.69 |
| | | $F^{\mathrm{cont}}(\boldsymbol{h}_u)$ | 69.24 | -3.90 |
| | | $\boldsymbol{h}_u$ | 68.83 | -4.29 |

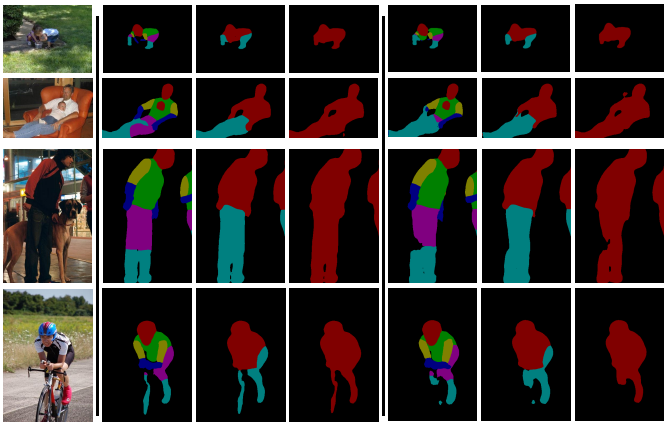


Fig. 3. Comparisons between our PRHP model with three part relations and base model with initial node embedding on PASCAL-Person-Part `test` [1]. Given an input image (a), results of our full model: fine-grained human parts (b), upper/lower body (c), and full body (d) predictions; results from the initial node embedding: fine-grained human parts (e), upper/lower body (f), and full body (g) predictions.

## REFERENCES


[1] F. Xia, P. Wang, X. Chen, and A. L. Yuille, "Joint multi-person pose estimation and semantic part segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6769–6778.

[2] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2402–2414, 2015.

[3] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 932–940.

[4] X. Luo, Z. Su, J. Guo, G. Zhang, and X. He, "Trusted guidance pyramid network for human parsing," in *ACM MM*, 2018, pp. 654–662.

[5] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2648–2655.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[9] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3640–3649.

[10] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," in *NIPS-workshop*, 2016.

[11] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-micro adversarial network for human parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.

[12] J. Zhao, J. Li, X. Nie, F. Zhao, Y. Chen, Z. Wang, J. Feng, and S. Yan, "Self-supervised neural aggregation networks for human parsing," in *CVPR-workshop*, 2017, pp. 7–15.

[13] T. Liu, T. Ruan, Z. Huang, Y. Wei, S. Wei, Y. Zhao, and T. Huang, "Devil in the details: Towards accurate single and multiple human parsing," *arXiv preprint arXiv:1809.05996*, 2018.

[14] X. Liu, M. Zhang, W. Liu, J. Song, and T. Mei, "BraidNet: Braiding semantics and details for accurate human parsing," in *ACM MM*, 2019, pp. 338–346.

[15] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3570–3577.

[16] K. Yamaguchi, M. Hadi Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 3519–3526.

[17] S. Liu, X. Liang, L. Liu, X. Shen, J. Yang, C. Xu, L. Lin, X. Cao, and S. Yan, "Matching-cnn meets knn: Quasi-parametric human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1419–1427.

[18] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic

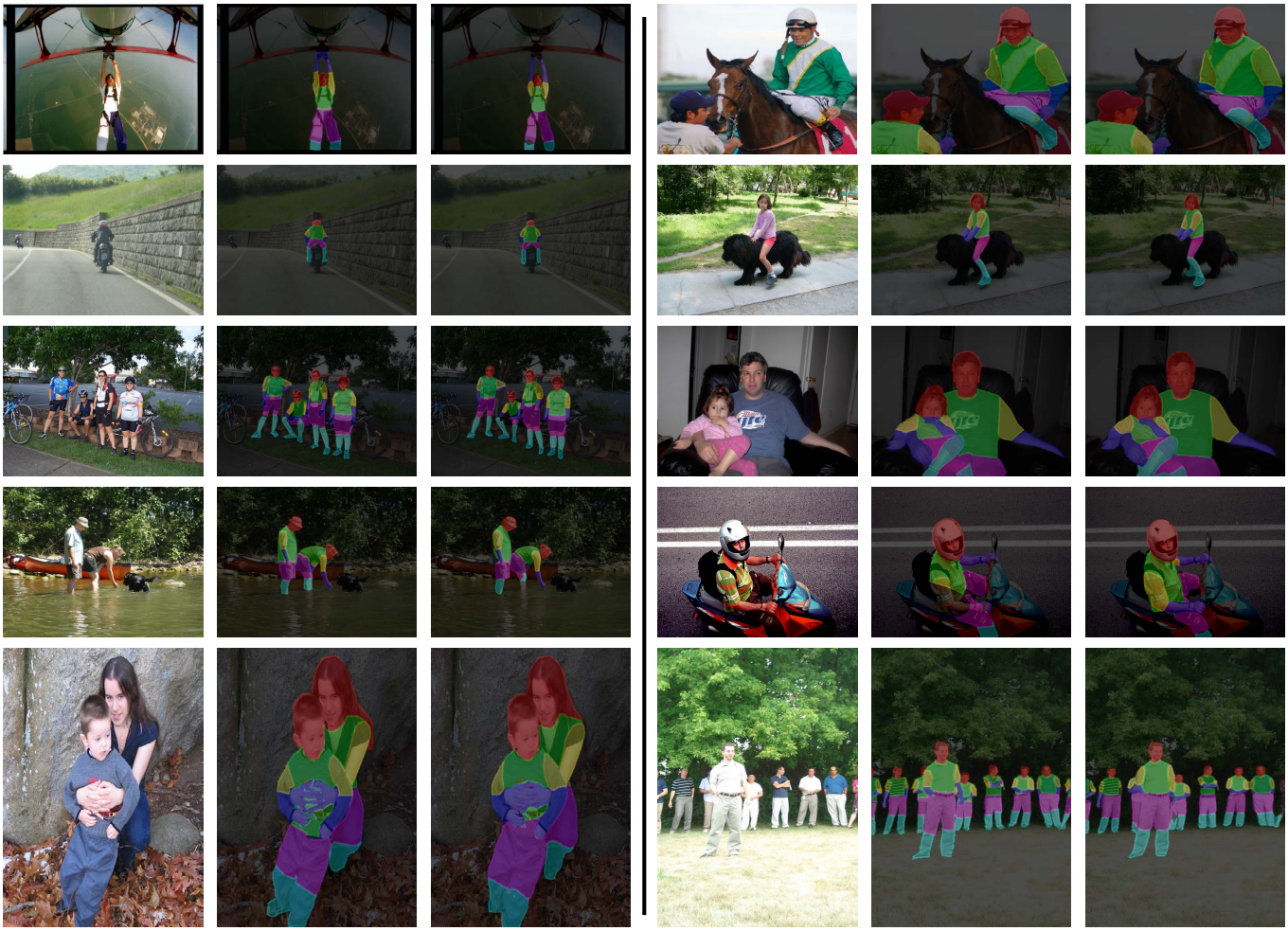


Fig. 4. Segmentation results of our CNIF model on the PASCAL-Person-Part `test` set [1].

image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[20] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan, "Human parsing with contextualized convolutional neural network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1386–1394.


Fig. 5. Segmentation results of our CNIF model on the LIP `val` set [3].


Fig. 6. Segmentation results of our CNIF model on the ATR `test` set [2].

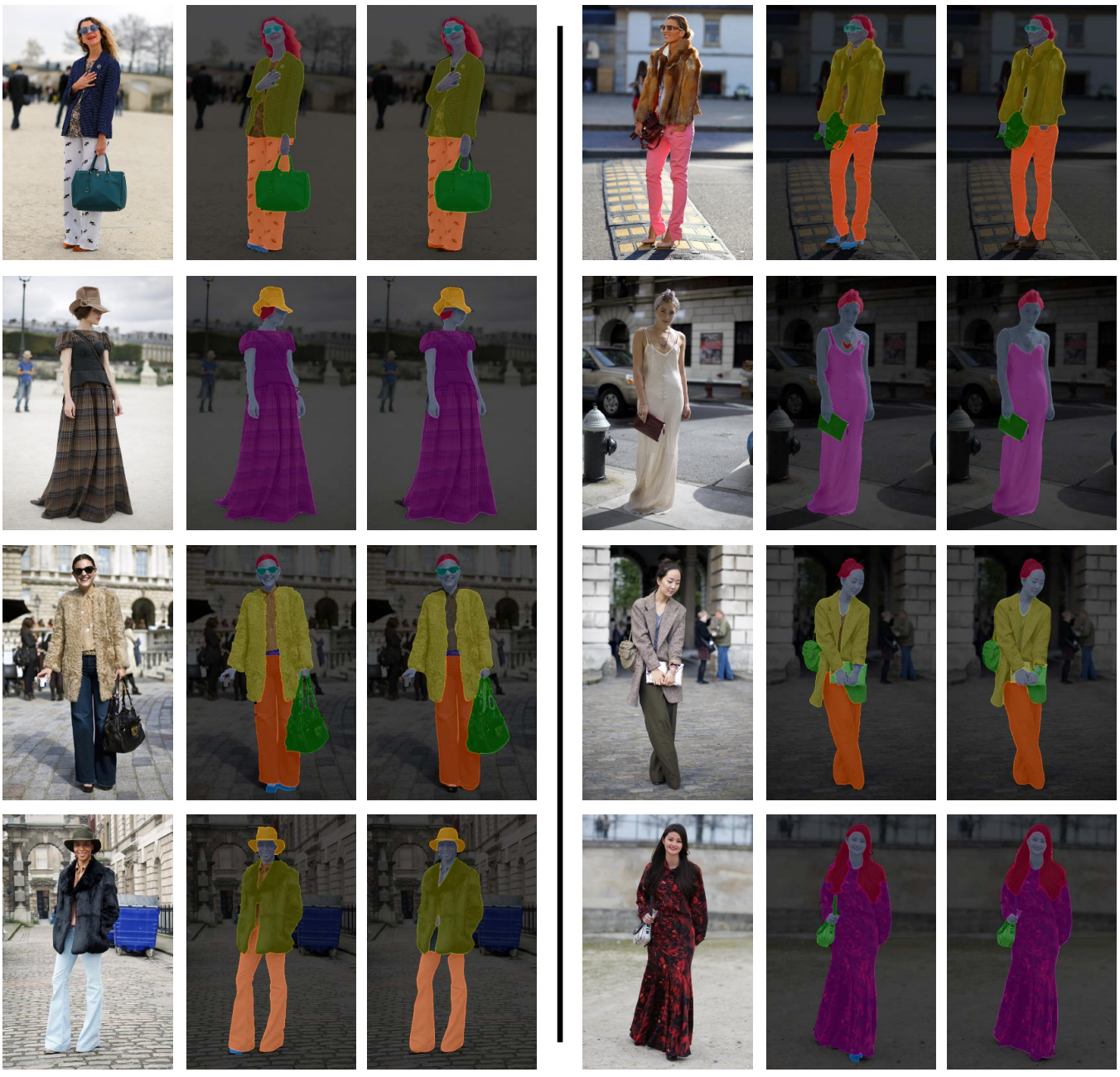


Fig. 7. Segmentation results of our CNIF model on the Fashion Clothing `test` set [4].


Fig. 8. Segmentation results of our CNIF model on the PPSS `test` set [5].

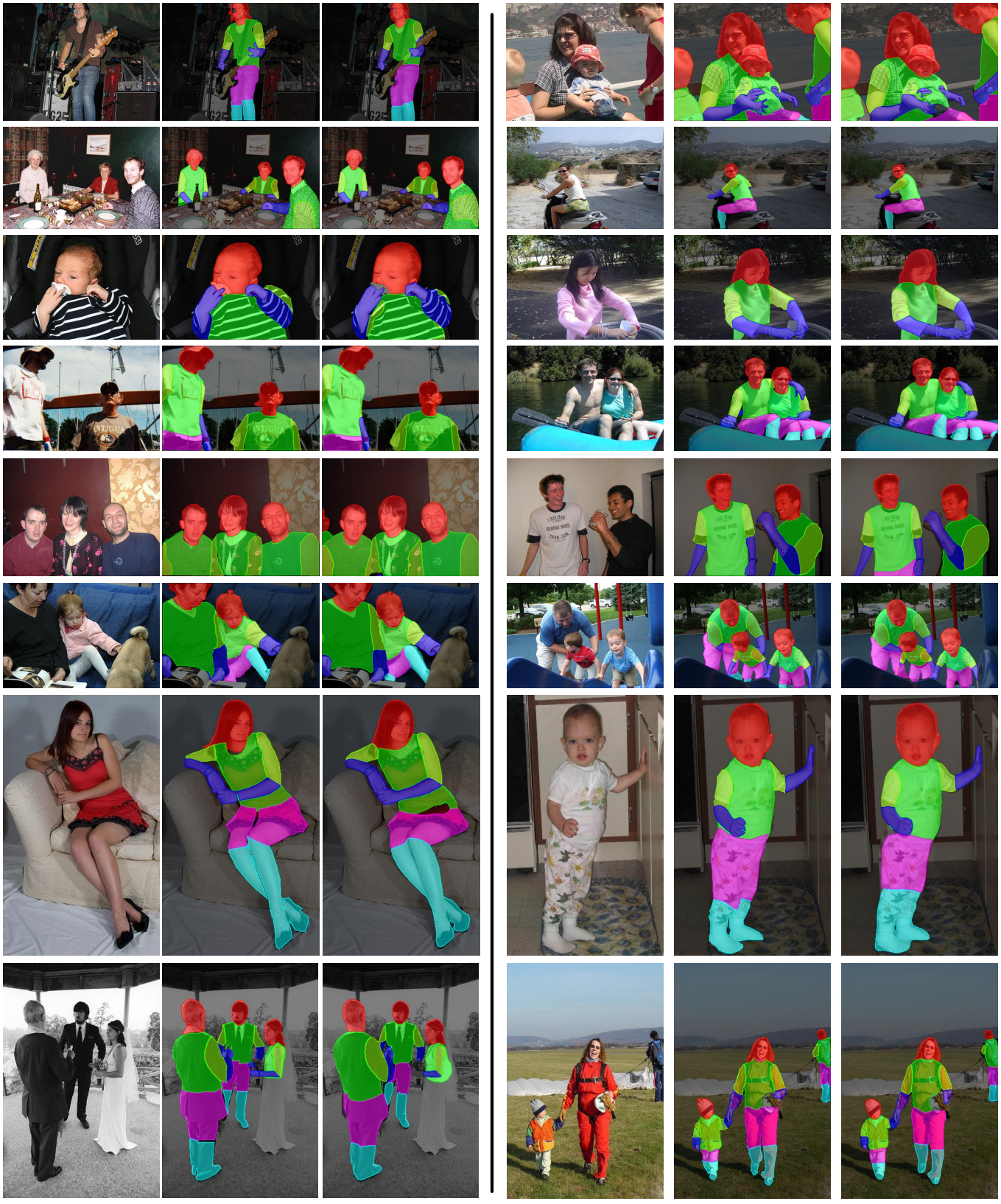


Fig. 9. Segmentation results of our PRHP model on PASCAL-Person-Part `test` set [1].

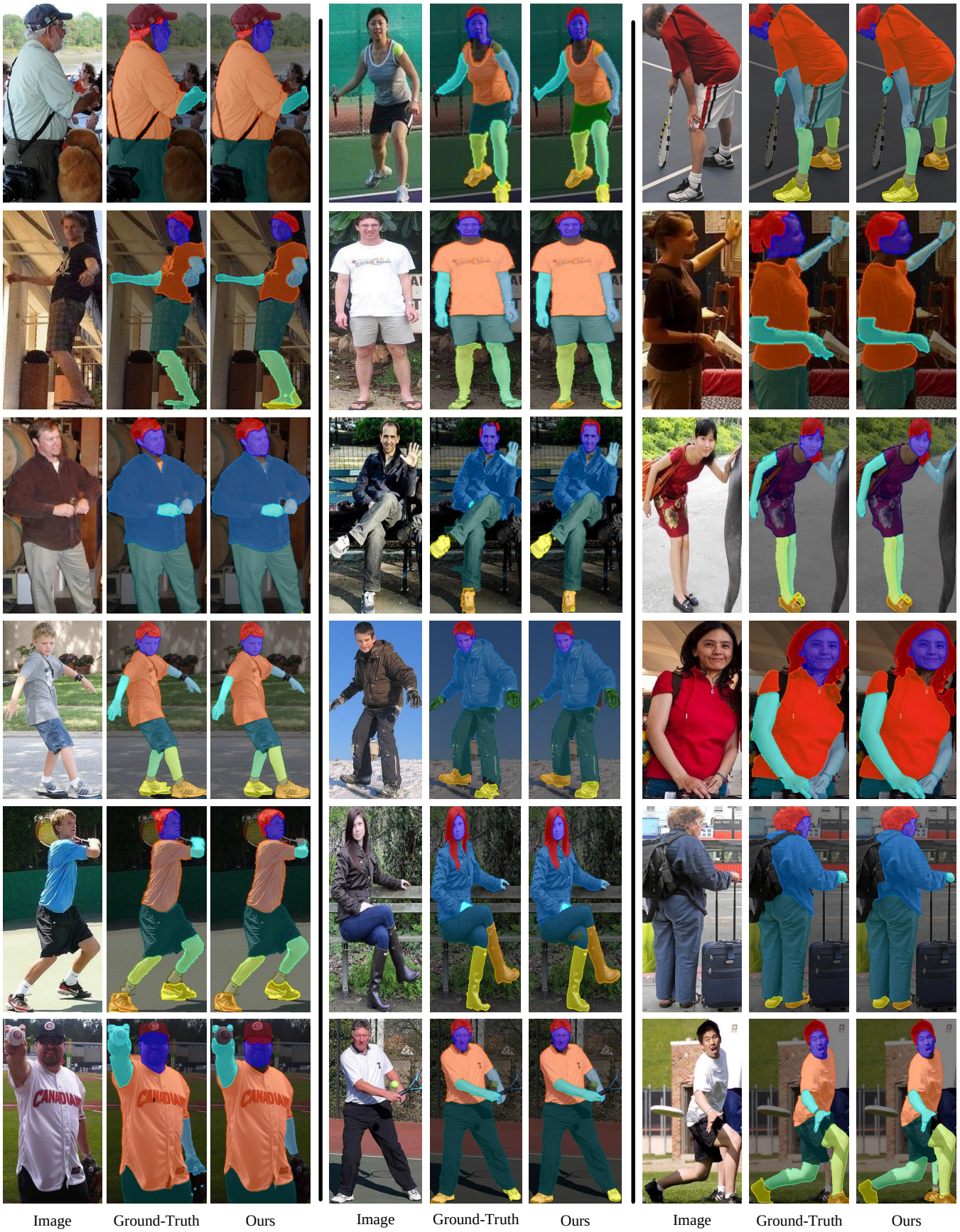


Fig. 10. Segmentation results of our PRHP model on LIP `val` set [3].


Fig. 11. Segmentation results of our PRHP model on ATR `test` set [2].

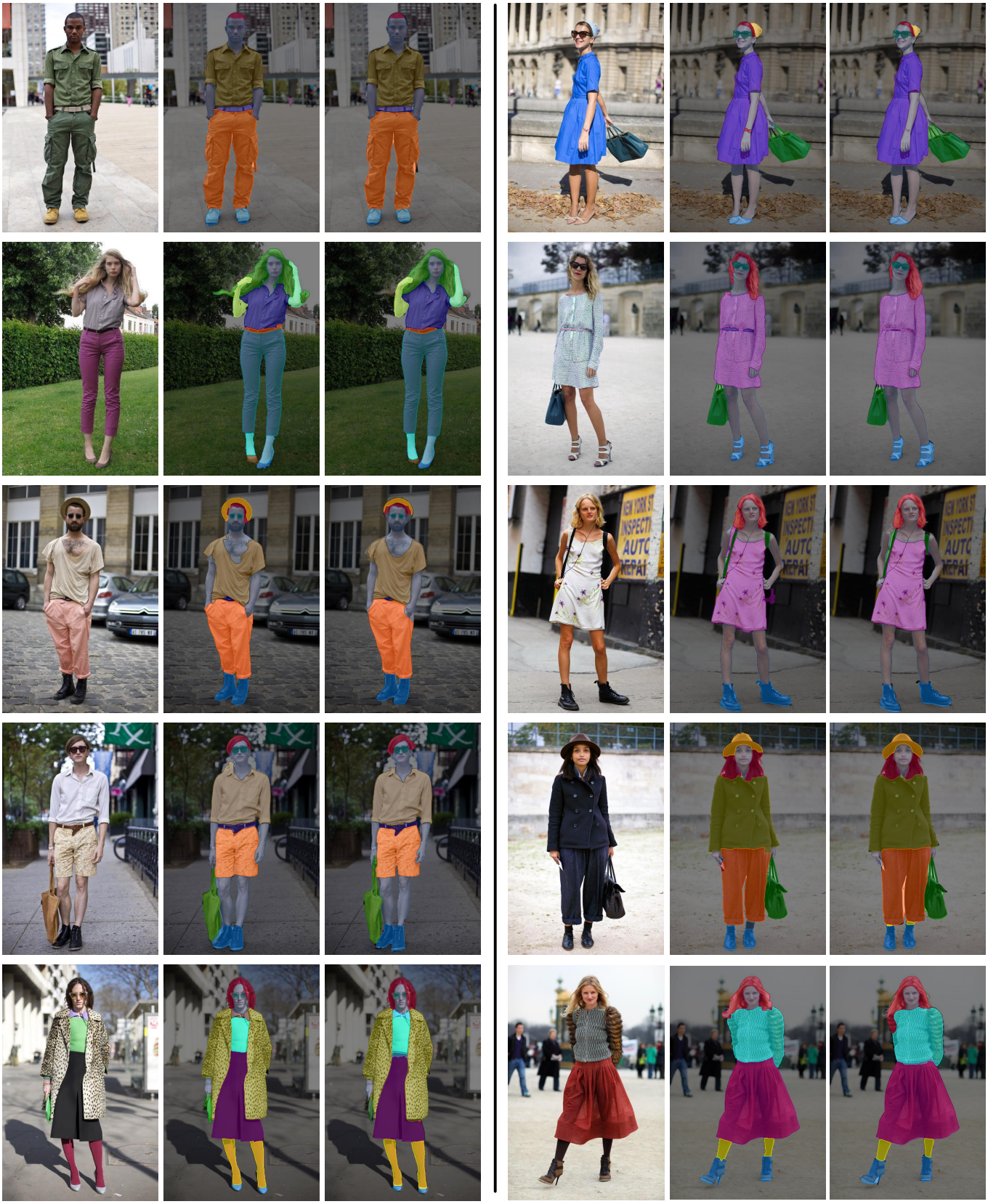


Fig. 12. Segmentation results of our PRHP model on Fashion Clothing `test` set [4].


Fig. 13. Segmentation results of our PRHP model on PPSS `test` set [5].