Quality-Aware Network for Human Parsing

Lu Yang ©, Qing Song ©, Zhihui Wang, Zhiwei Liu, Songcen Xu, and Zhihao Li ©

Abstract—How to estimate the quality of the network output is an important issue, and currently there is no effective solution in the field of human parsing. To solve this problem, this work proposes a statistical method based on the output probability map to calculate the pixel classification quality, which is called pixel score. In addition, the Quality-Aware Module (QAM) is proposed to fuse the different quality information, the purpose of which is to estimate the quality of human parsing results. We combine QAM with a concise and effective network design to propose Quality-Aware Network (QANet) for human parsing. Benefiting from the superiority of QAM and QANet, we achieve the best performance on three multiple and one single human parsing benchmarks, including CIHP, MHP-v2, Pascal-Person-Part, ATR and LIP. Without increasing the training and inference time, QAM improves the AP criterion by more than 10 points in the multiple human parsing task. QAM can be extended to other tasks with good quality estimation, e.g., instance segmentation. Specifically, QAM improves Mask R-CNN by -1% mAP on COCO and LVISv1.0 datasets. Based on the proposed QAM and QANet, our overall system wins 1st place in CVPR2021 L2ID High-resolution Human Parsing (HRHP) Challenge, and 2nd in CVPR2021 PIC Short-video Face Parsing (SFP) Challenge. Code and models are available at https://github.com/soeaver/QANet.

Index Terms—Computer vision, image segmentation, multimedia computing.

I. INTRODUCTION

PREDICTING the semantic category of each pixel on the human body is a fundamental task in multimedia and computer vision, which is usually known as human parsing [1], [2], [3], [4], [5], [6]. According to the number of human in the image, it can be divided into single human parsing [2], [5], [7] and multiple human parsing [3], [4], [8]. Due to the diverse human appearance, semantic ambiguity of different human parts, and complex background, the practical application of human parsing is still difficult.

In recent years, many efforts have been made to improve the performance of human parsing. Structured hierarchy [5], [7], [9], edge and pose supervisions [10], [11], [12], [13], attention mechanism [4], [8], [14], [15] and self-correction [16] have made remarkable progress. However, one important aspect has not been paid enough attention to, the quality estimation of human parsing results. The purpose of quality estimation is to score the prediction results of the system, thereby ranking outputs [17], [18], filtering out low-quality results [19] or eliminating data noise [16], [20]. Quality estimation has great potential in performance analysis and practical application for human parsing. There are a large number of network outputs that have a low-score with high-quality, or a high-score with low-quality. In practical application, the parsing scores of most methods can only reflect the quality of detection results, but can not reflect the quality of parsing results [4], [11], [21], [22], which makes it difficult to effectively filter out low-quality results by threshold. Moreover, in some works, parsing results will be output in the form of a structured hierarchy [5], [7], [9], [23]. In these methods, the score of each human part is the same and does not reflect the quality of semantic segmentation, so it is difficult to select the optimal level as the output.

The purpose of this work is to accurately estimate the quality of human parsing results. We use the box score predicted by the detector as the discriminant quality information. We extract the high confidence region from the probability map of network output, and take the average confidence of this region as the pixel classification quality information, which is recorded as the pixel score (Fig. 1). Referring [24] and [8], we also construct a lighter FCN to regress the IoU between the network output and the ground-truth, which is used to express the task quality information of the network output and record it as...
the IoU score. On this basis, we present Quality-Aware Module (QAM) as a way of building comprehensive and concise quality estimation for human parsing. QAM fuses the different quality information to generate a quality score to represent the quality of human parsing by using a flexible exponential weighting way. QAM is a post-processing mechanism independent of network structure, which can be combined with various methods. In this paper, we follow the simple and effective network design concept [25], combined with QAM to propose a network suitable for both single and multiple human parsing, called Quality-Aware Network (QANet). QANet takes ResNet-FPN [26], [27] or HR-Net [28] as the backbone, generates high-resolution semantic features through semantic FPN [29], and predicts parsing result and IoU score by two different branches. Finally, QANet uses the quality score to express the quality of human parsing results.

We extensively evaluate our approach on three multiple and one single human parsing benchmarks, including CIHP [3], MHP-v2 [21], Pascal-Person-Part [30] and LIP [2]. QANet has achieved state-of-the-art performances on these four benchmarks. Especially in terms of AP criterion, QANet is ahead of [4], [8], [9], [11] by more than 10 points on CIHP val. set. In addition, QAM is a post-processing quality estimation mechanism, which can be combined with other advanced methods. Experiments show that QAM can be directly integrated into the region-based multiple human parsing methods such as Parsing R-CNN [4] and RP R-CNN [8], and can significantly improve the performance of both methods. QAM also has good quality estimation ability in instance segmentation task, which can improve Mask R-CNN by 1% mAP on COCO [31] and LVISv1.0 [32] datasets. The performance improvement effect of QAM is equivalent to [24], but QAM does not require any additional training and inference time. These results show that QAM is a simple, modularized, low-cost and easily extensible module. We have reason to believe that it can be used in more tasks [33], [34], [35], [36], [37].

Benefiting from the superiority of QANet, our overall system wins 1st place in CVPR2021 L2ID High-resolution Human Parsing (HRHP) Challenge1, and 2nd place in CVPR2021 PIC Short-video Face Parsing (SFP) Challenge2.

II. RELATED WORK

Human Parsing: Thanks to the progress of convolutional neural network [26], [38], [39], [40] and open sourcing of large-scale human parsing datasets [2], [3], [21], human parsing has made great progress. For single human parsing, most of the work is devoted to introducing human body structure prior and improving network design. Wang et al [5], [7], [41] use deep graph networks and hierarchical human structures to exploit the human representational capacity. [11], [12] enhance the semantic information of human body by introducing edge or pose supervisions. Li et al [16] introduce a purification strategy to tackle the problem of learning with label noises, to promote the reliability of the supervised labels as well as the learned models. In the aspect of multiple human parsing, many effective top-down frameworks have been proposed. According to whether the human detector [39], [42], [43] and human parser are end-to-end trained, the top-down methods can be divided into one-stage and two-stage. One-stage top-down is a multi-task learning system based on R-CNN [44]. Yang et al present Parsing R-CNN [4] and RP R-CNN [8], which introduce multi-scale features [45], [46], attention mechanism [47], [48] and global semantic information to enhance the human visual perception ability of convolutional neural network, making breakthroughs in multiple human parsing and dense pose estimation [49] tasks. Two-stage top-down is a more flexible solution and the best-performing currently. Generally speaking, a single human parsing network can be used as the parser for two-stage top-down method [11], [16].

Quality Estimation: The purpose of quality estimation (or quality assessment) is to score or rank the prediction results of a system. The cross-entropy loss function can be regarded as an intuitive quality estimation, and most discriminant models use it to express the confidence of their results, such as image classification [26], [38], object detection [39], [50] and semantic segmentation [40]. However, more and more systems need customized quality estimation methods that are closer to their problems to accurately score or rank the outputs. In the research of recommendation system, [18], [51] use the list-wise loss function to compare the relationship between objects to achieve the ranking of the results. In order to build a high-performance face recognition system, [19] proposes an unsupervised estimation method to filter out low-quality face images. Huang et al [24] use a mask IoU head to learn the quality of the predicted instance masks. Similarly, Yang et al [8] present a parsing re-scoring network to sense the quality of human parsing maps and gives appropriate scores.

However, for human parsing, due to complex backgrounds, confusing categories, long-tailed phenomena [52] and etc., the existing methods can not accurately express the quality of results, and can not effectively estimate the quality of each human part independently.

III. METHODOLOGY

A. Motivation

The quality estimation of human parsing is an important but long-neglected issue. Inappropriate scoring or ranking the network outputs results in a large number of low-score with high-quality or high-score with low-quality outputs. Meanwhile, it is difficult for us to filter out low-quality results by threshold. Therefore, our goal is to propose a simple, low-cost solution, which can effectively score the results of human parsing, and can also give the independent score of each human part.

B. Estimating the Quality With Probability Map

Formally, the output of the human parsing network is a feature map \( f \in \mathbb{R}^{N \times C \times H \times W} \), where \( N \) denotes the number of human instances, \( C \) denotes the number of parsing categories, \( H \) and \( W \) are height and width of feature maps. For the feature map \( f \), we take the largest probability prediction among the

---

1[Online]. Available: https://l2id.github.io/challenge_localization.html

2[Online]. Available: https://maadaa.ai/cvpr2021-short-video-face-parsing-challenge/
C categories for each spatial position, and record it as probability map $p \in \mathbb{R}^{N \times H \times W}$. Each position in $p$ expresses the classification confidence of corresponding pixel. Therefore, the probability map can be regarded as the quality estimation of each pixel. By making full use of the pixel classification quality information of the probability map, we can estimate the quality of the whole human body or part. A naive solution is to average the confidence of all positions in the probability map to represent the quality of the parsing result. However, this scheme is inaccurate. Since the mechanism of neural network is still difficult to explain, the output of some difficult samples is always full of randomness [53], [54]. As shown in Fig. 2, some low-confidence regions on the probability map often correspond to the boundary of the instance and some easily confused areas, but this does not mean that the prediction is wrong. For example, for a human instance with many boundaries or a human part with small area, the average of the probability map will be lower no matter whether the prediction is accurate or not. Therefore, if these areas are included in the average calculation, it will bring serious deviation to the quality estimation.

To solve the above issue, we adopt a simple and effective scheme. We only calculate the average value of regions with confidence greater than threshold $T$, and the default $T = 0.2$. More specifically, we calculate the high confidence mask ($hcm \in \mathbb{R}^{H \times W}$) on the probability map according to the threshold $T$. Then the average confidence of $hcm$ is calculated to express the pixel classification quality information of the human parsing network output, which is recorded as the instance pixel score. In addition, we also use the same method to calculate the pixel information quality of each predicted human part, which is recorded as a category pixel score. Algorithm 1 provides the PyTorch-like style pseudo-code for calculating instance pixel score and category pixel score. This calculation process is model-independent and has less computational overhead.

### C. Quality Aware Module

In addition to the pixel classification quality information of the probability map, some other information can be used to estimate the quality of human parsing results. Detector not only outputs the location information of human instance, but also gives the confidence of classification, which is called discriminant quality information. Mean intersection over union (mIoU) [40] is the basic index to measure the parsing performance. Referring [24] and [8], we can regress the mIoU between the predicted parsing result and the ground-truth through a lightweight FCN, which represents the task quality information. Discriminant quality information, task quality information and pixel classification quality information all express the output quality of convolution neural networks in some aspects, and they are complementary to each other.

We propose Quality-Aware Module (QAM), which fuses the different quality information to estimate the output quality. In human parsing task, the box score $S_b$ represents discriminant quality information, IoU score $S_i$ represents task quality information and pixel score $S_p$ represents pixel classification quality information. We use the exponential weighting to calculate the final quality score $S_q$:

$$S_q = (S_b^\alpha \times S_i^\beta \times S_p^\gamma)^{1/\alpha + \beta + \gamma},$$

where $\alpha$, $\beta$, and $\gamma$ are the corresponding weight of each category quality information, respectively.
where $\alpha$, $\beta$, and $\gamma$ are used to adjust the weight of different quality information, $\alpha = \beta = \gamma = 1.0$ by default. When using the default weights, QAM can significantly improve the effectiveness of human parsing quality estimation, and can get better results by adjusting the weights. The idea of QAM is to flexibly fuse the different quality information to generate a more comprehensive quality score, which is a simple, low-cost and easily extensible method. QAM can be used not only for human parsing, but also for instance segmentation [55], [56] and dense pose estimation [4] after simple adaptation.

D. Architecture

To verify the effect of QAM, we propose a two-stage top-down human parsing network, called Quality-Aware Network (QANet). Our design principle starts from SimpleNet [25], a strong and concise pose estimation baseline. As illustrated in Fig. 3, we adopt a ResNet [26] with FPN [27] or HRNet [28] as backbone, which produces multi-scale features $P_2$ - $P_5$. Each scale feature is upsampled by convolutions and bilinear upsampling until it reaches 1/4 scale, these outputs are integrated into a high-resolution feature by element-wise summed [29]. Two branches are attached to the high-resolution feature to predict the parsing results and IoU scores respectively. The parsing loss is a combination of the cross-entropy loss ($L_c$) and the lovasz loss ($L_z$) [57]. The IoU branch is a lightweight FCN, and uses MSE loss ($L_m$) to regress the mIoU between the predicted parsing result and the ground-truth. Thus, the loss function of the whole network is:

$$L = L_c + L_z + L_m.$$  \hspace{1cm} (2)

Since the QAM module does not directly participate in the optimization of the parsing map, it has no impact on the quality of human semantic segmentation (mIoU criterion), but is used to predict and calculate the human parsing scores ($AP^p$ and $AP^r$ criterions). Specifically, in the training phase, only the IoU branch participates in the joint optimization of the whole parsing network. In the inference phase, QAM exponentially weights $S_i$ predicted by the IoU branch, $S_b$ predicted by the detector (non-end-to-end) and $S_p$ calculated by Algorithm 1, to calculate the final quality score $S_q$ according to Formula 1.

IV. Experiments

A. Experimental Settings

Datasets: Three multiple [3], [21], [30] and two single [1], [2] human parsing benchmarks are used for performance evaluation. CIHP [3] is a popular multiple human parsing dataset that has 28,280 images for training, 5,000 images for validation and 5,000 images for testing with 20 categories. MHP-v2 [21] includes 25,403 elaborately annotated images with 58 fine-grained semantic category labels. The validation set and test set have 5,000 images respectively. The rest 15,403 are provided as the training set. PASCAL-Person-Part [30] contains 3,535 annotated images distributed to 1,717 for training and 1,818 for testing, among which only 7 categories are labeled. ATR [1] is a classical single human parsing dataset with 18-class semantic annotations, which contains 7,700 images. These images are split into 6,000/700/1,000 for training/validating/testing. LIP [2] is a single human parsing dataset and contains 50,462 images, which are collected from realistic scenarios and divided into 30,462 images for training, 10,000 for validation and 10,000 for testing.

Evaluation Metrics: The standard mIoU criterion is adopted for evaluation on human part segmentation, both for single\(^3\) and multiple human parsing. We also use Average Precision based on part ($AP^p$), Average Precision based on region ($AP^r$) and

\(^3\)The quality estimation can not be reflected in the mIoU criterion, so we use QANet without QAM for single human parsing.
TABLE I
ABSTRACTION OF QUALITY WEIGHTS. ALL MODELS ARE TRAINED ON CIHP train SET AND EVALUATED ON CIHP val SET

| Quality Weights | α, β, γ | AP\textsuperscript{p} | AP\textsuperscript{50} | AP\textsuperscript{r} | AP\textsuperscript{5}\textsuperscript{0} |
|-----------------|---------|-----------------|-----------------|------------------|------------------|
| (a) (1.0, 0.0, 0.0) | 55.6 | 67.7 | 40.1 | 45.0 |
| (b) (0.0, 1.0, 0.0) | 56.9 (+1.3) | 70.7 (+3.0) | 44.3 (+4.2) | 50.0 (+5.0) |
| (c) (0.0, 0.0, 1.0) | 49.0 (−6.6) | 60.8 (−6.9) | 49.7 (−9.6) | 55.4 (−10.4) |
| (d) (1.0, 1.0, 0.0) | 60.2 (+4.6) | 74.5 (+6.8) | 45.6 (−5.8) | 51.4 (+6.4) |
| (e) (1.0, 0.0, 1.0) | 57.0 (+1.4) | 70.2 (+2.5) | 55.0 (−14.9) | 62.3 (+17.9) |
| (f) (0.0, 1.0, 1.0) | 57.0 (+1.4) | 70.9 (+3.2) | 52.5 (+12.4) | 59.6 (+14.4) |
| (g) (1.0, 1.0, 1.0) | 60.2 (+4.6) | 74.5 (+6.8) | 54.0 (−13.9) | 61.2 (−16.2) |
| (h) (1.0, 0.5, 3.0) | 60.1 (+4.5) | 74.3 (+6.6) | 56.2 (+14.1) | 63.5 (−16.5) |

Percentage of Correctly parsed semantic Parts (PCP\textsubscript{50}) to measure the performance of multiple human parsing, following [3], [4], [21].

Training: We implement the QAM and QANet based on Pytorch on a server with 8 NVIDIA RTX GPUs. We use ResNet50, ResNet101 and HRNet-W48 as the backbones to compare different methods fairly. In particular, we only conduct HRNet-W48 experiments on LIP. Following [25], the ground truth human box is made to a fixed aspect ratio (4 : 3) by extending the box in height or width. The default input size of QANet is 512 × 384. Data augmentation includes scale [−30%, +30%], rotation [−40°, +40°] and horizontal flip. For optimization, we use the ADAM [58] solver with 256 batch size. All models are trained for 140 epochs with 2e-3 base learning rate, which drops to 2e-4 at 90 epochs and 2e-5 at 120 epochs. No other training techniques are used, such as warm-up [59], syncBN [60], learning rate annealing [45], [61].

Inference: For multiple human parsing, a two-stage top-down paradigm is applied [25], [62]. By default we use a FCOS [42] with ResNet50 as detector. The QANet result is the combination of the original input and the flipped image. For evaluating on LIP, we also use multi-scale test augmentation, following [5], [7].

B. Ablation Studies

In this sub-section, we assess the effects of different settings on QAM and QANet by details ablation studies.

Threshold of Pixel Score: Threshold T is an important hyper-parameter for calculating the pixel score (see Algorithm 1). The table below shows QANet-ResNet50 with different thresholds:

| threshold T | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 |
|-------------|-----|-----|-----|-----|-----|
| AP\textsuperscript{p}/AP\textsuperscript{r} | 47.5/42.2 | 60.1/56.3 | 60.1/56.2 | 60.0/56.0 | 59.9/55.1 |

when T = 0.0, the effect of pixel score is negative. Increasing T to filter out low confidence region can significantly improve the effect. Therefore, we choose T = 0.2. As illustrated in Fig. 1, (category) pixel score can well reflect the quality of parsing results in different human parts.

Quality Weights: Quality weights are used to adjust the proportion of box score, IoU score and pixel score in quality estimation. Table I shows the ablation study of quality weights on CIHP val set. All experiments are based on ResNet50. Row (a), (b) and (c) indicate the performance of human parsing when only box score, IoU score or pixel score is used. It can be seen that the IoU score improves both AP\textsuperscript{p} and AP\textsuperscript{r}, but pixel score improves AP\textsuperscript{r} and degrades AP\textsuperscript{p}. Row (d), (e) and (f) indicate any combination of two scores will improve the AP\textsuperscript{p} and AP\textsuperscript{r}. Especially when the pixel score is used, the improvement of AP\textsuperscript{p} is very significant. This proves pixel score is better at representing local (human part) quality. When box score, IoU score and pixel score are used with equal weights (Row (g)), AP\textsuperscript{p} and AP\textsuperscript{r} are increased by 4.6 and 13.9 points respectively. This fully proves that the pixel score and QAM proposed by this work are effective. After grid search, we use (1.0, 0.5, 3.0) as the quality weights to get the best performance on CIHP. We also use this weight on both MHP-v2 and PASCAL-Person-Part. In addition, we use ground-truth boxes (GT-box) to replace the predicted boxes. The experiments (see Table II) shows that QAM can still significantly improve the performance of human parsing when using GT-box. To further reveal the effect of QAM, we visualize two samples from CIHP val set, as shown in Fig. 4. It can be clearly observed that the box score does not accurately reflect the parsing quality. For example, the human instance on the right has a high box score (S\textsubscript{b} = 0.96), but due to the complexity of the clothing, the mIoU between the prediction and ground-truth is very low (GT = 0.56). The quality score S\textsubscript{p} = 0.96 is very close to GT, which shows that the QAM outputs can better reflect the quality of parsing results than the box scores.

From SimpleNet to QANet: We adopt SimpleNet as our baseline, which uses ResNet50 as the backbone and followed by three deconvolutional layers [25]. By default, the input size of SimpleNet is 256 × 192. Table III shows the ablations of CIHP, the SimpleNet achieves 54.6% mIoU, 51.8% AP\textsuperscript{p} and 32.5% AP\textsuperscript{r} on CIHP val set. Semantic FPN, lovasz loss, IoU score and 512 × 384 input size outperform the baseline by 8.3% mIoU, 8.5% AP\textsuperscript{p} and 13.1% AP\textsuperscript{r}. This performance has exceeded the best method on CIHP [8], [9], [12], which shows that our network architecture is concise but effective. Furthermore, our proposed QANet (with QAM) achieves 62.9% mIoU, 60.1% AP\textsuperscript{p} and 56.2% AP\textsuperscript{r}. QAM brings more than 10 points AP\textsuperscript{p} improvement. For ablations of LIP, Table IV illustrates a roadmap similar to Table III, the SimpleNet achieves 51.10% mIoU on LIP val set. Combining with Semantic FPN, lovasz loss, IoU score
TABLE II
THE EFFECT OF QAM WITH GT-BOX OR PREDICTED BOX. ALL MODELS ARE TRAINED ON CIHP train SET AND EVALUATED ON CIHP val SET.

| Methods          | GT-box | QAM | mIoU | AP<sup>p</sup> | AP<sup>o</sup> | PC<sup>0.5</sup> | AP<sup>+</sup> | AP<sup>0.5</sup> |
|------------------|--------|-----|------|---------------|---------------|-----------------|-------------|-----------------|
| QANet-RS0        | ✓      | ✓   | 65.0(±2.1) | 55.6          | 67.7          | 68.9            | 40.1        | 43.0            |
|                  | ✓      | ✓   | 62.9(±0.6)  | 61.9(±2.8)    | 69.0(±0.1)    | 38.0(±2.1)      | 42.8(±2.2)  |

TABLE III
ABLATION STUDY OF QANET ON CIHP DATASET. 'S-FPN' IS SEMANTIC FPN, 'LOV ASZ' IS LOVASZ LOSS, 'IOU' DENOTES A LIGHTWEIGHT FCN TO PREDICT IOU SCORE, 'LARGE-RES' DENOTES 512 × 384 INPUT RESOLUTION.

| Methods          | S-FPN | Lovasz | IOU | Large-Res | QAM | mIoU | AP<sup>p</sup> | AP<sup>o</sup> | PC<sup>0.5</sup> | AP<sup>+</sup> | AP<sup>0.5</sup> |
|------------------|-------|--------|-----|-----------|-----|------|---------------|---------------|-----------------|-------------|-----------------|
| SimpleNet        | ✓     | ✓      | ✓   | ✓         | ✓   | 56.2(±1.6)  | 53.6(±2.0)    | 63.8(±3.7)    | 62.0(±2.9)     | 34.7(±2.2)  | 38.4(±2.3)      |
|                  | ✓     | ✓      | ✓   | ✓         | ✓   | 58.9(±4.3)  | 53.9(±2.1)    | 64.3(±3.2)    | 65.9(±6.8)     | 37.2(±4.7)  | 41.9(±6.8)      |
|                  | ✓     | ✓      | ✓   | ✓         | ✓   | 59.3(±4.7)  | 58.3(±6.5)    | 71.9(±11.8)   | 66.0(±6.9)     | 41.6(±9.3)  | 47.1(±11.0)     |
| QANet            | ✓     | ✓      | ✓   | ✓         | ✓   | 62.9(±8.3)  | 60.3(±8.5)    | 74.5(±14.4)   | 68.9(±9.8)     | 45.6(±13.1)| 51.4(±15.3)     |

TABLE IV
ABLATION STUDY OF QANET ON LIP DATASET. 'S-FPN' IS SEMANTIC FPN, 'LOV ASZ' IS LOVASZ LOSS, 'IOU' DENOTES A LIGHTWEIGHT FCN TO PREDICT IOU SCORE, 'LARGE-RES' DENOTES 512 × 384 INPUT RESOLUTION. THE SYMBOL † MARKS THE MULTI-SCALE TESTING RESULT.

| Methods          | S-FPN | Lovasz | IOU | Large-Res | QAM | mIoU |
|------------------|-------|--------|-----|-----------|-----|------|
| SimpleNet        | ✓     | ✓      | ✓   | ✓         | ✓   | 51.10|
|                  | ✓     | ✓      | ✓   | ✓         | ✓   | 53.60(±2.50)|
|                  | ✓     | ✓      | ✓   | ✓         | ✓   | 56.53(±1.93)|
|                  | ✓     | ✓      | ✓   | ✓         | ✓   | 55.97(±0.44)|
| QANet            | ✓     | ✓      | ✓   | ✓         | ✓   | 58.79(±7.69)|
| QANet†           | ✓     | ✓      | ✓   | ✓         | ✓   | 58.61(±8.82)|

TABLE V
COMPARISON OF PIXEL ACCURACY, F.G. ACCURACY AND F-SCORE ON ATR TEST SET. OUR BEST SINGLE MODEL ACHIEVES STATE-OF-THE-ART WITH SMALLER INPUT SIZE (512 × 384 < 473 × 473).

| Methods          | Pub. | Input Size | pixAcc. | F.G. Acc. | mIoU |
|------------------|------|------------|---------|-----------|------|
| Yamaguchi [63]   | CVPR2015 | 512 × 384  | 84.38   | 55.59     | 41.80|
| Paperdoll [64]   | ICCV2013 | 512 × 384  | 88.96   | 62.18     | 44.76|
| M-CNN [65]       | CVPR2015 | 512 × 384  | 89.57   | 73.98     | 62.81|
| ATR [1]          | TPAMI2015 | 227 × 227 | 91.11   | 71.04     | 64.38|
| Co-CNN [66]      | ICCV2015 | 150 × 100  | 95.23   | 80.90     | 76.95|
| LG-LST7 [67]     | CVPR2016 | 321 × 321  | 96.18   | 84.79     | 80.97|
| TGPNet [68]      | MM2018  | 321 × 321  | 96.45   | 87.91     | 81.76|
| CNIP [7]         | ICCV2019 | 473 × 473  | 96.26   | 87.91     | 85.51|
| SCHIP [16]       | TPAMI2020 | 473 × 473 | 96.25   | 87.97     | 85.55|
| CorrPM [12]      | CVPR2020 | 384 × 384  | 97.12   | 90.40     | 86.12|
| HHP [5]          | CVPR2020 | 473 × 473  | 96.84   | 89.23     | 87.25|
| QANet (ours)     | –     | 512 × 384  | 96.91   | 90.54     | 87.53|

and larger input resolution, QANet outperforms the baseline by 7.69% mIoU. Ablations of CIHP and LIP show the huge performance improvement of QANet compared with the baseline, which proves that our network is concise and effective.

C. Quantitative and Qualitative Results

ATR [1]: ATR is a classical benchmark for single human parsing. Table V exhibits QANet (ResNet101) and comparisons with the 11 state-of-the-arts models in terms of three metrics on ATR test set. Our approach achieves 96.81% pixAcc., 90.54% F.G. Acc. and 87.53% mIoU, which sets new state-of-the-arts on F.G. Acc. and mIoU without bells and whistles.

LIP [2]: LIP is a standard benchmark for single human parsing. Table VI shows QANet (HRNet-W48) with other state-of-the-arts on LIP val set. To ensure that the number of pixels of the input image is almost the same, we have carried out the experiments of 512 × 384 input size, which is smaller than others. In terms of mean Acc. and mean IoU, our QANet surpasses the best performing method HHP† [5] by 0.59 and 0.25 points, respectively.

CIHP [3]: CIHP is the most popular benchmark for multiple human parsing. As shown in Table VII, QANet with ResNet50 is better than the current best methods [8], [9] in terms of mIoU, AP<sup>p</sup> and AP<sup>o</sup>. In terms of AP<sup>r</sup>, QANet outperforms RP R-CNN [8] by 13.9 points, outperforms SemaTree [8] by 12.2 points. Such a performance gain is particularly impressive considering that improvement on multiple human parsing task is

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
very challenging. Combined with the larger capacity backbone HRNet-W48, QANet achieves state-of-the-art on CIHP val set, which is 66.1\% mIoU, 64.5\% AP\_p, 75.7\% PCP\_50 and 60.8\% AP\_r.

**MHP-v2 [21]:** In the middle part of Table VII also reports QANet performance on MHP-v2 val set. Considering that the previous methods do not provide AP\_r criterions, we only compare mIoU and AP\_p. QANet with ResNet50 outperforms RP R-CNN [8] by 3.8 points in mIoU. With ResNet101, QANet outperforms CE2P [11] by 2.0 points in mIoU and 6.5 points in AP\_p.

**PASCAL-Person-Part [30]:** In the lower part of Table VII, we compare our method against recent methods on PASCAL-Person-Part test set. In term AP\_p and AP\_r, our improvement is remarkable. QANet surpasses PGN [3], Parsing R-CNN [4] and RP R-CNN [8] more than 10.0 points in AP\_p and 15.0 points in AP\_r.

**Qualitative Results:** Some qualitative comparison results on CIHP val set are depicted in Fig. 5. Qualitative comparison mainly shows the semantic segmentation ability of different methods. Compared with Parsing R-CNN [4] (first row) and RP R-CNN [8] (second row), QANet (third row) has better performance in occlusion, human contour, complex background and confusing categories.

In some cases, the proposed QANet fails to obtain ideal parsing results, as shown in Fig. 6. For example, when the calculated quality score $S_q$ is too low, it will lead to missing parsing.

### Table VII

| Datasets | Methods | Pub | Backbones | Epochs | mIoU | AP\_p | AP\_r | PCP\_50 | AP\_p | AP\_r |
|----------|---------|-----|-----------|--------|------|-------|-------|--------|-------|-------|
| CIHP [3] | PGN\[^{[3]}\] | ECCV\_2018 | ResNet101 | ~80 | 55.8 | 39.0 | 61.0 | 33.6 | 35.8 |
|         | Graphonomy [23] | CVPR\_2019 | ResNet101 | 100 | 58.6 | - | - | - | - |
|         | GPM [80] | AAAI\_2020 | ResNet101 | 100 | 60.3 | - | - | - | - |
|         | Grapy-ML [80] | AAAI\_2020 | ResNet101 | 200 | 60.6 | - | - | - | - |
|         | CorPM [12] | CVPR\_2020 | ResNet101 | 150 | 60.2 | - | - | - | - |
|         | PCNet [76] | CVPR\_2020 | ResNet101 | 120 | 61.0 | - | - | - | - |
|         | HTConvM [81] | TPAMI\_2021 | HRNet-W48 | 180 | 62.0 | - | - | - | - |
|         | PRM [79] | TMID\_2022 | HRNet-W48 | 120 | 62.5 | - | - | - | - |
|         | **Bold** | **Numbers** are **State-of-the-Art on Each Dataset**, and **gray Numbers** are Based on the **Official Open Sourcing Implementation** by us [4], [8]. |
|          | **PPP** [30] | | | | | | | | |
|         | Parsing R-CNN [4] | CVPR\_2019 | ResNet50 | 75 | 56.3 | 53.9 | 63.7 | 60.1 | 36.5 | 40.9 |
|         | Unified [22] | BMVC\_2019 | ResNet101 | ~37 | 55.2 | 48.0 | 51.0 | 38.6 | 44.0 |
|         | RP R-CNN [8] | ECCV\_2020 | ResNet50 | 150 | 60.2 | 59.5 | 74.1 | 64.9 | 42.3 | 48.2 |
| MHP-v2 [21] | Parsing R-CNN [4] | CVPR\_2019 | ResNet50 | 75 | 36.2 | 39.5 | 24.5 | 37.2 | - | - |
|          | RP R-CNN [8] | ECCV\_2020 | ResNet50 | 150 | 38.6 | 46.8 | 45.3 | 43.8 | - | - |
|          | **Bold** | **Numbers** are **State-of-the-Art on Each Dataset**, and **gray Numbers** are Based on the **Official Open Sourcing Implementation** by us [4], [8]. |
|          | **PPP** [30] | | | | | | | | |
|         | Parsing R-CNN [4] | CVPR\_2019 | ResNet50 | 75 | 62.7 | 49.8 | 58.2 | 48.7 | 40.4 | 43.7 |
|         | RP R-CNN [8] | ECCV\_2020 | ResNet50 | 75 | 63.3 | 50.1 | 58.9 | 49.1 | 40.9 | 44.1 |
|          | **Bold** | **Numbers** are **State-of-the-Art on Each Dataset**, and **gray Numbers** are Based on the **Official Open Sourcing Implementation** by us [4], [8]. |
especially in densely occluded scenes. In addition, QANet occasionally makes category prediction errors in some easily confused categories.

### D. Extensibility

**Region-based Human Parsing:** QAM can be regarded as a post-processing module, so it can be easily combined with other studies. Table VIII shows the performance of QAM combined with two region-based multiple human parsing methods on CIHP val set. QAM improves Parsing R-CNN [4] about 2.9 points in APp and 10.7 points in APr. The IoU score has been used in RP R-CNN [8], so QAM can improves 6.0 APp and achieve similar APr. This is consistent with the experimental results in Row (d) and (e) of Table I.

**Instance Segmentation:** Instance segmentation and multiple human parsing are highly similar in the processing flow. Therefore, we believe that QAM (especially pixel score) is beneficial to the instance segmentation task. Table IX shows that results of some methods on COCO [31] and LVISv1.0 [32] with or without QAM. On COCO val set, Mask R-CNN [55] with QAM is better than the original counterpart: up to +1.0 APp.

QAM can also slightly improve SOLOv2 [56]. On LVISv1.0 val set, we can also observe steady improvements. In addition, we also apply QAM to the video instance segmentation task [86]. Extending the pixel score from single image masks to sequential masks yields significant improvements on Mask2Former [87], [88] (from 40.6% to 43.1%). This improvement is astonishing, revealing the great potential of QAM in the temporal dense pixel prediction tasks [89], [90], [91].

**High-resolution Human Parsing Challenge:** Benefiting from the superiority of QANet, we win 1st place in CVPR2021 L2ID High-resolution Human Parsing (HRHP) Challenge [92] (see Table X) with better human detector [42], [93] and test-time techniques. HRHP aims to recognize human parts (19 semantics) within high-resolution images (about 4,000 × 4,000). QANet achieves 49.09% EIoU (edge IoU) and 79.38% mIoU on HRHP test set, outperforms the 2nd place by 1.99 points and 0.62 points, respectively. The results show that QANet is not only suitable for low-resolution images, but also can be extended to extreme high-resolution images, and can segment the edges of parts very well (as shown in Fig. 7).

---

### TABLE VIII

**Comparison of Region-Based Multiple Human Parsing on CIHP val set With or Without QAM. The Quality Weights of Parsing R-CNN [4] is (1.0, 0.0, 1.0), and of RP R-CNN [8] is (3.0, 1.0, 1.0).**

| Methods       | QAM | APp | APs | APr |
|---------------|-----|-----|-----|-----|
| Parsing R-CNN [4] | ✔  | 53.9 | 63.7 | 38.5 |
|               |     | 55.5(+2.6) | 67.7(+4.0) | 47.2(+10.7) |
| RP R-CNN [8]   | ✔  | 59.5(+5.6) | 74.2(+10.1) | 48.3(+4.0) |
|               |     | 55.5(+2.5) | 63.7(+10.0) | 48.2 |

---

### TABLE IX

**Comparison of Instance Segmentation on COCO / LVISv1.0 / YTVIS-2021 val sets With or Without QAM. The Quality Weights is (1.0, 0.0, 1.0).**

| Dataset       | Methods         | QAM | APp  | APs  |
|---------------|-----------------|-----|------|------|
| COCO          | Mask R-CNN [55] | ✔  | 35.2 | 66.2 |
|               |                 |     | 36.2(+1.0) | 55.5(-0.7) |
|               | SOLOv2 [56]     | ✔  | 34.8 | 54.0 |
|               |                 |     | 35.2(+0.4) | 53.5(-0.6) |
| LVISv1.0      | Mask R-CNN [55] | ✔  | 20.1(+1.0) | 31.2(+1.0) |
|               |                 |     | 20.1(+1.0) | 31.2(+1.0) |
| YTVIS-2021    | Mask2Former [87]| ✔  | 40.6 | 65.3 |
|               |                 |     | 43.1(+2.8) | 65.3(+1.4) |

---

4emAll experiments are based on Detectron2 [85].
Fig. 6. Failure cases of QANet on CIHP val set. The left images are ground-truths, and the rights are the visual results of QANet. Upper: missing parsing; Lower: category error.

Table X
QANet Wins 1st Place in CVPR2021 L2ID High-Resolution Human Parsing (HRHP) Challenge. Our Result Ensembles Two QANet Models

| Team          | EIoU   | mIoU  | Rank |
|---------------|--------|-------|------|
| soeaver (ours)| 49.09  | 79.38 | 1    |
| DeepBlueAI    | 47.10  | 78.76 | 2    |
| DISL          | 45.33  | 78.68 | 3    |
| ligror        | 44.96  | 80.16 | 4    |
| hzcxq         | 44.50  | 77.83 | 5    |

Fig. 7. Visual results of QANet on HRHP val set. From left to right are: high-resolution images, locally enlarged images, high-resolution parsing results and locally enlarged parsing results. It can be seen that QANet has very good parsing results for small targets such as earrings, rings and hair.

Table XI
QANet Wins 2nd Place in CVPR2021 PIC Short-Video Face Parsing (SFP) Challenge. Our Result Does Not Use Model Ensembles

| Team          | Score | Rank |
|---------------|-------|------|
| TCParser      | 86.95 | 1    |
| BUPT-CASIA (ours) | 86.84 | 2    |
| rat           | 86.16 | 3    |
| SVFPCCC       | 85.94 | 4    |
| hust sb       | 85.87 | 5    |

V. Conclusion

The quality estimation of human parsing is an important but long-neglected issue. In this work, we propose a network that can generate high-quality parsing results and a simple, low-cost quality estimation method, named QANet. Its core components are pixel score based on probability map and the quality-aware module (QAM) which integrates different quality information. QANet has verified its effectiveness and advancement on several challenging benchmarks, and won 1st place and 2nd place in two popular challenges, including CVPR2021 L2ID High-resolution Human Parsing (HRHP) Challenge and CVPR2021 L2ID High-resolution Human Parsing (HRHP) Challenge. Beyond the human parsing task, it is possible to adopt QANet and QAM for other instance-level tasks, like instance segmentation and dense pose estimation. We hope to promote a breakthrough in the form of quality estimation and help with more visual tasks.

References

[1] X. Liang et al., “Deep human parsing with active template regression,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 12, pp. 2402–2414, Dec. 2015.
Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 4900–4909.

B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in Proc. Eur. Conf. Comput. Vis., 2018, pp. 466–481.

K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

T.-Y. Lin et al., "Feature pyramid networks for object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2117–2125.

J. Wang et al., "Deep high-resolution representation learning for visual recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

A. Kirillov, R. Girshick, K. He, and P. Dollar, "Panoptic feature pyramid networks," in Proc. IEEE/CVF Comput. Vis. Pattern Recognit., 2019, pp. 6399–6408.

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.

T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.

A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5356–5364.

S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4724–4732.

A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, "Panoptic segmentation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 9404–9413.

G. Gkioxari, J. Malik, and J. Johnson, "Mesh R-CNN," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 9785–9795.

L. Yang, Q. Song, Z. Wang, M. Ha, and C. Liu, "Hier R-CNN: Instance-level human parts detection and a new benchmark," IEEE Trans. Image Process., vol. 30, no. 39, pp. 39–54, 2021.

L. Li, T. Zhou, W. Wang, J. Li, and Y. Yang, "Deep hierarchical semantic segmentation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 1246–1257.

A. Kirillov, J. Tittekevers, and G. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017.

S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2016.

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.

W. Wang, T. Zhou, S. Qi, J. Shen, and S.-C. Zhu, "Hierarchical human semantic parsing with comprehensive part-relation modeling," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 7, pp. 3508–3522, Jul. 2022.

Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional onestage object detection," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 9627–9636.

B. Zhu et al., "CPM R-CNN: Calibrating point-guided misalignment in object detection," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., 2021, pp. 3248–3257.

R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 580–587.

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, Apr. 2018.

L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proc. Eur. Conf. Comput. Vis., 2018, pp. 801–818.

X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7794–7803.

L. Yang, Q. Song, Y. Wu, and M. Hu, "Attention inspiring receptive-fields networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 580–587.

R. A. Guler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7297–7306.

W. Liu et al., "SSD: Single shot multibox detector," in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.

C. Pei, Y. Zhang, Y. Zhang, F. Sun, and D. Pei, "Personalized re-ranking for recommendation," in Proc. 13th ACM Conf. Recommender Syst., 2019, pp. 3–11.

L. Yang, H. Jiang, Q. Song, and J. Guo, "A survey on long-tailed visual recognition," Int. J. Comput. Vis., vol. 130, pp. 1837–1872, 2022.

B. Zhou, D. Bau, A. Oliva, and A. Torralba, "Interpreting deep visual representations via network dissection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 9, pp. 2131–2145, Sep. 2019.
