Hierarchical Information Passing Based Noise-Tolerant Hybrid Learning for Semi-Supervised Human Parsing

Yunan Liu[1], **Shanshan Zhang**[1*], **Jian Yang**[1], **PongChi Yuen**[2]
[1]PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education,
Jiangsu Key Lab of Image and Video Understanding for Social Security,
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[2]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
{liuyunan, shanshan.zhang, csjyang}@njust.edu.cn, pcyuen@comp.hkbu.edu.hk

Abstract

Deep learning based human parsing methods usually require a large amount of training data to reach high performance. However, it is costly and time-consuming to obtain manually annotated high quality labels for a large scale dataset. To alleviate annotation efforts, we propose a new semi-supervised human parsing method for which we only need a small number of labels for training. First, we generate high quality pseudo labels on unlabeled images using a hierarchical information passing network (HIPN), which reasons human part segmentation in a coarse to fine manner. Furthermore, we develop a noise-tolerant hybrid learning method, which takes advantage of positive and negative learning to better handle noisy pseudo labels. When evaluated on standard human parsing benchmarks, our HIPN achieves a new state-of-the-art performance. Moreover, our noise-tolerant hybrid learning method further improves the performance and outperforms the state-of-the-art semi-supervised method (i.e. GRN) by 4.47 points w.r.t mIoU on the LIP dataset.

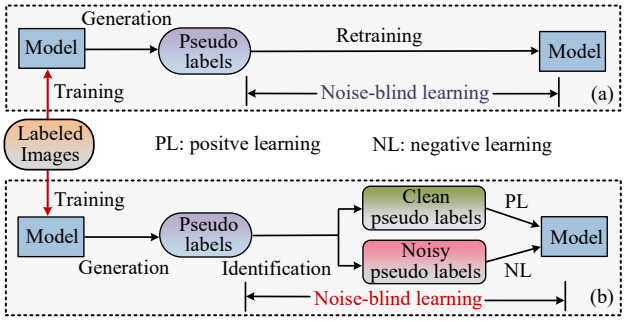

Figure 1: Comparison of different semi-supervised methods for human parsing. (a) Previous methods directly use generated pseudo labels to retrain the model. In contrast, (b) we identify noisy and clean labels, and further develop a noisy-tolerant hybrid learning strategy to make full use of correct information from clean pseudo labels while eliminate incorrect information from noisy pseudo labels.

Introduction

Humans exist commonly and are the most interesting objects in daily visual data, e.g. public surveillance videos, private photographs, etc. Therefore, it is of great interests for the computer vision community to analyze humans in visual data. Human parsing (i.e. partitioning human body to multiple semantically consistent regions), is a crucial yet challenging task for fine-grained human body understanding, which is helpful for many other high-level computer vision tasks such as pedestrian attribute recognition (Li et al. 2019b), human pose estimation (Liu et al. 2021), and person re-identification (Guo et al. 2019), etc.

With the development of convolutional neural networks (CNNs), human parsing has witnessed dramatic progress. Typically, CNN-based approaches require a large number of labeled training samples to boost performance. But it is rather expensive to collect high-quality and fine-grained annotations for human body. In order to reduce annotation costs, a promising solution is to take advantage of semi-supervised learning. Most existing methods first use a human parsing model trained on annotated data to generate pseudo labels for unlabeled images and then retrain the model using the augmented training samples, following the standard pipeline shown in Fig 1(a). Some of them (Fang et al. 2018; Lin et al. 2019) adopt the skeleton key points as auxiliary information to generate pseudo labels. However, these algorithms require extra computing resources of human pose, which are usually unavailable in real cases and inevitably introduce new errors. Some other works (Li et al. 2019a, 2020) adopt rectification strategy to correct some structure error and consistency error of pseudo labels. All of these works aim to design a powerful and robust model to generate high-quality pseudo labels. Nevertheless, the primary model trained on a small amount of data is somehow weak, and thus the generated pseudo labels are inevitably noisy. Therefore, it is necessary to take care of the noise during the second round of training, which has not been considered by previous works in this filed so far.

In this paper, we make the first attempt to train a ro-

*The corresponding author is Shanshan Zhang.

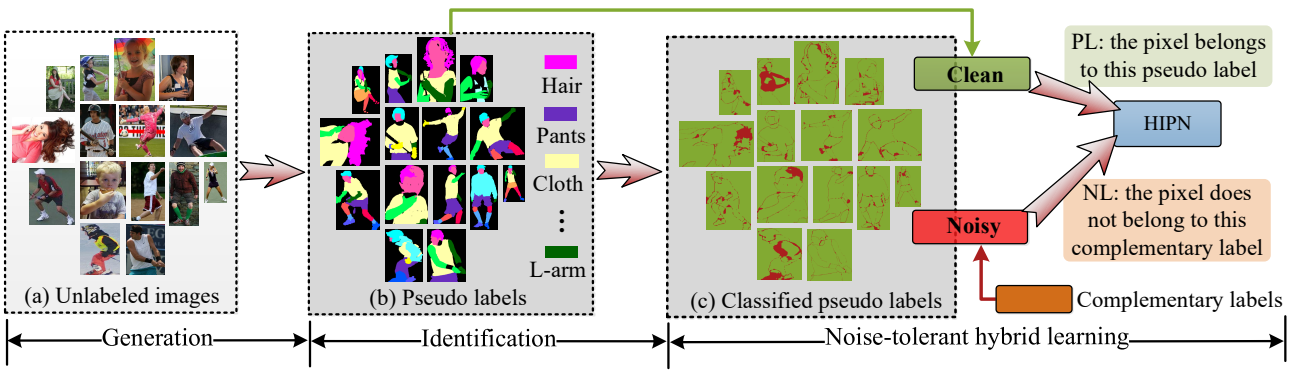

Figure 2: An overview of our proposed semi-supervised human parsing method. We first use the trained HIPN model to generate pseudo labels for unlabeled images. Then, we classify pseudo labels as noisy and clean ones. Finally, we develop a noise-tolerant hybrid learning strategy to retrain the HIPN model.

.

bust human parsing model against noisy pseudo labels. The pipeline is shown in Fig. 1(b). Compared to the traditional pipeline (Fig. 1(a)), we replace the noise-blind learning with a new noise-tolerant one. This noise-tolerant learning consists of noisy label identification and a hybrid learning strategy, which handles clean labels with traditional positive learning and handles noisy labels via humble negative learning (Kim et al. 2019). First, noisy label identification is performed based on the assumption that the higher the certainty of the prediction, the sharper the probability distribution is. Then we measure the certainty of each predicted pseudo label as the distance between the produced probability distribution and its sharpest counterpart. For those uncertain predictions, we apply a negative learning method to alleviate the negative effects from label noise. The negative learning method (Kim et al. 2019) sets a humble hypothesis that the input image does not belong to some semantic categories beside the predicted pseudo label. Once the pseudo label is wrong, negative learning has a smaller chance to make errors than positive learning. But the drawback is that the provided knowledge also becomes weaker when the pseudo label is correct. Therefore, in this work, we first identify noisy labels and apply negative learning on those labels only.

In summary, the main contributions of this paper include:

• We propose a hierarchical information passing network (HIPN), which exploits physiological structure of human body to explicitly model human part relations. Based on the HIPN, we obtain high-quality pseudo labels for unlabeled images and classify them as clean and noisy pseudo labels accurately.

• We develop a noise-tolerant hybrid learning strategy to take full advantage of positive and negative learning, which better handles noisy pseudo labels. To the best of our knowledge, this is the first work that handles noisy pseudo labels for semi-supervised human parsing.

• The experimental results on the challenging human parsing datasets demonstrate the effectiveness of the proposed semi-supervised method. In particular, our method achieves mIoU score of 60.81 on the large scale benchmark (i.e. LIP), which outperforms the state-of-the-art semi-supervised method by 4.47 pp.

Related Works

Fully-supervised human parsing. Human parsing is a subtask of semantic segmentation, but with the particular structure constraint on the human body. Aiming for better performance, some recent works propose to use additional cues to assist human parsing. For example, Zhao et al. (Zhao et al. 2018, 2020) use saliency maps as a basic prior to facilitate human parsing. Li et al. (Li et al. 2017a, 2018) tackles multiple human parsing by generating instance masks for multiple persons. The works of (Gong et al. 2017; Zhao et al. 2017; Xia et al. 2016; Fang et al. 2018) jointly train pose estimation and human parsing networks to improve the performance of both tasks. Considering the correlations between human parsing and edge detection, some methods (Gong et al. 2018; Ruan et al. 2019; Zhao et al. 2019) leverage edge details to sharpen human parsing predictions via an encoder-decoder structure. In (Wang et al. 2019a, 2020b; Li et al. 2020; He et al. 2020), the hierarchical structure of human body is adopted to facilitate human part reasoning. Most of these methods ignore to explicitly model human part relations, easily suffering from weak expressive ability and risk of sub-optimal results.

Semi-supervised human parsing. To reduce annotation costs, some semi-supervised methods are proposed. A straight forward way is to generate pseudo labels on unlabeled images and then include them for training. Gong et al. (Gong et al. 2019) propose to generate labels on a different human parsing dataset via transferring unmatched labels based on a graph model, and then add this dataset to the training pool. Fang et al. (Fang et al. 2018) aim to employ those images with labeled human body keypoints for training; the parsing labels can be automatically generated by finding the most similar skeleton in the labeled human parsing dataset. Some other works (Li et al. 2019a, 2020) attempt to refine pseudo labels via a better performed model for a more robust re-training. Nevertheless, one can hardly generate perfect pseudo labels, and the inevitable noises may result in error amplification and accumulation during

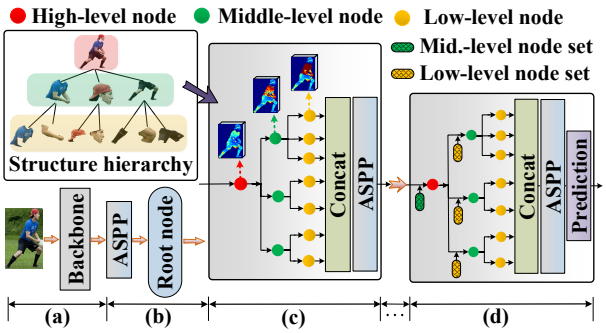

Figure 3: Illustration of our hierarchical information passing network (HIPN). Specifically, based on the hierarchy of human body, we design a cascaded pipeline for progressive refinement, which consists of K stages. (a) We adopt backbone network to extract features from an input image. And then (b) we use the ASPP module as projection function to obtain root node representations. (c) At the 1st stage, we initialize node features by passing information in a top-down manner. (d) At each later stage, we update each node by fusing the information from both its parent and child nodes.

.

the second round of training. A new training strategy tolerant to noisy labels is necessary, but has not been investigated before for human parsing.

Learning with noisy labels. The solutions of learning with noisy labels can be roughly classified into two types. The first one aims to rectify potential noisy labels allowing for easier optimization. Some of them achieve this by modeling the noises through graphical models (Xiao et al. 2015) or knowledge graphs (Li et al. 2017b). However, these methods either require a large number of clean annotations to estimate the noise model or become ineffective at high noise regimes. To solve this problem, Tanaka et al. (Tanaka et al. 2018) propose a joint optimization framework to alternatively update network parameters and labels. Zhang et al. (Zhang et al. 2020) propose a meta re-labeling method, which leverages a small trusted set to estimate pseudo labels for noisy data in order to reuse them in the following supervised training. In contrast, the second group of methods design some noise-robust models, which can be fed with noisy labels. In some works (Jiang et al. 2018; Huang et al. 2019), each training sample is re-weighted based on the reliability of the given label. However, it is challenging to estimate effective weights. An alternative way is to modify the loss functions to compensate for the incorrect guidance provided by noisy labels. To this end, some noise-robust loss functions are designed, such as generalized cross entropy loss (Zhang and Sabuncu 2018), symmetric cross entropy loss (Wang et al. 2019b), and negative loss (Kim et al. 2019), etc. These methods only require minimal intervention to existing algorithms and architectures. In this work, we propose a noise-tolerant hybrid learning strategy, where noisy labels are identified based on our hierarchical information passing network and the negative loss (Kim et al. 2019) is used to alleviate the negative effects from noisy labels.

Method

In this section, we first give an overview of our proposed method. After that, we provide detailed descriptions to our Hierarchical Information Passing Network (HIPN), and noise-tolerant hybrid learning strategy.

Overview

The pipeline of our semi-supervised human parsing is shown in Fig. 2. We first propose a hierarchical information passing network (HIPN) and train a primary model using a small set of images with clean labels. By utilizing the primary HIPN model, we then generate pseudo labels for a large number of unlabeled images, and we classify these pseudo labels as noisy and clean ones. Finally, we develop a noise-tolerant hybrid learning strategy to retrain the HIPN. In order to better handle noisy pseudo labels, we apply positive learning on clean pseudo labels that "the pixel belongs to this pseudo label". In contrast, for each noisy pseudo label, we first replace it with a complementary label, and then apply negative learning for training the model that "the pixel does not belong to this complementary label". In this manner, we make sure the correct information from clean labels are fully exploited yet the risk of using incorrect information from noisy labels is decreased.

Hierarchical Information Passing Network

As shown in Fig. 3, our Hierarchical Information Passing Network (HIPN) consists of K stages. Specifically, we begin with the backbone network for feature extraction. Next, we adopt the atrous spatial pyramid pooling (ASPP) module (Chen et al. 2017) as projection function to obtain enriched features with various receptive fields.

Let us denote the features from the HRNet (Wang et al. 2020a) backbone as X, and the projection function as F, then the root representation h_r can be represented as $h_r^{W\times H\times C} = F(X)$, where W, H, and C are the width, height and number of channels of h_r, respectively. We also define a hierarchical human body structure, which consists of three levels, as shown in Fig. 3. The principle is that each parent node covers its child nodes.

At the first stage, we only allow information pass from higher level nodes to lower level ones and we use a node information passing (NIP) module to model the relation between the parent node and its child nodes. For a parent node p with representation h_p, let us denote its child node as v. By utilizing the NIP module, the representation h_v of v can be initialized as:

$$h_v = h_p \otimes F_{NIP}(h_p) + h_p, \tag{1}$$

where $\otimes$ indicates the element-wise multiplication operation, and F_{NIP} represents the function of NIP module, which applies on h_p to extract attention map for h_v. Here, we serve the global root representation extracted from ASPP as the parent node of high-level node. In each NIP module, we first use one self-calibrated block (Liu et al. 2020a) to generate more discriminative representations by explicitly incorporating richer information. After that, we use one convolutional layer with kernel size 1×1 and one softmax

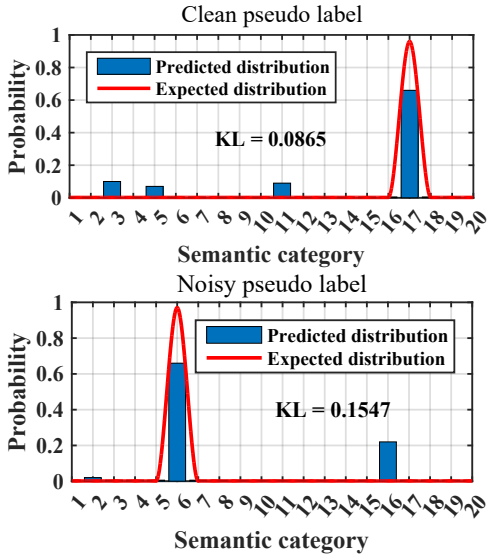

Figure 4: Measuring certainty of pseudo labels. Two examples have the same highest probability, and thus cannot be distinguished by applying a threshold on it (Kim et al. 2019). In contrast, we use the KL-divergence to measure the distance between predicted and expected probability distributions, which exploits the probability distribution across different categories, and thus is more robust.

.

layer to output the attention map $R_{u,v} = F_{NIP}(h_p)$ for node v. Notably, the channels of $R_{u,v}$ consists of foreground and background channels, where $R_{u,v} \in (0, 1)$. To this end, we supervise $R_{u,v}$ based on the labels of pre-set virtual categories. Finally, we combine all the feature of each low-level node and use ASPP to extract global root representation for the later stage.

At the second and subsequent stages, we consider additional bottom-up relations, which transfers the information of human body structure from lower level to higher level nodes. For each node u^k at k-th stage, its representation h_u^k can be represented as:

$$h_u^k = h_p^k \otimes F_{NIP}(F_{1\times1}([H_{u,v}^{k-1}, h_p^k]) + h_p^k, \qquad (2)$$

where $[\cdot]$ indicates the concatenation operation; $F_{1\times1}$ represents the mapping function implemented by a 1×1 convolutional layer; and $H_{u,v}^{k-1} = [h_{u,1}^{k-1}, ..., h_{u,V}^{k-1}]$ indicates the features of u's child nodes from the preceding stage, where V indicates the number of its child nodes. In this way, we allow information pass throughout the entire tree structure in both top-down and bottom-up directions, so as to obtain more discriminative features.

At the end of the last stage, we use a prediction module to predict the final parsing results. Specifically, we first use a 1×1 convolutional layer to reduce the dimension of feature maps according to the number of semantic category, and then we use a softmax layer to estimate a probability distribution for each pixel as final human parsing prediction.

Noise-Tolerant Hybrid Learning with Pseudo Labels

Human parsing is a pixel-wise classification task. A typical way of training CNNs for human parsing with given pixels and the corresponding labels is called positive learning (PL), i.e. for each pixel, the information provided to CNNs is that "it belongs to this given label". However, when the the labels are noisy, wrong information will be provided to the model. Recently, Kim et al. (Kim et al. 2019) use the negative learning strategy (NL) to decrease the risk of providing incorrect information from noisy labels. With negative learning, the CNNs are trained that "the pixel does not belong to this complementary label"; the complementary label is selected among those categories except the most likely one with the highest probability. Compared to PL, NL provides humble information for training, while reducing the chance of making errors. When the pseudo labels are clean, it is preferable to use PL to provide more information to the optimization procedure; once the pseudo labels are noisy, NL is more suitable to reduce errors. Therefore, we propose a hybrid learning strategy, where we first distinguish noisy labels from clean ones and then apply PL for clean labels and NL for noisy ones, respectively.

In order to identify noisy pseudo labels, a simple way is to make the decision based on the predicted confidences for the most likely category, i.e. when the confidence is below a predefined threshold, this label is identified as noisy (Kim et al. 2019). However, it is challenging to choose a suitable threshold; and it is not sufficient only looking at the highest probability. To solve this problem, we propose a robust method to distinguish noisy labels from clean ones, which takes advantage of multi-level predictions of HIPN. Noisy labels are identified via the following two criterion:

(1) Uncertainty of the final prediction. It is assumed that the more concentrate the probability distribution is, the more certain the prediction is. The ideal probability distribution should be like a Dirac delta distribution with a probability of 1 at the most likely category and 0 for others. The probabilities usually become more distributed when the prediction is less certain. Based on the above assumption, we measure the distance between the obtained probability distribution and the Dirac delta distribution as the prediction's uncertainty. We use the KL-divergence as the distance measure. In principle, a larger distance value indicate to higher uncertainty. Compared to the previous work (Kim et al. 2019) that only considers the highest probability to judge the reliability of each prediction, we exploit the probability distribution across different categories, and thus is more robust. We show two examples in Fig. 4. We can see that the two samples have the same highest probability, but belong to clean and noisy, respectively; the difference lies in their probability distributions. Thus it is more robust to consider the whole probability distribution rather than the highest value only. Those labels with a higher distance than a given threshold is considered as noisy ones. The threshold is determined by the average KL-divergence values inside each image.

(2) Inconsistency of predictions at different levels. In prin-

Variant			mIoU	Δ
Stage K	Information passing dirction			
	Top-down	Bottom-up		
Baseline			69.45	-/-
1	✓		71.81	2.36
2	✓		72.67	3.22
	✓	✓	72.94	3.49
3	✓		73.15	3.70
	✓	✓	73.51	4.06
4	✓		73.58	4.13
	✓	✓	73.66	4.21

Table 1: Comparison of different variants on the Extended PASCAL-Person-Part dataset. "Δ" indicates the improvement w.r.t mIoU compared to baseline method (i.e. HRNet).

Method	Accuracy	mIoU
Baseline	-/-	74.13
NLNL (Kim et al. 2019)	60.59	74.27
Our method (certainty)	63.81	74.57
Our method (certainty & consistency)	64.91	74.91

Table 2: Comparison of different noise identification methods on the Extended PASCAL-Person-Part dataset. The performance is evaluated in terms of clean/noisy classification accuracy and mIoU, respectively. "Baseline" indicates the noise-blind model without noisy label identification.

ciple, the predictions at different levels should be consistent. For example, if one pixel is predicted at the middle level as upper body, it is unreasonable to be predicted as leg at the low level. Once inconsistency is found across levels, the pseudo label should be considered as unreliable, i.e. noisy.

After distinguishing between clean and noisy pseudo labels, we develop a noise-tolerant hybrid learning strategy to retrain the HIPN, in which the positive and negative learning are applied on clean and noisy pseudo labels, respectively. We consider the problem of c-class classification for each pixel. Let x represent a pixel of input image, y and $\hat{y}$ indicate its pseudo label and complementary label, respectively. Suppose the proposed HIPN $f(x:\theta)$ maps the input pixel to a c-dimensional probability distribution p, where θ indicates network parameter set. When applying PL on clean pseudo labels, the cross entropy loss function is defined as follows:

$$L_{PL}(f, y) = -\sum_{k=1}^{c} y_k \log p_k, \tag{3}$$

where p_k is the k-th element of p. Eq. 3 aims at optimizing the probability value corresponding to the given pseudo label as 1. In contrast, NL optimizes the predicted probability corresponding to the complementary label to be far from 1. Therefore, the cross entropy loss function of NL becomes:

$$L_{NL}(f, \hat{y}) = -\sum_{k=1}^{c} \hat{y}_k (1 - \log p_k), \tag{4}$$

DS	PL	HL	Pixel Acc.	Mean Acc.	mIoU	ΔmIoU
1/8			80.48	53.67	39.93	-/-
	✓		82.66	55.94	42.91	2.98
		✓	83.25	56.35	43.54	3.61
1/2			84.37	62.00	49.32	-/-
	✓		86.92	65.36	53.97	4.65
		✓	87.67	66.10	54.69	5.37
1			88.21	67.43	55.90	-/-

Table 3: Semi-supervised experiments on the LIP dataset. "DS" denotes the labeled data size; "PL" and "HL" represent the standard positive learning and our noise-tolerant hybrid learning, respectively.

For every iteration during training, the complementary label $\hat{y}$ is randomly selected from the labels of all semantic categories except for the given noisy pseudo label y. Eq. 4 enables the probability value of the complementary label to be optimized as zero, resulting in an increase in the probability values of other classes. Considering Eqs. 3 and 4, the total loss is defined as:

$$L = \alpha L_{PL} + (1 - \alpha) L_{NL}, \tag{5}$$

where the α is a weighting parameter to balance two losses.

Experiment

In this section, we first introduce the datasets used in our experiments and implementation details. Then, we conduct ablation studies to validate the effectiveness of each main component of our method. Furthermore, we compare our proposed method with state-of-the-art methods.

Datasets

We conduct experiments on two standard benchmarks for human parsing: LIP dataset (Zhao et al. 2017), and extended PASCAL-Person-Part dataset (Xia et al. 2017). The LIP dataset includes 50,462 images for single human parsing. Each image contains one full or partial human body with pixel-wise annotations of 19 part categories. The whole dataset is split into training, validation and test sets, which consist of 30,462 images, 10,000 images and 10,000 images, respectively. All methods are trained on the training set and evaluated on the validation set, as the test set is held by the authors for the LIP challenge. The extended PASCAL-Person-Part contains multiple humans per image with diverse poses and occlusion patterns. Each person is annotated with 6 semantic parts. In total, there are 3,533 images, among which 1,716 are used for training and another 1,817 are used for testing. For LIP, following its standard protocol (Ruan et al. 2019), we report pixel accuracy, mean accuracy and mean IoU (mIoU). For Extended PASCAL-Person-Part, following conventions (Nie, Feng, and Yan 2018), the performance is evaluated in terms of mIoU.

Implementation Details

We implemented the proposed framework in PyTorch. All models are trained on two NVIDIA 2080Ti GPUs (8 im-

Methods	Head	Torso	Upper Arm	Lower Arm	Upper Leg	Lower Leg	Background	mIoU
PCNet (Zhu et al. 2018)	86.81	69.06	55.35	55.27	50.21	48.54	96.07	65.90
WSHP (Fang et al. 2018)	87.15	72.28	57.07	56.21	52.43	50.36	97.72	67.60
HRNet (Wang et al. 2020a)	87.58	72.03	61.91	61.82	53.81	52.86	96.13	69.45
CNIF (Wang et al. 2019a)	88.02	72.91	64.31	63.52	55.61	54.96	96.02	70.76
DTCF (Liu et al. 2020b)	88.32	73.54	64.19	63.91	55.01	54.34	96.25	70.80
DPC (Chen et al. 2018)	88.81	75.54	63.85	63.73	57.24	54.55	96.66	71.34
CDCL (Lin et al. 2019)	86.39	74.70	68.32	65.98	59.86	58.70	95.79	72.82
LSNT (Ji et al. 2020)	89.01	74.63	62.90	64.70	57.53	54.62	**97.74**	71.59
HHP (Wang et al. 2020b)	**89.73**	75.22	66.87	66.21	58.69	58.17	96.94	73.12
Our HIPN	89.18	76.07	66.66	66.86	60.15	59.03	96.64	73.51
Our HIPN(+HL)	89.64	**76.23**	**68.38**	**68.64**	**61.42**	**63.26**	96.76	**74.91**

Table 4: Comparison with state-of-the-art methods on the extended Pascal-Person-part test set.

Method	Pixel acc.	Mean acc.	mIoU
DeepLab (Chen et al. 2017)	82.66	51.64	41.64
MuLA (Nie et al. 2018)	88.50	60.50	49.30
JPPNet (Liang et al. 2019)	86.39	62.32	51.37
CE2P (Ruan et al. 2019)	87.37	63.20	53.10
BraidNet (Liu et al. 2019)	87.60	66.09	54.42
HRNet (Wang et al. 2020a)	88.21	67.43	55.90
LSNT (Ji et al. 2020)	88.10	70.41	54.86
CNIF (Wang et al. 2019a)	88.03	68.80	57.74
DTCF (Liu et al. 2020b)	88.61	68.89	57.82
GRN (Li et al. 2020)	88.33	66.53	56.34
HHP (Wang et al. 2020b)	89.05	70.58	59.25
Our HIPN	89.14	71.09	59.61
Our HIPN(+HL)	**89.54**	**71.85**	**60.81**

Table 5: Comparison on the LIP val set.

ages per GPU). We adopt the HRNet (Wang et al. 2020a) that is pre-trained on the ImageNet dataset as the backbone network. After an ASPP module (Chen et al. 2017), the dimension of each image representation is reduced to $W \times H \times C$, where $W = 119$, $H = 119$, and $C = 64$. For our joint loss function, we empirically find that comparable results can be achieved by setting $\alpha \in [0.5, 0.7]$. And in our method, we set the weight as $\alpha = 0.5$. Following the standard protocol (Nie, Feng, and Yan 2018; Wang et al. 2019a), we adopt data argumentation at both the first training and retraining, such as randomly augmenting samples with a scaling factor in $[0.5, 2.0]$, crop size of 473×473, and horizontal flip. We use the SGD optimizer with a base learning rate of 0.007, the momentum of 0.9 and the weight decay of 0.0005. The poly learning rate policy with power of 0.9 is used for decreasing the learning rate. The training process is terminated within 150 epochs.

Ablation Study

Evaluation on key settings of HIPN. We study the influence of two important settings in our HIPN on the LIP dataset. (1) *Number of stages K*. As presented in Table 1, we find that the performance improves as we increase the number of stages from 1 to 4, but the gain starts to saturate at 4 stages. Considering efficiency, we use 3 stages in our experiments unless otherwise specified. (2) *Information passing directions*. As presented in Table 1, compared to only allowing for top-down information passing, we obtain consistent improvements by using additional bottom-up passing across the second and subsequent stages. This gain is mainly due to that richer information of human body structure is exploited and more powerful features are obtained.

Effectiveness of our noisy label identification method. We compare our method with a recently proposed method (Kim et al. 2019) for noisy label identification, and explore the impact of noise identification on the human parsing performance. Experiments are implemented on the Extended PASCAL-Person-Part dataset, and we use the training set as labeled samples, and test set as unlabeled samples. Within our noise-tolerant hybrid learning framework, we adopt different methods to distinguish noisy labels from clean ones. An evaluation of different noise identification methods in terms of classification accuracy and human parsing mIoU is shown in Table 2. We have the following findings: (1) *Our proposed hybrid learning strategy is effective and robust.* Compared to the baseline, all methods using our hybrid learning obtain better results w.r.t. mIoU. (2) *Our noisy label identification method using two criterion gives the best results.* Compared to previous method (Kim et al. 2019), our method that take into account the whole probability distribution better identify noisy labels, with a gain of 3.22%. By further checking the inconsistency, the classification further improves by $\sim 1\%$. (3) *Better noisy label identification leads to higher human parsing performance.* As the clean/noisy classification accuracy increases from 60.59% to 64.91%, the performance of human parsing consistently improves from 74.27% to 74.91% w.r.t. mIoU.

Impact of our noise-tolerant hybrid learning. We con-

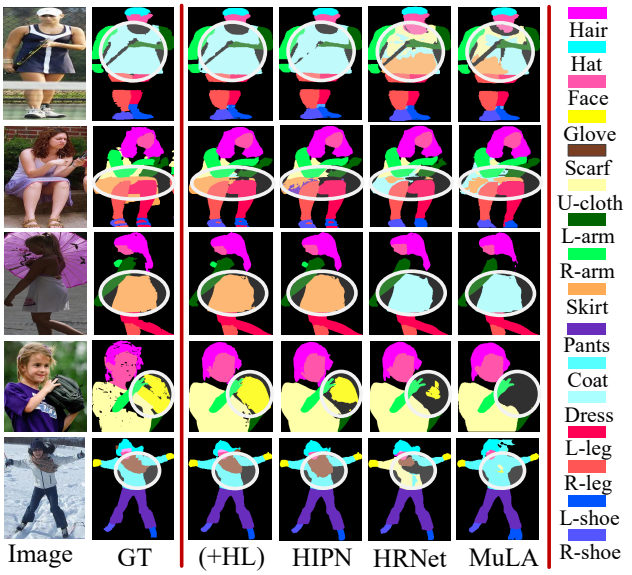

Figure 5: Visualization of human parsing results produced by state-of-the-art methods and ours on the LIP validation set. Our HIPN and HIPN(+HL) output more accurate predictions compared to previous state-of-the-art competitors, i.e. HRNet (Wang et al. 2020a) and MuLA (Nie, Feng, and Yan 2018). We highlight the most notable differences for each sample by a white circle.

.

duct the semi-supervised learning experiments on the LIP dataset under different settings: {1/8, 1/2} labeled samples and other unlabeled ones on LIP training set are used for training. The experimental results evaluated on the LIP validation set are shown in Table 3, where we study the impact of our proposed noise-tolerant hybrid learning. We have the following observations: (1) *When the provided labeled data is limited, it is helpful to employ extra unlabeled data via semi-supervised learning.* Either PL or HL brings extra gains via using some amount of unlabeled data for training; the improvements range from 3% to 5%. (2) *Our proposed noise-tolerant hybrid strategy consistently outperforms the standard positive learning under various settings.* When we use 1/8 or 1/2 labeled data, our hybrid learning brings an improvement of 3.61%/ 5.37% w.r.t. mIoU to the baseline, larger than that from positive learning. (3) *Our semi-supervised method using only half annotations achieves comparable performance to the fully supervised method.* When we only use half labeled data for training, our method is able to achieve comparable results to the fully supervised method using the whole set of labeled training data, 54.69% vs. 55.90%. These results indicate that our proposed semi-supervised method helps to reduce annotation burdens.

Comparisons with the State-of-the-Arts

We compare our proposed method with other state-of-the-art methods on two challenging datasets, i.e. LIP and extended PASCAL-Person-Part. We conduct the experiments with two training settings. Specifically, we denote our method as "HIPN" and "HIPN(+HL)", respectively. (1) "HIPN": we use all the training images with labels in a fully-supervised manner for a fair comparison with other methods. (2) "HIPN(+HL)": we use all the training images with labels and the test images without labels in a semi-supervised manner.

In Table. 4, we compare our method with 9 state-of-the-art methods on the extended PASCAL-Person-Part test set. We have the following observations. (1) Our HIPN outperforms previous methods by a large margin w.r.t mIoU, establishing a new state-of-the-art performance. (2) Our HIPN(+HL) further improves mIoU to 74.95%, and ranks 1st on five semantic categories, demonstrating the effectiveness of our proposed hybrid learning method. In Table. 5, we compare our method with 11 state-of-the-art methods on the LIP validation set. We have the following observations. (1) Our HIPN outperforms well established baseline approach (i.e. HRNet (Wang et al. 2020a)) by 0.93% w.r.t. pixel accuracy, 3.66% w.r.t mean accuracy, and 3.71% w.r.t. mIoU. (2) Our HIPN(+HL) further improves the performance when using additional unlabeled images, and outperforms state-of-the-art semi-supervised approach (i.e. GRN (Li et al. 2020)) by 1.21% w.r.t. pixel accuracy, 5.32% w.r.t mean accuracy, and 4.47% w.r.t. mIoU. These results demonstrate the effectiveness of our semi-supervised human parsing method.

In Fig. 5, we show some qualitative results on the LIP validation set. Our HIPN and HIPN(+HL) yield more precise predictions, while other methods (Wang et al. 2020a; Nie, Feng, and Yan 2018) sometimes mix up confusing components, resulting in wrong boundaries, e.g. the dress in the 1st row, and the skirt in the 2nd row. Our model also predict more accurate categories, for example, in the 3rd row, our method correctly labels the skirt, while other methods (Wang et al. 2020a; Nie, Feng, and Yan 2018) mix it up with dress. For the last two rows, our model can successfully predict the complete structure for each component, while other methods may lose some components in detail, such as the glove and scarf.

Conclusion

To reduce annotation costs for human parsing, we propose a semi-supervised learning framework, which consists of high-quality pseudo label generation and noise-tolerant hybrid learning. Specifically, we first propose a HIPN to generate high-quality pseudo label for unlabeled images. Then, considering both the certainty and consistency of multi-level predictions of HIPN, we distinguish noisy pseudo labels from clean ones and develop a noise-tolerant hybrid learning strategy to retrain the HIPN, which better handles noisy pseudo labels. The experimental results demonstrate the effectiveness of our method. In the future, we would like to extend our method to multiple-person human parsing and video human parsing tasks.

Acknowledgments

This work was supported by the Funds for International Cooperation and Exchange of the National Natural Science Foundation of China (Grant No. 61861136011), the Natural Science Foundation of Jiangsu Province, China (Grant No.

BK20181299), the National Science Fund of China (Grant Nos. 61702262, U1713208), the Fundamental Research Funds for the Central Universities (Grant No.30920032201), the National Key Research and Development Program of China under Grant 2017YFC0820601, and "111" Program B13022.

References

Chen, L.; Collins, M.; Zhu, Y.; Papandreou, G.; Zoph, B.; Schroff, F.; Adam, H.; and Shlens, J. 2018. Searching for efficient multi-scale architectures for dense image prediction. In *NeurIPS*, 8699–8710.

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. 2017. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(4): 834–848.

Fang, H.; Lu, G.; Fang, X.; Xei, J.; Tai, Y.; and Lu, C. 2018. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. In *CVPR*, 70–78.

Gong, K.; Gao, Y.; Liang, X.; Shen, X.; Wang, M.; and Lin, L. 2019. Graphonomy: universal human parsing via graph transfer learning. In *CVPR*, 7450–7459.

Gong, K.; Liang, X.; Li, Y.; Chen, Y.; Yang, M.; and Lin, L. 2018. Instance-level human parsing via part grouping network. In *ECCV*, 770–785.

Gong, K.; Liang, X.; Zhang, D.; Shen, X.; and Lin, L. 2017. Look into person: self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 932–940.

Guo, J.; Yuan, Y.; Huang, L.; Zhang, C.; Yao, J.; and Han, K. 2019. Beyond human parts: dual part-aligned representations for person re-identification. In *ICCV*, 3642–3651.

He, H.; Zhang, J.; Zhang, Q.; and Tao, D. 2020. Grapy-ML: graph pyramid mutual learning for cross-dataset human parsing. In *AAAI*, 10949–10956.

Huang, J.; Qu, L.; Jia, R.; and Zhao, B. 2019. O2U-Net: a simple noisy label detection approach for deep neural networks. In *ICCV*, 3326–3334.

Ji, R.; Du, D.; Zhang, L.; Wen, L.; Wu, Y.; Zhao, C.; Huang, F.; and Lyu, S. 2020. Learning semantic neural tree for human parsing. In *ECCV*, 205–221.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.; and Li, F. 2018. Mentornet: learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*.

Kim, Y.; Yim, J.; Yun, J.; and Kim, J. 2019. NLNL: negative learning for noisy labels. In *ICCV*, 101–110.

Li, J.; Zhao, J.; Chen, Y.; Roy, S.; Yan, S.; Feng, J.; and Sim, T. 2018. Multi-human parsing machines. In *ACM Multimedia*, 45–53.

Li, J.; Zhao, J.; Wei, Y.; Lang, C.; Li, Y.; T., S.; Yan, S.; and Feng, J. 2017a. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206* .

Li, P.; Xu, Y.; Wei, Y.; and Yang, Y. 2019a. Self-correction for human parsing. *arXiv preprint arXiv:1910.09777* .

Li, Q.; Zhao, X.; He, R.; and Huang, K. 2019b. Pedestrian attribute recognition by joint visual-semantic reasoning and knowledge distillation. In *IJCAI*, 833–839.

Li, T.; Liang, Z.; Zhao, S.; Gong, J.; and Shen, J. 2020. Self-learning with rectification strategy for human parsing. In *CVPR*, 9263–9272.

Li, Y.; Yang, J.; Song, Y.; Cao, L.; Luo, J.; and Li, L. 2017b. Learning from noisy labels with distillation. In *ICCV*, 1910–1918.

Liang, X.; Gong, K.; Shen, X.; and Lin, L. 2019. Look into person: joint body parsing & pose estimation network and a new benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(4): 871–885.

Lin, K.; Wang, L.; Luo, K.; Chen, Y.; Liu, Z.; and M, S. 2019. Cross-domain complementary learning using pose for multi-person part segmentation. *arXiv preprint arXiv:1907.05193* .

Liu, J.; Hou, Q.; Cheng, M.; Wang, C.; and Feng, J. 2020a. Improving convolutional networks with self-calibrated convolutions. In *CVPR*, 10093–10102.

Liu, X.; Zhang, M.; Liu, W.; Song, J.; and Mei, T. 2019. BraidNet: braiding semantics and details for accurate human parsing. In *ACM MM*, 338–346.

Liu, Y.; Zhang, S.; Xu, J.; Yang, J.; and Tai, Y.-W. 2021. An Accurate and Lightweight Method for Human Body Image Super-Resolution. *IEEE Transactions on Image Processing* 30: 2888–2897.

Liu, Y.; Zhao, L.; Zhang, S.; and Yang, J. 2020b. Hybrid Resolution Network Using Edge Guided Region Mutual Information Loss for Human Parsing. In *ACM Multimedia*, 1670–1678.

Nie, X.; Feng, J.; and Yan, S. 2018. Mutual learning to adapt for joint human parsing and pose estimation. In *ECCV*, 519–534.

Ruan, T.; Liu, T.; Huang, Z.; Wei, Y.; Wei, S.; and Zhao, Y. 2019. Devil in the details: towards accurate single and multiple human parsing. In *AAAI*, 4814–4821.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint optimization framework for learning with noisy labels. In *CVPR*, 5552–5560.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; and Mu, Y. 2020a. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* .

Wang, W.; Zhang, Z.; Qi, S.; Shen, J.; Pang, Y.; and Shao, L. 2019a. Learning compositional neural information fusion for human parsing. In *ICCV*, 5703–5713.

Wang, W.; Zhu, H.; Dai, J.; Pang, Y.; Shen, J.; and Shao, L. 2020b. Hierarchical human parsing with typed part-relation reasoning. In *CVPR*, 8929–8939.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019b. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 322–330.

Xia, F.; Wang, P.; Chen, X.; and Yuille, A. 2017. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, 6769–6778.

Xia, F.; Zhu, J.; Wang, P.; and Yuille, A. L. 2016. Pose-guided human parsing by an AND/OR graph using pose-context features. In *AAAI*, 3632–3640.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *CVPR*, 2691–2699.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 8778–8788.

Zhang, Z.; Zhang, H.; Arik, S. O.; Lee, H.; and Pfister, T. 2020. Distilling effective supervision from severe label noise. In *CVPR*, 9294–9303.

Zhao, J.; Li, J.; Liu, H.; Yan, S.; and Feng, J. 2020. Fine-grained multi-human parsing. *Int. J. Comput. Vis.* 128(8): 2185–2203.

Zhao, J.; Li, J.; Nie, X.; Zhao, F.; Chen, Y.; Wang, Z.; Feng, J.; and Yan, S. 2017. Self-supervised neural aggregation networks for human parsing. In *CVPR Workshops*, 7–15.

Zhao, L.; Li, J.; Cheng, Y.; Zhou, L.; Sim, T.; Yan, S.; and Feng, J. 2018. Understanding humans in crowded scenes: deep nested adversarial learning and a new benchmark for multi-human parsing. In *ACM MM*, 792–800.

Zhao, Y.; Li, J.; Zhang, Y.; and Tian, Y. 2019. Multi-class part parsing with joint boundary-semantic awareness. In *ICCV*, 9177–9186.

Zhu, B.; Chen, Y.; Tang, M.; and Wang, J. 2018. Progressive cognitive human parsing. In *AAAI*, 7607–7614.