This ECCV 2018 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ECCV 2018
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/eccv

Mutual Learning to Adapt for Joint Human Parsing and Pose Estimation

Xuecheng Nie[1][0000−0003−2433−5983], Jiashi Feng[1], and Shuicheng Yan[1,2]

[1] ECE Department, National University of Singapore, Singapore
niexuecheng@u.nus.edu, elefjia@nus.edu.sg
[2] Qihoo 360 AI Institute, Beijing, China
yanshuicheng@360.cn

Abstract. This paper presents a novel *Mutual Learning to Adapt* model (MuLA) for joint human parsing and pose estimation. It effectively exploits mutual benefits from both tasks and simultaneously boosts their performance. Different from existing post-processing or multi-task learning based methods, MuLA predicts dynamic task-specific model parameters via recurrently leveraging guidance information from its parallel tasks. Thus MuLA can fast adapt parsing and pose models to provide more powerful representations by incorporating information from their counterparts, giving more robust and accurate results. MuLA is implemented with convolutional neural networks and end-to-end trainable. Comprehensive experiments on benchmarks LIP and extended PASCAL-Person-Part demonstrate the effectiveness of the proposed MuLA model with superior performance to well established baselines.

Keywords: Human Pose Estimation · Human Parsing · Mutual Learning

1 Introduction

Human parsing and pose estimation are two crucial yet challenging tasks for human body configuration analysis in 2D monocular images, which aim at segmenting human body into semantic parts and allocating body joints for human instances respectively. Recently, they have drawn increasing attention due to their wide applications, *e.g.*, human behavior analysis [22,9], person-identification [29,20] and video surveillance [14,30]. Although analyzing human body from different perspectives, these two tasks are highly correlated and could provide beneficial clues for each other. Human pose can offer structure information for body part segmentation and labeling, and on the other hand human parsing can facilitate localizing body joints in difficult scenarios. Fig. 1 gives examples where considering such mutual guidance information between the two tasks can correct labeling and localization errors favorably, as highlighted in Fig. 1 (b), and improve parsing and pose estimation results, as shown in Fig. 1 (c).

Motivated by the above observation, some efforts [26,12,8,25,24,10] have been made to extract and use such guidance information to improve performance of

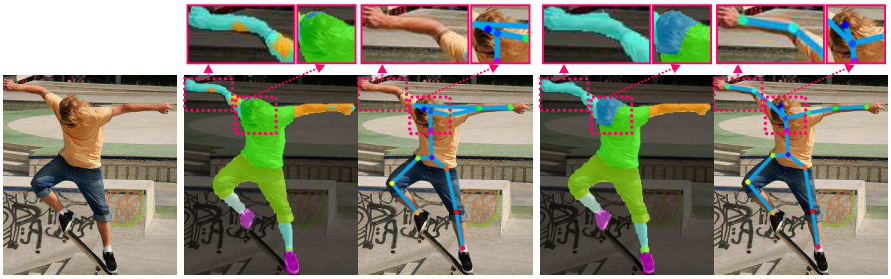

Fig. 1. Illustration of our motivation for joint human parsing and pose estimation. (a) Input image. (b) Results from independent models. (c) Results of the proposed MuLA model. MuLA can leverage mutual guidance information between human parsing and pose estimation to improve performance of both tasks, as shown with highlighted body parts and joints. Best viewed in color

the two tasks mutually. However, existing methods usually train the task-specific models separately and leverage the guidance information for post-processing, suffering several drawbacks. First, they heavily rely on hand-crafted features extracted from outputs of one task to assist the other, in an *ad hoc* manner. Second, they only utilize guidance information in inference procedure and fail to enhance model capacity during training. Third, they are one-stop solutions and too rigid to fully utilize enhanced models and iteratively improve the results. Last but not least, the models are not end-to-end learnable.

Targeting at these drawbacks, we propose a novel *Mutual Learning to Adapt* (MuLA) model to sufficiently and systematically exploit mutual guidance information between human parsing and pose estimation. In particular, our MuLA has a carefully designed interweaving architecture that enables effective between-task cooperation and mutual learning. Moreover, instead of simply fusing learned features from two tasks as in existing works, MuLA introduces a *learning to adapt* mechanism where the guidance information from one task can be effectively transferred to modify model parameters for the other parallel task, leading to augmented representation and better performance. In addition, MuLA is capable of recurrently performing model adaption by transforming estimation results to the representation space and thus can continuously refine semantic part labels and body joint locations based on enhanced models in the previous iteration.

Specifically, the MuLA model includes a representation encoding module, a mutual adaptation module and a classification module. The representation encoding module encodes input images into preliminary representations for human parsing and pose estimation individually, and meanwhile provides guidance for model adaptation. With such guidance information, the mutual adaptation module learns to dynamically predict model parameters for augmenting representations by incorporating useful prior learned from the other task, enabling effective between-task interaction and cooperation in model training. Introduc-

ing such a mutual adaptation module improves the learning process of one task towards benefiting the other, providing easily transferable information between tasks. In addition, these dynamic parameters are efficiently learned in a one-shot manner according to different inputs, leading to fast and robust model adaptation. MuLA fuses mutually-tailored representations with the preliminary ones in a residual manner to produce augmented representations for making final prediction, through the classification modules. MuLA also allows for iterative model adaption and improvement by transforming estimation results to the representation space, which serve as enhanced input for the next stage. The proposed MuLA is implemented with deep Convolutional Neural Networks and is end-to-end learnable.

We evaluate the proposed MuLA model on Look into Person (LIP) [10] and extended PASCAL-Person-Part [24] benchmarks. The experiment results well demonstrate its superiority over existing methods in exploiting mutual guidance information for joint human parsing and pose estimation. Our contributions are summarized in four aspects. First, we propose a novel end-to-end learnable model for jointly learning human parsing and pose estimation. Second, we propose a novel mutual adaptation module for dynamic interaction and cooperation between two tasks. Third, the proposed model is capable of iteratively exploiting mutual guidance information to consistently improve performance of two tasks. Fourth, we achieve new state-of-the-art on LIP dataset, and outperform the previous best model for joint human parsing and pose estimation on extended PASCAL-Person-Part dataset.

2 Related Work

Due to their close correlations, recent works have exploited human parsing (human pose estimation) to assist human pose estimation (human parsing) or leveraged their mutual benefits to jointly improve the performance for both tasks.

In [12], ladicky *et al.* proposed to utilize body parts as additional constraint for the pose estimation model. Given locations of all joints, they introduced a body part mask component to predict labels of pixels belonging to each body part, which can be optimized with the overall model together. In [25], Xia *et al.* proposed to exploit pose estimation results to guide human parsing by leveraging joint locations to extract segment proposals for semantic parts, which are selected and assembled using an And-Or graph to output a parse of the person. In [10], Gong *et al.* proposed to improve human parsing with pose estimation in a self-supervised structure-sensitive manner through weighting segmentation loss with joint structure loss. Similar to [10], Zhao *et al.* [28] proposed to improve human parsing via regarding human pose structure from a global perspective for feature aggregation considering the importance of different positions. Yamaguchi *et al.* [26] proposed to optimize human parsing and pose estimation and improve the performance of two tasks in an alternative manner: utilizing pose estimation results to generate body part locations for human parsing and then exploiting human parsing results to update appearance features in the pose estimation model

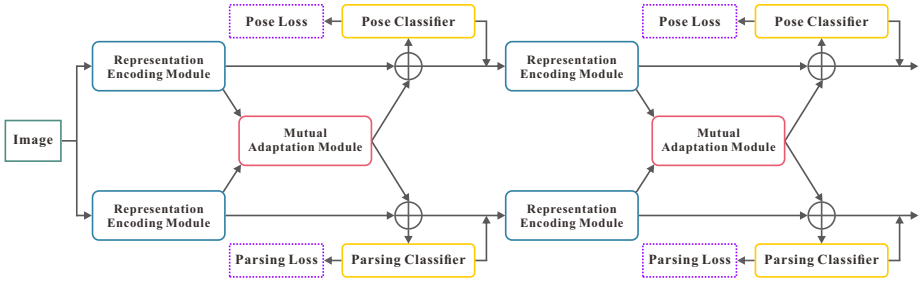

Fig. 2. Illustration of overall architecture of the proposed Mutual Learning to Adapt model (MuLA) for joint human parsing and pose estimation. Given an input image, MuLA utilizes the novel *mutual adaptation* module to build dynamic interaction and cooperation between parsing and pose estimation models in an iterative way for fully exploiting their mutual benefits to simultaneously improve their performance

for refining joint locations. Dong *et al.* [8] proposed a Hybrid Parsing Model for unified human parsing and pose estimation under the And-Or graph framework. They utilized body joints to assist human parsing via constructing the mixture of joint-group templates for body part representation, and exploited body parts to improve human pose estimation through forming parselets to constrain the position and co-occurrence of body joints. In [24], Xia *et al.* proposed to utilize deep learning models for joint human parsing and pose estimation. They utilized parsing results for hand-crafted features to assist pose estimation by considering relationships of body joints and parts, and then exploited the generated pose estimation results to construct joint label maps and skeleton maps for refining human parsing. With the powerful deep learning models, they achieved superior performance over previous methods.

Despite previous success, existing methods suffer from limitations of hand-crafted features relying on estimation results for exploiting guidance information to improve the counterpart models. In contrast, the proposed Mutal Learning to Adapt model can mutually learn to fast adapt the model of one task conditioned on representations of the other for specific inputs. In addition, MuLA utilizes the guidance information in both training and inference phases for joint human parsing and pose estimation. Moreover, it is end-to-end learnable via implementation with CNNs.

3 The Proposed Approach

3.1 Formulation

For an RGB image $I \in \mathbb{R}^{H \times W \times 3}$ with height H and width W, we use $S=\{s_i\}_{i=1}^{H \times W}$ to denote the human parsing result of I, where $s_i \in \{0, \ldots, P\}$ is the semantic part label of the ith pixel and P is the total number of semantic part categories. Specially, 0 represents the background category. We use $J=\{(x_i, y_i)\}_{i=1}^{N}$

to denote body joint locations of the human instance in I, where (x_i, y_i) represents the spatial coordinates of the ith body joint and N is the number of joint categories. Our goal is to design a unified model for simultaneously predicting human parsing S and pose J via fully exploiting their mutual benefits to boost performance for both tasks.

Existing methods for joint human parsing and pose estimation usually extract hand-crafted features from the output of one task to assist the other task at post-processing. They can neither extract powerful features nor strengthen the models. Targeting at such limitations, we propose a *Mutual Learning to Adapt* (MuLA) model to substantially exploit mutual benefits from human parsing and pose estimation towards effectively improving performance of the counterpart models, through learning to adapt model parameters. In the following, we use $g_{[\psi,\psi_*]}(\cdot)$ and $h_{[\phi,\phi_*]}(\cdot)$ to denote the parsing and pose models respectively, with parameters specified in the subscripts. Specifically, ψ_* and ϕ_* denote parameters that are adaptable to the other task. Then, our proposed MuLA is formulated as following recurrent learning process:

$$\begin{aligned} S^{(t)} &= g_{[\psi^{(t)},\psi_*^{(t)}]}(F_S^{(t)}), \text{ where } \psi_*^{(t)} = h'(F_J^{(t)}, \hat{J}), \\ J^{(t)} &= h_{[\phi^{(t)},\phi_*^{(t)}]}(F_J^{(t)}), \text{ where } \phi_*^{(t)} = g'(F_S^{(t)}, \hat{S}), \end{aligned} \tag{1}$$

where t is the iteration index, $\hat{S}$ and $\hat{J}$ are parsing and pose annotations for the input image I, and $F_S^{(t)}$ and $F_J^{(t)}$ denote the extracted features for parsing and pose prediction respectively. Note, at the beginning, $F_S^{(1)}{=}F_J^{(1)}{=}I$.

The above formulation in Eqn. (1) highlights the most distinguishing feature of MuLA from existing methods: MuLA explicitly adapts some model parameters of one task (*e.g.* parsing model parameter ψ_*) to the guidance information of the other task (*e.g.* pose estimation) via adapting functions $h'(\cdot,\cdot)$ and $g'(\cdot,\cdot)$. In this way, the adaptive parameters $\psi_*^{(t)}$ and $\phi_*^{(t)}$ encode useful information from the parallel tasks. With these parameters, the MuLA model can learn complementary representations and boost performance for both human parsing and pose estimation tasks, by more flexibly and effectively exploiting interaction and cooperation between them. In addition, MuLA bases $\psi_*^{(t)}$ and $\phi_*^{(t)}$ on the input images. Different inputs would modify the model parameters dynamically, making the model robust to various testing senarios. Moreover, MuLA has the ability to iteratively exploit mutual guidance information between two tasks via the recurrent learning process and thus continuously improves both models.

The overall architecture of MuLA is shown in Fig. 2. Concretely, MuLA presents an interweaving architecture and consists of three components: a *representation encoding* module, a *mutual adaptation* module and a *classification* module. The representation encoding module consists of two encoders $E^S_{\psi_e^{(t)}}(\cdot)$ and $E^J_{\phi_e^{(t)}}(\cdot)$ for transforming inputs $F_S^{(t)}$ and $F_J^{(t)}$ into high-level preliminary representations for human parsing and pose estimation.

The mutual adaptation module targets at adapting parameters $\psi_*^{(t)}$ and $\phi_*^{(t)}$ to augment preliminary representations from $E^S_{\psi_e^{(t)}}(\cdot)$ and $E^J_{\phi_e^{(t)}}(\cdot)$ by leveraging

auxiliary guidance information from the parallel tasks. Inspired by the "Learning to Learn" framework [2], for achieving fast and effective adaptation, within functions $g'(\cdot,\cdot)$ and $h'(\cdot,\cdot)$, we design two learnable adapters $A_{\psi_a^{(t)}}(\cdot)$ and $A_{\phi_a^{(t)}}(\cdot)$ to learn to predict these adaptive parameters. For reliable and robust parameter prediction, we take the highest-level representation from $E^S_{\psi_e^{(t)}}(\cdot)$ and $E^J_{\phi_e^{(t)}}(\cdot)$ as mutual guidance information. Namely, $A_{\psi_a^{(t)}}(\cdot)$ and $A_{\phi_a^{(t)}}(\cdot)$ take $E^S_{\psi_e^{(t)}}(F_S^{(t)})$ and $E^J_{\phi_e^{(t)}}(F_J^{(t)})$ as inputs and output $\phi_*^{(t)}$ and $\psi_*^{(t)}$. Formally,

$$\begin{aligned}\psi_*^{(t)} &= h'(F_J^{(t)}, \hat{J}) := A_{\phi_a^{(t)}}\left(E^J_{\phi_e^{(t)}}(F_J^{(t)})\right),\\ \phi_*^{(t)} &= g'(F_S^{(t)}, \hat{S}) := A_{\psi_a^{(t)}}\left(E^S_{\psi_e^{(t)}}(F_S^{(t)})\right).\end{aligned} \tag{2}$$

Here $\psi_*^{(t)}$ and $\phi_*^{(t)}$ can tailor preliminary representations extracted by $\psi_e^{(t)}$ and $\phi_e^{(t)}$ for better human parsing and pose estimation via leveraging their mutual guidance information. We utilize the tailored representations extracted by $\psi_e^{(t)}$ and $\phi_e^{(t)}$ together with $\psi_*^{(t)}$ and $\phi_*^{(t)}$ for making final predictions, and use $E^S_{[\psi_e^{(t)},\psi_*^{(t)}]}(\cdot)$ and $E^J_{[\phi_e^{(t)},\phi_*^{(t)}]}(\cdot)$ to denote the derived adaptive encoders in MuLA. The mutual adaptation module allows for dynamic interaction and cooperation between two tasks within MuLA for fully exploiting their mutual benefits.

MuLA uses two classifiers $C^S_{\psi_w^{(t)}}(\cdot)$ and $C^J_{\phi_w^{(t)}}(\cdot)$ following the mutual adaptation module for predicting human parsing $S^{(t)}$ and pose $J^{(t)}$. Specifically, $[\psi_e^{(t)}, \psi_w^{(t)}]$ and $[\phi_e^{(t)}, \phi_w^{(t)}]$ together instantiate parameters $\psi^{(t)}$ and $\phi^{(t)}$ in Eqn. (1), respectively. For iteratively exploiting mutual guidance information, we design two mapping modules $M^S_{\psi_m^{(t)}}(\cdot,\cdot)$ and $M^J_{\phi_m^{(t)}}(\cdot,\cdot)$ to map representations from $E^S_{[\psi_e^{(t)},\psi_*^{(t)}]}(\cdot)$ and $E^J_{[\phi_e^{(t)},\phi_*^{(t)}]}(\cdot)$ together with prediction results $S^{(t)}$ and $J^{(t)}$ into inputs $F_S^{(t+1)}$ and $F_J^{(t+1)}$ for the next stage. Namely,

$$F_S^{(t+1)} = M^S_{\psi_m^{(t)}}\left(E^S_{[\psi_e^{(t)},\psi_*^{(t)}]}(F_S^{(t)}), S^{(t)}\right) \text{ and } F_J^{(t+1)} = M^J_{\phi_m^{(t)}}\left(E^J_{[\phi_e^{(t)},\phi_*^{(t)}]}(F_J^{(t)}), J^{(t)}\right). \tag{3}$$

By the definition in Eqn. (3), $F_S^{(t)}$ and $F_J^{(t)}$ provide preliminary representations at the start of the next stage and avoid learning from scratch at each stage. In addition, $S^{(t)}$ and $J^{(t)}$ offer additional guidance information for generating better prediction results and alleviate learning difficulties in subsequent stages [23,15].

To train MuLA, we add groundtruth supervision $\hat{S}$ and $\hat{J}$ for human parsing and pose estimation at each stage, and define the following loss function:

$$\mathcal{L} = \sum_{t=1}^{T}\left(\mathcal{L}^S\left(C^S_{\psi_w^{(t)}}\left(E^S_{[\psi_e^{(t)},\psi_*^{(t)}]}(F_S^{(t)})\right), \hat{S}\right) + \beta\mathcal{L}^J\left(C^J_{\phi_w^{(t)}}\left(E^J_{[\phi_e^{(t)},\phi_*^{(t)}]}(F_J^{(t)})\right), \hat{J}\right)\right) \tag{4}$$

where T denotes the total number of iterations in MuLA, $\mathcal{L}^S(\cdot,\cdot)$ and $\mathcal{L}^J(\cdot,\cdot)$ represent loss functions for human parsing and pose estimation, respectively, and β is a weight coefficient for balancing $\mathcal{L}^S(\cdot,\cdot)$ and $\mathcal{L}^J(\cdot,\cdot)$. In next subsection, we will provide details on implementation of MuLA.

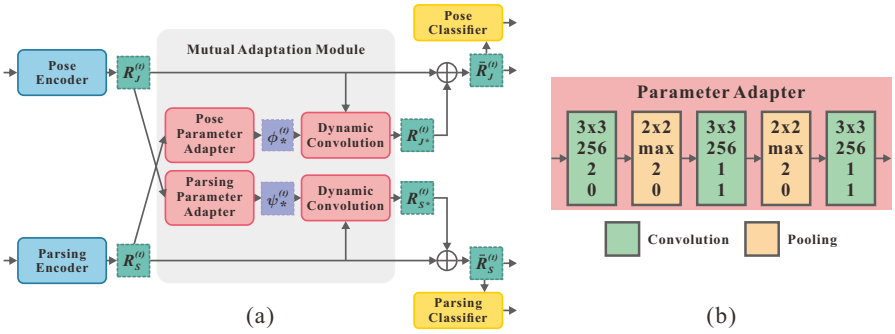

Fig. 3. (a) The CNN implementation of MuLA for one stage. Given inputs $F_S^{(t)}$ and $F_J^{(t)}$ at stage t, the parsing and pose encoders generate preliminary representations $R_S^{(t)}$ and $R_J^{(t)}$. Then, the parameter adapters predict dynamic parameters $\psi_*^{(t)}$ and $\phi_*^{(t)}$ for learning complementary representations $R_{S*}^{(t)}$ and $R_{J*}^{(t)}$ via dynamic convolutions, which are exploited to tailor preliminary representations via addition in a residual manner for producing refined representations $\bar{R}_S^{(t)}$ and $\bar{R}_J^{(t)}$. Finally, MuLA feeds $\bar{R}_S^{(t)}$ and $\bar{R}_J^{(t)}$ to classifiers for parsing and pose estimation, respectively. (b) The network architecture of parameter adapter, consisting of three convolution and two pooling layers. For each layer, the kernel size, the number of channel/pooling types, stride and padding size are specified from top to bottom

3.2 Implementation

We implement MuLA with deep Convolutional Neural Networks (CNNs), and show architecture details in Fig. 3 (a).

Representation Encoding Module This module is composed of two encoders $E^S_{\psi_e^{(t)}}(\cdot)$ and $E^J_{\phi_e^{(t)}}(\cdot)$, targeting at encoding inputs $F_S^{(t)}$ and $F_J^{(t)}$ into discriminative representations $R_S^{(t)}$ and $R_J^{(t)}$ for estimating parsing and pose results, as well as for predicting adaptive parameters. We implement $E^S_{\psi_e^{(t)}}(\cdot)$ and $E^J_{\phi_e^{(t)}}(\cdot)$ with two different state-of-the-art architectures: the VGG network [19] and Hourglass network [15]. VGG network is a general architecture widely applied in various vision tasks [18,5]. We utilize its fully convolutional version with 16 layers, denoted as VGG16-FCN, for both tasks. In addition, we modify VGG16-FCN to reduce the total stride from 32 to 8 via removing the last two max-pooing layers, aiming to enlarge feature maps for improving part labeling and joint localization accuracy. The Hourglass network has a U-shape architecture which is initially designed for human pose estimation. We extend it to parsing by making the output layer aim for semantic part labeling instead of joint confidence regression. Other configurations of Hourglass network exactly follow [15]. Note that paring and pose encoders need not have the same architecture as they are independent from each other.

Mutual Adaptation Module This module includes two adapters $A_{\phi_a^{(t)}}(\cdot)$ and $A_{\psi_a^{(t)}}(\cdot)$ to predict adaptive parameters $\psi_*^{(t)}$ and $\phi_*^{(t)}$ which are used to tailor preliminary representations $R_S^{(t)}$ and $R_J^{(t)}$. In particular, we implement $A_{\psi_a^{(t)}}(\cdot)$ and $A_{\phi_a^{(t)}}(\cdot)$ with the same small CNN for predicting convolution kernels of counterpart models, as shown in Fig. 3 (b). The adapter networks take $R_S^{(t)}$ and $R_J^{(t)}$ as inputs and output tensors $\phi_*^{(t)} \in \mathbb{R}^{h\times h\times c}$ and $\psi_*^{(t)} \in \mathbb{R}^{h\times h\times c}$ as convolution kernels, where h is the kernel size and $c{=}c_i{\times}c_o$ is the number of kernels with input and output channel number c_i and c_o, respectively.

However, it is not feasible to directly predict all the convolution kernels due to their large scale. To reduce the number of kernels to predict by adapters $A_{\psi_a^{(t)}}(\cdot)$ and $A_{\phi_a^{(t)}}(\cdot)$, we follow [2] to use a way analogous to SVD for decomposing parameters $\psi_*^{(t)}$ and $\phi_*^{(t)}$ via

$$\psi_*^{(t)} = U_S^{(t)} \otimes \widetilde{\psi}_*^{(t)} \otimes_{\mathrm{c}} V_S^{(t)} \text{ and } \phi_*^{(t)} = U_J^{(t)} \otimes \widetilde{\phi}_*^{(t)} \otimes_{\mathrm{c}} V_J^{(t)}, \tag{5}$$

where $\otimes$ denotes convolution operation, $\otimes_{\mathrm{c}}$ denotes channel-wise convolution operation, $U_S^{(t)}/U_J^{(t)}$ and $V_S^{(t)}/V_J^{(t)}$ are auxiliary parameters and can be viewed as parameter bases, and $\widetilde{\psi}_*^{(t)} \in \mathbb{R}^{h\times h\times c_i}$ and $\widetilde{\phi}_*^{(t)} \in \mathbb{R}^{h\times h\times c_i}$ are the actual parameters to predict by $A_{\phi_a^{(t)}}(\cdot)$ and $A_{\psi_a^{(t)}}(\cdot)$. In this way, the number of predicted parameters can be reduced by an order of magnitude.

For tailoring preliminary represenations with adaptive parameters, we utilize *dynamic convolution layers* for direclty applying $\psi_*^{(t)}$ and $\phi_*^{(t)}$ to conduct convolution operations on $R_S^{(t)}$ and $R_J^{(t)}$, which is implemented by just replacing static convolution kernels with the predicted dynamic ones in the traditional convolution layer:

$$\begin{aligned} R_{S*}^{(t)} &= \psi_*^{(t)} \otimes R_S^{(t)} = U_S^{(t)} \otimes \widetilde{\psi}_*^{(t)} \otimes_{\mathrm{c}} V_S^{(t)} \otimes R_S^{(t)}, \\ R_{J*}^{(t)} &= \phi_*^{(t)} \otimes R_J^{(t)} = U_J^{(t)} \otimes \widetilde{\phi}_*^{(t)} \otimes_{\mathrm{c}} V_J^{(t)} \otimes R_J^{(t)}, \end{aligned} \tag{6}$$

where $R_{S*}^{(t)}$ and $R_{J*}^{(t)}$ are dynamic representations learned from the guidance information of task counterparts, overcoming drawbacks of existing methods with hand-crafted features from estimation results. In addition, $R_{S*}^{(t)}$ and $R_{J*}^{(t)}$ are efficiently generated in a one-shot manner, avoiding the time-consuming iterative updating scheme utilized by traditional methods for representation learning. We implement $U_S^{(t)}/U_J^{(t)}$ and $V_S^{(t)}/V_J^{(t)}$ with 1×1 convolutions and apply them together with $\widetilde{\psi}_*^{(t)}/\widetilde{\phi}_*^{(t)}$ sequentially on $R_S^{(t)}/R_J^{(t)}$ to produce $R_{S*}^{(t)}/R_{J*}^{(t)}$.

Through leveraging mutual benefits between human parsing and pose estimation, $R_{S*}^{(t)}$ and $R_{J*}^{(t)}$ can provide powerful complementary cues to tailor $R_S^{(t)}$ and $R_J^{(t)}$ for better labeling semantic parts and localizing body joints. We fuse complementary representations and preliminary ones via addition in a residual manner for generating tailored representations $\bar{R}_S^{(t)}$ and $\bar{R}_J^{(t)}$ for final predictions:

$$\bar{R}_S^{(t)} = R_S^{(t)} + R_{S*}^{(t)} \text{ and } \bar{R}_J^{(t)} = R_J^{(t)} + R_{J*}^{(t)}. \tag{7}$$

Classification Module Given representations $\bar{R}_S^{(t)}$ and $\bar{R}_J^{(t)}$, we apply two linear classifiers $C^S_{\psi_w^{(t)}}(\cdot)$ and $C^J_{\phi_w^{(t)}}(\cdot)$ for predicting semantic part probability maps $S^{(t)}$ and body joint confidence maps $J^{(t)}$, respectively. In particular, we implement classifiers with 1×1 convolution layers.

After getting $S^{(t)}$ and $J^{(t)}$, the mapping modules $M^S_{\psi_m^{(t)}}(\cdot,\cdot)$ and $M^J_{\phi_m^{(t)}}(\cdot,\cdot)$ transform them and tailored representations $\bar{R}_S^{(t)}$ and $\bar{R}_J^{(t)}$ into inputs $F_S^{(t+1)}$ and $F_J^{(t+1)}$ for the next stage. Following [15], we use 1×1 convolutions on $S^{(t)}$ and $J^{(t)}$ to map predictions into the representation space. We also apply 1×1 convolutions on $\bar{R}_S^{(t)}$ and $\bar{R}_J^{(t)}$ to map highest-level representations of the previous stage into preliminary representations for the following stage. We integrate these two representations via addition for obtaining $F_S^{(t+1)}$ and $F_J^{(t+1)}$.

Training and Inference As exhibited in the loss function in Eqn. (4), we apply both parsing and pose supervision at each mutual learning stage for training the MuLA model. In particular, we utilize CrossEntropy loss and Mean Square Error loss for parsing and pose models respectively. MuLA is end-to-end trainable by gradient back propagation.

At the inference phase, MuLA simultaneously estimates parsing and pose for an input image in one forward pass. The semantic part probability maps $S^{(T)}$ and body joint confidence maps $J^{(T)}$ from the last stage of MuLA are used for final predictions. In particular, for human parsing, the category with maximum probability at each position of $S^{(T)}$ is output as the semantic part label. For pose estimation, in the single-person case, we take the position with maximum confidence for each confidence map in $J^{(T)}$ as the location of each type of body joints; in the multi-person case, we perform Non-Maximum Suppression (NMS) on each confidence map in $J^{(T)}$ for generating joint candidates.

4 Experiments

4.1 Experimental Setup

Datasets We evaluate the proposed MuLA model on two benchmarks for simultaneous human parsing and pose estimation: the Look into Person (LIP) dataset [10] and extended PASCAL-Person-Part dataset [24]. The LIP dataset includes 50,462 single-person images collected from various realistic scenarios, with pixel-wise annotations provided for 19 categories of semantic parts and location annotations for 16 types of body joints. In particular, LIP images are split into 30,462 for training, 10,000 for validation and 10,000 for testing. The extended PASCAL-Person-Part is a challenging multi-person dataset, containing annotations for 14 body joints and 6 semantic parts. In total, there are 3,533 images, which are split into 1,716 for training and 1,817 for testing.

Data Augmentation We conduct data augmentation strategies commonly used in previous works [28,3] for both human parsing and pose estimation, including random rotation in $[-40°, 40°]$, random scaling in $[0.8, 1.5]$, random cropping

Table 1. VGG16-FCN based ablation studies on LIP validation set

Methods	PCK	mIOU
VGG16-FCN	69.1	34.5
VGG16-FCN-Add	69.7	36.5
VGG16-FCN-Multi	69.4	35.8
VGG16-FCN-Concat	69.5	36.1
VGG16-FCN-MTL	65.3	31.2
VGG16-FCN-Self	69.8	36.1
VGG16-FCN-LA-Pose	75.0	32.1
VGG16-FCN-LA-Parsing	66.5	40.0
VGG16-FCN-MuLA	**76.0**	**40.2**

Table 2. Hourglass network based ablation studies on LIP validation set

Methods	PCK	mIOU
HG-0s-1u-MuLA	78.8	38.5
HG-1s-1u-MuLA	82.2	43.5
HG-2×1u	80.8	41.3
HG-2s-1u-MuLA (1st Stage)	82.8	45.5
HG-2s-1u-MuLA (2nd Stage)	83.1	45.6
HG-2s-1u-MuLA	84.4	46.9
HG-3s-1u-MuLA	85.0	47.8
HG-4s-1u-MuLA	85.1	48.9
HG-5s-1u-MuLA	**85.4**	**49.3**

based on the person center with translational offset in $[-40\text{px}, 40\text{px}]$, and random horizontally mirroring. We resize and pad augmented training samples into 256×256 as input to CNNs.

Implementation We train MuLA from scratch for LIP and extended PASCAL-Person-Part datasets with their own training samples, separately. For multi-person pose estimation on extended PASCAL-Person-Part dataset, we follow the method proposed in [16]. It partitions joint candidates into corresponding persons via a dense regression branch in the pose model of MuLA for transforming joint candidates into the centroid embedding space. We implement MuLA with PyTorch [17] and use RMSProp [21] as the optimizer. We set the initial learning rate as 0.0025 and drop it with multiplier 0.5 at the 150th, 170th, 200th and 230th epochs. We train MuLA for 250 epochs in total. We perform multi-scale testing to produce final predictions for both human parsing and pose estimation. Our codes and pre-trained models will be made available.

Metrics Following conventions, Mean Intersection-over-Union (mIOU) [10] is used for evaluating human parsing performance. We use PCK [27] and Mean Average Precision (mAP) [11,16] for measuring accuracy of single- and multi-person pose estimation, respectively.

4.2 Results on LIP Dataset

Ablation Analysis We evaluate the proposed MuLA model with two kinds of backbone architectures, *i.e.*, the VGG16-FCN and Hourglass networks, for both human parsing and pose estimation as mentioned in Sec. 3.2.

Firstly, we conduct ablation experiments on LIP validation set with VGG16-FCN based model, denoted as VGG16-FCN-MuLA, to investigate efficacy of MuLA on leveraging mutual guidance information to simultaneously improve parsing and pose performance. The results are shown in Table 1. To demonstrate effectiveness of the adaptive representations learned by MuLA, we compare with prevalent strategies that directly fuse representations from parallel models, including addition, multiplication, concatenation. We denote these baselines as VGG16-FCN-Add/Multi/Concat respectively. To evaluate the advantages of the

interweaving architecture of MuLA, we also compare it with traditional multi-task learning framework for joint human parsing and pose estimation, implemented by adding both parsing and pose supervision on a single VGG16-FCN, denoted as VGG16-FCN-MTL. To investigate effects of the residual architecture followed by the adaptation modules, we wipe off mutual interaction between tasks through replacing dynamic convolution layers with traditional convolution layers. Such a variant is denoted as VGG16-FCN-Self. To validate advantages of bidirectionally utilizing guidance information between two tasks, we simplify MuLA by alternatively removing parsing and pose adapters, resulting in single-direction adaptation models, denoted as VGG16-FCN-LA-Pose and VGG16-FCN-LA-Parsing.

From Table 1, we can see that the proposed VGG16-FCN-MuLA significantly improves performance of baseline VGG16-FCN by a large margin on both human parsing and pose estimation, from 34.5% to 40.2% mIoU and 69.1% to 76.0% PCK, respectively. These results clearly show efficacy of MuLA on exploiting mutual benefits to jointly enhance model performance. We can also observe direct fusion of representations from both models as VGG16-FCN-Add/Multi/Concat cannot sufficiently utilize guidance information, resulting in very limited performance improvement. In contrast to these naive fusion strategies, VGG16-FCN-MuLA can learn more powerful representations via dynamically adapting parameters. Traditional multi-task learning framework VGG16-FCN-MTL suffers performance decline for both parsing and pose estimation, due to limitations brought by its tied architecture trying to learn single representation for both tasks. In contrast, MuLA learns separate representations for each task, providing a flexible and effective model for multi-task learning. Adding a residual architecture to the adaptation modules only slightly improves performance for both tasks, revealing performance gain is not simply from network architecture engineering. Instead, MuLA indeed learns useful complementary representations.

Single-direction learning to adapt variants VGG16-FCN-LA-Pose/Parsing can successfully leverage parsing (or pose) information to adapt pose (or parsing respectively) models, leading to performance improvement. This verifies effectiveness of our proposed learning to adapt module in exploiting guidance information from parallel models. However, we can also observe such single-direction learning harms performance of "source" tasks, due to over-concentration on the "target" tasks. It demonstrates the necessity of mutual learning for simultaneously boosting performance of human parsing and pose estimation.

To evaluate the power of MuLA on iteratively exploiting mutual benefits between human parsing and pose estimation, we further perform ablation studies with the Hourglass based model. The results are summarized in Table 2. We use HG-ms-nu-MuLA to denote the model containing m stages each with n-unit depth (32-layer per unit depth per Hourglass module is the basic configuration in [15]). Specially, HG-0s-1u-MuLA denotes independent Hourglass networks (without mutual learning to adapt) are utilized for the two tasks. We purposively make all stages have the same architecture for disentangling effects of architecture variations on performance. In particular, HG-2s-1u-MuLA (1st/2nd

Table 3. Comparison with state-of-the-arts on LIP for human pose estimation task

Methods	PCK
Hybrid Pose Machine	77.2
BUPTMM-POSE	80.2
Pyramid Stream Network	82.1
Chou *et al.* [7]	87.4
Our model	**87.5**

Table 4. Comparison with state-of-the-arts on LIP for human parsing task

Methods	PixelAcc	MeanAcc	mIoU
SegNet [1]	69.0	24.0	18.2
FCN-8s [13]	76.1	36.8	28.3
DeepLabV2 [4]	82.7	51.6	41.6
Attention [5]	83.4	54.4	42.9
Attention+SSL [10]	84.4	54.9	44.7
SS-NAN [28]	87.6	56.0	47.9
Our model	**88.5**	**60.5**	**49.3**

Stage) denotes ablation cases of HG-2s-1u-MuLA where only the 1st or 2nd stage contains the module for mutual-learning to adapt. We use HG-$k\times n$u to denote standard Hourglass network with k stacked Hourglass modules of n-unit depth.

From Table 2, we can observe that increasing the number of stages in MuLA from 0 to 5 can continuously improve the performance for both tasks, from 38.5% to 49.3% mIoU for human parsing and 78.8% to 85.4% PCK for pose estimation. Comparing HG-2s-1u-MuLA with HG-2×1u, we can find the proposed MuLA model can learn valuable representations from model counterparts rather than benefiting from stacking Hourglass modules. Comparing HG-2s-1u-MuLA with HG-2s-1u-MuLA (1st/2nd Stage), we can see that removing mutual-learning process at any stage will always harm the performance for both parsing and pose estimation, demonstrating that the proposed adaptation module is effective at leveraging mutual guidance information and necessary to be applied for all the stages in MuLA. In addition, we find using more than 5 stages for MuLA will not bring observable improvement. Hence, we set T=5 for efficiency.

Comparisons with State-of-the-arts We compare our model HG-5s-1u-MuLA with state-of-the-arts for both human parsing and pose estimation on LIP dataset. The results are shown in Table 3 and 4.

For human pose estimation, the method in [7] wins the first place in Human Pose Estimation track in the 1st LIP Challenge. It extensively exploits adversarial training strategies. The pyramid stream network introduces top-down pathway and lateral connections to combine features of different levels for recurrently refining joint confidence maps. BUPTMM-POSE and Hybrid Pose machines are from combining the Hourglass network and Convolutional Pose Machines. From Table 3, we can find our model achieves superior accuracy over all these strong baselines. It achieves new state-of-the-art 87.5% PCK on the LIP dataset.

Table 4 shows comparison with state-of-the-arts for human parsing. In addition to mIoU, we also report pixel accuracy and mean accuracy, following conventions [10,28,5]. In particular, the methods in [10,28] utilize human pose information as extra supervision to assist human parsing via introducing a structure-sensitive loss based on body joint locations. We can observe that our model outperforms all previous methods consistently for all the evaluation metrics. It gives new state-of-the-art 88.5% pixel accuracy, 60.5% mean accuracy and 49.3% mIoU. This demonstrates our learning to adapt module indeed provides a more

Table 5. Results on the PASCAL-Person-Part dataset for Human Pose Estimation

Methods	mAP
Chen and Yuille [6]	21.8
Insafutdinov *et al.* [11]	28.6
Xia *et at.* [24]	39.2
Our baseline (w/o MuLA)	38.6
Our model	**39.9**

Table 6. Results on the PASCAL-Person-Part dataset for Human Parsing

Methods	mIoU
Attention+SSL [10]	59.4
SS-NAN [28]	62.4
Xia *et al.* [24]	64.4
Our baseline (w/o MuLA)	62.9
Our model	**65.1**

effective way for exploiting human pose information to guide human parsing than the other sophisticated strategies like structure-sensitive loss in [10,28].

Qualitative Results Fig. 4 (a) shows qualitative results to visually illustrate the efficacy of MuLA in mutually boosting human parsing and pose estimation. We can observe that MuLA can exploit body part information from human parsing to constrain body joint locations, *e.g.*, from the 1st and 2nd examples. On the other hand, MuLA can use human pose to provide structure information to benefit human parsing by improving accuracy of semantic part labeling, as shown in the 3rd and 4th examples. Moreover, we can see that MuLA simultaneously improves both parsing and pose quality for all the examples.

4.3 Results on PASCAL-Person-Part Dataset

Different from LIP dataset, the extended PASCAL-Person-Part dataset presents more challenging pose estimation problems due to existence of multiple persons. As mentioned in Sec. 4.1, we utilize the model in [16] as the pose model in MuLA for partitioning joint candidates to corresponding person instances. We exploit Hourglass network based MuLA with 5 stages for experiments. The results are shown in Table 5 and 6.

We can see that our baseline models achieves 38.6% mAP and 62.9% mIoU for multi-person pose estimation and human parsing. With the proposed MuLA model, the performance for two tasks can be improved to 39.9% mAP and 65.1% mIoU, respectively. We also observe that our model achieves superior performance over previous methods for both tasks. In particular, [24] presents the state-of-the-art model for joint human parsing and pose estimation via exploiting hand-crafted features from estimation results as post-processing. The superior performance of our model over [24] further demonstrates the effectiveness of learning to adapt with mutual guidance information for enhancing models for joint human parsing and pose estimation.

We visualize human parsing and multi-person pose estimation results in Fig. 4 (b). We can see that MuLA can use body joint information to recover missing detected parts, *e.g.*, left arm of left person in the 1st example and right arm of right person in the 2nd example. In addition, MuLA can also utilize semantic part information to constrain body joint location, *e.g.*, right knee of the right person in the 1st example and left ankle of the left person in the 2nd example.

(a) (b)

Fig. 4. Qualitative results on (a) LIP and (b) extended PASCAL-Person-Part dataset. For each column, the first two rows are results of the baseline model HG-5×1u without exploiting mutual guidance information and the last two rows are results of the proposed model HG-5s-1u-MuLA. Best viewed in color

5 Conclusion

In this paper, we present a novel Mutual Learning to Adapt (MuLA) model for solving the challenging joint human parsing and pose estimation problem. MuLA uses a new interweaving architecture to leverage their mutual guidance information to boost their performance simultaneously. In particular, MuLA achieves dynamic interaction and cooperation between these two tasks by mutually learning to adapt parameters of parallel models for tailoring their preliminary representations by injecting information from the other one. MuLA can iteratively weave mutual guidance information for continuously improving performance for both tasks. It effectively overcomes limitations of previous works that exploit mutual benefits between two tasks through using hand-crafted features in the post-processing. Comprehensive experiments on benchmarks have clearly verified the efficacy of MuLA for joint human parsing and pose estimation. In particular, MuLA achieved new state-of-the-art for both human parsing and pose estimation tasks on the LIP dataset, and outperformed all previous methods devoted to jointly performing these two tasks on PASCAL-Person-Part dataset.

Acknowledgement

Jiashi Feng was partially supported by NUS IDS R-263-000-C67-646, ECRA R-263-000-C87-133 and MOE Tier-II R-263-000-D17-112.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017) 12
2. Bertinetto, L., Henriques, J.F., Valmadre, J., Torr, P., Vedaldi, A.: Learning feed-forward one-shot learners. In: NIPS (2016) 6, 8
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017) 9
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In: ICLR (2015) 12
5. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: CVPR (2016) 7, 12
6. Chen, X., Yuille, A.: Parsing occluded people by flexible compositions. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on (2015) 13
7. Chou, C.J., Chien, J.T., Chen, H.T.: Self adversarial training for human pose estimation. In: CVPR Workshops (2017) 12
8. Dong, J., Chen, Q., Shen, X., Yang, J., Yan, S.: Towards unified human parsing and pose estimation. In: CVPR (2014) 1, 4
9. Gan, C., Lin, M., Yang, Y., de Melo, G., Hauptmann, A.G.: Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition. In: AAAI (2016) 1
10. Gong, K., Liang, X., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: CVPR (2017) 1, 3, 9, 10, 12, 13
11. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: ECCV (2016) 10, 13
12. Ladicky, L., Torr, P.H., Zisserman, A.: Human pose estimation using a joint pixel-wise and part-wise formulation. In: CVPR (2013) 1, 3
13. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015) 12
14. Lu, Y., Boukharouba, K., Boonært, J., Fleury, A., Lecoeuche, S.: Application of an incremental svm algorithm for on-line human recognition from video surveillance using texture and color features. Neurocomputing (2014) 1
15. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016) 6, 7, 9, 11
16. Nie, X., Feng, J., Xing, J., Yan, S.: Generative partition networks for multi-person pose estimation. arXiv preprint arXiv:1705.07422 (2017) 10, 13
17. Paszke, A., Gross, S., Chintala, S.: Pytorch (2017) 10
18. Ren, S., Kaiming, H., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS (2015) 7
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. ICLR (2015) 7
20. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: CVPR (2017) 1
21. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning (2012) 10

22. Wang, C., Wang, Y., Yuille, A.L.: An approach to pose-based action recognition. In: CVPR (2013) 1
23. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016) 6
24. Xia, F., Wang, P., Chen, X., Yuille, A.: Joint multi-person pose estimation and semantic part segmentation. In: CVPR (2017) 1, 3, 4, 9, 13
25. Xia, F., Zhu, J., Wang, P., Yuille, A.L.: Pose-guided human parsing by an and/or graph using pose-context features. In: AAAI (2016) 1, 3
26. Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L.: Parsing clothing in fashion photographs. In: CVPR (2012) 1, 3
27. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2878–2890 (2013) 10
28. Zhao, J., Li, J., Nie, X., Zhao, F., Chen, Y., Wang, Z., Feng, J., Yan, S.: Self-supervised neural aggregation networks for human parsing. In: CVPR Workshops (2017) 3, 9, 12, 13
29. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR (2013) 1
30. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3d human pose estimation from monocular video. In: CVPR (2016) 1