Cross-domain Human Parsing via Adversarial Feature and Label Adaptation

Si Liu[1,4,5], **Yao Sun**[1,*], **Defa Zhu**[1], **Guanghui Ren**[1], **Yu Chen**[2], **Jiashi Feng**[3], **Jizhong Han**[1]
[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] JD.com [3] Department of ECE, National University of Singapore
4 Jiangsu Key Laboratory of Big Data Analysis Technology /B-DAT, Nanjing University of Information Science & Technology
5 Collaborative Innovation Center of Atmospheric Environment and Equipment Technology
Nanjing University of Information Science and Technology, Nanjing, China
{liusi, sunyao, zhudefa, renguanghui, hanjizhong}@iie.ac.cn, chenyu6@jd.com, elefjia@nus.edu.sg
∗ corresponding author

Abstract

Human parsing has been extensively studied recently (Yamaguchi et al. 2012; Xia et al. 2017) due to its wide applications in many important scenarios. Mainstream fashion parsing models (i.e., parsers) focus on parsing the high-resolution and clean images. However, directly applying the parsers trained on benchmarks of high-quality samples to a particular application scenario in the wild, e.g., a canteen, airport or workplace, often gives non-satisfactory performance due to domain shift. In this paper, we explore a new and challenging cross-domain human parsing problem: taking the benchmark dataset with extensive pixel-wise labeling as the source domain, how to obtain a satisfactory parser on a new target domain *without requiring any additional manual labeling?* To this end, we propose a novel and efficient cross-domain human parsing model to bridge the cross-domain differences in terms of visual appearance and environment conditions and fully exploit commonalities across domains. Our proposed model explicitly learns a feature compensation network, which is specialized for mitigating the cross-domain differences. A discriminative feature adversarial network is introduced to supervise the feature compensation to effectively reduces the discrepancy between feature distributions of two domains. Besides, our proposed model also introduces a structured label adversarial network to guide the parsing results of the target domain to follow the high-order relationships of the structured labels shared across domains. The proposed framework is end-to-end trainable, practical and scalable in real applications. Extensive experiments are conducted where LIP dataset is the source domain and 4 different datasets including surveillance videos, movies and runway shows without any annotations, are evaluated as target domains. The results consistently confirm data efficiency and performance advantages of the proposed method for the challenging cross-domain human parsing problem.

Introduction

Recently human parsing (Liu et al. 2015) has been receiving increasing attention owning to its wide applications, such as person re-identification (Zhao, Ouyang, and Wang 2014), people search (Li et al. 2017), fashion synthesis (Zhu et al.).

Existing human parsing algorithms can be divided into following two categories. The first one is **constrained human parsing**. More specifically, the clean images of well-posed persons are collected from some fashion sharing websites, e.g., Chictopia.com, for training and testing. Representative datasets include Fashionista (Yamaguchi et al. 2012) with 685 images, Colorful-Fashion dataset (Liu et al. 2014) with $2,682$ images and ATR dataset (Liang et al. 2015a) with $7,700$ dataset. Each image in these datasets contains only one person, with relatively simple poses (mostly standing), against relatively clean backgrounds. The human parsers trained in such strictly constrained scenario often fail when applied to images captured under the real-life, more complicated environments. The second category is **unconstrained human parsing**. Representative datasets include Pascal human part dataset (Chen et al. 2014b) with $3,533$ images and LIP dataset (Gong et al. 2017) with $50,462$ images. The images in these dataset present humans with varying clothing appearances, strong articulation, partial (self-) occlusions, truncation at image borders, diverse viewpoints and background clutters. Although they are closer to real environments than the constrained datasets, when applying the human parser trained on these unconstrained datasets to a real application scenario, such as shop, airport, the performance is still worse than the parser trained on that particular scenario even with much less training samples, due to domain shift on visual features.

In this paper, we explore a new **cross-domain human parsing** problem: taking the unconstrained benchmark

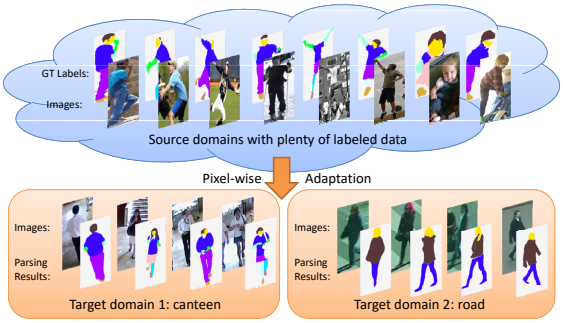

Figure 1: Cross-domain human parsing: the upper panel is the source domain with a large amount of training data, e.g., the LIP dataset; the lower panel shows the target domain, e.g, canteen and road, without any manual labeling.

dataset with rich pixel-wise labeling as the source domain, how to obtain a satisfactory parser for a totally different target domain without any additional manual labeling? As shown in Figure 1, the source domain (shown in the upper panel of Figure 1) is a set of labeled data. The target domain training set (shown in the lower panel of Figure 1) is as a set of images without any annotations. We believe investigation on this challenging problem will push human parsing models toward more practical applications.

From Figure 1, we observe the following differences and commonality across two domains, e.g., the source domain and the first target domain, canteen. On the one hand, they have very different illumination, view points, image scale, resolution and degree of motion blur etc. For example, the lighting condition in the canteen scenario is much darker than the source domain. On the other hand, the persons to parse from both domains share the intrinsic commonality such as the high-order relationships among labels (reflecting human body structure) are similar. For example, in both domains, the arms are below the head, but above the legs. Therefore, the cross-domain human parsing problem can be solved by *minimizing the differences* of the features and *maximizing the commonality* of the structured labels.

A typical semantic segmentation network (Long, Shelhamer, and Darrell 2015; Chen et al. 2014a) is composed of a feature extractor and a pixel-wise labeler. In this work, we propose to introduce a new and learnable feature compensation network that transforms the features from different domains to a common space where the cross-domain difference can be effectively alleviated. In this way, the pixel-wise labeler can be readily applied to perform parsing on the compensated features. The feature compensation network is trained under the joint supervision from a feature adversarial network and a structured label adversarial network. More specifically, the feature adversarial network serves as a supervisor and provides guidance on the feature compensation learning like the discriminator of the Generative Adversarial Network (GANs) (Goodfellow et al. 2014; Radford, Metz, and Chintala 2015). It is trained to differentiate target and compensated source feature representations. Similarly, the structured label adversarial network differentiates the groundtruth structural labels and the predicted target domain labels. Supervised by these two level information, the cross-domain shift issues can be effectively addressed. We evaluate our approach using LIP (Gong et al. 2017) as source domain and 4 datasets as target domains. Extensive experiments demonstrate the effectiveness of our method on all domain shifts adaptation tasks.

The contributions of the paper can be summarized as following. Firstly, we are the first to explore the cross-domain human parsing problem. Since no manual labeling in the target domain is needed, it is very practical. Secondly, we propose a cross-domain human parsing framework with the novel feature adaptation and structured label adaptation network. It is the first cross-domain work to consider both feature invariance and label structure regularization. Thirdly, we will release the source code of our implementation to the academic to facilitate future studies.

Related Work

Human parsing and cross-domain feature transformation have been studied for decades. However, they generally develop independently. There are few works consider solving the cross-domain human parsing by considering these directions jointly. In this section, we briefly review recent techniques on human parsing as well as feature adaption.

Human parsing: Yamaguchi *et al.* (Yamaguchi, Kiapour, and Berg 2013) tackle the clothing parsing problem using a retrieval based approach. Simo-Serra *et al.* (Simo-Serra et al. 2014) propose a Conditional Random Field (CRF) model that is able to leverage many different image features. Luo *et al.* (Luo, Wang, and Tang 2013) propose a Deep Decompositional Network for parsing pedestrian images into semantic regions. Liang *et al.* (Liang et al. 2015b) propose a Contextualized Convolutional Neural Network to tackle the problem and achieve very impressing results. Xia *et al.* (Xia et al. 2015) propose the "Auto-Zoom Net" for human paring. Wei *et al.* (Wei et al. 2016; 2017) propose several weakly supervised parsing methods to reduce the human labeling burden. Existing human parsing methods work well in the benchmark datasets. However, when applied in the new application scenarios, the performances are unsatisfactory. The cross-domain human parsing problem becomes a significant problem for making the technology practical.

Feature Adaptation: There have been extensive prior works on domain transfer learning (Gretton et al. 2009). Recent works have focused on transferring deep neural network representations from a labeled source dataset to a target domain where labeled data is sparse. In the case of unlabeled target domains (the focus of this paper) the main strategy has been to guide feature learning by minimizing the differences between the source and target feature distributions (Ganin and Lempitsky 2015; Liu and Tuzel 2016; Long et al. 2015). Different from existing feature adaptation methods, we explicitly consider the cross-domain differences via a feature compensation network.

Structured Label Adaptation: There are few works to consider the label structure adaptation during domain adaption. Some pioneer pose estimation works take the geometric constraints of human joints connectivity into consideration. For example, Chen *et al.* (Chen et al. 2017) propose Adversarial PoseNet, which is a structure-aware convolutional network to implicitly take such priors into account during training of the deep network. Chou *et al.* (Chou, Chien, and Chen 2017) employ GANs as pose estimator, which enables learn plausible human body configurations Our proposed cross-domain human parsing method differs from existing domain adaptation methods in that we consider both feature and structured label adaption simultaneously.

Proposed Cross-domain Adaptation Model

Suppose the source domain images and labels are denoted as S_x and S_y respectively. The target domain images are represented as T_x. Typical human parsing models are composed of a feature extractor $E(\cdot)$ and a pixel-wise labeler $L(\cdot)$. However, the parsing model trained on the source domain

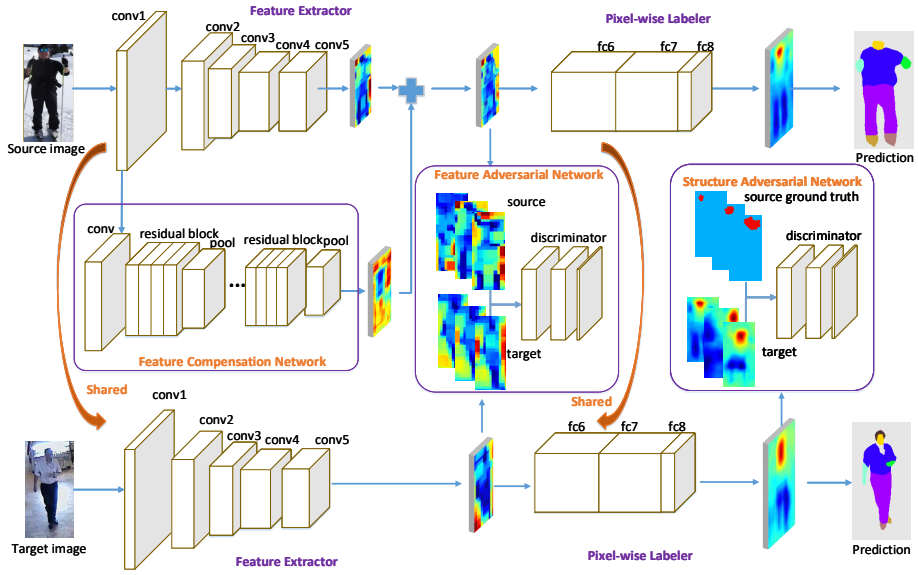

Figure 2: The cross-domain parsing model. It contains a feature adaptation component to minimize the feature differences between two domains, and a structured label adaptation component to maximize the label map commonalities across the domains.

does not perform well in the target domain in presence of significant domain shift.

Our proposed cross-domain adaptation model to address this issue is shown in Figure 2. It includes a novel feature compensation component supervised by two components, namely adversarial feature adaptation and adversarial structured label adaptation components. The feature adaptation component aims to minimize the feature differences between different domains, while the structured label adaptation is used to maximize the label map commonalities across the domains. Therefore, the whole model introduces three novel components (shown in purple rectangular) on top of conventional human parsing models: feature compensation network $C(\cdot)$, feature adversarial network $A_f(\cdot)$ and structured label adversarial network $A_l(\cdot)$ to address the cross-domain human parsing problem. Next, we will introduce the two adversarial learning components and explain how they help feature compensation to mitigate the domain difference.

Adversarial Feature Adaptation

The feature compensation network and feature adversarial network collaboratively contribute to the feature adaptation. $C(\cdot)$ maps the feature representation of the source domain toward the target domain under the supervision of $A_f(\cdot)$. Alternatively updating them gradually narrows the cross-domain feature differences.

The *feature compensation network*, as shown in Figure 2, takes as input the extracted features $E_1(S_x)$ from source domain, where $E_1(\cdot)$ is a part of the feature extractor $E(\cdot)$. The output $C(E_1(\mathcal{S}_x))$ is the feature differences (shift) between source and target domains. $E(\cdot)$ is composed of 5 convolutional layers of VGG-16 net (Simonyan and Zisserman 2014), from *conv1* till *pool5* in VGG-16.The first several layers (from *conv1* till *pool1*) forms $E_1(\cdot)$. The structure of the feature compensation network is a ResNet-like (He et al. 2016) network with a 7×7 convolution filters and then 6 residual blocks with the identical layout consisting of two 3×3 convolution filters followed by batch-normalization layer and ReLU activation layer. Every three blocks follows a max pool layer and a 3×3 convolution filter to reduce feature maps' sizes. The result of the feature compensation network is pixel-wisely added to that of the feature extractor to produce the compensated source domain feature $E(\mathcal{S}_x) + C(E_1(\mathcal{S}_x))$.

The *feature adversarial network* $A_f(\cdot)$ is introduced to guide the cross-domain feature adaptation. Different from traditional adversarial learning models (e.g., vanilla GAN (Goodfellow et al. 2014)) that performs judgment over raw images, our proposed feature adversarial network is defined upon the high-level feature space (*pool5*) which incorporates essential feature information for human parsing. It can accelerate the training and inference. The architecture of $A_f(\cdot)$ is composed of the same fc6-fc7 layers of the Atrous Spatial Pyramid Pooling (ASPP) scheme in DeepLab (Chen et al. 2016). Then a convolution layer with 3×3 convolution filters is appended to create a 1-channel probability map, which is used to calculate the pixel-wise least square feature adversarial loss, like the local LSGANs (Shrivastava et al. 2016).

The *optimization* of $A_f(\cdot)$ and $C(\cdot)$ are iterative. More specifically, the objective for updating $A_f(\cdot)$ is:

$$\min_{A_f} \mathcal{E}_{A_f} = \tfrac{1}{2} E_{T_x \sim p_{target}(T_x)} \left[\left(A_f\left(E\left(T_x\right)\right) - \mathbf{1} \right)^2 \right] + \tfrac{1}{2} E_{S_x \sim p_{source}(\mathrm{S}_x)} \left[\left(A_f \left(\underbrace{\mathrm{E}\left(S_x\right) + C\left(E_1\left(S_x\right)\right)}_{compensated\ feature} \right) \right)^2 \right], \tag{1}$$

where $\mathbf{1}$ is an all-one tensor. The feature adversarial network adopts the least squares loss function, regressing the feature of the target domain $E(T_x)$ to $\mathbf{1}$ while regressing the features of the compensated source domain $E(\mathcal{S}_x) + C(E_1(\mathcal{S}_x))$ to $\mathbf{0}$. It distinguishes the target feature and the

compensated source domain feature, while the feature compensation network tries to transform them into indistinguishable ones.

The learning target of the feature compensation network is to mitigate the difference between source and target features. It is trained by optimizing the following objective function:

$$\min_{C} \mathcal{E}_c = \frac{1}{2} E_{S_x \sim p_{source}(\mathrm{S}_x)} \left[\left(A_f \left(\underbrace{\mathrm{E}\left(S_x\right) + C\left(E_1\left(S_x\right)\right)}_{compensated\ feature} \right) - \mathbf{1} \right)^2 \right]. \tag{2}$$

The target of $C\left(\cdot\right)$ is to transform the source domain features to the one similar to target domain by trying to confuse $A_f\left(\cdot\right)$. Or more concretely, the $C\left(\cdot\right)$ tries to generate features that persuade the $A_f\left(\cdot\right)$ to predict the feature is from target domain (output binary prediction of **1**). It implicitly maps the source domain features toward the target domain by encoding lighting conditions, environment factors. By iteratively boosting the abilities of $A_f\left(\cdot\right)$ and $C\left(\cdot\right)$ through alternative training, the gap between the two domains are gradually narrowed down.

Algorithm 1: Training details of the integrated cross-domain human parsing framework.

Input: Source domain images S_x; source domain labels S_y; target domain images T_x; feature extractor E; feature compensation network C; feature adversarial network A_f; structured label adversarial network A_l; pixel-wise labeler L; number of training iterations N; a constant K_C.

for $t = 1, \cdots, N$ **do**
  sample $\{S_x^i\}$, $\{S_y^i\}$, $\{T_x^i\}$, $i = 1, \cdots, n$.
  update E, L by minimizing $\mathcal{P}_{E,L}^{(1)}$.
  update C by minimizing Equation (2).
  update A_f by minimizing Equation (1).
  if $\mathrm{mod}(t, K_C) == 0$ **then**
    update E, L by minimizing Equation (4).
    update A_l by minimizing Equation (3).
  end
  update E, L by minimizing $\mathcal{P}_{E,L}^{(2)}$.
end

Adversarial Structured Label Adaptation

Only feature compensation cannot fully utilize the valuable information about human body structure and leads to sub-optimal parsing performance. Therefore, we also propose a structured label adversarial network that learns to capture the commonalities of parsing labels from different domains. Such information is learnable from the source domain data because of the following reasons. Firstly, the labels have very strong spatial priors. For example, in daily-life scenarios, the head always lies on the top, while the shoes appear in the bottom in most cases. Moreover, relative positions between the labels are consistent across domains. For example, the arms lie on both sides of the body, while the head is at the top of the body. Finally, the part shapes of certain labels are basically similar on both domains. For example, the faces are usually round or oval while the legs are often long striped. The pixel-wise labeler and the structured label adversarial network collaboratively adapt the structured label prediction.

The *pixel-wise labeler* is composed of the $fc6$, $fc7$ and $fc8$ layers of DeepLab (Chen et al. 2016), which is a fully convolutional variant of the VGG-16 net (Simonyan and Zisserman 2014) by modifying the atrous (dilated) convolutions to increase the field-of-view. Depending on the properties of the input, two losses are defined upon the network.

1 $\mathcal{P}_{E,L}^{(1)}$: The pixel-wise cross entropy loss defined upon the source domain images $\mathrm{E}\left(S_x\right)$ and S_y.

2 $\mathcal{P}_{E,L}^{(2)}$: The pixel-wise cross entropy loss defined upon the compensated source domain features $\mathrm{E}\left(S_x\right) + C\left(E_1\left(S_x\right)\right)$ and S_y.

The *structured label adversarial network* is used to distill the high-order relationships of the labels from the source domain groundtruth pixel-wise labels S_y and *transfer* to guide parsing target domain images. The architecture of $A_l\left(\cdot\right)$ is as follows. LeakyReLU activations and batch normalization are used for all layers except the output. All layers contain stride = 2 convolution filter except the last layer, which just contains one stride = 1 convolution filter to produce the confidence map. All convolution filter used in the network is 5×5 convolution filter.

The *optimization* is conducted jointly through a minimax scheme that alternates between optimizing the parsing network and the adversarial network. $A_l\left(\cdot\right)$ takes either the ground truth label or the prediction parsing result, and output the probability estimate of the input is the ground truth (with training target **1**) or the segmentation network prediction (with training target **0**). The learning target is:

$$\begin{aligned} \min_{A_l} \mathcal{E}_{A_l} &= \frac{1}{2} E_{S_y \sim p_{source}(S_y)} [\left(A_l\left(S_y\right) - \mathbf{1}\right)]^2 \\ &+ \frac{1}{2} E_{T_x \sim p_{target}(\mathrm{T}_x)} \left[\left(A_l\left(L\left(\mathrm{E}\left(T_x\right)\right)\right)\right)^2 \right]. \end{aligned} \tag{3}$$

The A_l can help refine the feature extractor and pixel-wise labeler via:

$$\min_{E,L} \mathcal{E}_{E,L} = \frac{1}{2} E_{T_x \sim p_{target}(\mathrm{T}_x)} \left[\left(A_l\left(L\left(\mathrm{E}\left(T_x\right)\right) - \mathbf{1}\right)\right)^2 \right]. \tag{4}$$

Both $E\left(\cdot\right)$ and $L\left(\cdot\right)$ collaboratively confuse A_f to produce the output **1**, which means that the parsing results are drawn from the ground truth labels.

Model Learning and Inference

Training details of the integrated cross-domain human parsing framework are summarized in Algorithm 1. Generally speaking, all the model parameters are alternatively updated. Note that before every update of A_l, the network E, L, C and A_f are updated for 5 times. Experiments show that the different updating scheduling between A_l and the remaining network facilitate the model convergence.

During inference, the parsing label of the test sample is obtained by $L\left(E\left(\mathrm{S}_x\right)\right)$. Note that the feature compensation

Table 1: From LIP to Indoor dataset. (%).

Methods	Avg. acc	Fg. acc	Avg. pre	Avg. rec	Avg. F1
Target Only	89.50	74.53	60.07	59.55	59.75
Source Only	86.84	68.87	51.12	50.97	49.70
DANN	88.04	71.74	52.23	50.73	50.50
Feat. Adapt	**88.16**	72.56	**53.63**	52.23	51.59
Lab. Adapt	88.14	72.82	53.21	51.54	50.95
Feat. + Lab. Adapt	87.98	**73.86**	50.84	**54.49**	**51.73**

Table 2: From LIP to Daily Video dataset. (%).

Methods	Avg. acc	Fg. acc	Avg. pre	Avg. rec	Avg. F1
Target Only	85.96	62.64	58.63	61.07	59.73
Source Only	87.11	63.47	62.05	63.93	62.41
DANN	87.56	63.20	**64.28**	62.73	62.84
Feat. Adapt	87.64	64.83	63.95	64.25	63.40
Lab. Adapt	87.52	**66.53**	62.64	65.62	63.62
Feat. + Lab. Adapt	**87.88**	65.87	64.08	**65.97**	**64.36**

network, tow adversarial networks are not involved in the inference stage. Therefore, the complexity of our algorithm is the same with conventional human parsing method.

Discussions

In terms of the architecture of the adversarial networks, we originally tried DCGANs (Radford, Metz, and Chintala 2015). However, we found it difficult to optimize (issue of convergence) and performs not so well. Therefore, we borrow the architecture Least Squares Generative Adversarial Networks (LSGANs) (Mao et al. 2016) to build our adversarial learning networks, which adopts least squares loss function for the discriminator. It performs more stable during learning. For the feature adversarial network, the adversarial loss is defined pixel-wisely on the 2-dim feature maps. The *local LSGANs* structure (Shrivastava et al. 2016) can hence the capacity of the network. The situation is similar for structured label adversarial network.

Experiments

We conduct extensive experiments to evaluate performance of our model for 4 cross-domain human parsing scenarios.

Experimental Setting

Source Domain : We use *LIP* dataset (Gong et al. 2017) as the source domain that contains more than $50,000$ images with careful pixel-wise annotations of 19 semantic human parts. These images are collected from real-world scenarios and the persons present challenging poses and views, heavily occlusions, various appearances and low-resolutions. The original 19 labels are merged to 12 labels or 4 labels by discarding or combining to be consistent with target domains.

Target Domain: The following *four* target domains are investigated in this paper. Some example images from these target domains are shown in Figure 3.

Indoor dataset (Liu et al. 2016) contains $1,900$ labeled images with 12 semantic human part labels and $15,436$ unlabeled images. The images are captured in the canteen by surveillance cameras and have dim lights.

Daily Video dataset is a newly collected dataset, containing $1,584$ labeled images with 12 semantic human part labels and $19,964$ unlabeled images. These images are collected from a variety of scenes including shop, road, etc.

Table 3: From LIP to PridA dataset. (%).

Methods	Avg. acc	Fg. acc	Avg. pre	Avg. rec	Avg. F1
Target Only	89.90	81.44	81.38	82.96	82.12
Source Only	86.10	78.39	72.54	80.60	76.00
DANN	86.17	**81.99**	73.51	82.18	76.99
Feat. Adapt	86.63	81.51	73.41	**82.88**	77.39
Lab. Adapt	87.01	79.55	73.74	81.81	77.14
Feat. + Lab. Adapt	**87.24**	80.81	**74.76**	82.32	**77.92**

Table 4: From LIP to PridB dataset. (%).

Methods	Avg. acc	Fg. acc	Avg. pre	Avg. rec	Avg. F1
Target Only	88.50	79.71	79.83	82.28	81.00
Source Only	84.46	80.01	72.85	80.01	75.63
DANN	83.91	**83.06**	71.55	**82.83**	75.83
Feat. Adapt	85.63	82.30	74.47	81.69	77.28
Lab. Adapt	84.62	80.54	73.13	80.42	75.89
Feat. + Lab. Adapt	**86.26**	82.39	**75.20**	81.62	**77.89**

PridA and *PridB* datasets are selected from camera view A and camera view B of Person Re-ID 2011 Dataset (Roth et al. 2014). Person Re-ID 2011 Dataset consists of images extracted from multiple person trajectories recorded from two different, static surveillance cameras.

Baseline & Evaluation We compare the proposed method is compared with following baseline methods.

Target Only: Since all of our target domains have pixel-level annotations, we train and test the parsing model directly on the target domains. We take the results, derived from accessing the full supervision, as performance upper bound for the cross-domain parsing models. In the following experiments, the basic model is the same as our feature extraction network and label predicting network.

Source Only: We apply the model trained on the source domain directly to the target domain, without any fine-tuning on the target domain datasets. It is a valid performance lower bound of the cross-domain methods.

DANN: There are a few works investigating cross-domain learning problems following the adversarial learning strategy. Here, we take the most competitive one proposed in (Ganin et al. 2016). It resolves the cross-domain problems on classifications. DANN uses an adversary network to make the features extracted from the source domain and target domains undistinguishable. The feature extraction network are shared for images from both domains. We adapt this method for the human parsing problem.

For ablation studies, we consider three variants of our method, to evaluate the contribution of each sub-network. **Feat. Adapt:** Our method with the Feature Adversarial network alone. **Lab. Adapt:** Our method with the Structured Label Adversarial network alone. **Feat. + Lab. Adapt:** Our method with both Feature Adversarial network and Structured Adversarial network.

We adopt five popular evaluation metrics, i.e., accuracy, foreground accuracy, average precision, average recall, and average F-1 scores over pixels (Yamaguchi, Kiapour, and Berg 2013). All these scores are obtained on the testing sets of the target domains. The annotations of target domains are

Table 5: F-1 Scores of each category from LIP to Indoor. (%).

Methods	bg	face	hair	U-clothes	L-arm	R-arm	pants	L-leg	R-leg	dress	L-shoe	R-shoe
Target Only	95.05	66.46	77.30	81.35	50.79	50.29	80.95	38.28	39.342	63.15	37.285	36.68
Source Only	94.17	58.89	59.10	77.51	**43.35**	**43.39**	75.06	35.16	**32.53**	26.55	24.11	26.54
DANN	94.48	**61.38**	65.26	78.41	42.01	41.74	78.83	32.84	25.53	35.56	23.76	26.19
Feat. Adapt	**94.53**	58.92	62.99	78.27	41.14	40.11	**79.42**	**41.49**	22.90	45.15	**26.53**	**27.69**
Lab. Adapt	94.48	57.71	63.32	78.60	41.20	41.22	79.06	38.99	22.64	45.52	25.90	22.70
Feat. + Lab. Adapt	94.49	56.73	**67.86**	**78.81**	42.79	42.64	78.97	36.25	22.86	**47.00**	25.32	27.00

Table 6: F-1 Scores of each category from LIP to Daily Video dataset. (%).

Methods	bg	face	hair	U-clothes	L-arm	R-arm	pants	L-leg	R-leg	dress	L-shoe	R-shoe
Target Only	94.50	69.06	57.37	68.16	46.33	42.37	65.01	59.97	60.35	67.06	41.85	44.72
Source Only	95.15	70.28	59.54	69.91	55.25	50.72	72.95	61.52	61.82	60.32	45.55	45.88
DANN	95.18	70.98	58.87	71.13	54.73	50.64	73.23	61.84	61.16	64.16	46.55	45.61
Feat. Adapt	95.35	**72.13**	55.99	73.01	56.55	52.38	73.08	60.60	62.91	61.72	**48.77**	48.24
Lab. Adapt	95.37	70.99	**59.66**	72.18	55.94	52.33	72.76	62.68	63.60	63.08	46.51	48.31
Feat. + Lab. Adapt	**95.38**	70.88	57.11	**73.04**	**57.05**	**53.92**	**73.34**	**64.80**	**64.73**	**64.80**	48.34	**48.97**

Table 7: F-1 Scores of each category from LIP to PridA dataset. (%).

Methods	bg	head	U-body	L-body
Target Only	93.90	76.63	81.83	76.11
Source Only	92.06	69.05	74.49	68.38
DANN	92.01	71.49	75.65	68.80
Feat. Adapt	92.28	71.61	76.71	68.96
Lab. Adapt	92.78	70.24	76.11	69.43
Feat. + Lab. Adapt	**92.83**	**72.01**	**76.72**	**70.14**

Table 8: F-1 Scores of each category from LIP to PridB dataset. (%).

Methods	bg	head	U-body	L-body
Target Only	92.88	77.23	80.38	73.51
Source Only	90.75	71.79	74.56	65.41
DANN	90.16	**72.73**	74.79	65.66
Feat. Adapt	91.30	72.59	76.95	68.29
Lab. Adapt	90.88	71.92	74.69	66.07
Feat. + Lab. Adapt	**91.59**	72.71	**78.27**	**68.99**

only used in the "Target Only" method.

Implementation details: The feature extractor and the pixel-wise labeler use the DeepLab model, with pre-trained models on PASCAL VOC. The other networks are initialized with "Normal" distribution.

During training of the feature adversarial adaption component, "Adam" optimizer is used with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate is $1e$-5. When training the structured label adaptation component, we use "Adam" optimizer with $\beta_1 = 0.5$, and $\beta_2 = 0.999$, while the learning rate is $1e$-8. The remaining networks are optimized via "SGD" optimizer with momentum of 0.9, learning rate $1e$-8 and weight decay of 0.0005. The whole framework is trained on PyTorch with a mini-batch size of 10. The input image size is 241×121. The experiments are done on a single NVIDIA GeForce GTX TITAN X GPU with 12GB memory. The constant K_C is 5 in our experiment.

Quantitative Results

Table 1 to 4 show the quantitative comparison of the proposed method with baseline methods. The best scores except those performed by "Target Only" (upper bound) are shown in black bold.

From these results, we can observe that the "Feat. + Lab. Adapt" method always outperforms about $1\% \sim 2\%$ than the method "Source Only" and "DANN" in the value "Avg. F-1", which verifies the effectiveness of the proposed cross-domain method. Note that, the "Avg. F-1" score of "Feat. + Lab. Adapt" is even higher than those of "Target Only" on the Daily Video dataset. We believe the reason is that the number of images in the training set is quite limited in this dataset and our proposed model is effective at transferring useful knowledge to address the sample-insufficiency issue. Besides, "Feat. Adapt" performs better than "Lab. Adapt" on the dataset Indoor, PridA, and PridB. This is from the fact that the features output by the "pool5" layer contain more sufficient characteristics, so the adversary network on these features has more influence on the whole performance.

The detailed "F-1" scores of each category are shown in Table $5 \sim 8$, which verify the effects of our method.

Qualitative Results

Some qualitative comparisons on the four target domains are shown in Figure 3.

For the dataset Indoor, back-view persons appear more frequently, and the illuminations are poor. Therefore, the predictions of left and right arms/shoes are often incorrect, and the hairs may be mis-predicted as backgrounds as well. For the persons in the 1st and 3rd rows of the dataset Indoor, the left and right arms are confused by "Source Only". The DANN performs slightly better, but our model is able to predict the left and right arms correctly. The hair of the second person is missed in both the "Source Only" and "DANN" methods, due to the dim lights of the image. The dress of the 4th person looks smaller because the camera is much higher

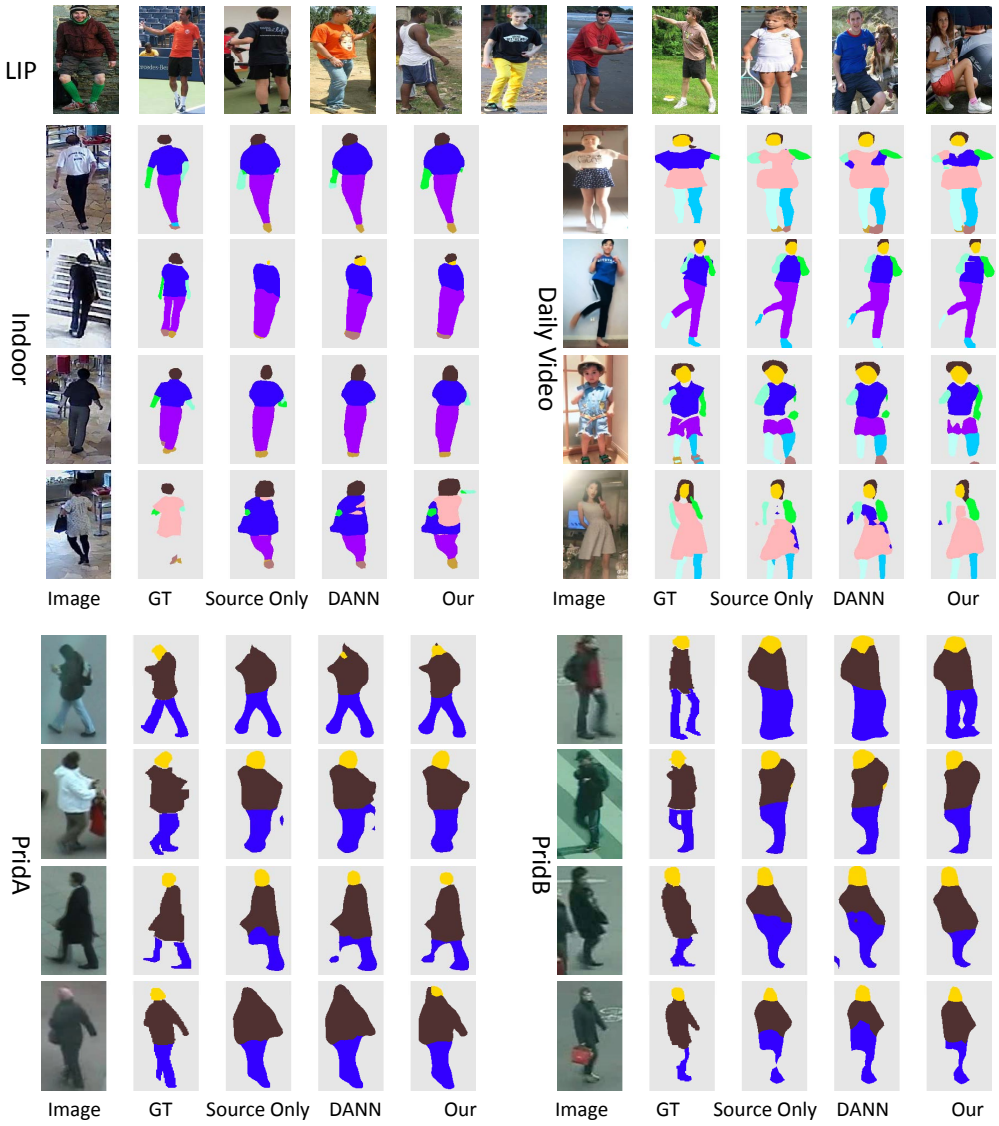

Figure 3: Qualitative Results on 4 target domains. "GT" stands for the groundtruth labels.

than the person. So "Source Only" and "DANN" methods wrongly predict them as upper clothes.

For the dataset Daily Video, cameras are put at general positions but the poses of persons are more challenging. People usually appear in frontal view, but they are often moving fast, e.g. the 2nd person, or in nonuniform illuminations, e.g. 3th and 4th persons. In these cases, the proposed model performs better, benefiting from the structure adversary network. Our method also performs better in predicting the classes of clothes, e.g. the 1st person.

The resolution of the dataset PridA and PridB are very low. As shown in Figure 3, our model and its variants also win in predicting details of the persons.

Conclusion

In this paper, we explored a new cross-domain human parsing problem: making use of the benchmark dataset with extensive labeling, how to build a human parsing for a new scenario without additional labels. To this end, an adversarial feature and structured label adaptation method were developed to learn to minimize the cross-domain feature differences and maximize the label commonalities across the two domains. In future, we plan to explore unsupervised domain adaptation when the target domain are unsupervised videos. The videos provide rich temporal context and can facilitate cross-domain adaptation. Moreover, we would like to try other types of GANs, such as WGAN (Arjovsky, Chintala, and Bottou 2017) in our network.

Acknowledgments

This work was supported by Natural Science Foundation of China (Grant U1536203, Grant 61572493, Grant 61572493), the Open Project Program of the Jiangsu Key Laboratory of Big Data Analysis Technology, Fundamental theory and cutting edge technology Research Program of Institute of Information Engineering, CAS(Grant No. Y7Z0241102) and Grant No. Y6Z0021102.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv:1701.07875*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2014a. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv:1412.7062*.

Chen, X.; Mottaghi, R.; Liu, X.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014b. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2016. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*.

Chen, Y.; Shen, C.; Wei, X.-S.; Liu, L.; and Yang, J. 2017. Adversarial posenet: A structure-aware convolutional network for human pose estimation. *arXiv:1705.00389*.

Chou, C.-J.; Chien, J.-T.; and Chen, H.-T. 2017. Self adversarial training for human pose estimation. *arXiv:1707.02439*.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR*.

Gong, K.; Liang, X.; Shen, X.; and Lin, L. 2017. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. *arXiv:1703.05446*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

Gretton, A.; Smola, A. J.; Huang, J.; Schmittfull, M.; Borgwardt, K. M.; and Schölkopf, B. 2009. Covariate shift by kernel mean matching.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person search with natural language description. *arXiv:1702.05729*.

Liang, X.; Liu, S.; Shen, X.; Yang, J.; Liu, L.; Dong, J.; Lin, L.; and Yan, S. 2015a. Deep human parsing with active template regression. *TPAMI*.

Liang, X.; Xu, C.; Shen, X.; Yang, J.; Liu, S.; Tang, J.; Lin, L.; and Yan, S. 2015b. Human parsing with contextualized convolutional neural network. *ICCV*.

Liu, M.-Y., and Tuzel, O. 2016. Coupled generative adversarial networks. In *NIPS*.

Liu, S.; Feng, J.; Domokos, C.; Xu, H.; Huang, J.; Hu, Z.; and Yan, S. 2014. Fashion parsing with weak color-category labels. *TMM*.

Liu, S.; Liang, X.; Liu, L.; Shen, X.; Yang, J.; Xu, C.; Lin, L.; Cao, X.; and Yan, S. 2015. Matching-cnn meets knn: Quasi-parametric human parsing. *CVPR*.

Liu, S.; Wang, C.; Qian, R.; Yu, H.; and Bao, R. 2016. Surveillance video parsing with single frame supervision. *arXiv:1611.09587*.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *ICML*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. *CVPR*.

Luo, P.; Wang, X.; and Tang, X. 2013. Pedestrian parsing via deep decompositional network. In *ICCV*.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Smolley, S. P. 2016. Least squares generative adversarial networks. *arXiv:1611.04076*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*.

Roth, P. M.; Hirzer, M.; Koestinger, M.; Beleznai, C.; and Bischof, H. 2014. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*.

Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J.; Wang, W.; and Webb, R. 2016. Learning from simulated and unsupervised images through adversarial training. *arXiv:1612.07828*.

Simo-Serra, E.; Fidler, S.; Moreno-Noguer, F.; and Urtasun, R. 2014. A high performance crf model for clothes parsing. In *ACCV*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.

Wei, Y.; Xia, W.; Lin, M.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; and Yan, S. 2016. Hcp: A flexible cnn framework for multi-label image classification. *TPAMI*.

Wei, Y.; Liang, X.; Chen, Y.; Shen, X.; Cheng, M.-M.; Feng, J.; Zhao, Y.; and Yan, S. 2017. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *TPAMI*.

Xia, F.; Wang, P.; Chen, L.-C.; and Yuille, A. L. 2015. Zoom better to see clearer: Human part segmentation with auto zoom net. *arXiv:1511.06881*.

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

Yamaguchi, K.; Kiapour, M. H.; Ortiz, L. E.; and Berg, T. L. 2012. Parsing clothing in fashion photographs. In *CVPR*.

Yamaguchi, K.; Kiapour, M. H.; and Berg, T. 2013. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*.

Zhao, R.; Ouyang, W.; and Wang, X. 2014. Learning mid-level filters for person re-identification. In *CVPR*.

Zhu, S.; Fidler, S.; Urtasun, R.; Lin, D.; and Loy, C. C. Be your own prada: Fashion synthesis with structural coherence.