Pacific Graphics (2017)
J. Barbic, W.-C. Lin, and O. Sorkine-Hornung (Editors)

Short Paper

Transferring Pose and Augmenting Background Variation for Deep Human Image Parsing

Takazumi Kikuchi Yuki Endo Yoshihiro Kanamori Taisuke Hashimoto Jun Mitani

University of Tsukuba

Abstract

Human parsing is a fundamental task to estimate semantic parts in a human image such as face, arm, leg, hat, and dress. Recent deep-learning based methods have achieved significant improvements, but collecting training datasets of pixel-wise annotations is labor-intensive. In this paper, we propose two solutions to cope with limited dataset. First, to handle various poses, we incorporate a pose estimation network into an end-to-end human parsing network in order to transfer common features across the domains. The pose estimation network can be trained using rich datasets and feed valuable features to the human parsing network. Second, to handle complicated backgrounds, we increase the variations of background images automatically by replacing the original backgrounds of human images with those obtained from large-scale scenery image datasets. While each of the two solutions is versatile and beneficial to human parsing, their combination yields further improvement.

CCS Concepts

•***Computing methodologies*** → *Image segmentation; Image processing;*

1. Introduction

Human parsing is an image-processing task to assign semantic labels to human body parts and clothing regions, including face, arm, leg, hat, dress, etc. This task plays a crucial role in various applications in the fields of computer graphics as well as computer vision, e.g., virtual fitting systems [YHK*14], clothing retrieval [WCA*13], and recommendation [HYD15, KKL13].

Recent human-parsing methods using deep learning have exhibited significant improvements. Such methods require sufficiently-large training dataset in order to cope with various human poses and complicated background images. The problem here is that, if sufficient training data cannot be obtained, this approach is difficult to deal with various human poses and complicated backgrounds, and thus its performance degrades. A straightforward solution is to manually annotate pixel-wise labels for increasing training dataset, which is quite tedious and costly if we employ crowd sourcing. This fact leads us to the following research question: *"Can we improve human parsing using limited training dataset?"*

In this paper, we answer the research question through the following two solutions. First, to handle various poses, we exploit transfer learning with human pose estimation. For pose estimation, the required data is joint-wise annotations, which are easier to collect than pixel-wise annotations for human parsing. The key idea is to integrate human pose estimation into an end-to-end network model for human parsing in order to transfer information of human pose estimation to the human parsing network across domains that share a common feature space. While this idea can be accomplished in various ways, as a proof of concept, we adopt the relatively-simple, state-of-the-art convolutional neural networks (CNNs) for human pose estimation [WRkS16] and human parsing [LLS*16]. Although other deep-learning-based methods for human parsing do not consider pose information explicitly, the explicit integration of this human-specific knowledge is actually beneficial to human parsing. Second, we propose a simple yet effective data-augmentation method for human parsing. To handle various background images, we automatically replace the backgrounds of existing labeled data with new background images obtained from public large-scale datasets for scene recognition, e.g., [QT09]. While each of our solutions boosts the accuracy of human parsing, combination of both yields further improvement. We demonstrate the effectiveness of our approach by comparing existing CNN-based methods quantitatively and qualitatively.

2. Related Work

The early methods for human parsing use conditional random fields (CRFs). Yamaguchi et al. proposed a seminal work for human parsing, which learns human pose and segmentation mutually [YKOB12]. They also improved the performance of human parsing by using tag information of similar images retrieved by k-nearest neighbor search [YKOB14]. Simo-Serra et al. improved Yamaguchi et al.'s method [YKOB12] by considering the position and shape of superpixels [SSFMNU14]. Instead of using CRFs, Dong et al. presented a novel framework called the Hybrid Parsing Model [DCS*14], which unifies human image parsing

DOI: 10.2312/pg.20171317

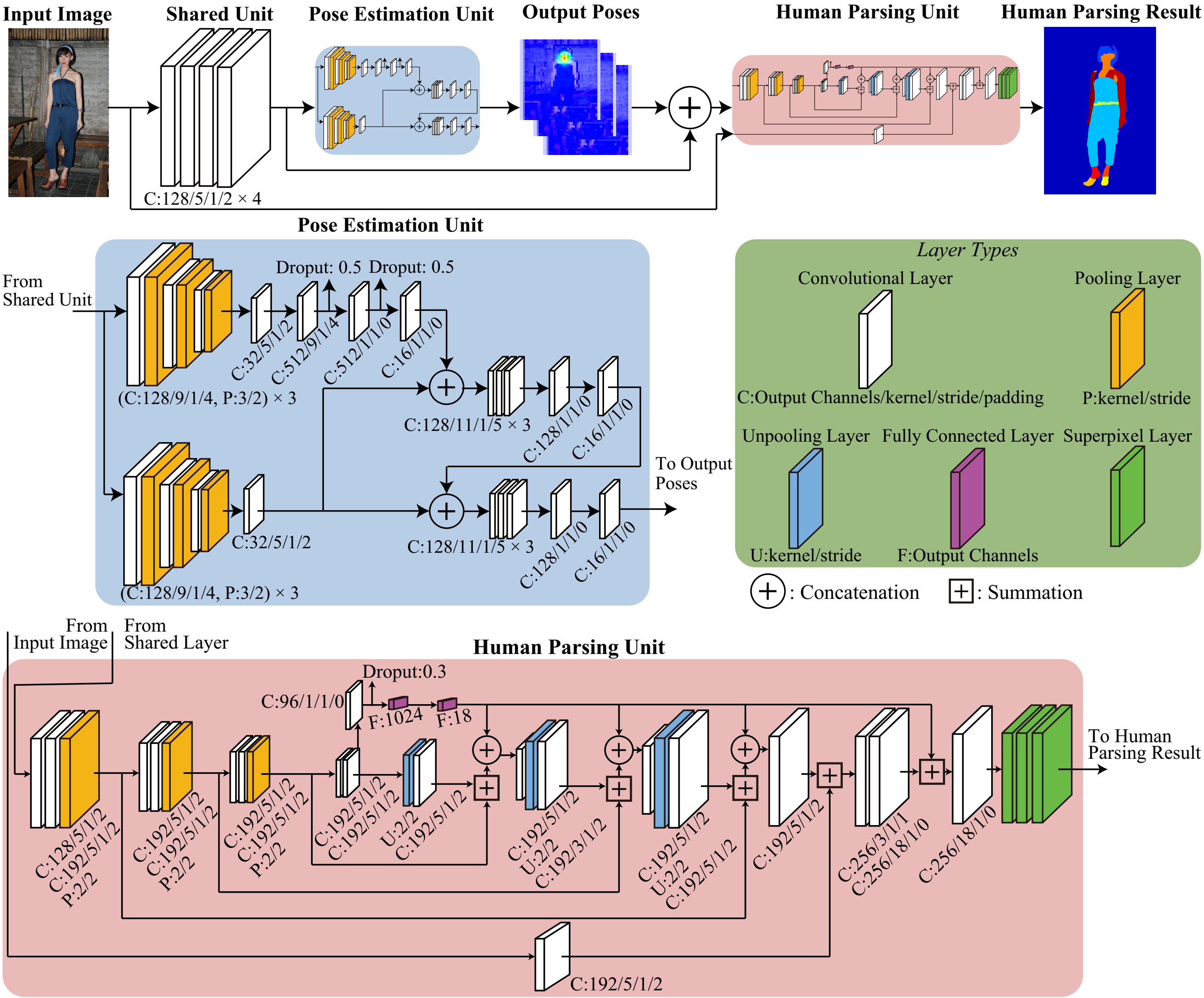


Figure 1: *Illustration of our network model. Given an input image, image features are extracted in the shared unit. The human pose is then estimated as joint-wise heatmaps in the pose estimation unit. The outputs of the shared and pose estimation units are concatenated. Finally, the human parsing unit outputs a labeled result from the concatenated features.*

and pose estimation. Such a unified approach was also applied to videos [LLL*14].

In recent years, deep learning-based methods have achieved significant improvements. Liang et al. first used a CNN for human parsing [LLS*15], and then they developed a novel network called Contextualized CNN (Co-CNN) which crosses the output of each layer and global image features [LLS*16]. Liu et al. proposed Matching CNN, which uses as input a target image and its similar image retrieved by k-nearest neighbor search [LLL*15].

Human parsing is a specific type of semantic object segmentation, for which various CNN-based methods have been proposed [GJL16, GF16, LSvdHR16, ROMYR16, LSX*16, LSF*16]. In particular, there are CNN-based methods that use training datasets in different domains. Dai et al. proposed a network called multi-task network cascades (MNCs), which processes multiple tasks (object detection, mask extraction, and semantic labeling) in a single network [DHS16]. Besides, Hong et al. proposed to learn semantic segmentation and image classification in the same network [HOLH16]. Papandreou et al. developed an Expectation-Maximization method [PCMY15] for training with many weak annotation training data such as many bounding boxes, image level labels, and small number of pixel-level semantic segmentation data.

As for pose estimation, there have been several methods that use CNNs, for example, methods of estimating poses by a simple model consisting of convolution and pooling layers [WRkS16], by incorporating geometric prior knowledge of the body into a

CNN framework [WWHX16], and by inferring correlation between joints [XWHX16].

The main contributions of this paper are to integrate human pose estimation into human parsing in the context of CNN as well as to increase background image variations automatically. Both approaches can be easily integrated into the existing deep-learning methods to improve human parsing even if only small dataset of pixel-wise annotations is available. Although human poses have previously been exploited in the CRF-based methods [YKOB12, YKOB14] and other methods [DCS*14, LLL*14], ours is the first attempt to explicitly integrate such information into deep neural networks, to the best of our knowledge.

3. Background

This section reviews the existing methods for human pose estimation [WRkS16] and human image parsing [LLS*16] , which we adopted in our architecture.

3.1. Convolutional Pose Machines

Convolutional pose machines [WRkS16] define a partial network consisting of the convolutional layer and pooling layer as one stage to obtain a heatmap for each joint. This stage is repeated multiple times to improve output human poses represented as heat maps. For example, the pose estimation unit in Fig. 1 has three stages. The network is learned by minimizing loss functions for the multiple stages to avoid the vanishing gradient problem due to its deep architecture. This network structure can be easily integrated into our framework because it simply consists of the convolutional and pooling layers and enables end-to-end learning.

3.2. Contextualized CNN (Co-CNN)

Co-CNN [LLS*16] is a neural network for human image parsing. To improve performance of human parsing, it learns global as well as local features based on cross-layer context and global image-level context. As illustrated in the human parsing unit in Fig. 1, the cross-layer context is captured by using the skip connections between down-convolution and up-convolution layers from fine to coarse scales. On the other hand, the global image-level context is captured by using the fully-connected layers which predict image-level labels on the entire image. The predicted image-level labels are subsequently concatenated with each input to be unpooled. In addition, Co-CNN accounts for the local superpixel context. To capture this context, it has three layers at the end of the network: within-superpixel smoothing, cross-superpixel neighbor voting, and pixel-wise prediction. These layers can retain local label consistency by using superpixel over-segmentation map.

4. Proposed Method

This section describes our network that transfers information of human pose estimation to the human parsing domain as well as our approach for background augmentation.

4.1. Transferring Information of Pose Estimation

To deal with various human poses, our approach first estimates human poses before human parsing, and then assigns labels to each pixel of an input image using the result of pose estimation. Fig. 1 shows our network model. First, the input image is fed to the shared unit, and low and mid-level features are extracted. The shared unit consists of four convolutional layers with the same kernel size of 5×5, stride of 1, padding of 2, and output channels of 128. Next, features extracted in the shared unit are fed to the pose estimation unit. The network structure of the pose estimation unit is built based on the network by Wei et al. [WRkS16]. In this network, the partial network consisting of the convolutional layer and the pooling layer is defined as one stage, and human pose estimation is improved gradually by repeating this stage multiple times. After that, the pose estimated on the pose estimation unit is concatenated with the output of the shared unit. The concatenated features are then fed to the human parsing unit, and finally the network outputs a labeled image. In the human parsing unit, we use the Co-CNN model [LLS*16]. The Co-CNN model outputs a global distribution of labels through the fully connected layers after the convolutional layers. On the other hand, the human parsing result is calculated via the deconvolutional layers, and the final result is obtained by superpixel-based smoothing.

4.1.1. Learning

We train the proposed network using pose estimation and human parsing datasets. For the pose estimation dataset, the parameters θ_s and θ_p of the shared unit and pose estimation unit are optimized by minimizing the following error function:

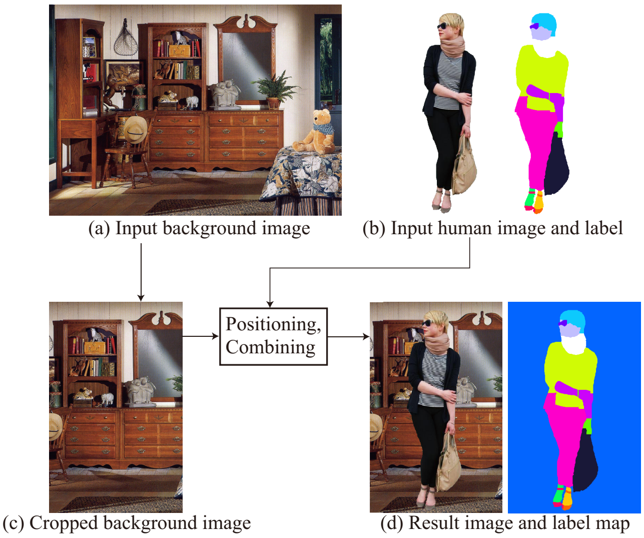
$$E_p = \sum_{\{\mathbf{b}_i, \mathbf{b}_l\} \in \mathcal{B}} \sum_{t=1}^{T} \sum_{j=1}^{J} \|\mathbf{b}_l^j - \mathbf{B}_t^j(\mathbf{b}_i; \theta_s, \theta_p)\|_2^2, \qquad (1)$$

where $\mathcal{B}$ is the dataset of pose estimation containing the input image $\mathbf{b}_i$ and ground-truth joint heatmap $\mathbf{b}_l$. T is the number of repeating stages, J is the number of joints to be estimated, and $\mathbf{B}$ is the joint heatmap estimated by the pose estimation unit. The ground-truth joint heatmaps are generated by a Gaussian function $\exp(-\|\mathbf{x} - \mu_j\|^2/\sigma^2)$ for position $\mathbf{x}$, where μ_j is the position of joint j and $\sigma = 2$.

For the human parsing dataset, instead of the error function (1) defined for pose estimation, the parameter θ of the entire network is optimized by minimizing the following error function:

$$\begin{aligned} E_l &= E_l^{orig} + E_l^{accel}, \\ E_l^{orig} &= -\sum_{\{\mathbf{d}_i, \mathbf{d}_l\} \in \mathcal{D}} \sum_{j}^{M} \sum_{k}^{L} \mathbf{d}_{l_{jk}} \ln(\mathbf{F}_{jk}(\mathbf{d}_i; \theta)) \\ &\quad + \sum_{\{\mathbf{d}_i, \mathbf{d}_{l'}\} \in \mathcal{D}} ||\mathbf{d}_{l'} - \mathbf{H}(\mathbf{d}_i; \theta)||^2, \\ E_l^{accel} &= -\sum_{\{\mathbf{d}_i, \mathbf{d}_l\} \in \mathcal{D}} \sum_{j}^{N} \sum_{k}^{L} \mathbf{d}_{l_{jk}} \ln(\mathbf{G}_{jk}(\mathbf{d}_i; \theta)), \end{aligned} \qquad (2)$$

where E_l^{orig} is similar to the error function used in [LLS*16]. By adding E_l^{accel}, we found that the convergence is accelerated. $\mathcal{D}$ is the human parsing dataset containing the input image $\mathbf{d}_i \in \mathbf{R}^{h \times w \times c}$,



(a) Input background image (b) Input human image and label

(c) Cropped background image (d) Result image and label map

Figure 2: *Procedure of background augmentation.*

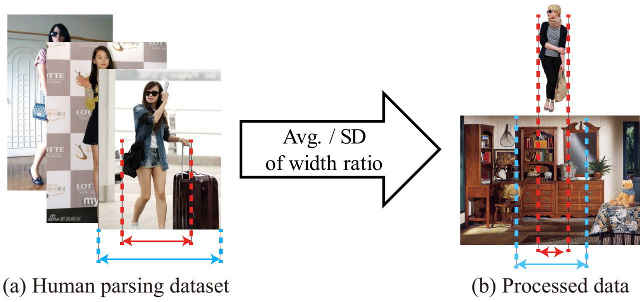
ground-truth labeled image $\mathbf{d}_l \in R^{h\times w\times L}$, and global class distribution $\mathbf{d}_{l'} \in \mathbf{R}^L$ in the entire image. w and h are the width and height of an input image, c is the number of channels of the input image, M is the number of superpixels in the image, N is the number of pixels in the image, and L is the number of class labels (i.e., $L = 18$, the same as [LLS*16]). $\mathbf{F}$ is the output of the human parsing unit, $\mathbf{G}$ is the output before superpixel processing of the human parsing unit, and $\mathbf{H}$ is the output after the fully connected layers.

For training the network, we divide one epoch of the learning procedure into two steps. In the first step, we optimize the model parameters of the shared unit and pose estimation unit on the basis of E_p by using the pose estimation dataset. In the second step, we optimize the model parameters of the "entire" network on the basis of E_l by using the human parsing dataset. We used the MomentumSGD optimizer of a learning rate of 0.001, momentum term of 0.9, and weight decay term of 0.0005.

4.2. Augmenting Background Variations

To make human parsing robust against various backgrounds, we augment background patterns in a training dataset. Specifically, we cut out foreground human regions from labeled images and paste them on new background images obtained from scenery image datasets.

Fig. 2 illustrates how to augment the dataset. Inputs are a pair of a cut-out human image and its corresponding label map (Fig. 2(b)) as well as a new background image (Fig. 2(a)). Because most background images are horizontally long, we trim background images so that the width ratios between cut-out human images and background images become consistent to those in the original dataset for human parsing (Fig. 2(c)). Fig. 3 shows the detailed procedure. First, in the original dataset for human parsing, we calculate statistics (i.e., the means and standard deviations) of the relative width and relative position of the human region in each image. We then determine the new background width for trimming as well as the new position of the cut-out human image according to normal distributions defined by the original statistics. With the determined width, we crop the left and right sides of the input background image. The position of cropping can also be determined randomly. Finally, we paste the cut-out human image onto the cropped background while placing the human label map at the same position (Fig. 2 (d)). This technique reasonably scales human images. Our data augmentation plays an important role to increase background variations and to improve the performance of human parsing, as demonstrated in the evaluation section.



(a) Human parsing dataset (b) Processed data

Figure 3: *Details of how to trim background images.*

5. Experiments

This section describes evaluation experiments that compare the proposed approaches with the baseline method Co-CNN [LLS*16].

5.1. Experimental Settings

In the pose estimation unit, the number of stages (Section 3.1) is six in the original method [WRkS16] but it is set to three in our unit in order to reduce computational time and GPU memory usage. For the human parsing network, the existing method [LLS*16] uses several types of features to calculate the similarity between superpixels. However, we only use the RGB feature because the implementation details on other features, e.g., the HOG feature on each superpixel, are not clearly presented in their paper and their source codes are not publicly available. We implemented our method and the baseline method using Python language and Chainer library and ran the codes on a PC with NVIDIA GeForce GTX 1080. The calculation time of the model as a whole was about 0.028 seconds when it was averaged with 1,000 test data.

For the human parsing dataset, we used the ATR dataset [LLS*15]. This dataset contains 7,702 images, and we used 6,000 images for training, 702 images for validation, and 1,000 images for testing. For the background dataset used for data augmentation, we randomly selected 6,000 images from the indoor scene recognition dataset [QT09] and doubled the 6,000 training images of the ATR dataset by synthesizing these background images. For the pose estimation dataset, we used the MPII Human Pose Dataset [APGS14]. This dataset contains 24,984 images, and we used 10,298 images assigned annotation of training data and including only one human in dataset for learning.

Table 1: *Comparisons of F1-score of each class for each method.*

Method	bg	Hat	Hair	glass	U-cloth	Skirt	Pants	Dress	Belt	L-shoe	R-shoe	Face	L-leg	R-leg	L-arm	R-arm	bag	scarf
Co-CNN1000	93.89	**4.17**	52.46	4.08	51.40	9.63	37.41	26.66	4.14	25.44	25.57	61.42	42.66	41.32	31.22	27.72	12.81	0.46
DA1000	94.76	3.11	57.70	**9.39**	55.02	9.11	32.32	**32.48**	**4.32**	30.33	30.95	64.23	47.41	46.55	33.03	34.19	15.30	**1.03**
PE1000	95.54	0.29	61.34	0.52	60.96	21.48	40.65	30.49	0.00	**38.26**	35.75	72.23	48.85	50.18	**41.94**	39.14	28.93	0.00
DA+PE1000	**96.18**	0.50	**63.06**	0.00	**62.88**	**36.31**	**49.50**	16.23	0.46	36.41	**38.86**	**73.22**	**54.51**	**54.64**	41.65	**43.45**	**34.54**	0.00
Co-CNN6000	95.73	18.15	66.37	14.04	64.09	23.83	49.39	37.26	7.05	39.77	40.59	74.08	58.13	58.12	48.27	47.39	35.90	**3.56**
DA6000	95.93	0.15	68.28	8.00	63.89	28.76	50.83	36.67	4.50	35.96	39.70	73.62	57.82	57.54	47.50	47.01	36.99	0.37
PE6000	97.20	40.30	74.71	18.87	69.64	41.57	**61.55**	50.75	**21.56**	44.85	45.09	80.54	65.39	64.31	62.16	61.70	48.58	0.03
DA+PE6000	**97.55**	**45.58**	**77.22**	**31.31**	**74.46**	**47.49**	61.40	**51.67**	16.73	**45.72**	**46.09**	**82.44**	**67.11**	**66.89**	**65.07**	**63.25**	**53.32**	0.10

Table 2: *Performance comparisons for each method by learning 1,000 and 6,000 training data.*

Method	Acc.	Pre.	Rec.	F1
Co-CNN1000	82.07	79.14	82.07	80.19
DA1000	83.27	81.64	83.28	81.81
PE1000	84.77	83.06	84.77	83.49
DA+PE1000	**85.18**	**84.67**	**85.18**	**84.43**
Co-CNN6000	86.15	84.79	86.15	84.95
DA6000	86.16	84.78	86.16	85.15
PE6000	88.31	88.82	89.00	88.41
DA+PE6000	**89.73**	**89.46**	**89.73**	**89.37**

We used 100 × 150 images as input similarly to [LLS*16], in the baseline method and when we use only the proposed data augmentation method. In the case of using the proposed network including the pose estimation part, we used 256 × 256 images as input because the size of the input image must be power of two so that the size of the output image of the pose estimation does not change. All generated results are finally resized to their original size.

5.2. Evaluation Methods

We compared the baseline method (Co-CNN) [LLS*16], our data augmentation method (DA), and the proposed network, which uses pose estimation information (PE). As evaluation metrics, we used accuracy, precision, recall, and F1. To verify the effectiveness of the proposed method depending on the amount of training data, we conducted experiments by training different number of the training data of human parsing, that is, 1,000 and 6,000 images. We stopped learning when the error function in Eq. (2) converges and used the models with the maximum accuracy for validation.

Note that faithful reproduction of the Co-CNN performance [LLS*16] is almost impossible for anyone but the authors of [LLS*16]; first, their source codes are unavalable. Second, the choices of test data, training data or validation data are not revealed. Third, several implmentation details are missing, as mentioned in Section 5.1. Nonetheless, our goal here is to answer our research question; we demonstrate that our method designed for small dataset outperforms the baseline.

5.3. Results

Tab. 2 shows the performance of each method for the test data. The numbers after the name of each method indicate the number of training data. As can be seen in the results of data augmentation, the performance improved over the result of Co-CNN when the number of training data was 1,000. On the other hand, the performance difference was marginal when the number of training data was 6,000. This is natural because the more training images, the more variations in background images. Our purpose is to improve the performance of human parsing in case of limited training data, and our background augmentation approach is effective for this purpose. Moreover, in the case of transferring pose estimation information to the human parsing part, the performance improved in both of the cases of using 1,000 and 6,000 training data. Furthermore, as shown in Tab. 1, a similar tendency was confirmed for F1 of each class. In particular, when the number of training data was small, our data augmentation method outperformed the baseline for the multiple classes including the background (bg). Also, even when the number of the training data was large, the proposed network based on pose estimation significantly outperformed the baseline for all labels except scarf. As shown in Fig. 4, we qualitatively compared the results. In these results, we can confirm that our data augmentation method successfully classified the background and foreground, and the proposed network based on pose estimation accurately extracted the human body parts.

6. Conclusion and Future Work

In this paper, we have proposed a novel data augmentation method and a novel neural network that transfers pose estimation information to the human parsing domain. We also demonstrated comparisons with previous work and verified the data augmentation method and pose estimation-based network were effective for human parsing. Although the proposed method improved the accuracies for most of the classes, the accuracies for the classes that exist in small regions in images like scarf were low. In the future, we would like to improve performance for specific classes with small training data. As shown in [HBGL08], we want to be able to deal with even less data by sampling biased data to be even.

References

[APGS14] ANDRILUKA M., PISHCHULIN L., GEHLER P., SCHIELE B.: 2d human pose estimation: New benchmark and state of the art analysis. In *Proc. of CVPR 2014* (2014), pp. 3686–3693. 4

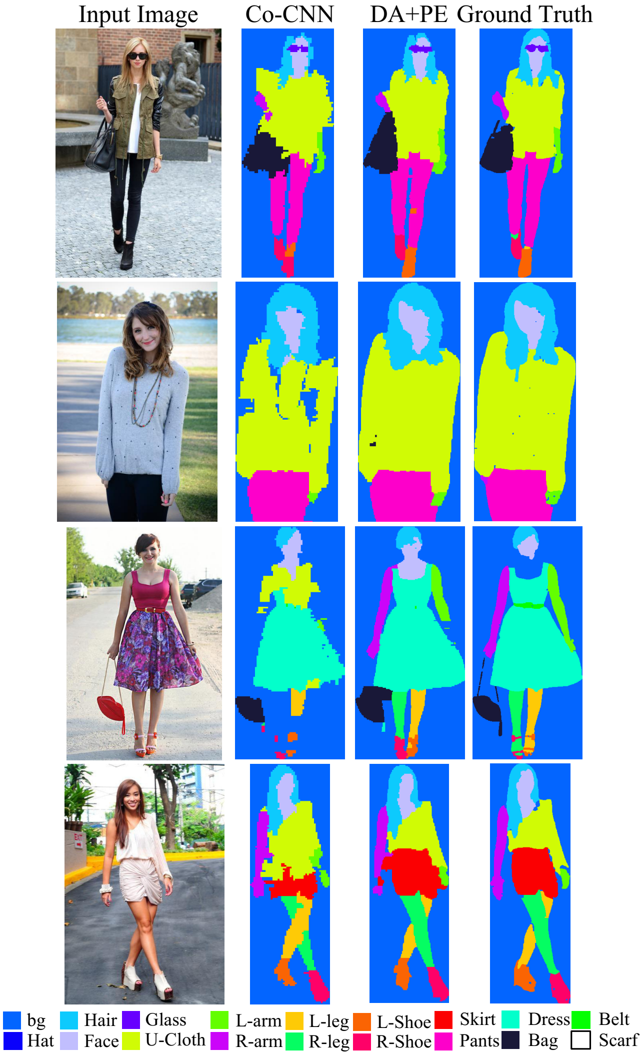


Figure 4: *Comparisons of the resultant images obtained by each method.*

[DCS*14] Dong J., Chen Q., Shen X., Yang J., Yan S.: Towards unified human parsing and pose estimation. In *Proc. of CVPR 2014* (2014), pp. 4321–4328. 1, 3

[DHS16] Dai J., He K., Sun J.: Instance-aware semantic segmentation via multi-task network cascades. In *Proc. of CVPR 2016* (2016), pp. 3150–3158. 2

[GF16] Ghiasi G., Fowlkes C. C.: Laplacian pyramid reconstruction and refinement for semantic segmentation. In *Proc. of ECCV 2016* (2016), pp. 519–534. 2

[GJL16] Gedas B., Jianbo S., Lorenzo T.: Semantic segmentation with boundary neural fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 3602–3610. 2

[HBGL08] He H., Bai Y., Garcia E. A., Li S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Proc. of IEEE International Joint Conference on Neural Networks (IJCNN) 2008* (2008), pp. 1322–1328. 5

[HOLH16] Hong S., Oh J., Lee H., Han B.: Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *Proc. of CVPR 2016* (2016), pp. 3204–3212. 2

[HYD15] Hu Y., Yi X., Davis L. S.: Collaborative fashion recommendation: A functional tensor factorization approach. In *Proc. of ACM international conference on Multimedia* (2015), pp. 129–138. 1

[KKL13] Kalantidis Y., Kennedy L., Li L.-J.: Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proc. of ACM conference on international conference on multimedia retrieval* (2013), pp. 105–112. 1

[LLL*14] Liu S., Liang X., Liu L., Lu K., Lin L., Yan S.: Fashion parsing with video context. In *Proc. of ACM Multimedia 2014* (2014), pp. 467–476. 2, 3

[LLL*15] Liu S., Liang X., Liu L., Shen X., Yang J., Xu C., Lin L., Cao X., Yan S.: Matching-CNN meets KNN: Quasiparametric human parsing. In *Proc. of CVPR 2015* (2015), pp. 1419–1427. 2

[LLS*15] Liang X., Liu S., Shen X., Yang J., Liu L., J. Dong L. L., Yan S.: Deep human parsing with active template regression. *IEEE Trans. on PAMI 37*, 12 (2015), 2402–2414. 2, 4

[LLS*16] Liang X., Liu S., Shen X., Yang J., Liu L., Dong J., Lin L., Yan S.: Human parsing with contextualized convolutional neural network. *IEEE Trans. on PAMI 39*, 1 (2016), 115–127. 1, 2, 3, 4, 5

[LSF*16] Liang X., Shen X., Feng J., Lin L., Yan S.: Semantic object parsing with graph lstm. In *Proc. of ECCV 2016* (2016). 2

[LSvdHR16] Lin G., Shen C., van den Hengel A., Reid I.: Efficient piecewise training of deep structured models for semantic segmentation. In *Proc. of CVPR 2016* (2016), pp. 3194–3203. 2

[LSX*16] Liang X., Shen X., Xiang D., Feng J., Lin L., Yan S.: Semantic object parsing with local-global long short-term memory. In *Proc. of CVPR 2016* (2016). 2

[PCMY15] Papandreou G., Chen L., Murphy K. P., Yuille A. L.: Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proc. of ICCV 2015* (2015), pp. 1742–1750. 2

[QT09] Quattoni A., Torralba A.: Recognizing indoor scenes. In *Proc. of CVPR 2009* (2009), pp. 413–420. 1, 4

[ROMYR16] Raviteja V., Oncel T., Ming-Yu L., Rama C.: Gaussian conditional random field network for semantic segmentation. In *Proc. of CVPR 2016* (2016), pp. 3224–3233. 2

[SSFMNU14] Simo-Serra E., Fidler S., Moreno-Noguer F., Urtasun R.: A high performance crf model for clothes parsing. In *Proc. of ACCV 2014* (2014), pp. 869–877. 1

[WCA*13] Wei D., Catherine W., Anurag B., Robinson P., Neel S.: Style finder: Fine-grained clothing style detection and retrieval. In *Proc. of CVPR Workshops* (2013), pp. 8–13. 1

[WRkS16] Wei S., Ramakrishna V., kanade T., Sheikh Y.: Convolutional pose machines. In *Proc. of CVPR 2016* (2016), pp. 4724–4732. 1, 2, 3, 4

[WWHX16] Wei Y., Wanli O., Hongsheng L., Xiaogang W.: End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *Proc. of CVPR 2016* (2016), pp. 3073–3082. 3

[XWHX16] Xiao C., Wanli O., Hongsheng L., Xiaogang W.: Structured feature learning for pose estimation. In *Proc. of CVPR 2016* (2016), pp. 4715–4723. 3

[YHK*14] Yamada H., Hirose M., Kanamori Y., Mitani J., Fukui Y.: Image-based virtual fitting system with garment image reshaping. In *Cyberworlds 2014* (2014), pp. 47–54. 1

[YKOB12] Yamaguchi K., Kiapour M., Ortiz L., Berg T.: Parsing clothing in fashion photographs. In *Proc. of CVPR 2012* (2012), pp. 3570–3577. 1, 3

[YKOB14] Yamaguchi K., Kiapour M., Ortiz L., Berg T.: Retrieving similar styles to parse clothing. *IEEE Trans. on PAMI 37*, 5 (2014), 1028–1040. 1, 3