A Human Body Part Semantic Segmentation Enabled Parsing for Human Pose Estimation

Aditi Verma*
International Institute of Information Technology
Naya Raipur, India
aditi21300@iiitnr.edu.in

Vivek Tiwari
International Institute of Information Technology
Naya Raipur, India
vivek@iiitnr.edu.in

Mayank Lovanshi
International Institute of Information Technology
Naya Raipur, India
mayank@iiitnr.edu.in

Rahul Shrivastava
Sagar Institute of Science, Technology & Research
Bhopal, India
rahul.vidishaa@gmail.com

ABSTRACT

Human Body Part Semantic Segmentation and Human Pose estimation are considered to be essential for understanding human behaviours. Both of these tasks are correlated with each other. Employing them together in a unified framework to perform two distinct Human Centric Visual Analysis tasks simultaneously allows benefiting from each other. Taking advantage of the correlation between Human Body Part Semantic Segmentation and Human Pose Estimation, this paper proposes a unified framework that explores efficient context modelling. The framework simultaneously predicts the human body part semantic segmentation and pose estimation with high-quality results. The results extracted from the segmentation are used to predict the pose estimation task. An experimental analysis of the proposed framework is done on the benchmark LIP Dataset. The analysis of the results shows that the proposed framework outperforms the state-of-the-art by 7.3% when evaluated on mean IoU. Moreover, Mean Accuracy, Pixel Accuracy and PCKh are the other metrics used for the evaluation of the framework.

CCS CONCEPTS

• **Computing methodologies → Image segmentation**; • **Human-centered computing → Visualization**.

KEYWORDS

Human part semantic segmentation, Look into person, Human pose detection, Human parsing

ACM Reference Format:
Aditi Verma, Vivek Tiwari, Mayank Lovanshi, and Rahul Shrivastava. 2023. A Human Body Part Semantic Segmentation Enabled Parsing for Human Pose Estimation. In *2023 5th International Conference on Image, Video and Signal Processing (IVSP 2023), March 24–26, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3591156.3591162

1 INTRODUCTION

Human Body Part Semantic Segmentation is also known as Human Parsing. This technique segments the human body into fine-grained components like hat, hair, face, pants, coat, right and left legs, left and right arm and many more, along with the background. Due to factors like intricate patterns and textures of clothing, changeable positions of people, and the scale variation of various semantic pieces, human part semantic segmentation falls under scene parsing, where pixel categorization is carried out for particular images. With the development of Deep Convolution Neural Networks (DCNN) and large-scale as well as comprehensive datasets, human-part semantic segmentation has recently gained significant interest and made remarkable strides. Most earlier research, such as dilated convolution, LSTM [23] structure, encoder-decoder architecture, and human posture restrictions, concentrate on creating new structures and providing additional information guidance to enhance general feature representation. However, they directly use one flat prediction layer to classify all labels, ignoring the inherent semantic relationships between concepts and wastefully using the annotations, even though these methods produce promising results on each human part semantic segmentation dataset. The human part semantic segmentation approach is comparable to semantic segmentation [4, 27], which anticipates the labelling of each pixel in the scene.

The goal of Human Pose Estimation (HPE) is to create a skeleton-like illustration of the human body, further processing it for task-specific applications. HPE presents geometric and motion information that is used to apply to a wide range of applications like human-computer interaction, motion analysis, virtual reality, healthcare, augmented reality, and many more. Human pose estimation consists of three different approaches Volume Based, Contour Based, and Skeleton Based [19] approach. This paper includes work implemented on the Skeleton-Based Model. Using deep learning techniques in HPE jobs has already resulted in notable advancements in performance. Occlusion, a lack of training data, and depth ambiguity are obstacles that must be solved. With the aid of deep learning approaches, high performance has been attained for the human pose estimate[10, 14] of a single person utilizing 2D HPE from pictures and videos with 2D pose annotations. Recently, highly occluded multi-person HPE in complicated situations has received some attention. In comparison, it is significantly harder to generate correct 3D pose annotations for 3D HPE than for 2D HPE. Motion

IVSP 2023, March 24–26, 2023, Singapore, Singapore

ACM ISBN 978-1-4503-9838-1/23/03…$15.00
https://doi.org/10.1145/3591156.3591162

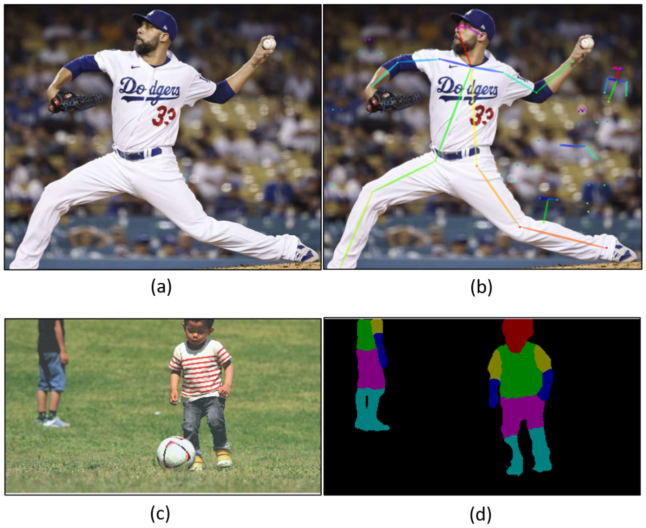

Figure 1: Example of pose estimation and Human Body Part Semantic Segmentation. Figures (a) and (c) are the original images, while (b) and (d) are the predicted result of pose estimation and human body part semantic segmentation, respectively.

capture devices can gather 3D position annotation in controlled lab settings but are less effective in natural settings.

Pose estimation and human body semantic part segmentation are two fundamental challenges for studying human behaviour. These two jobs are connected on a visual level. The fact that key points are frequently present inside the various semantic zones suggests that the semantic data gleaned from human parsing might be used to pinpoint these points and provide precise posture estimates. On the other hand, the keypoint group comes with abundant structural data that can help create semantic pieces. Convolutional Neural Networks (CNN) are increasingly used to handle these two tasks because of the deep neural network's [2, 17] improved learning capabilities [9]. Most extant models share a basic encoder-decoder design. However, their learning objectives may change. They [38] intend to transfer the up-sampled output to pixel-wise annotations for human parsing [18], whereas the pose estimate ground truth [8] correlates to the heatmaps of sparse key points.

Recent research [15, 20, 31, 39] attempts to perform joint inference via neural networks from the standpoint of multi-task learning in light of the correlation between two tasks. These models use a single shared encoder-decoder structure [15] or two distinct encoder-decoder structures [20, 39], and they hand-design the modules [20] to interact with high-level features derived for two tasks. It can be difficult to build the ideal network architecture and feature interactions for cooperative learning. On the one hand, despite having outward similarities, the two activities nonetheless have their own unique qualities. While human parsing requires investigating the pixel-by-pixel context information, the general techniques for posture estimation concentrate on aggregating information into small joint areas. Therefore, while tackling both jobs at once, it is difficult to extract discriminative characteristics. However, it is challenging to model the relationship between the two jobs. By manually building the fusion modules to interface with the high-level features from the two branches, existing works [20, 39]address this issue. But given that different levels of characteristics call for more precise interaction, it is fairly rigid and ignores the variety of intermediate features with multi-scale information.

This paper proposes a combined framework that integrates two independent works, human body part semantic segmentation and human pose estimation. The framework is fabricated by achieving success on smaller tasks like feature extraction, context modelling, semantic and pose subnet, and refining the maps by passing through integration and refinement network. Now, for the feature extraction task, a deep residual network is shared for both semantic and pose estimation tasks. Afterwards, the features are carried forward to two separate small networks for encoding and predicting the contextual information and results. An efficient integration and refinement network is constructed for both semantic segmentation and pose prediction to explore coherent context modelling that leads to semantic and pose estimation tasks mutually beneficial. This framework works to integrate multi-scale feature combinations and iterative location refinement, which are frequently posed as two different coarse-to-fine methodologies that are extensively researched for human parsing and pose estimation individually. The summary of the contribution can be understood in the following manner:

- Introduction of a unified framework that performs efficient context modelling and simultaneously predicts the human body part semantic segmentation and pose estimation with high-quality results.
- Experimental analysis of human body part semantic segmentation and human pose estimation. The framework is evaluated on the validation set of the benchmark LIP Dataset.
- Qualitative and quantitative analysis of the proposed model with other benchmark methods. Results are validated using four parameters Pixel Accuracy, Mean Accuracy, Mean IoU and PCKh values. Results are represented in visual as well as graphical manner.

Section II of this paper discusses the state-of-the-art methods in the Human Centric Visual Analysis domain. Section III explains the proposed methodology with well-labelled architectures. Experimental Analysis and results are shown in section IV. This section contains visual, tabular, and graphical outputs of the proposed framework. Lastly, section V contains the conclusion followed by references.

2 RELATED WORKS

Human body part semantic segmentation and human pose estimation are two correlated tasks. The fact that key points are frequently present inside the various semantic zones suggests that the semantic data gleaned from human parsing might be used to pinpoint these points and provide precise posture estimates. This section covers the most recent studies performed in the respective field of study.

Human Body Part Semantic Segmentation: In the area of human body part semantic segmentation, numerous deep-learning-based networks have achieved outstanding benchmarks. Ruan et al.

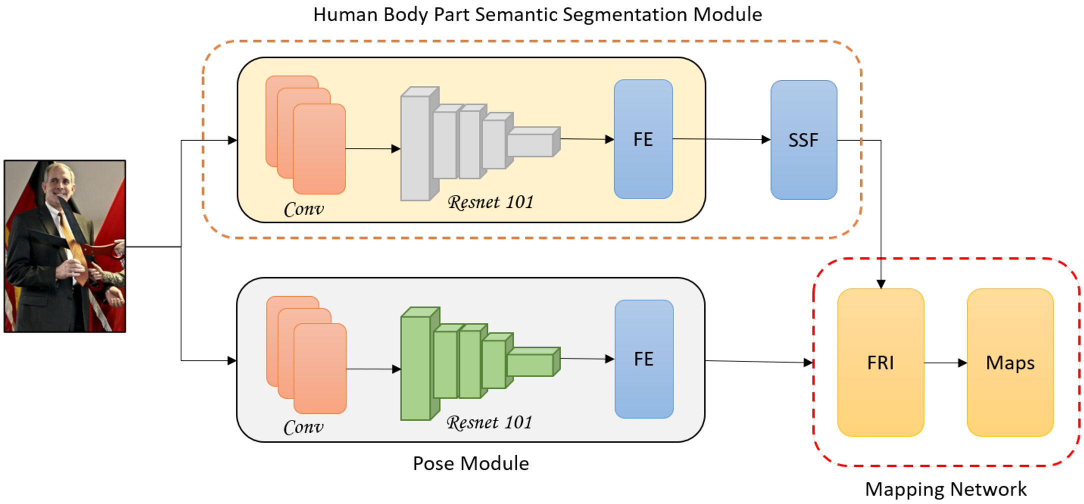

Figure 2: This figure represents the overview of the proposed framework. Here is the explanation of some of the acronyms used in the figure. FE: Feature Extraction, SSF: Semantically Segmented Features, FRI: Feature Refinement and Integration.

[24] constructed CE2P [24] framework built on Resnet-101 heavily utilising edge detail, global spatial context information, and feature resolution and won the LIP challenge held in 2019. Pixel's label indicates the class of the object to which the particular pixel fits; Yuan et al. [34] constructed a representation technique for semantic segmentation based on the objects' context that performed altruistically on LIP. Human prior knowledge has also been combined in some experiments for human parsing. Wang et al. [28] studied the hierarchy of the human body. The structure was set up to allow for effective, thorough human processing. Designing a new neural network tree for the purpose of segmentation and human parsing Ji et al. [13] also took advantage of the natural anatomical shape of the human body. Using the natural hierarchical construction of the human body and its relationship with various human organs, Zhang et al. [37] obtained good human parsing [18] outcomes by applying rules of grammar in a parallel and cascaded fashion. Similarly, Zhang et al. [39] combined keypoint positions with human semantic boundaries to improve human parsing. When multiple humans are present or unexpected occlusions cover specific human body parts, these strategies rely on the unique human posture or the preceding human hierarchical structure, making it impossible to guarantee generality.

Human Pose Estimation: Recognising the position and location of the human body in a particular scenario is the goal of human pose estimation. Now, human pose estimation can be performed for a single person when there is only one person in the frame, and multi-person human pose estimation when there is more than one person in the frame. So, in Single Person Pose Estimation, the conventional method for estimating an articulated human pose involves performing reasoning over a set of circumscribed observations on different components of the body along with the spatial relationships between them. The spatial model for an interlocked pose is based on two different methods, tree-structured graphical methods and non-tree methods. The tree-structured graphical models [1, 30, 36] encode the spatial relationships parametrically with adjoining parts along with the kinematic chain. Whereas the non-tree structure methods [10, 14] extend the tree's construction with additional edges to record symmetry, occlusion and long-range relationships. Convolutional Neural Networks (CNNs) have been broadly utilised to gather trustworthy local observations of the parts of the body with significantly improved accuracy of the body posture estimation [19, 21, 26, 29]. Tompson et al. developed a deep neural network architecture, including a graphical model whose parameters could learn collaboratively with the network. The sequential prediction framework [22] was the foundation for the pose machine architecture, including convolutions, introduced by Wei et al. [29]. They were iteratively applying overall context to update partial confidence maps along with maintaining the multimodal variability from earlier repetitions. To combat the issue of vanishing gradients during training, intermediary supervision is mandated after the conclusion of each distinct stage [3].In Multi-Person Human Pose Estimation, the majority of the methods [5, 8] have employed a top-down approach, first detecting individuals and then evaluating each person's position individually on every particular region detected. Although this approach directly applies the techniques created for the single-person situation, it also needs a prior dedication to detect a person and misses the spatial connections among various persons that call for global inference. In some cases, Human Pose Estimation is combined with Human Activity Recognition [25, 33] technique to predict the movement of the human body. **Combined Human Body Part Semantic Segmentation and Human Pose Estimation:** Taking into account how both the tasks, i.e. Human Body Part Semantic Segmentation and Human Pose Estimation, are correlated to each other, some research is done on this collaborative training paradigm. The framework developed for collaborative work was found to be challenging and complicated. Dong et al. in [11] offers an And-Or graph model

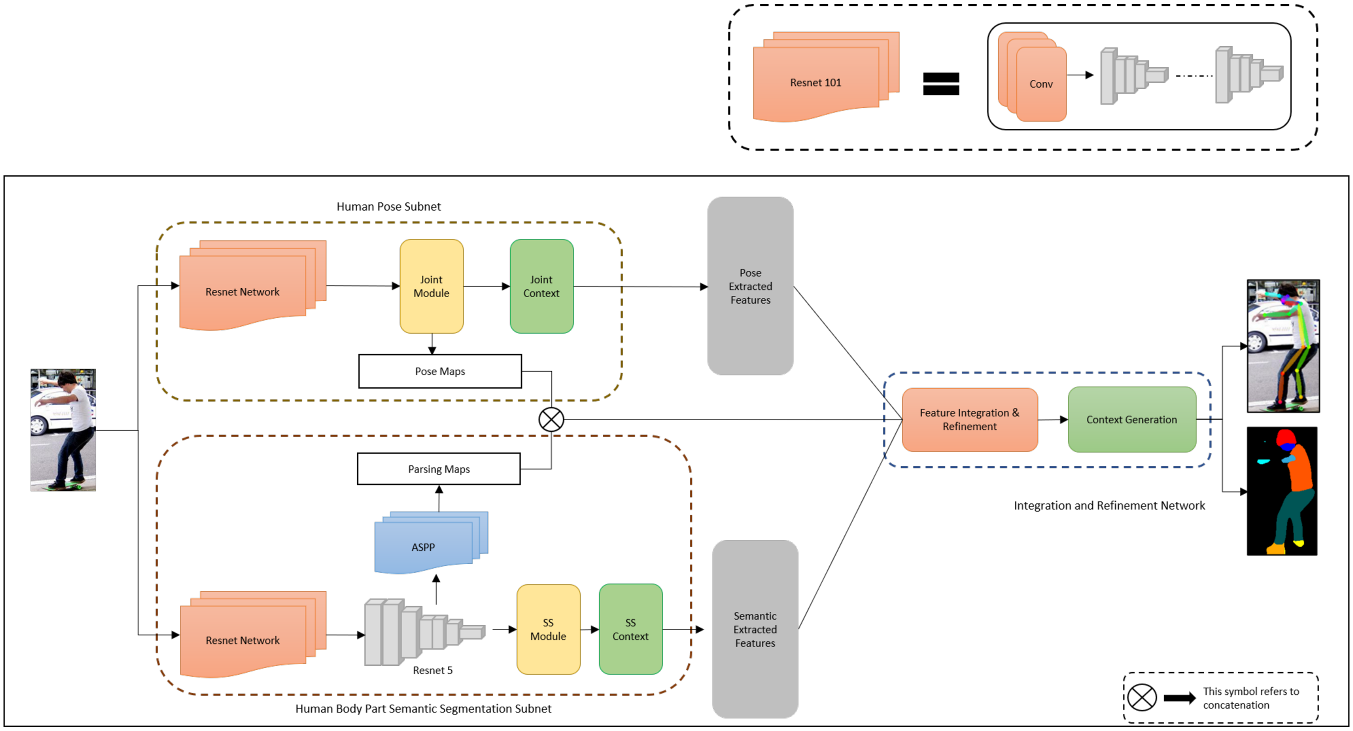

Figure 3: The above figure represents the architecture of the proposed model. In the architecture, SS represents Human Body Part Semantic Segmentation Module. Resnet Network is explained in the upper right corner of the image. Resnet Network is formed by using an Atrous Convolution then, followed by four ResNet 101 layers.

that constrains keypoint positions to the regions produced by the parsing and takes advantage of the spatial information of individual keypoints to encourage human body parsing. The earliest neural network-based method [32] connects high-level features from two different jobs via a) Convolutions and b) Conditional Random Fields. For this, the corresponding framework is taught gradually. Also, for both jobs in [15], standard features are extracted using a common backbone. Pose estimation's performance, however, needs to improve in terms of quality. In [20], two distinct sections are employed to obtain the features for parsing and pose. To combine the extracted features, hand-designed modules are provided to interact with these two branches. However, in contrast, [39] utilises a module built on attention to integrate the edge, pose, and parsing features. Even though these techniques concentrate on clearly modelling the interdependence between pose and parsing tasks, the manually planned exchange is insufficient because the outcomes of the two tasks frequently need to be more consistent.

3 PROPOSED METHODOLOGY

This section comprises an elaborate explanation of the suggested framework in the paper. The overview of the proposed framework can is illustrated in Figure 2. Human Body Part Semantic Segmentation and Human Pose Estimation deal with labelling each distinct image with finer details. These details include different granularities, which are pixel-wise semantic labelling against joint-wise structure prediction. The pixel-to-pixel labelling can produce elaborate information, while the joint-wise structure gives a more high-level structure. This makes both tasks complimentary. The coarse-to-fine scheme is used to improve accuracy in the already existing methods. Semantic Segmentation and pose tasks use this approach to improve the efficiency of networks. However, for both tasks, the meaning of coarse-to-fine-scheme varies. For Semantic Segmentation refers to pixel-wise classification, whereas for pose tasks, it indicates iterative displacement refinement. These two distinct definitions illustrated that it is reasonable to combine both tasks as they correlate with each other.

3.1 Unified Network

In order to use the logical and systematic illustration of human body part semantic segmentation and human pose to support each task, we put forward a joint human body part semantic segmentation and human pose estimation framework that integrates two different coarse-to-fine strategies, namely multi-scale feature and iterative refinement strategies. The detailed architecture of the introduced framework is explained in fig 3. In the proposed framework, the primary network of the Semantic Segmentation framework is the residual network [12]. In contrast, the Pose Estimation framework is constructed taking Stacked Hourglass Network [19] as its foundation. When combined with both network, the proposed framework share the residual network to obtain human image features. As a result, the architecture has two different networks for producing

Table 1: Quantitative comparison result of the proposed with state-of-the-art

Method	Pixel Accuracy	Mean Accuracy	Mean IoU
CorrPM [39]	87.68	67.21	55.33
MuLA [20]	88.50	60.50	49.30
JPPNet [15]	86.39	62.32	51.37
Ours	78.13	64.50	**58.67**

parser and pose features and findings. Following that is an integration and refinement network, which uses features and results as input to build more precise segmentation and joint localization.

Feature extraction: In the architecture, the convolution layer with upsampled filters, also known as "Atrous Convolutions" [6], is used to remodel ResNet-101 for dense predictions. Atrous convolution enables us to deliberately adjust the resolution at which feature responses are generated within DCNNs. Without increasing the number of parameters or the amount of calculation, it also successfully broadens the field of view of filters to include a broader context. There are four stages of ResNet-101 introduced in the framework. Further, deeper layers in the framework are designed for learning different tasks accordingly.

Parsing Subnet and Pose Subnet: The parsing subnet is built based on ResNet-101. This subnet is built using atrous convolution along with ResNet-101 acting as the fundamental unit. Now the output layer that gives the segmented results is constructed with ASPP [7] that segments object at various scales reliably.ASPP, with the help of filters, performs probing to capture objects as well as visual contexts at different scales that come from the convolutional feature layer through various effective fields of view and sampling rates. Two convolutions are performed after Res-5 to produce the context needed for the refining stage. Numerous 3 x 3 convolutional layers are added to ResNet-101's fourth stage (Res-4) to create pose features along with heatmaps for the pose subnet.

Integration and Refinement Network: Integration and refinement networks can simultaneously enhance both body part semantic segmentation and pose results. The intermediate body part semantic segmentation and prediction of the pose are concatenated again by mapping them together to a more significant number of channels and adding another 1x1 convolution layer into a feature space. This network is constructed by combining many convolutional layers with incremental kernel sizes ranging from 1 to 9 to collect enough semantic segmentation and pose features from the last stage. For the final semantic segmentation prediction, the ASPP module is used after concatenating the initially generated remapped results.

3.2 Modules of the architecture

The architecture is made from different small modules combined step by step to form a bulky framework.

SS Module SS Module is made up of only two convolutional layers. The kernel size of both layers is 3×3 with 512 and 256 channels, respectively. First convolutional Layer takes the output of ResNet-5 as input and sends its result to the second convolutional layer. This module generates SS context, extracting semantically segmented human body features.

Joint Module The joint module in the architecture contains multiple convolutional layers for kernel sizes of 3×3 and 1×1. The first layer of the module takes the output of ResNet4 as its input and has 512 channels. The size of the channels in the convolutional layer decreases after Conv2. The last two layers of the module have kernel size 1×1 with 512 and 16 channels, respectively.

Pose Refinement In this module, remap-1, remap-2 and pose context are concatenated. Remap1 and remap2 both are convolutional layers of kernel size 1×1 and 128 channels. It takes pose maps and part semantic segmentation maps as inputs, respectively. The pose refinement module is made of 6-size convolutional Layers with incremental kernel sizes ranging from 3 to 9. Although the channel size of the pose refinement network is 512, then decreases to 256, and the last layer has a channel size of 16.

Semantic Segmentation Refinement Similar to the pose refinement module, in the semantic segmentation module, the concatenation is performed between remap 1, remap2 and SS context, which is generated after extracting the feature from ResNet5 and proceeding it from SS Module. The results of the concatenation will be sent into a convolutional layer of size 3×3 and channel 512. The module is made up of 5 convolutional Layers and one ASPP module. The ASPP module will take input the result of the Conv-5 layer as input and reduce its channel size from 256 to 20.

4 EXPERIMENTAL ANALYSIS AND RESULTS

This section includes the quantitative and qualitative analysis of the results produced by the proposed framework. Separate results for the Human Body Semantic Part segmentation and Pose Estimation are discussed in the below section, along with the dataset description.

4.1 Dataset

The Look Into Person (LIP) [15] is an extensive image dataset that focuses on the task of segmenting human parts semantically from the given images, called Human Parsing. In other words, the LIP dataset is fabricated for the semantic interpretation of human body parts with diverse, intriguing features. The dataset contains around 50,462 human images with detailed pixel-wise annotations of approx 19 labels enlighting semantic human part labels, a background label from Human Parsing, and 2-Dimensional Human poses, including 16 key points for Human Pose Estimation. The images included in the dataset are assembled using real-world scenarios. People with challenging poses, substantial occlusions, extensive resolution, different viewpoints and appearances were included as a necessary obstacle in the dataset. Images contained by the LIP dataset are constructed by cropping instances of a person shown in the image from the Microsoft COCO [16] training and validation datasets. The dataset contains about 19 labels of human body parts clothes for the purpose of annotation. These labels are upper clothes, face, sunglasses, dress, coat, jumpsuit, arms, and many more, along with the background. This dataset also provides detailed annotations for 16 main body joints, including their visibility and positions for human pose estimation. Joint annotation of LIP Images is done in a "Person-Centric" approach implying that the right/left joints

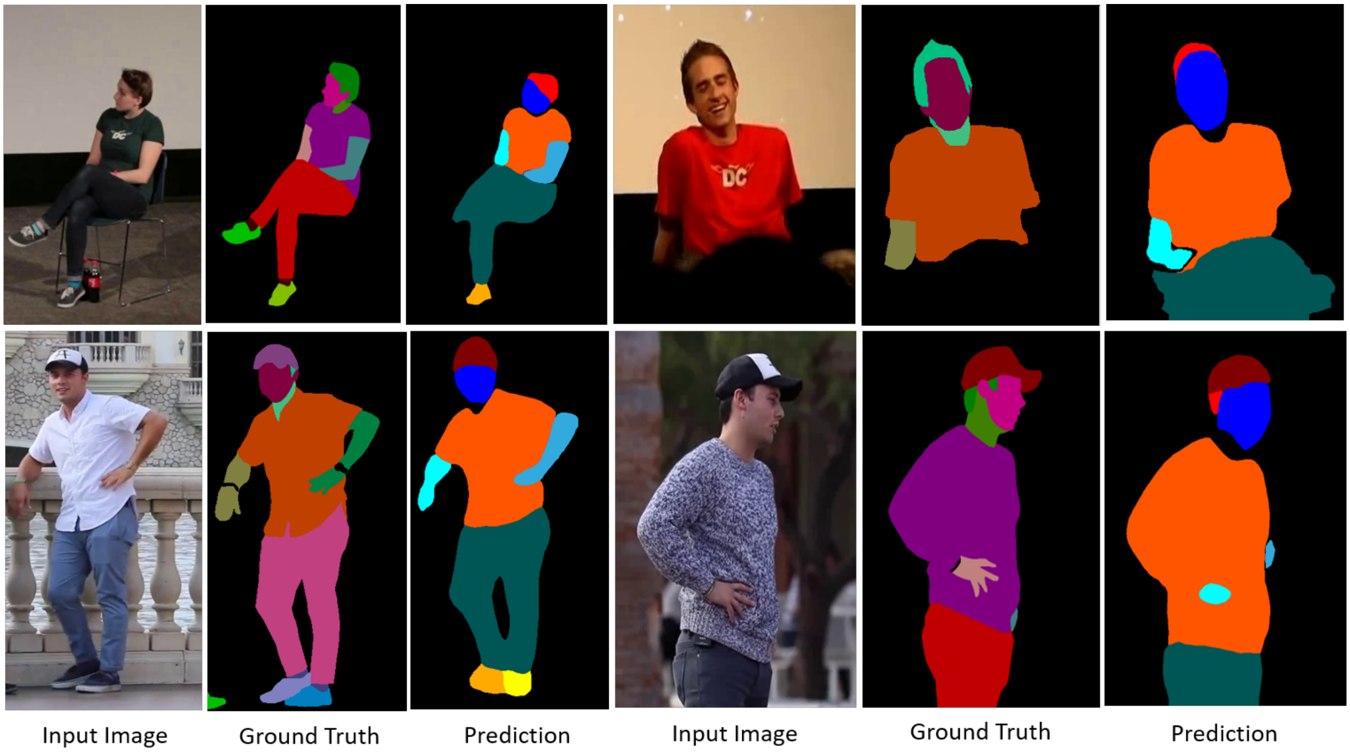

Figure 4: The above figure represents the visualized result of the proposed work when evaluated on the validation set. The figure contains the input image, its ground truth and the predicted out that is generated after the evaluation of the model. The generated output is closely similar to its ground truth.

refer to the respective right/left limbs of the person in the image. To expedite the annotation process, multi-scale superpixels of the photos were constructed using an annotation tool.

4.2 Evaluation Metric

PCKh is an accuracy measurement metric that predicts whether the key point and the actual joint are within a defined distance threshold. The PCKh is usually decided according to the scale of the object enclosed in a bounding box. The threshold of the PCKh can be of a different type. If PCKh is denoted in PCK@0.5, it refers to the constraint when the threshold distance is 50% of the head bone link. Similarly, PCK@0.2 is denoted when the distance between the predicted and actual joint is less than 0.2 times the diameter of the torso. Sometimes, the threshold is taken as 150 mm as the default value. This evaluation metric can also be used to evaluate 3D pose estimation.

4.3 Result of Human Body Part Semantic Segmentation

Table 1 contains the overall results of the framework, including the results of the current state-art-of-the-art methods. The experiment was performed on the LIP dataset, which consists of 50,462 images. When analysed on the test and validation set of the dataset, our proposed framework outperformed all the existing work. In terms

Table 2: 16 Joint values of the human body. Y and X represent the Y and X-axis where the joints are located. Frames are the images whose results are generated.

Joints	Frame 1		Frame 2		Frame 3		Frame 4		Frame 5	
	y	x	y	x	y	x	y	x	y	x
Joint 1	17	124	32	161	31	249	59	228	53	332
Joint 2	43	94	38	161	28	206	16	180	51	276
Joint 3	16	72	101	161	27	150	18	140	40	198
Joint 4	42	70	204	156	57	150	46	140	97	193
Joint 5	43	94	251	95	58	203	46	180	98	274
Joint 6	41	128	234	159	59	251	55	228	100	330
Joint 7	41	64	174	158	41	145	33	132	71	187
Joint 8	47	45	164	88	40	89	40	85	66	95
Joint 9	51	27	182	50	37	53	39	56	65	57
Joint 10	60	15	194	27	28	18	44	27	67	19
Joint 11	33	77	132	66	7	126	49	87	33	132
Joint 12	62	56	84	148	7	103	10	112	6	123
Joint 13	47	32	123	55	12	66	9	56	30	71
Joint 14	42	28	187	61	67	66	59	60	106	71
Joint 15	58	56	241	140	74	102	72	93	124	128
Joint 16	81	68	234	15	74	136	64	136	100	135

Table 3: Class-wise Quantitative Comparison of mean intersection over union (IoU) with other methods on the LIP validation set. Here each class represents the mIoU value of each human part predicted by our proposed method.

Method	background	hat	hair	upperclothes	coat	pants	face	l-arm	r-arm	l-leg	r-leg	Avg
CorrPM [39]	87.77	66.20	71.56	70.20	57.95	75.19	74.36	66.53	68.61	62.80	62.81	55.33
MuLA [20]	-	-	-	-	-	-	-	-	-	-	-	49.30
JPPNet [15]	86.25	63.55	70.20	68.15	55.65	72.19	73.36	61.97	63.88	58.21	75.99	51.37
NPPNet [35]	88.38	66.43	72.34	71.88	60.54	77.07	75.31	71.04	71.50	71.75	70.55	58.56
Ours	**92.61**	57.72	68.95	**72.25**	50.52	70.35	**78.81**	44.32	47.04	34.39	35.46	**58.97**

of overall mean IoU, the proposed method gives the best result by outperforming JPPNet [15] by 7.6%, MuLA [20] by 9.67%, and CorrPM [39] by 3.64%. Observing the other two evaluation factors, i.e. Pixel Accuracy and Mean Accuracy, the proposed framework was not able to outperform MuLA [20] in pixel accuracy and CorrPM [39] in mean accuracy but still shows a satisfactory performance by giving pixel accuracy of 78.13 % and Mean Accuracy of 64.52% which is 2.30% more than JPPNet [15].

Classwise values of mIoU, together with comparative results of the state-of-the-art methods, are discussed in Table 3. Values represent the mIoU values generated for each class, like face, hair, and hat, when assessed on the validation of the LIP dataset. For ease of comparison, 11 classes are considered in the analysis table. Observing the discussed result, the proposed framework notably improves the performance of classes background, upper clothes, and face by 4.23%, 0.37% and 3.5%, respectively. Also, the overall average mean IoU value outperforms the current state-of-the-art value by 0.47%. The improved results establish that the proposed framework works efficiently. Also, the framework improved the best performance from 58.56% to 58.97%.

The visual output of the proposed model is shown in Figure 4. The mentioned figure demonstrates the input image, the ground truth of the input image and the predicted output. Predicted images are very similar to the ground truth. Although LIP Dataset mainly contains single-person images, we believe our proposed model can perform well for multi-human or multi-person images.

4.4 Result of Human Pose Estimation

Figure 5 and Figure 6 represent the PCKh results of the proposed network. The tables mentioned above also exhibit the comparative analysis of the prior benchmark works. PCKh metric is an evaluation metric that determines how well joints in the human body are localised. After completing the pose estimation task, the results are produced in a ".txt" file for each image of the validation set. The text file contains the locations of the 16 joints. Now the value of these 16 joints are computed according to locations in Y and X-axis, respectively. The total number of values that will be printed in the text file will be 32. The text file is demonstrated in Table 2.

5 CONCLUSION

In this paper, we proposed a unified human body part semantic segmentation and human pose estimation framework that explores and takes advantage of both tasks' intrinsic connection. The results demonstrated by extensive experiments on the LIP dataset

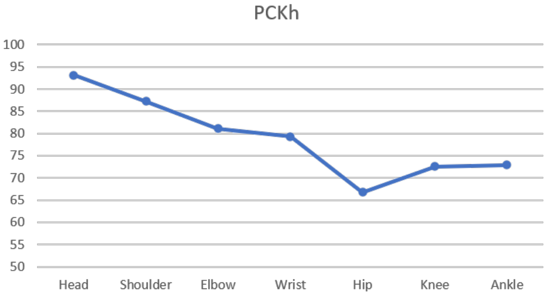

Figure 5: Graphical representation of PCKh values of the proposed framework

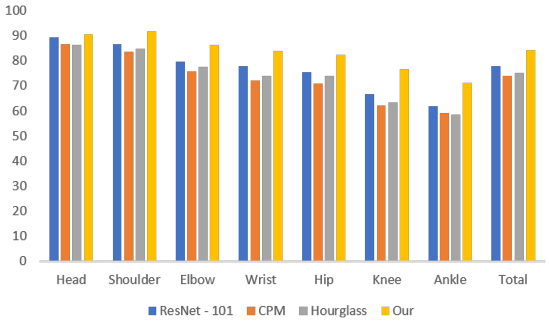

Figure 6: Graphical representation of the comparison of proposed work with state of the art.

validate that the proposed framework works efficiently. The framework of the traditional human pose estimation technique utilizes the concept of object detection on the image frame of the given input. However, object detection is being replaced with human body part semantic segmentation through this proposed work to identify human body parts accurately. This framework fulfils our goal of creating a unified semantic segmentation framework and pose estimation for our further task. The experiments can be more challenging for future aspects by utilizing a multi-person dataset, Crowd Instance-level Human Parsing(CIHP). Also, the results generated after performing human pose estimation can be considered as features and applied for Human Activity Recognition.

REFERENCES

[1] Israr Akhter, Ahmad Jalal, and Kibum Kim. 2021. Pose estimation and detection for event recognition using Sense-Aware features and Adaboost classifier. In *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*. IEEE, 500–505.

[2] Anam Arshad, Vivek Tiwari, Mayank Lovanshi, and Rahul Shrivastava. 2023. Role Identification from Human Activity Videos using Recurrent Neural Networks. In *proceedings of the 8th IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*.

[3] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5, 2 (1994), 157–166.

[4] Kunal Bose, Kumar Shubham, Vivek Tiwari, and Kuldip Singh Patel. 2023. Insect Image Semantic Segmentation and Identification Using UNET and DeepLab V3+. In *ICT Infrastructure and Computing*. Springer, 703–711.

[5] Markus Braun, Qing Rao, Yikang Wang, and Fabian Flohr. 2016. Pose-rcnn: Joint object detection and pose estimation using 3d object proposals. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 1546–1551.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[8] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7103–7112.

[9] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. 2017. Multi-context attention for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1831–1840.

[10] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool. 2013. Human pose estimation using body parts dependent joint regressors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3041–3048.

[11] Jian Dong, Qiang Chen, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. 2014. Towards unified human parsing and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 843–850.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] Ruyi Ji, Dawei Du, Libo Zhang, Longyin Wen, Yanjun Wu, Chen Zhao, Feiyue Huang, and Siwei Lyu. 2020. Learning semantic neural tree for human parsing. In *European Conference on Computer Vision*. Springer, 205–221.

[14] Leonid Karlinsky and Shimon Ullman. 2012. Using linking features in learning non-parametric part models. In *European Conference on Computer Vision*. Springer, 326–339.

[15] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. 2018. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence* 41, 4 (2018), 871–885.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[17] Mayank Lovanshi and Vivek Tiwari. 2023. Human Pose Estimation: Benchmarking Deep Learning-based Methods. In *proceedings of the IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation*.

[18] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2018. Macro-micro adversarial network for human parsing. In *Proceedings of the European conference on computer vision (ECCV)*. 418–434.

[19] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*. Springer, 483–499.

[20] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. 2018. Mutual learning to adapt for joint human parsing and pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 502–517.

[21] Wanli Ouyang, Xiao Chu, and Xiaogang Wang. 2014. Multi-source deep learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2329–2336.

[22] Varun Ramakrishna, Daniel Munoz, Martial Hebert, James Andrew Bagnell, and Yaser Sheikh. 2014. Pose machines: Articulated pose estimation via inference machines. In *European Conference on Computer Vision*. Springer, 33–47.

[23] Mrigank Rochan et al. 2018. Future semantic segmentation with convolutional lstm. *arXiv preprint arXiv:1807.07946* (2018).

[24] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. 2019. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 4814–4821.

[25] Rahul Shrivastava, Vivek Tiwari, Swati Jain, Basant Tiwari, Alok Kumar Singh Kushwaha, and Vibhav Prakash Singh. 2022. A role-entity based human activity recognition using inter-body features and temporal sequence memory. *IET Image Processing* (2022).

[26] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. 2015. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 648–656.

[27] Chengjia Wang, Tom MacGillivray, Gillian Macnaught, Guang Yang, and David Newby. 2018. A two-stage 3D Unet framework for multi-class segmentation on full resolution image. *arXiv preprint arXiv:1804.04341* (2018).

[28] Wenguan Wang, Hailong Zhu, Jifeng Dai, Yanwei Pang, Jianbing Shen, and Ling Shao. 2020. Hierarchical human parsing with typed part-relation reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8929–8939.

[29] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 4724–4732.

[30] Yongpeng Wu, Dehui Kong, Shaofan Wang, Jinghua Li, and Baocai Yin. 2022. HPGCN: Hierarchical poselet-guided graph convolutional network for 3D pose estimation. *Neurocomputing* 487 (2022), 243–256.

[31] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. 2017. Joint multi-person pose estimation and semantic part segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6769–6778.

[32] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. 2017. Joint multi-person pose estimation and semantic part segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6769–6778.

[33] Yang Xu and Ting Ting Qiu. 2021. Human activity recognition and embedded application based on convolutional neural network. *Journal of Artificial Intelligence and Technology* 1, 1 (2021), 51–60.

[34] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. 2018. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916* (2018).

[35] Dan Zeng, Yuhang Huang, Qian Bao, Junjie Zhang, Chi Su, and Wu Liu. 2021. Neural Architecture Search for Joint Human Parsing and Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11385–11394.

[36] Feng Zhang, Xiatian Zhu, and Mao Ye. 2019. Fast human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3517–3526.

[37] Xiaomei Zhang, Yingying Chen, Bingke Zhu, Jinqiao Wang, and Ming Tang. 2020. Blended grammar network for human parsing. In *European Conference on Computer Vision*. Springer, 189–205.

[38] Xiaomei Zhang, Yingying Chen, Bingke Zhu, Jinqiao Wang, and Ming Tang. 2020. Part-aware context network for human parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8971–8980.

[39] Ziwei Zhang, Chi Su, Liang Zheng, and Xiaodong Xie. 2020. Correlating edge, pose with parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8900–8909.