Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20094-1

Edge devices friendly multi-human parsing with lightweight encoding and multi-scale self-attention based decoding

Md Imran Hosen[1] · **Tarkan Aydin**[1]

Received: 16 March 2024 / Revised: 3 August 2024 / Accepted: 14 August 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
Multi-human parsing has received considerable research attention in recent years. Deep learning-based Multi-human parsing methods demonstrated promising results. In reality, most methods suffer while running on edge devices due to their extensive network architecture and low inference speed. Moreover, the inadequacies in modeling long-range feature dependencies have led to suboptimal representations of discriminative features across semantic classes. To address these challenges and facilitate real-time implementation on edge devices, we design a deep yet lightweight Encoder and a Multi-Scale Self-Attention based Decoder to capture long-range dependencies and spatial relationships. Furthermore, we have optimized our model through half-precision quantization, enhancing efficiency for edge devices. Experiments on publicly available Crowd Instance-level Human Parsing (CIHP) and Look into Person (LIP) datasets show the efficacy of our framework to parse multi-human with high inference speed at 55.6 FPS. Additionally, real-world testing on Jetson Nano edge devices showcases competitive performance. An extensive ablation study on different modules validates our network.

Keywords Multi-human parsing · Edge devices · Self-Attention · Inverted residual block

1 Introduction

In the realm of computer vision, the intricate task of multi-class human parsing involves a meticulous examination of pixel-level human-centric interpretation. This endeavor aims to precisely delineate each instance into distinct human components, discerning the diverse manifestations of the human form. The significance of multi-class human parsing lies in its ability to furnish specific details about human instances, thereby proving indispensable for a spectrum of human-related applications, including human-object interaction, dense pose estimation, fashion editing, person re-identification, fashion landmark recognition, and complex human-centric video analysis in challenging scenarios [1–8].

✉ Md Imran Hosen
md.hosen@bahcesehir.edu.tr

Tarkan Aydin
tarkan.aydin@bau.edu.tr

[1] Department of Computer Engineering, Bahcesehir University, Yıldız, 34349 Istanbul, Turkey

Published online: 27 August 2024

The landscape of multi-human parsing has witnessed remarkable advancements through the lens of deep learning-based methodologies, particularly within detection-based two-stage techniques. Pioneering this domain, Faster r-CNN [9] emerged as a front-runner, employing a two-stage human instance detector to extract proposal regions and execute fine-grained parsing of human parts. Despite its outperformance over predecessors, Faster r-CNN encountered misalignment issues at the pixel level between network inputs and outputs. In response to this challenge, Mask r-CNN [10] was introduced, leveraging RoIAlign to uphold precise spatial relationships. While Mask r-CNN exhibited substantial progress, it fell short of an end-to-end solution.

Addressing the quest for end-to-end solutions, Zhu et al. proposed a parsing network with a component-aware convolution structure [11]. Although achieving end-to-end functionality, its computational demands and dependency on detection approaches hindered efficiency, particularly in edge devices. Introducing a groundbreaking detection-free approach, Gong et al. presented a part grouping network [12] that reframed instance-level human parsing as two collaborative sub-tasks. Despite its innovation, the addition of parallel branches for human parsing and edge prediction lacked effective modeling of their connections. The subsequent introduction of Graphonomy [13] aimed to overcome this limitation through a hierarchical graph structure, yet faced challenges in generalizability due to dataset-specific constraints.

In recent advancements in human parsing, self-attention mechanisms have become essential. For example, Song et al. [14] introduced the Global Transformer Module (GTM) to capture long-range dependencies and improve contextual information extraction. Guan et al. [15] addressed multi-human parsing using a graph transformer module to infer semantic correlations. Yang et al. [16] proposed mask2former parsing, a transformer-based framework for human parsing. Despite its accuracy, its computational demands limit its applicability. Inspired by these methods, we incorporate self-attention mechanisms into our human parsing framework.

In response to these challenges, we present our novel end-to-end low-cost multi-human parsing model. Harnessing the power of a deep yet lightweight encoder inspired by the squeeze and excitation network [17] and a Vision Transformer-based Decoder employing a Multi-scale Self-Attention (MSSA) mechanism for capturing long-range dependencies, our approach achieves a delicate balance. Drawing inspiration from UNet [18], our model seamlessly integrates context and global information through skip connections, facilitating rapid convergence. Figure 1 showcases a representative result generated by our approach on edge devices. Following is a summary of the strengths and significant contributions of this paper.

- Propose a novel end-to-end edge device-friendly multi-human parsing network optimized for edge devices, enhanced with a light encoder and a self-attention based decoder.

Fig. 1 Illustration of multi-human parsing using our method on Edge devices (Jetson Nano)

- Design a deep yet lightweight encoder based on a squeeze and excitation network to provide the decoder with rich feature representations while preserving real-time implementation feasibility. We enriches the decoder performances through utilizing the multi-scale self-attention mechanism and Path Aggregation Network.
- We optimize our model through half-precision quantization and conduct extensive ablation studies to validate the proposed network on the Crowd Instance-level Human Parsing (CIHP) and Look into Person (LIP) datasets, and implement the model on Jetson Nano edge devices.

The remaining part of this paper has been arranged as follows. In Section 2, we have discussed related literature. We have presented the details of the proposed methodology in Section 3. The experiment of our proposed method takes place in Section 4, and we have analyzed the results in Section 5. We have analyzed the effect of our changes in Section 6. In the end, we have concluded our work in Section 7.

2 Related works

The field of human parsing has garnered considerable attention in recent times, fueled by its practical applications and commercial significance. Despite the strides made in this domain, determining the intricate structure of the human body remains a formidable challenge. Extensive research efforts have been documented in the literature, with a predominant categorization into two key domains: single-human parsing and multi-human parsing.

2.1 Single human parsing

The domain of single human parsing, often referred to as per-pixel classification, is primarily concerned with the comprehensive understanding of individual human images. In addressing this task, several innovative approaches have been introduced, each leveraging unique modules and methodologies. The Mutual Learning to Adapt framework, as introduced by Nie et al. [19], is tailored for simultaneous pose estimation and human parsing tasks. It facilitates rapid adaptation of parsing and pose models by amalgamating insights from their respective models. A comparable strategy was pursued in the study by Liang et al. [20]. Furthermore, Wang et al. [21] proposed the High-Resolution Network to address visual recognition challenges, employing a mechanism that iteratively integrates multi-resolution representations to enhance feature richness. Zhang et al, introduced a part-aware context network [22], specifically designed for single human parsing. This network incorporates a part class module, a dispersion module, and a relational module. The synergy between these components enables effective communication of human structure information, contributing to improved parsing accuracy. Hierarchical Human Parsing, proposed by Wang et al., [23], adopts the representational capacity of Graph Convolutional Networks. This approach aids in the comprehension of hierarchical human layouts, providing a holistic understanding of the relationships between different anatomical components. In another vein, Zhang et al., proposed a heterogeneous non-local block in [24]. This block, operating based on crucial points locations and human semantics, facilitates the thorough exploration of connections among parser, posture, and edge. By considering these diverse elements in tandem, the model achieves a more comprehensive grasp of contextual information.

Liu et al., introduced the Human Kinematic Skeleton Graph Layer in [25], leveraging the human kinematic skeleton to enhance the original neural networks. This augmentation

addresses challenges related to occlusions, varied positions, and the differentiation between right and left components, enhancing the model's robustness.

Zhang et al. introduced HTCorrM [26], an extension of CE2P [27], which leverages part edges to enhance the modeling of inter-part relationships. This augmentation contributes to a more comprehensive understanding of the complex interactions between different body parts. In a parallel development, Liu et al. proposed HIPN [28], a hybrid learning framework that combines denoising with semi-supervised learning techniques. By incorporating positive and negative learning mechanisms, HIPN exhibits improved resilience to noise, making it better equipped to handle noisy pseudo labels.

A notable addition is the Region-Level Parsing Refiner module proposed by Zhou et al. [29]. This module, integrated into the parsing pipeline, focuses on refining parsing performance by incorporating region-level parsing learning. This targeted enhancement contributes to more precise and contextually informed parsing outcomes. However, it is important to note that the aforementioned models are tailored for scenarios involving single human parsing. Real-world problems often present more complex scenarios, necessitating the extension of these models' applicability to multi-human parsing scenarios. This highlights the ongoing challenge in broadening the scope of these models for increased versatility.

2.2 Multi-human parsing

Multi-human parsing (MHP) transcends the limitations of single-human parsing, encompassing the intricate task of partitioning crowd scene images into semantically consistent regions associated with body parts or clothing items, while discerning diverse identities. Two primary paradigms characterize MHP: top-down (comprising two-stage and one-stage approaches) and bottom-up.

Top-Down Approaches:

In the realm of top-down approaches, a bounding box-based detector identifies humans, extracts information, and generates regions of interest (ROIs) from the original image. The parsing results are then obtained through comprehensive segmentation of these ROIs. The two-stage top-down technique, exemplified by Faster r-CNN [9] and Mask r-CNN [10], leverages an initial object detection stage followed by a dedicated parsing module. Context Embedding [27] introduces a unique method involving two single human parsing training models, integrating ground truth instances and predicted examples, followed by a global fusing method. BraidNet [30] incorporates a braiding network with local structure capturing and semantic knowledge learning networks. A part decomposition and refinement network (PDRNet) based on part-wise semantic prediction was proposed in [31]. PDRNet utilizes part-wise mask prediction to decompose the human body into different semantic parts and employs a refinement module to obtain accurate masks for each part. Despite achieving state-of-the-art accuracy, two-stage methods face challenges in flexibility and real-world application inference time [32].

In the one-stage top-down paradigm, such as Nested Adversarial Network [33] and the unified framework proposed by Qin et al. [34], the end-to-end process concurrently recognizes human instances and employs attention modules for parsing human aspects. However, these methods struggle with parsing small regions due to limited contextual information. Renovating Parsing r-CNN [35] and AI-Parsing [36] address this limitation through global semantic representation and instance-level parsing, respectively. While achieving high accuracy, these detection-based approaches heavily rely on the detection module, incurring high computational costs.

Bottom-Up Approaches:

Bottom-up approaches focus on faster inference, treating global fine-grained semantic segmentation as an instance-level task. Techniques like Atrous Spatial Pyramid Pooling (ASPP) [37] and Macro-structure Parsing [38] leverage atrous convolution and ASPP frameworks for multi-scale context capture. Part Grouping Network [12] generates human instance outcomes using additional edge data. Hierarchical graph-based methods like Graphonomy [13] and Grapy-ml [39] explore label semantic relations at the global level but face challenges in instance-level performance due to encoder limitations.

A novel approach presented in [40] addresses instance-aware human part parsing by simultaneously learning category-level human semantic segmentation and multi-person pose estimation in a joint, end-to-end fashion. By incorporating a dense-to-sparse projection field and formulating joint association as maximum-weight bipartite matching, the framework achieves robustness and end-to-end trainability. Building upon this work, [41] introduces an efficient solver for differentiable joint association, enhancing training efficiency compared to the previous approach, despite marginal sacrifices in parsing accuracy. Although this approach is efficient, it is not optimized for implementation on edge devices.

Customized approaches, like fusion [42], blend top-down and bottom-up strategies to leverage their respective merits. Incorporating insights from the previous work [42], the authors proposed a part-relation-aware human parser (PRHP) in their subsequent study [43]. PRHP precisely delineates three types of human part relations-decomposition, composition, and dependency-by employing three distinct relation networks, enhancing the model's generality and efficacy. While effective, such fusion increases implementation complexity.

Multi-human parsing has been a subject of extensive study, with approaches falling into the categories of top-down and bottom-up methods. Existing methods, while making significant strides, often grapple with complexities, high computational costs, and dependencies on parallel branches, such as human detection or edge prediction. This poses challenges for real-time deployment on edge and resource-constrained devices.

To overcome these limitations, our proposed approach adopts a bottom-up strategy to streamline the multi-human parsing process for efficient inference. In contrast to traditional methods, our solution is anchored in a deep yet lightweight encoder, ensuring computational efficiency without compromising on feature richness. The core of our innovation lies in the decoder, which incorporates a vision transformer-based multi-scale self-attention. This module enables the capturing of long-range dependencies, providing a more holistic understanding of the global context in crowd scenes.

One distinctive feature of our network is its departure from the conventional reliance on parallel branches, such as detection and edge prediction. Instead, our approach emphasizes an end-to-end paradigm, aiming to efficiently learn both local and global features without the need for auxiliary models. This not only simplifies the parsing pipeline but also enhances the adaptability of the network for real-time deployment on edge devices.

In the subsequent sections, we delve into the details of our proposed framework, highlighting the key components and their contributions to addressing the challenges posed by existing multi-human parsing methodologies.

3 Proposed method

Our objective is to predict a pixel-wise label map with dimensions $H \times W \times C_{out}$ based on an input image with dimensions $H \times W \times C_{in}$. Here, H and W correspond to the height

and width of the image, respectively, while C_{in} and C_{out} represent the number of input channels and the number of output classes, respectively. The schematic representation of our comprehensive method is illustrated in Fig. 2, which delineates the two primary components: the Encoder and the Decoder.

3.1 Encoder

Within the encoder of our architecture, we employ a diverse set of building blocks, including the Convolutional Block (Conv), Efficient Block (EffiBlock), Residual Convolutional Block (RCB), and Spatial Pyramid Pooling (SPP), to encode the input image into feature representations at multiple hierarchical levels. The input image initially undergoes a Conv block, followed by ten consecutive EffiBlocks. This sequential application facilitates deep feature extraction while managing computational complexity effectively. Subsequently, the input traverses another Conv block, a Residual Convolutional Block (RCB) [44], and a Spatial Pyramid Pooling (SPP) block. This combination of operations enables the encoder to learn residual characteristics and merge features with varying resolutions, furnishing our decoder with superior feature representations for segmentation.

Conv Block. A Conv Block comprises three fundamental operations, which include convolution (Conv2d), batch normalization (BN), and activation using the SiLU function [45]. Initially, the input data is subjected to convolution, which serves to extract relevant features. Subsequently, the application of batch normalization (BN) not only regularizes the process but also expedites learning. Finally, the SiLU activation function plays a crucial role in determining whether a neuron should be activated by computing weights and adding bias to the result.

EffiBlock. Adding extra layers is one of the primary methods for enhancing deep learning performance [46]. However, it leads to the model becoming over-fitted. We utilize EffiBlock based on the inverted residual block [47, 48] and Squeeze and Excitation Network [17] shown in Fig. 3 to improve information flow and offer better features encoding without increasing computational load. The Inverted Residuals Block leverages the concept of inverted residuals [47] to enhance gradient propagation within the feature extraction network while minimizing memory usage during inference. EffiBlock employs squeeze-and-excitation attention mech-

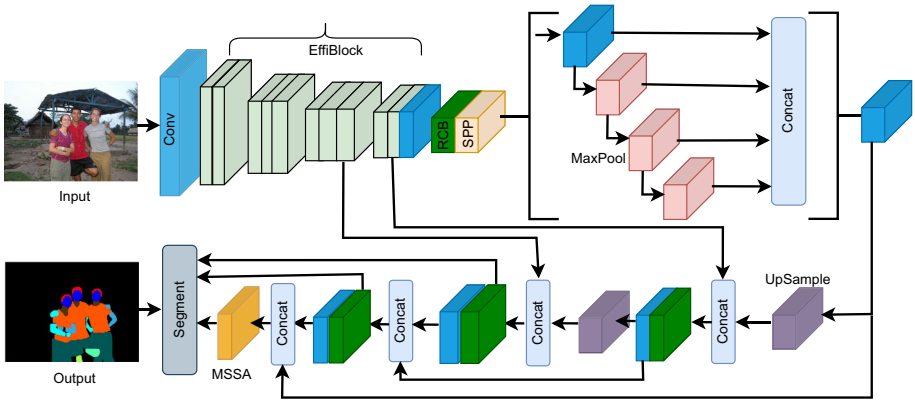

Fig. 2 The proposed architecture for edge devices friendly multi-human parsing network

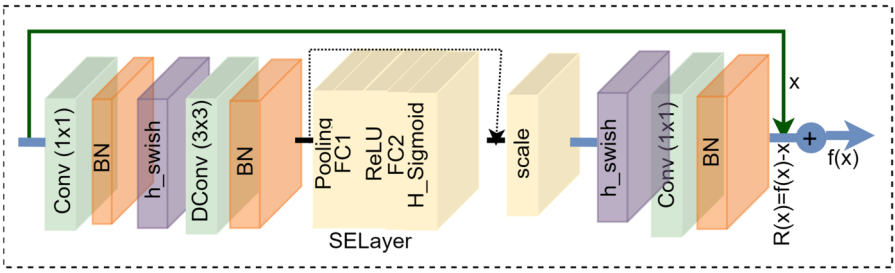

Fig. 3 EffiBlock based on the inverted residual block and Squeeze and Excitation Network

anisms in the channel dimension, prioritizing informative channel features and suppressing less relevant ones. We primarily utilized EffiBlock to expedite and streamline feature extraction, thus enhancing segmentation speed.

3.2 Decoder

The encoder's output is gradually extended in the decoder. With upsampling and convolutional operation, exact localization is made possible by the decoder [49]. Within the decoder, high-level features are fused with spatial information and feature maps obtained from the encoder. The inputs to the decoder consist of data from the immediate preceding decoder layer, alongside the corresponding encoder features. Furthermore, we incorporate a residual connection approach in the decoder, inspired by the path aggregation network [45, 50], as illustrated in Fig. 2. Towards the conclusion of our decoder, we employ a Multi-scale Self-Attention Module to better understand and segment complex patterns [51, 52], followed by the segmentation block for the final processing.

Multi-scale Self-Attention (MSSA) Module.

The Module for Multi-Scale Self-Attention (MSSA) emerges as a pivotal component, facilitating the model's comprehension of the intricate dynamics within various symmetrical regions of the human body. Its strategic focus on pertinent segments within the input sequence enables the model to accentuate crucial attributes while efficiently filtering out extraneous data. The significance of multi-scale features predates the advent of deep learning, as highlighted in [53]. In the domain of deep segmentation networks, the amalgamation of multi-scale features has exhibited remarkable efficacy, as evidenced by the works [54, 55]. Drawing inspiration from these advancements, our approach leverages learned features across multiple scales, synergistically integrated with self-attention mechanisms. This fusion empowers the encoding of both global and local intricacies, thereby enhancing the model's discriminative capabilities [56].

In the MSSA framework (shown in Fig. 4), input is derived from a 2D feature map F characterized by dimensions (H, W, C), where H and W denote height and width, respectively, and C signifies the number of channels. Initially, the input channels undergo division via conventional convolutional operations, succeeded by max-pooling procedures employing kernel sizes of 5, 9, and 13. Subsequently, the outputs from these max-pooling operations are concatenated with the original data, facilitating the amalgamation of features across diverse resolutions and yielding refined feature representations denoted as X. These processed features are then subjected to group normalization and projection into a lower-dimensional space via learnable weight matrices (W_q, W_k, and W_v). This projection yields sequences of query

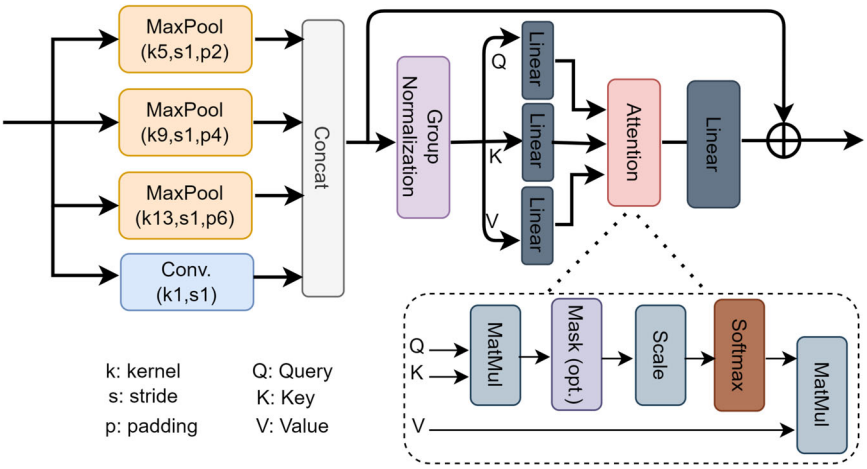

Fig. 4 The architecture of Multi-Scale Self-Attention Module

(Q), key (K), and value (V) vectors, which are subsequently utilized within an attention mechanism to capture contextual dependencies across the input feature map.

$$A(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

The attention mechanism operates through an (1) wherein the query, key, and value vectors are orchestrated to compute attention scores, which are further normalized via softmax. Following this, the output of the self-attention mechanism undergoes transformation via a position-wise feedforward layer, applying non-linear operations to individual elements within the sequence.

$$FF(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{2}$$

This feedforward layer, characterized by (2), encompasses learnable weight matrices and biases (W_1, b_1, W_2, and b_2), facilitating the extraction of intricate features. The resulting output from the feedforward layer is then integrated with the output of multi-scale features X via residual connections, culminating in a comprehensive representation enriched with contextual insights and discriminative features.

Segment. The segment block takes on the crucial task of generating the ultimate segmented output. Its function involves the integration of decoder network features at varying scales. This process is achieved through the implementation of a three-layer Fully Connected Network (FCN) equipped with sigmoid-weighted linear units (SiLU). The outcome of this stage comprises prototype masks of image size that are agnostic to individual instances, as inspired by the work of Bolya et al. [57].

3.3 Model compression

In our efforts to enhance the efficiency of our model for human parsing, we encountered the familiar dilemma of striking a balance between accuracy and computational speed within the realm of deep learning. While recent strides have significantly improved accuracy levels, the demand for faster model processing has prompted the exploration of techniques such as

quantization. Quantization emerges as a strategy to reduce the precision of a model's weights and activations, thereby alleviating memory and computational requirements. This technique leads to reduced model sizes, expedited inference speeds, diminished memory usage, and decreased computational power consumption.

To tackle these challenges, we undertook post-training quantization (PTQ), a process involving the transition of our parsing model to half-precision quantization [58]. Unlike methods requiring additional training, PTQ solely involves calibrating the model's weights using a small calibration dataset after the initial training phase to achieve satisfactory quantization outcomes.

3.4 Activation and loss function

The choice of activation function plays a pivotal role in determining the model's performance. For instance, replacing SiLU with ReLU [59] can enhance inference speed but may lead to notable decrease in performance. While alternative activation functions like ELU [60], FReLU [61], and AconC [62] exist, our experimental investigation suggests that these alternatives do not consistently offer significant performance improvements and may also exhibit memory inefficiencies. The choice of the loss function is crucial for any deep learning-based model, especially for complex image segmentation architectures such as human parsing, since it triggers the algorithm's learning process. Deep learning algorithms employ the stochastic gradient descent technique to optimize and learn the objective. We conducted a series of tests involving various loss functions, which included Focal loss, Varifocal loss, and the loss method introduced in a recent research paper [63]. After thorough evaluation, it became evident that cross-entropy with logits emerged as the most efficient choice for training our model.

4 Dataset and experiment

4.1 Dataset

To evaluate the efficacy of our approach in comparison to edge techniques, we conducted experiments using the CIHP [12] and LIP [64] datasets.

The CIHP dataset comprises around 38,280 diverse images categorized into training (28,000), validation (5,000), and test (5,000) sets. This extensive human parsing dataset includes pixel-wise semantic part annotations for 20 categories and instance-level identification.

The LIP dataset, or Look into Person, is a large-scale dataset focused on semantic-level annotation of people. It comprises 50,000 photos with detailed pixel-by-pixel annotations for 20 classes. The dataset is divided into training (30,000), validation (10,000), and test (10,000) sets.

4.2 Experimental setup

Our experiments were conducted on two distinct environments: Environment 1 and Environment 2. For the training and ablation study, we utilized Environment 1, which comprises a Windows 10 system with 64GB of RAM, leveraging a single NVIDIA GeForce RTX 3090 GPU with 24GB of memory. To ensure a fair comparison with state-of-the-art methods, we

employed Environment 2, consisting of a Windows 10 environment with 32GB of RAM and a NVIDIA GeForce RTX 2070 GPU with 8GB of memory.

The model underwent 100 epochs of training, utilizing a batch size of 16, and the input images were resized to dimensions of 640×640. In order to augment the dataset and enhance model generalization, we applied various data augmentation techniques, encompassing geometric and lighting distortions. The initial learning rate was set at 0.005, accompanied by a momentum value of 0.9 and a decay rate of 0.0001, strategically implemented to mitigate overfitting.

5 Results and discussion

5.1 Quantitative performance

To validate our network's effectiveness, we assessed its performance on two diverse datasets (CIHP and LIP) and conducted a quantitative comparison with state-of-the-art methods. Our evaluation included a comparison with state-of-the-art (SOTA) models across metrics such as mIoU [66], inference frames per second (FPS), the number of parameters, and the model size.

Table 1 illustrates the superior performance of our network when compared to various bottom-up and one-stage top-down methods. Notably, our approach exhibits significantly better results than both one-stage (Yang et al. [65]) and bottom-up methods, including Gong et al. (2018) [12], Graphonomy [13], and Grapy-ml [39]. In the same vein, one-stage top-down methods such as Aiparsing [36] (mIoU 59.7) and Yang et al. [35] (mIoU 58.4) approach our level of performance with mIoU scores of 59.8. However, in terms of computational efficiency, our method outshines the competition. It enables faster human parsing with a higher mIoU score. Aiparsing [36] and Yang et al. [35] require 50.2 million and 150 million parameters, respectively, and achieve inference speeds of 8.9 FPS and 5.0 FPS, respectively. In contrast, our method operates with just 21 million parameters and delivers inference speeds that are 6 times faster than Aiparsing and approximately 11 times faster than Yang et al. [35]. Among the bottom-up-based methods, Graphonomy [13] achieves an impressive inference speed of around 25.0 FPS but lags behind in mIoU performance. Furthermore, the weight size of our model is remarkably compact, totaling just 43 MB, whereas the closest competitors weigh in at 159 MB and 177 MB, respectively. This makes our model resource-efficient and well-suited for edge devices, setting it apart as a more accessible and lightweight solution.

Table 1 Examples of high-performing deep learning networks' efficiency on CIHP Validation sets

Method	Approach	Backbone	Parameters (M)	Model Size (MB)	mIoU	FPS
Yang et al. [65]	One stage top-down	ResNet50	54.3	213	56.3	7.4
Yang et al. [35]	One stage top-down	ResNet50	58.4	224	58.3	5.0
Aiparsing [36]	One stage top-down	ResNet50	50.2	393	59.7	8.9
Gong et al. [12]	Bottom-up	ResNet101	629.2	1228	54.4	4.2
Graphonomy [13]	Bottom-up	Xception	157.0	159	55.5	25.0
Graphy-ml [39]	Bottom-up	Xception	176.0	177	56.2	16.6
Our's	Bottom-up	EffiBlock	**21.0**	**43**	**59.8**	**55.6**

The **bold** text indicates the best result

We also report the quantitative score of our approach on the LIP dataset (validation) shown in Table 2. The results clearly demonstrate our method's significantly outperforms compared to the current state-of-the-art techniques. To illustrate, we attain an mIoU of 61.55, surpassing the closest competitors. Specifically, the bottom-up approach CDGNet [71] achieves 60.30, the hybrid approach HIPN [28] obtains 59.61, and the top-down approach M2FP [16] also reaches 59.86.

The key advantage of our approach lies in its end-to-end design with a deep yet lighter-weight encoder and MSSA based decoder which effectively captures the global context and significantly enhances parsing accuracy. Our encoder utilizes efficient blocks, which are notably lighter than Xception, ResNet100, and ResNet50. Additionally, our decoder incorporates a multi-scale self-attention module with an efficient design. This translates into a model with fewer parameters and faster inference times. Optimization further enhances the model's compatibility with edge devices. Our model can run at 55.6 FPS, making it both efficient and suitable for edge devices.

5.2 Performance on edge devices and resource constrained devices

Edge devices, characterized by their lightweight and compact nature, are adept at efficiently running computer vision models. The concept of edge computing involves bringing data processing and storage closer to the origin of the data. This proximity enables edge devices to perform data processing internally, thereby minimizing data transmission costs and enhancing security by reducing vulnerability.

The adoption of edge devices is on the rise in various industries, as highlighted by Imani et al. (2023) in their study on efficient usage [72]. Our implementation involves deploying models on the Jetson Nano (B01), as illustrated in Fig. 5. The Jetson Nano boasts impressive specifications, featuring a Quad-core ARM Cortex-A57 MPCore processor for the CPU, NVIDIA Maxwell architecture with 128 CUDA cores for the GPU, 4 GB 64-bit LPDDR4

Table 2 Examples of high-performing deep learning networks' performance on LIP Validation sets

Method	Approach	Backbone	mIoU
BraidNet [30]	Top-down	ResNet101	54.42
SemaTree [67]	Top-down	ResNet101	54.73
SCHP [68]	Top-down	ResNet101	59.33
QANet [69]	Top-down	HRNetW48	59.61
M2FP [16]	Top-down	ResNet101	59.86
CNIF [42]	Hybrid	ResNet101	57.74
HIPN [28]	Hybrid	ResNet101	59.61
MMAN [38]	Bottom-up	ResNet101	46.93
Chen et al. [37]	Bottom-up	ResNet101	44.80
HHP [23]	Bottom-up	ResNet101	59.25
HTCorrM [26]	Bottom-up	HRNetW48	56.85
PRM [70]	Bottom-up	ResNet101	58.86
CDGNet [71]	Bottom-up	ResNet101	60.30
Our's	Bottom-up	EffiBlock	**61.55**

The **bold** text indicates the best result

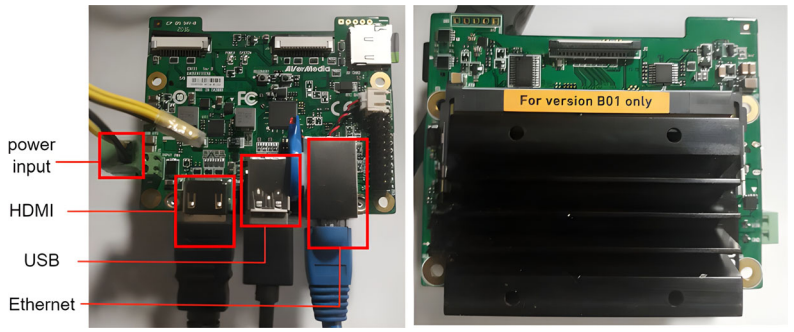

Fig. 5 The Visualization of the Jetson nano (B01) edge device

memory running at 1600MHz with a bandwidth of 25.6 GB/s, and a 16 GB eMMC 5.1 storage capacity.

Despite the inherent constraints of edge devices, our models demonstrated remarkable effectiveness on hardware with limited resources. We present comprehensive performance results for our model on image and real-time data (YouTube) in Table 3.

Our analysis revealed that employing smaller image sizes, such as 256, led to quicker inference times in comparison to larger image sizes (320). Additionally, the choice of dataset (LIP or CIHP) had a discernible impact on inference times. For instance, models operating on the LIP dataset demonstrated an inference time of 116.5 ms at an image size of 256, while for the CIHP dataset with same size, it was 129.9 ms.

To illustrate the real-time viability of our model, we conducted experiments using randomly selected YouTube videos on our Jetson Nano. The obtained results, as depicted in Fig. 6, showed high promise, even when considering the distinct lighting conditions and the absence of this data in our training and validation sets. Our model accurately predicted each frame, delivering an impressive inference speed of 8 frames per second.

A noteworthy aspect of our work is that we've managed to maintain the same level of accuracy and model quality despite the reduced inference speed on these edge devices. This robust compatibility of our approach opens the door to its potential application on various low-computing Internet of Things (IoT) devices and smartphones.

5.3 Qualitative performance

Our approach yields superior results when compared to state-of-the-art methods, exemplifying its remarkable performance. In Fig. 7 (first row), the approach by Chen et al. [37]

Table 3 Performance Comparison on Edge Devices

Dataset	Img-size	Inference
LIP	256	116.5
	320	144.4
CIHP	256	129.9
	320	149.8
Live Data	320	125.0

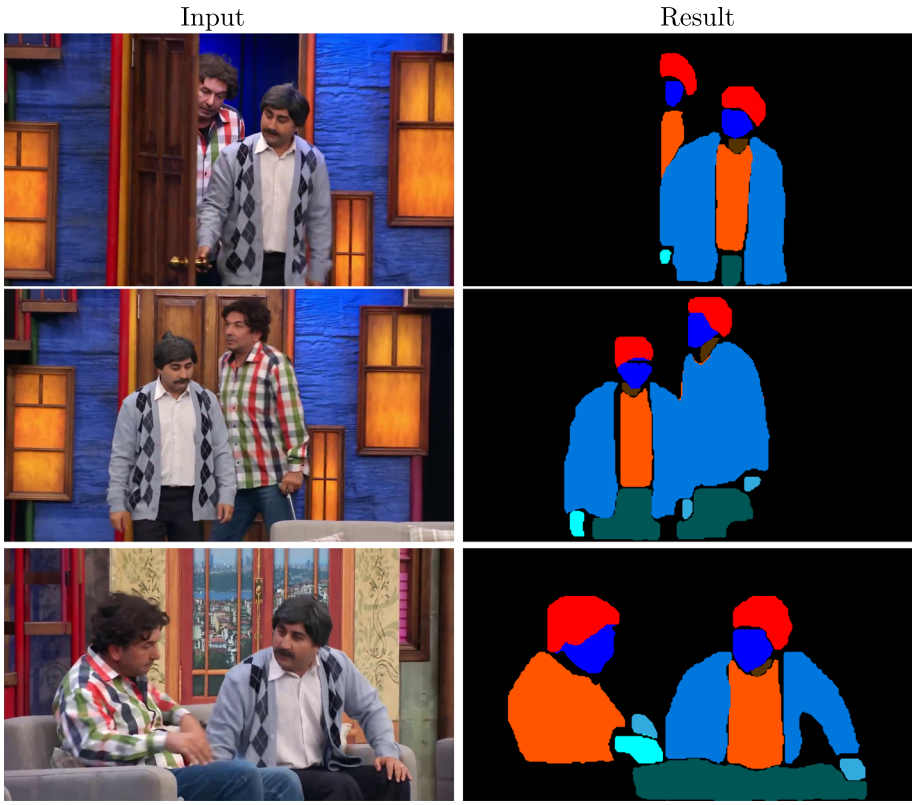

Fig. 6 Real-time Performance Evaluation on YouTube Videos with Edge Devices (Jetson Nano)

fails to accurately parse the head and hair regions, while Grapy-ml [39] falters in capturing the upper clothing of the subject. In contrast, our method excels in parsing each body part correctly, owing to its robust feature extraction capabilities, which effectively capture and represent correlations between body parts. In the second row of Fig. 7, both [37] and [39] methods struggle to parse overlapping individuals, whereas our network delivers complete and accurate results, aligning closely with ground truth annotations. Notably, for smaller regions, [37] and [39] exhibit ambiguity issues, for example, erroneously classifying a bald head as part of the background, as demonstrated in row 3. In contrast, our model correctly classifies such regions as part of the ground truth.

To ensure the broader applicability of our network, we subjected it to rigorous testing on the LIP dataset and compared it with state-of-the-art networks, further establishing its capabilities. Our approach seamlessly handles the challenges posed by the LIP dataset, as illustrated in Fig. 8. Notably, our method demonstrates superior performance across various human body parts. In the first row, our approach accurately parses the right arm, while competitors such as CorrPM [24] and CDGNet [71] struggle to do so. Similarly, in the second row, our method excels in parsing the right leg, whereas others falter. In the third row, while Chen et al. [73] encounter difficulties in parsing the left hand and CE2P [27] struggles with the upper body, our model performs with precision. Finally, in the last row, although [26] shows improvement in parsing gloves and shoes, our approach achieves even

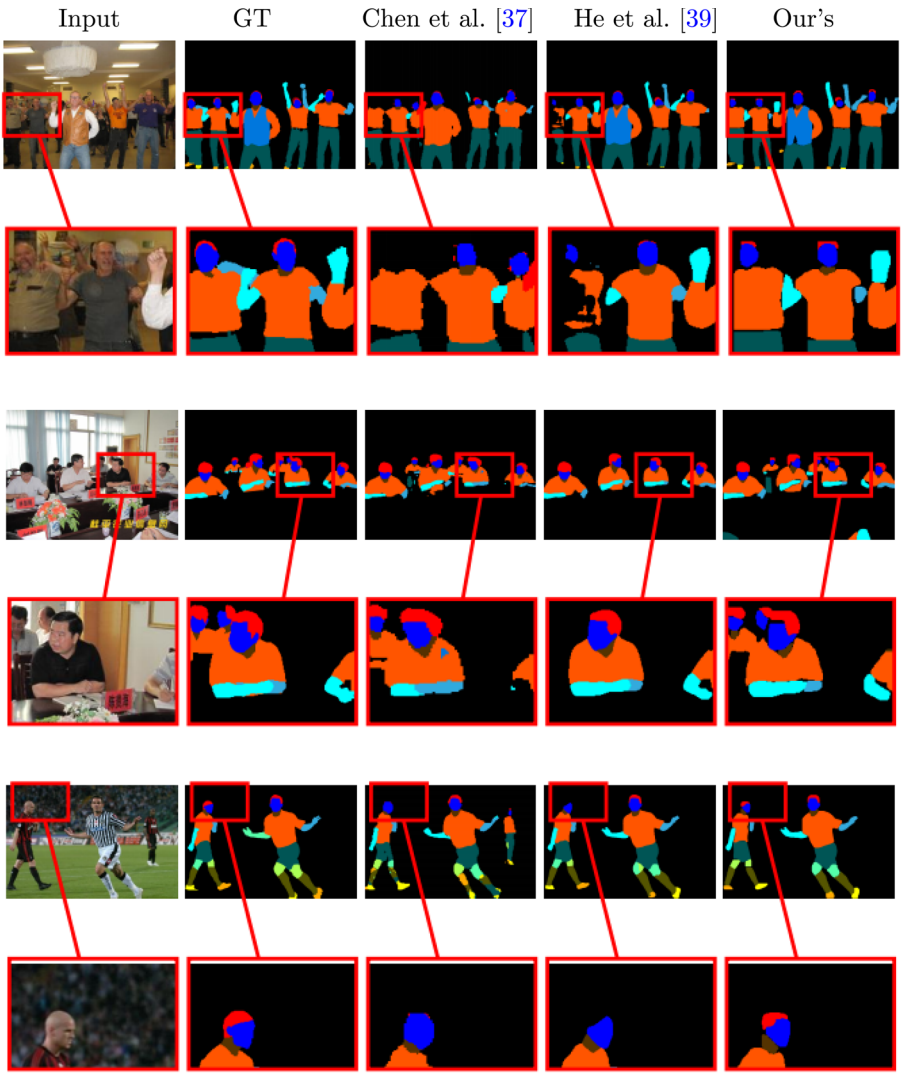

Fig. 7 Qualitative comparison of our multi-human parsing model with state-of-the-art methods on CIHP dataset

more comprehensive results. The robust and complete performance of our model can be attributed to its efficient design, particularly its enhanced feature extraction and multi-scale self-attention-based decoding mechanisms.

6 Extensive experiments

In this ablation study, we explore the effects of network modifications using the CIHP dataset. We trained the model for 10 epochs using Environment 1. To evaluate the impact of these modifications, we randomly selected 100 samples from the CIHP validation set

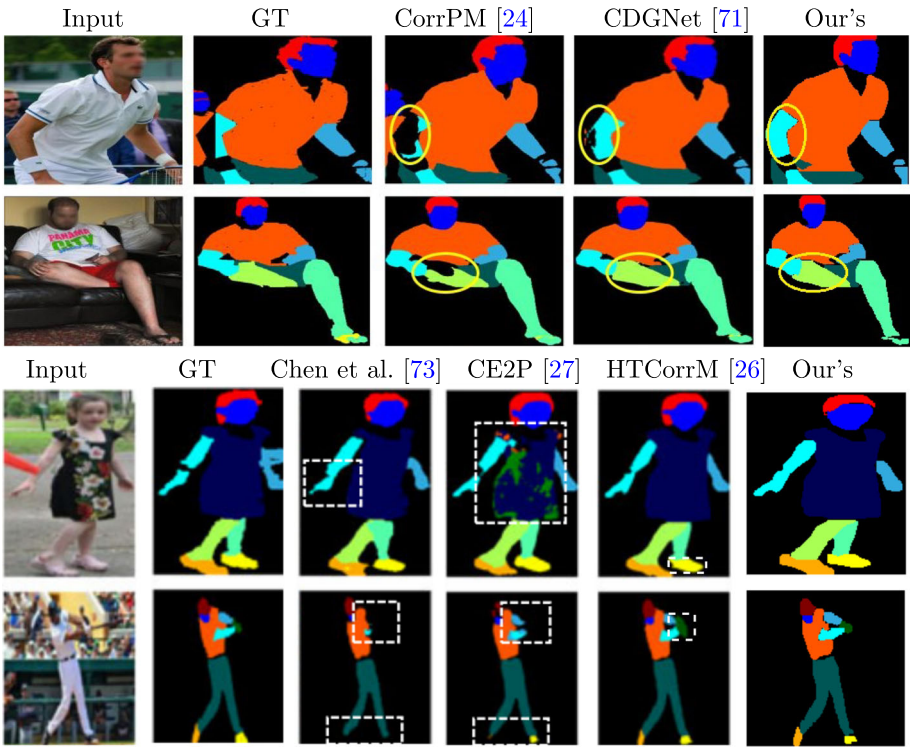

Fig. 8 Qualitative comparison of our human parsing model with state-of-the-art methods on LIP dataset

Effect of Encoder: We conducted comprehensive tests on various encoders, observing a trade-off between inference speed and accuracy. Some encoders yielded faster inference times but sacrificed accuracy, while others prioritized accuracy at the expense of speed. For instance, in our efforts to match the complexity of MobileNetV3, we streamlined our model by removing the last layer of EffiBlock (our standard model typically comprises 10 layers) and compared their performances. As illustrated in Table 4, MobileNetV3 [48] boasts fewer parameters than our streamlined version. However, despite having fewer layers, our model consistently outperforms MobileNetV3 [48] in terms of accuracy. For instance, while MobileNetV3 [48] achieves a mean Intersection over Union (mIoU) of 46.2 and takes 13.8 milliseconds to infer, our model completes inference in just 13.4 milliseconds and achieves a higher mIoU of 49.6.

Effect of Decoder: We performed a thorough comparative analysis between a conventional decoder design and our enhanced decoder featuring a Multi-scale Self-Attention. The findings reveal that our modified decoder excels in decoding features with heightened accuracy compared to the traditional UNet decoder. For instance, while the traditional decoder

Table 4 Extensive experiments on various encoders

Encoder	Layers	Parameters	mIoU	Inf. (ms)
MobileNetv3 [48]	354	6.3 M	46.2	13.8
Our's (light)	302	9.5 M	49.6	13.4

Table 5 Performance comparison between Swin Transformer and MSSA

Decoder	Layers	Parameters	mIoU	Inf. (ms)
Swin Transformer [74]	304	9.9 M	48.4	13.3
MSSA	302	9.5 M	49.6	13.4

design achieved a mean Intersection over Union (mIoU) of 47.0, employing the modified decoder with MSSA elevated the mIoU to an impressive 51.7. Although this enhancement resulted in a slight increase in inference time from 11.8 milliseconds to 14.2 milliseconds, the trade-off in time is marginal when weighed against the substantial gain in accuracy.

Effect of Multi-Scale Self-Attention (MSSA): The impact of MSSA was evaluated by replacing MSSA with Swin Transformer. Table 5 presents the performance comparison. MSSA significantly outperforms in terms of $mIoU$, while requiring fewer parameters and maintaining the same inference time.

We have also qualitatively compared the effects of MSSA and Swin Transformer attention, as shown in Fig. 9. The model trained with the Swin Transformer struggled to correctly parse each human instance, whereas the model using MSSA demonstrated comparatively better performance. For instance, in both images, the Swin Transformer misclassified upper clothes as background, while MSSA accurately identified them as the correct classes. Additionally, we analyzed the network's features with Swin and MSSA. The analysis revealed MSSA's superior ability to capture patterns, as illustrated in Fig. 9 (columns 5 and 6). It is noteworthy that our MSSA implementation utilized a single head, as opposed to a multi-head setup. Interestingly, experiments comparing single-head and multi-head configurations revealed that, for our specific task, the single-head approach slightly outperformed the multi-head approach. For example, while the multi-head configuration achieved an mIoU of 49.06%, the single-head configuration exhibited a slightly higher mIoU of 49.64%.

Effect of Activation Function: Table 6 presents the impact of different activation functions on the model's performance, evaluated through various metrics. Activation functions including SiLU, ReLU, AconC, ELU, and FReLU were examined. SiLU and ELU achieve the highest mIoU scores of 49.74% and 49.41%, respectively, indicating superior segmentation accuracy. Conversely, ReLU and FReLU exhibit lower mIoU scores of 47.24% and 47.26%, respectively. In terms of inference speed, ReLU, SiLU, and ELU perform similarly well, with inference times of 12.1 ms, 12.4 ms, and 12.4 ms, respectively. However, AconC and FReLU demonstrate slower inference speeds of 15.5 ms and 15.4 ms, respectively.

Regarding memory consumption during training, AconC requires the highest memory at 17.6 G, followed by FReLU at 16.8 G, SiLU and ELU both at 15.2 G, and ReLU at

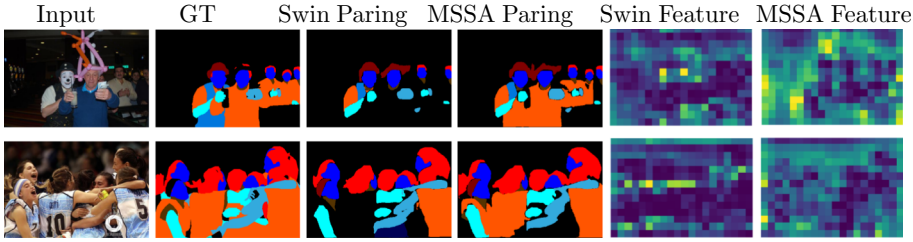

Fig. 9 Illustration of MSSA's ability to capture complex patterns in human parsing

Table 6 Effect of various activation function on our model

Activation	mIoU	Inference (ms)	Train Memory (G)
SiLU	49.74	12.4	15.2
ReLU	47.24	12.1	14.7
AconC	48.43	15.5	17.6
ELU	49.41	12.4	15.2
FReLU	47.26	15.4	16.8

14.7 G. Despite SiLU's slightly higher memory usage than ReLU, it offers better segmentation accuracy compared to ReLU, indicating a trade-off between memory efficiency and performance.

6.1 Failure cases

Our model exhibits occasional challenges in producing the desired outcomes, particularly when handling instances that are rare within the dataset or involve extreme overlap and crowding. Figure 10 illustrates instances where our approach encounters difficulties. In the first sample of this example, the model mistakenly segments a coat as the upper-cloth, while in the second sample, it becomes confused and predicts the upper-cloth as part of the background.

7 Conclusion

We present an end-to-end multi-human parsing network tailored specifically for edge devices. Our novel approach involves the creation of a deep yet lightweight encoder, strategically

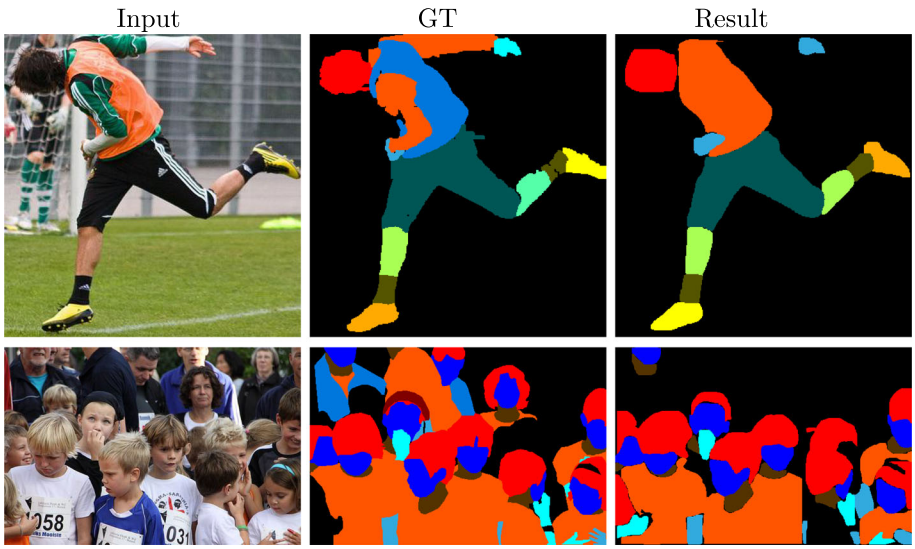

Fig. 10 An example of failure cases generated by our approach

incorporating a squeeze and excitation network to achieve a delicate equilibrium between computational efficiency and feature richness. The integration of the Multi-scale Self-Attention (MSSA) further elevates the model's ability to capture long-range dependencies, thereby significantly enhancing its representation of global features. The meticulous design of our decoder plays a crucial role in facilitating smooth information flow throughout the network. Optimization further enhances the model's compatibility with edge devices. While our model demonstrates proficiency in accurately parsing multiple humans, there are scenarios where it may encounter challenges, particularly in cases of extreme overlap, crowding, and rare occurrences within the dataset. Addressing these complexities may involve enhancing the model's capability to handle heavily overlapping instances and rare scenarios with greater agility.

Acknowledgements Not applicable.

Availability of data and materials The source code for this work is available upon request to the corresponding author.

Declarations

Ethics approval and consent Not applicable

Competing interests The authors declare that they have no competing interests.

References

1. Kikuchi T, Endo Y, Kanamori Y, Hashimoto T, Mitani J (2018) Transferring pose and augmenting background for deep human-image parsing and its applications. Computational Visual Media. 4:43–54
2. Zhou D, Zhang C, Tang Y, Li Z (2022) Fine-grained alignment network and local attention network for person re-identification. Multimedia Tools and Applications. 81(30):43267–43281
3. Tong Z, Xu P, Denoeux T (2021) Evidential fully convolutional network for semantic segmentation. Appl Intell 51:6376–6399
4. Sun Y, Hu J, Shi J, Sun Z (2020) Progressive decomposition: a method of coarse-to-fine image parsing using stacked networks. Multimedia Tools and Applications. 79(19):13379–13402
5. Liu M, Yan X, Wang C, Wang K (2021) Segmentation mask-guided person image generation. Appl Intell 51:1161–1176
6. Yang L, Song Q, Wang Z, Hu M, Liu C (2020) Hier r-cnn: Instance-level human parts detection and a new benchmark. IEEE Trans Image Process 30:39–54
7. Lin W, Liu H, Liu S, Li Y, Qian R, Wang T, Xu N, Xiong H, Qi G-J, Sebe N (2020) Human in events: A large-scale benchmark for human-centric video analysis in complex events. arXiv:2005.04490
8. Kumar P, Chauhan S, Awasthi LK (2022) Human pose estimation using deep learning: review, methodologies, progress and future research directions. International Journal of Multimedia Information Retrieval. 11(4):489–521
9. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28:801–818
10. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969
11. Zhu B, Chen Y, Tang M, Wang J (2018) Progressive cognitive human parsing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, pp. 7607–7614
12. Gong K, Liang X, Li Y, Chen Y, Yang M, Lin L (2018) Instance-level human parsing via part grouping network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 770–785
13. Gong K, Gao Y, Liang X, Shen X, Wang M, Lin L (2019) Graphonomy: Universal human parsing via graph transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7450–7459

14. Song J, Shi Q, Li Y, Yang F (2022) Enhanced context learning with transformer for human parsing. Appl Sci 12(15):7821
15. Guan H, Chen M, Su Z (2022) Graph transformer for human parsing. In: 2022 9th International Conference on Digital Home (ICDH), pp. 87–92
16. Yang L, Jia W, Li S, Song Q (2023) Deep learning technique for human parsing: A survey and outlook. arXiv:2301.00394
17. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141
18. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241
19. Nie X, Feng J, Yan S (2018) Mutual learning to adapt for joint human parsing and pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 502–517
20. Liang X, Gong K, Shen X, Lin L (2018) Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE Trans Pattern Anal Mach Intell 41(4):871–885
21. Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X et al (2020) Deep high-resolution representation learning for visual recognition. IEEE Trans Pattern Anal Mach Intell 43(10):3349–3364
22. Zhang X, Chen Y, Zhu B, Wang J, Tang M (2020) Part-aware context network for human parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8971–8980
23. Wang W, Zhu H, Dai J, Pang Y, Shen J, Shao L (2020) Hierarchical human parsing with typed part-relation reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8929–8939
24. Zhang Z, Su C, Zheng L, Xie X (2020) Correlating edge, pose with parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8900–8909
25. Liu J, Zhang Z, Shan C, Tan T (2020) Kinematic skeleton graph augmented network for human parsing. Neurocomputing 413:457–470
26. Zhang Z, Su C, Zheng L, Xie X, Li Y (2021) On the correlation among edge, pose and parsing. IEEE Trans Pattern Anal Mach Intell 44(11):8492–8507
27. Ruan T, Liu T, Huang Z, Wei Y, Wei S, Zhao Y (2019) Devil in the details: Towards accurate single and multiple human parsing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33(01), pp. 4814–4821
28. Liu Y, Zhang S, Yang J, Yuen P (2021) Hierarchical information passing based noise-tolerant hybrid learning for semi-supervised human parsing. Proceedings of the AAAI Conference on Artificial Intelligence 35:2207–2215
29. Zhou Y, Mok P (2023) Enhancing human parsing with region-level learning. IET Computer Vision
30. Liu X, Zhang M, Liu W, Song J, Mei T (2019) Braidnet: Braiding semantics and details for accurate human parsing. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 338–346
31. Yang L, Liu Z, Zhou T, Song Q (2022) Part decomposition and refinement network for human parsing. IEEE/CAA Journal of Automatica Sinica. 9(6):1111–1114
32. Yan M, Zhang G, Zhang T, Zhang Y (2021) Nondiscriminatory treatment: A straightforward framework for multi-human parsing. Neurocomputing 460:126–138
33. Zhao J, Li J, Cheng Y, Sim T, Yan S, Feng J (2018) Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 792–800
34. Qin15 H, Hong25 W, Hung W-C, Tsai Y-H, Yang35, M-H (2019) A top-down unified framework for instance-level human parsing. University of California Merced
35. Yang L, Song Q, Wang Z, Hu M, Liu C, Xin X, Jia W, Xu S (2020) Renovating parsing r-cnn for accurate multiple human parsing. In: European Conference on Computer Vision, pp. 421–437
36. Zhang S, Cao X, Qi G-J, Song Z, Zhou J (2022) Aiparsing: Anchor-free instance-level human parsing. IEEE Trans Image Process 31:5599–5612
37. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818
38. Luo Y, Zheng Z, Zheng L, Guan T, Yu J, Yang Y (2018) Macro-micro adversarial network for human parsing. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 418–434
39. He H, Zhang J, Zhang Q, Tao D (2020) Grapy-ml: Graph pyramid mutual learning for cross-dataset human parsing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34(07), pp. 10949–10956

40. Zhou T, Wang W, Liu S, Yang Y, Van Gool L (2021) Differentiable multi-granularity human representation learning for instance-aware human semantic parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1622–1631
41. Zhou T, Yang Y, Wang W (2023) Differentiable multi-granularity human parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence
42. Wang W, Zhang Z, Qi S, Shen J, Pang Y, Shao L (2019) Learning compositional neural information fusion for human parsing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5703–5713
43. Wang W, Zhou T, Qi S, Shen J, Zhu S-C (2021) Hierarchical human semantic parsing with comprehensive part-relation modeling. IEEE Trans Pattern Anal Mach Intell 44(7):3508–3522
44. Wang C-Y, Liao H-YM, Wu Y-H, Chen P-Y, Hsieh J-W, Yeh I-H (2020) Cspnet: A new backbone that can enhance learning capability of cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 390–391
45. Jocher G. ultralytics/yolov5: V6.0 - YOLOv5n 'Nano' Models, Roboflow Integration, TensorFlow Export, OpenCV DNN Support. https://doi.org/10.5281/zenodo.5563715
46. Liu F, Liu J, Fu J, Hanqing L (2018) Improving residual block for semantic image segmentation. In: 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), pp. 1–5
47. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520
48. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V et al (2019) Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324
49. Wang C, MacGillivray T, Macnaught G, Yang G, Newby D (2018) A two-stage 3d unet framework for multi-class segmentation on full resolution image. arXiv:1804.04341
50. Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768
51. Liu Y, Zhang D, Zhang Q, Han J (2021) Part-object relational visual saliency. IEEE Trans Pattern Anal Mach Intell 44(7):3688–3704
52. Chen C, Han J, Debattista K (2024) Virtual category learning: A semi-supervised learning method for dense prediction with extremely limited labels. IEEE Trans Pattern Anal Mach Intell 46(8):5595–5611
53. Arbelaez P, Maire M, Fowlkes C, Malik J (2010) Contour detection and hierarchical image segmentation. IEEE Trans Pattern Anal Mach Intell 33(5):898–916
54. Sinha A, Dolz J (2020) Multi-scale self-guided attention for medical image segmentation. IEEE J Biomed Health Inform 25(1):121–130
55. Chen S, Duan J, Zhang N, Qi M, Li J, Wang H, Wang R, Ju R, Duan Y, Qi S (2023) Msa-yolov5: Multi-scale attention-based yolov5 for automatic detection of acute ischemic stroke from multi-modality mri images. Comput Biol Med 165:107471
56. Duan H, Long Y, Wang S, Zhang H, Willcocks CG, Shao L (2023) Dynamic unary convolution in transformers. IEEE Trans Pattern Anal Mach Intell 45(11):12747–12759
57. Bolya D, Zhou C, Xiao F, Lee YJ (2019) Yolact: Real-time instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9157–9166
58. Yu X, Qin L, Chen X, Wu L, Zhang B (2023) Research on optimization of neural network model deployment for edge devices. In: 2023 4th International Conference on Computer Engineering and Intelligent Control (ICCEIC), pp. 130–134
59. Agarap AF (2018) Deep learning using rectified linear units (relu). arXiv:1803.08375
60. Clevert D-A, Unterthiner T, Hochreiter S (2015) Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289
61. Qiu S, Xu X, Cai B (2018) Frelu: flexible rectified linear units for improving convolutional neural networks. In: 2018 24th International Conference on Pattern Recognition (icpr), pp. 1223–1228
62. Ma N, Zhang X, Liu M, Sun J (2021) Activate or not: Learning customized activation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8032–8042
63. Köksal A, Tuzcuoğlu Ö, İnce KG, Ataseven Y, Alatan AA (2022) Improved hard example mining approach for single shot object detectors. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 3536–3540
64. Gong K, Liang X, Zhang D, Shen X, Lin L (2017) Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 932–940
65. Yang L, Song Q, Wang Z, Jiang M (2019) Parsing r-cnn for instance-level human analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 364–373

66. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S (2019) Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 658–666
67. Ji R, Du D, Zhang L, Wen L, Wu Y, Zhao C, Huang F, Lyu S (2020) Learning semantic neural tree for human parsing. In: European Conference on Computer Vision, pp. 205–221
68. Li P, Xu Y, Wei Y, Yang Y (2020) Self-correction for human parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence
69. Yang L, Song Q, Wang Z, Liu Z, Xu, S, Li Z (2022) Quality-aware network for human parsing. IEEE Transactions on Multimedia
70. Zhang X, Chen Y, Tang M, Wang J, Zhu X, Lei Z (2022) Human parsing with part-aware relation modeling. IEEE Transactions on Multimedia
71. Liu K, Choi O, Wang J, Hwang W (2022) Cdgnet: Class distribution guided network for human parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4473–4482
72. Imani H, Hosen MI, Feryad V, Akyol A (2023) Efficient object detection model for edge devices. In: International Conference on Advanced Engineering, Technology and Applications, pp. 83–94
73. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
74. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.