Ultra-light Multi-scale Human Parsing with Attention-guided Grouping

Abderrahim Ouza [1,*], Mohamed El Ghmary [2], and Ali Choukri [1]

[1] Faculty of Science, Ibn Tofail University, Morocco
[2] ENSAM, Mohammed V University in Rabat, Morocco
Email: Abderrahim.ouza@uit.ac.ma (A.O.); mohamed.elghmary@um5.ac.ma (M.E.G.); ali.choukri@uit.ac.ma (A.C.)
*Corresponding author

***Abstract*—Human parsing in large and dynamic environments is commonly handled by large-scale models that deliver high segmentation accuracy but require hundreds of millions of parameters and substantial computational resources. This level of complexity makes such approaches difficult to deploy on power- and resource-constrained devices, particularly in real-time edge Artificial Intelligence (AI) settings. In this paper, we introduce Fast Dilated Spatial Pyramid Pooling + Parsing Grouping Network + Attention (Fast DSPP + PGN + ATTN), an efficient human parsing framework designed to balance segmentation performance with computational efficiency. The model employs a MobileNetV2 encoder with DSPP module for contextual modeling and a PGN-style grouping decoder integrated with light-weight spatial attention and Squeeze-and-Excitation (SE) attention. With 2.14 M parameters and 5.70 Giga Floating Point Operations (GFLOPs), it attains 40.67% mean Intersection over Union (mIoU). In ablation results, we observe that DSPP has the most significant effect (0.0151 mIoU), while spatial attention and SE yield smaller improvements. The accuracy of the grouping decoder is retained while the number of computations can be reduced. The model achieves 0.2490 mIoU on Look into Person (LIP) dataset, with variants to 0.2640, thus confirming robustness over domain shift. This indicates that integrating structured pixel grouping and efficient contextual aggregation together can solve the task of human parsing, even with severe computational constraints, which could render the proposed method particularly appealing to real-world applications such as embedded vision systems/mobile perception and real-time surveillance.**

***Keywords*—human parsing, efficient architectures, contextual modeling, edge Artificial Intelligence (AI), multi-scale feature**

I. Introduction

Semantic segmentation is one of the main computer vision tasks that requires assigning a semantic label to every pixel in an image, and for understanding scenes [1, 2]. For human-centred scenarios, this task is extended into human parsing: localizing the body parts and clothing categories in complex environments. Such functionality is crucial for applications like surveillance, behavior analysis of people in a crowd and in immersive systems. Important advances in recent years have been made within convolutional architectures, exploiting encoder-decoder designs, e.g., U-Net and DeepLab, that merge high-level semantic information with recovered spatial details [3, 4].

A concurrent research line is in lightweight models that have small computational budgets (generally <5 M parameters and single-digit GFLOPs) to allow real-time inferencing on edge devices. Still, high-performing human parsing methods largely rely on large-scale backbones like ResNet-101 or High-Resolution Network (HRNet), or transformer-based architectures [5]. These architectures obtain impressive accuracy (typically 40–60% mIoU on difficult datasets such as Crowd Instance-level Human Parsing (CIHP) dataset, but they come at a high computational and time cost, with a common order of over 100 GFLOPs/image feed [6, 7]. This very requirement restricts their use in latency-sensitive or resource-constrained environments. This gap indicates that the new architectures have to keep competitive segmentation quality while decreasing computational cost for practical deployment on edge and mobile platforms.

On the other hand, lightweight models are often weak to large occlusions, overlapping people and complex body-part boundaries [5]. The feature prevents them from attending to multi-scale contexts, as they are limited to smaller receptive fields and excellence, which causes fuzzy or wrong predictions [8]. Thus, the real challenge is to build a fast yet reliable and practical human parsing system suitable for both context interference environment and low power hard-ware [9].

To address these issues, this paper proposes a simple human parsing framework. The goal is to enhance multi-scale feature extraction and boundary refining while maintaining the lowest computation cost. While recent deep network methods achieve impressive results, they inevitably produce high-frequency visual artefacts in the predicted segmentation [9]. The detailed proposed model is shown in Fig. 1, which seems 4 instantiation components: (i) a MobileNetV2 encoder that uses inverted residual blocks and depthwise separable convolutions to

Manuscript received December 28, 2025; revised April 23, 2026; accepted May 12, 2026; published July 28, 2026.

extract features efficiently; (ii) the DSPP module was introduced for aggregating contextual information across multiple receptive fields [9]; (iii) pixel-grouping based decoder inspired on Parsing Grouping Network (PGN) capability with respect to continue spatial coherence and improve boundaries under occlusion; (iv) a light weight dual attention mechanism consists of spatial attention and squeeze-and-excitation for discriminates human parts from highly confused background [10].

We also performed additional experiments on the Crowd Instance-level Human Parsing (CIHP) dataset, which demonstrates that the proposed architecture with only a model with 2.14 M parameters and a computational cost of 5.70 GFLOPs can achieve real-time performance at 51.9 Frames Per Second (FPS) while reaching an mIoU of 40.67% and a pixel accuracy 87.3%. The adaptation of multi-scale context modeling, pixel groupings and attention mechanisms results in a good trade-off between accuracy and efficiency that is suitable for dense scenes while maintaining edge deployment adaptability. The quantitative ablation study shows that the DSPP module contributes most to performance, while separating spatial attention and SE modules brings a consistent refinement, and a grouping decoder is presented to increase efficiency without losing accuracy. Furthermore, we also evaluate on the LIP dataset in domain shift condition and observe that our model still has competitive performance (0.2490 mIoU) while even our variant can reach 0.2640 mIoU, suggesting the robustness of the proposed architecture as extensive comparisons.

This work has the following key contributions:

- We propose a unified DSPP human parsing framework based on MobileNetV2 encoder-decoder, including multi-scale context modeling via Deep Separable Pyramid Pooling (DSPP), augmented affinity-based spatial grouping leveraging Partial Group Normalization (PGN), and complementary attention module.
- We introduce an encoder that is sensitive to boundaries, which leverages pixel affinities from multimodal data under the influence of occlusion, for enhancing structural consistency and better differentiating penumbras (overlapped body parts)—a concern many light models address poorly.
- We introduced a very good trade-off, the compromise between segmentation accuracy and computational cost that yields achieving 40.67% mIoU while maintaining a lightweight footprint of 2.14 M parameters and 5.70 GFLOPs (Table I). The proposed framework demonstrates consistent improvements over representative lightweight baselines while being at approximately 50 FPS, i.e., following the speed in real time.
- We perform a systematic analysis of our components at the component level (both quantitative and qualitative)—evaluation of training dynamics, per-class distributions, attention visualizations, multi-scale feature responses—so that we can ascertain how much each module contributes to overall performance.
- We conduct a quantitative ablation study and cross-dataset evaluation on LIP to further validate the effectiveness of each component as well as the generalization ability of our model.

In contrast to previous works that concentrate on augmenting model capacity, we explore how complementary mechanisms can be optimally co-located within unforgiving compute budgets. The novelty therefore, brings together the interaction of the multi-scale context modeling, grouping constraints and attention mechanism in an ultra-lightweight regime instead of any individual components.

The rest of the paper is organized as follows; Section II surveys related works. Section III outlines the proposed method. In Section IV, we present the research findings and a discussion. Related works and the conclusion with directions for future research are presented in Section V.

II. Related Works

Over the past decade, semantic segmentation has, however, achieved tremendous improvements in deep-learning methods [9]. The following segmentation approaches were mostly dependent on handcrafted features, and conventional machine learning approaches perform pixel-level interpretation using handcrafted features and statistical classifiers [11]. However, these techniques were groundbreaking for their time, but they struggled with complexities and ambiguities present in real-world scenes [12]. A major limitation of these approaches are their inability to capture interdependencies between various regions in an image as well as a limited modeling capacity of complex spatial context, which is critical for accurate scene understanding [6, 13].

Fully Convolutional Networks (FCNs) are another landmark discussion in the era of semantic segmentation [12]. FCNs enabled end-to-end learning directly from raw input data while patch-based Convolutional Neural Network (CNN) classifiers only learned within small local regions in isolation [12]. This change in architecture allowed models to learn much richer representations for performing dense prediction tasks, which increased segmentation accuracy on many benchmarks [2, 3]. However, FCNs failed when confronted with multi-scale reasoning problems, and loss of boundary precision surfaced widely as a result of their limited insufficient receptive field coverage and significant computational complexity in downsampling processes.

One approach is to set up a dilated (or atrous) convolutions [14] to address the bottleneck of receptive-field receptivity. Dilation convolutions gap the Convolution kernels, allowing to exponentially enlarge the area of reaction and reduce parameters [6]. Such a network is capable of capturing both fine-grained detail and long-range context at scale, which makes it essential in complex scenes. Together with the related architectures from the DeepLab family, they laid a foundation to combine pyramid pooling with convolutions having dilated coefficients, and thus an efficient multi-scale feature aggregation [2].

Explicit coupling upsampling and downsample path-ways (i.e. U-Net) such as developed in Section II also allowed for further technology advances [1, 15]. The skip connections allow merging the fine-grained spatial features fused with deeper semantic information and make it much easier to find boundaries [7]. These structures quickly became the foundation of diverse segmentation tasks, together with medical imaging, notably in human part segmentation applications [16]. Subsequent versions adopted depthwise separable convolutions to reduce computational complexity [17], inspired by building upon efficient architectural principles introduced by lightweight backbone architectures exemplified by MobileNet and ShuffleNet [18, 19], to achieve a better cost/performance trade-off.

Semantic segmentation does not directly predict the fine-grained polygons of objects, however, semantic segmentation is an additional human parsing challenge that contributes to difficulties. Mismatches often arise in image pairs with multiple interacting people, occlusions from posing, and similar or coincident appearances of legs. This data set from CIHP has become a de facto benchmark standard for parsing crowded human. To cope with diverse challenges posed by overlapping or occluded body parts, a number of specialized approaches were proposed, including two-stage networks based on Part Grouping Network (PGN), to learn about pixel affinities and instances to construct social acceptability relations among different parts [20, 21]. Recently, functional uses of contextual encoding as well as cross-scale feature fusion have been broadened; however, they heavily depend on enormous backbones and intricate decoders that incur extensive computational expense [22, 23].

But even the most accurate of these models is not fit for real-time or resource-constrained environments due to their heavy computational cost. On the other hand, adding lightweight architectures with efficient backbones like MobileNet is a more reasonable option for edge deployment [11]. In general, these designs use a very tight trade-off between all three critical dimensions of latency, accuracy, and memory footprint [24]. This balance is important in application domains including autonomous systems, mobile computing, and surveillance operations, which may have strict power consumption limits as well as real-time latency requirements.

To tackle this difficulty, this work uses a MobileNetV2 encoder with the addition of a DSPP-based context module and a PGN-inspired grouping decoder in a lightweight framework. The architecture we propose also verifies the hypothesis that a small network can retain reasonable segmentation performance on the CIHP dataset, while achieving fast inference time suitable for edge devices.

III. Method

A. Overview of the Proposed Framework

This work proposes the Fast Dilated Spatial Pyramid Pooling with Pixel Grouping Network and Dual Attention framework, which is a model of feature extraction and reconstruction architecture that aims to obtain highly accurate but computationally-constrained human parsing. In contrast to typical strategies that increase the depth or width of a network in search of performance gains, our contribution lies in the framework of structural efficiency, that is, simultaneously combining complementary mechanisms while maximizing representational capacity without increasing footing.

Let input image $X = \mathbb{R}^{H \times W \times 3}$ and dense prediction semantic segmentation map $Y = \mathbb{R}^{H \times W \times C}$, where C is the number of classes to predict. Overall transformation denoted by:

$$Y = D(G(C(\varepsilon(X)))) \quad (1)$$

where:

- ε is a lightweight feature encoder,
- C is a multi-scale context aggregation module (DSPP),
- G is a grouping-based decoder,
- D is the final prediction layer.

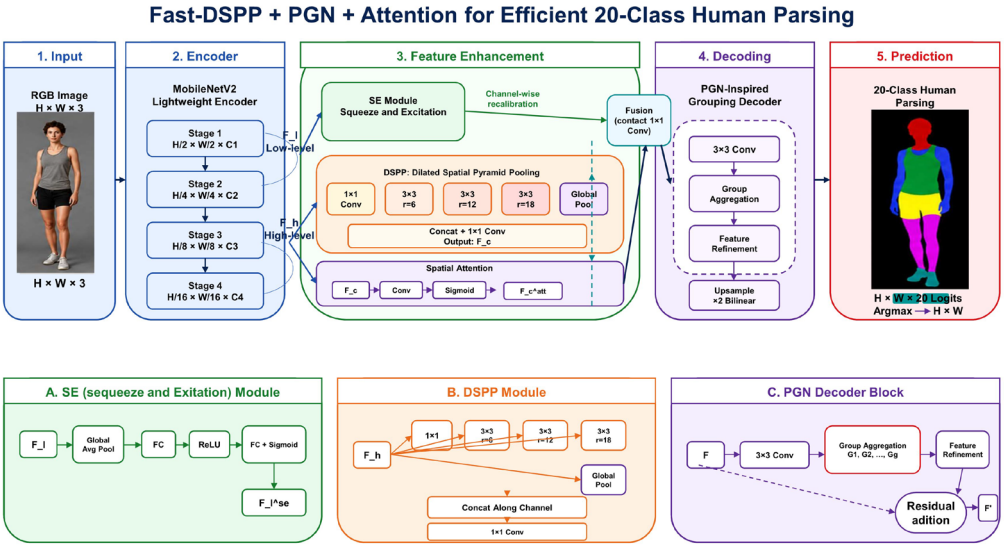


Fig. 1. Schematic representation of the proposed model architecture.

This work is novel, however, not due to the introduction of entirely new components, but rather because in this work we study context modeling, grouping constraints and attention mechanisms jointly as a design and optimization problem under extremely low-complexity settings. Such a design directly opposes the primary weaknesses of lightweight models in terms of limited receptive field, boundary localization and feature discriminability.

For reference, we provide the entire pipeline in Fig. 1, depicting how encoder features are involved in multi-scale context integration, grouping-based structural refinement, and attention-guided decoding.

B. Architecture Components and Design Justification

1) Lightweight encoder

The encoder ε, based on MobileNetV2 with inverted residual blocks and depthwise separable convolutions, aims to reduce a considerable computational cost. To keep efficiency, we truncate the network into output stride 16.

Two feature maps are extracted:

- Low-level features F_l (high resolution), without losing fine spatial details,
- High-level features F_h (low resolution), which contain semantic information.

Representation in this dual-scale way allows the direct fusion of spatial precision and semantic context, which is critical for human parsing problems with small structures and complex boundaries.

2) Multi-Scale Context Modeling (DSPP)

Receptive fields are generally small for lightweight encoders. To overcome this limitation, a Dilated Spatial Pyramid Pooling (DSPP) module is used on F_h.

So the output of the DSPP is defined as:

$$F_{DSPP} = \sum_{i=1}^{K} \phi_{d_i}(F_h) \tag{2}$$

where ϕ_{d_i} denotes a dilated convolution with dilation rate d_i.

The model, therefore utilizes a combination of dilation rates, which enables it to jointly represent local spatial structures and global contextual relationships at slightly higher parameter counts. It is especially important when the contextual relations among body parts have to be preserved with the presence of crowd such as scenes where people are close to each other.

3) Grouping-based decoder

The independent pixel treatment of standard segmentation models is one important reason for such inconsistent boundaries. To address this problem, the proposed decoder presents a grouping approach based on PGN.

The affinity between two pixels, i and j, is modelled as follows, where the given feature representations are denoted F:

$$A_{ij} = \sigma(f(F_i, F_j)) \tag{3}$$

where $f(\cdot)$ is a learnable similarity function and σ is the sigmoid activation.

This ensures similar representations between pixels that lie in a common semantic region. This allows the model to become structurally consistent and more robust against occlusion, especially for overlapping people.

4) Attention mechanisms

In addition, two lightweight attention modules are utilized in order to enhance the discrimination of features:

- Spatial attention, which is applied to DSPP features so that informative regions can be highlighted and background irrelevant (large empty area in those images) to our target location can also be suppressed,
- Squeeze-and-Excitation (SE), applied to skip connections, adaptively readjusts channel-wise feature importance.

Spatial attention improves localization, and SE makes the channel more selective (so they all compensate). Most significantly, both modules add tangible but very low computational overhead and thus enable ultra-lightweight design.

C. Training Strategy and Optimization

The complete training and inference pipeline is summarized in Algorithm 1.

Algorithm 1. Fast-DSPP + PGN + ATTN Forward Architecture

Require: Input image $X \in \mathbb{R}^{H\times W\times 3}$
Ensure: Pixel-wise prediction map $\hat{Y}$

1: Extract low-level features F_l and high-level features F_h using the MobileNetV2 encoder $\mathcal{E}$
2: Apply the DSPP module to F_h to obtain multi-scale contextual features:
$$F_c \leftarrow C(F_h)$$
3: Refine F_c using the spatial attention module to emphasize informative human regions
4: Recalibrate the skip features Fl using the squeeze-and-excitation (SE) block
5: Upsample the refined contextual features F_c
6: Fuse the upsampled features with the SE-enhanced low-level features
7: Apply the PGN-inspired grouping decoder G to enforce local spatial consistency
8: Generate class logits through the final prediction layer D
9: Compute pixel-wise predictions:
$$\hat{Y} \leftarrow \arg\max(D(G(C(\mathcal{E}(X)))))$$
10: **Return** $\hat{Y}$

The model is optimized using a composite loss function:

$$L = L_{seg} + \lambda L_{group} \tag{4}$$

where:

- L_{seg} is the pixel-wise cross-entropy loss,
- L_{group} is a binary cross-entropy loss applied to affinity predictions,
- $\lambda = 0.3$ controls the trade-off.

The grouping loss is essential for ensuring boundary consistency and enhancing segmentation robustness, particularly in complex scenes.

Train using the Adam optimizer with an initial value of the learning rate 5×10^{-4} and cosine-based learning rate decay scheduling. These considerations are reflected in our data augmentation techniques, of which random horizontal flipping, scale jittering in [0.75, 1.25], and random cropping are aimed towards improving generalization.

The implementation applies the mixed-precision training available in TensorFlow 2.13 to save memory and speed up computation. The model runs in real-time for edge deployment without multi-scale testing and post-processing (single forward pass input during inference).

Each module is made to solve a specific drawback of lightweight models: the DSPP captures additional context than prevailing methods with a clear and loose receptive area, grouping reduces boundary inconsistency, and attention enhances weak feature discrimination. The joint integration of both guarantees complementary behavior without greater computational expense.

IV. Experimental Evaluation and Discussion

The present section testifies the introduced Fast-DSPP+PGN+ATTN framework on the CIHP human parsing benchmark, aiming to give concern about not only quantitative performance, but also the interpretable behavior of the model. In an ultra-lightweight regime, the model shows signs of large architecture properties such as boundary consistency and contextual reasoning.

One of the key challenges in human parsing is to reduce ambiguity at object boundaries—a problem effectively addressed through the central multi-scale context aggregation DSPP grouping-based refinement. More than global statistics, the model has stable optimization dynamics and continues to have spatial coherence between scenes. It indicates that, at least under strict computational constraints, structural design choices have a greater impact on segmentation performance than the scale of the model.

The section is structured as follows. We start with a description of the experimental setup and evaluation metrics. Last, the key quantitative results are presented and evaluated against to those from state-of-the-art human parsing models of face. We investigate the learning behavior and error patterns then perform with qualitative evaluation through visual inspection grounded on analysis of attention heatmaps and multi-scale feature representations and two-to-two representative qualitative results. Afterwards, we stipulate the principal limitations and broader implications of the results.

A. Experimental Configuration and Performance Metrics

Starting from the established CIHP training and validation protocol, following conventional approaches [25], we partition the images such that the dataset is partitioned into 30,000 training images and 8000 validation images, spanning over 20 distinct semantic categories (19 human body part classes and one background category). All images are rescaled by scaling the shorter spatial dimension to 384 pixels; at train time we sample a random spatial crop of size 384×384, and at inference time take a single centrally extracted 384×384 crop. We do not make modifications such as multi-scale inference or flip-based augmentation at test time, which means our runtime measurements are directly applicable to the real-time setting.

All experimental evaluations are carried out in the Google Colab Pro+ environment on a NVIDIA GPU that is single and use TensorFlow 2.13. It is also employing mixed-precision training for improved computational efficiency. The model is optimized. It used Adam optimizer with an initial value of 5×10^{-4} and a cosine annealing learning rate schedule over 50 epochs. The batch size is 8. The optimization objective relies on sparse categorical cross-entropy [24].

The main segmentation prediction layer, whereas the PGN-inspired grouping branch, trained using an objective based on binary cross-entropy weighted $\lambda = 0.3$. Data augmentation: horizontal flip augmentation flips, scale jitter in [0.75, 1.25], random spatial cropping combined with light augmentation strategies, photometric augmentation (estimated that this gives robustness and covariance in similar setups [5]).

We evaluate the model using standard semantic segmentation evaluation metrics:

- pixel-level classification accuracy (PA): ratio of correctly predicted pixels over all image pixels.
- average per-class pixel accuracy (mPA).
- Intersection over Union (IoU) for class c:

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c}$$

where TP, FP, and FN are defined as True Positives, False Positives, and False Negatives, respectively.

- Mean IoU (mIoU): arithmetic mean of IoU over the 20 classes.

We report the mean per-image inference latency and Frames Per Second (FPS) on CIHP validation set with a mini-training batch size of 1 and 384×384 spatial input size. This facilitates a straightforward comparison under consistent evaluation settings with edge-centric practical real-world applications.

B. Comprehensive Quantitative Results

We first report the simple performance and efficiency metrics of the proposed model over CIHP, followed by direct comparisons against existing methods. The essential metrics refer to mIoU, pixel accuracy, computational complexity (FLOPs), number of parameters and inference speed (FPS). As shown in Table II, the proposed methodology has a property to reach a good trade-off between high predictive performance and computational efficiency.

For example, Table I lists the key indicators: pixel accuracy, mIoU, average medium time precision, together with their computational load measured as the number of parameters, means the main statistics threshold measure of

the Computational Compute Performance structure Fast DSPP+PGN+ATTNin (CIHP validation set). These indicators deliver a simple interpretation of how effectively the model maintains a trade-off between segmentation accuracy and efficiency with computational speed under the constraint of realistic deployment.

TABLE I. QUANTITATIVE PERFORMANCE AND EFFICIENCY METRICS OF THE MODEL

Metric	Value
PA	87.30%
mIoU	40.67%
Parameters (M)	2.14
GFLOPs	5.70
FPS	51.94
Latency (ms)	19.25

As shown in Table I, the proposed model achieves 40.67% mIoU, 87.3% pixel accuracy with only 2.14 M parameters and 5.70 GFLOPs.

More importantly, these results indicate a crucial finding: that performance deterioration of lightweight models is not merely due to a loss in capacity but more often attributed to poor context modeling and weak reasoning at the boundaries. Following the approach of DSPP, and enforcing grouping constraints, the proposed architecture is able to recoup most of the accuracy attributable to larger models, specifically addressing these two limitations.

This implies that well-designed structural parts can mitigate lower bean size, from capacity versus accuracy towards design against figure. This allows it to fit well with real edge implementations, where the computational ability may be restricted, but spatial accuracy is still important.

C. Comparison with Existing Methods

Not a minor architectural change, to prove that the proposed architecture contributes something processional, the proposed architecture is compared with a range of human-centric segmentation models, giving quantitative results on the CIHP benchmark and when available relevant information regarding the execution time during inference. The work is concentrated in terms of the trade-off between segmentation performance measured by mIoU, inference throughput measured in Frames Per Second (FPS) and model computational complexity (number of parameters), which are crucial assets for deployment of models on resource-constrained devices. This comparison is provided in Table II with respect to our simplified architecture, state-of-the-art methods, alongside edge-efficient baseline architectures.

TABLE II. QUANTITATIVE COMPARISON OF SEGMENTATION PERFORMANCE AND EFFICIENCY ON THE CIHP DATASET

Method	MIoU (%)	FPS	Params (M)	GFLOPs
SCHP [26]	59.36	-	29.4	-
CSENet [27]	66.80	-	42.1	-
WNet [28]	57.60	26.7	30.2	-
HRNetV2-W48 [29]	55.90	-	65.9	-
Edge-MHP [30]	38.10	55.6	3.1	~10
Ours	**40.67**	**51.94**	**2.14**	**5.70**

Table II shows a generic trend for the trade-off between segmentation performance and computational cost through-out all model families.

Models like HRNet, SCHP and CSENet are deep back-bones with a large receptive field to achieve higher mIoU.

However, their high computational cost (more than 150 GFLOPs) and slow inference speed hinder the popularity of their deployment on resource-constrained devices.

In contrast, although lightweight models can greatly reduce computation complexity and achieve high inference speed, generally at the cost of segmentation accuracy under difficult occlusion scenarios and scenes consisting of fine-grained structures.

The proposed model is a balance between these two extremes. It achieves 40.67% mIoU and has only a model size of 2.14 M parameters with a computational cost of 5.70 GFLOPs, providing real-time results (more than 50 FPS). The proposed model achieves over the lightweight baselines (Edge-devices MHP) a consistent improvement in accuracy performance while maintaining comparable efficiency.

It neither increases intermediate feature dimension nor aggregates multi-scale contexts as well, by improving their representation of features through context aggregation and per-feature refinement. These results suggest that architectural design can mitigate reduced capacity and improve performance subject to a strict computational bottleneck.

The results indicate the proposed model advances the accuracy-efficiency trade-off in lightweight human parsing. It retains a tiny architecture and real-time inference speed with competitive segmentation performance, making it deployable on resource-constrained edge devices.

D. Ablation Study

We perform the ablation study on our full model, removing one module at a time to measure its respective contribution. The components under consideration are the DSPP module, the grouping decoder, the spatial attention module and the SE block.

The segmentation performance and computational cost for each variant are reported in Table III.

TABLE III. ABLATION STUDY ON CIHP VALIDATION SET

Variant	mIoU	Params (M)	GFLOPs
Full model	**0.4067**	2.14	5.70
w/o grouping decoder	0.4036	2.53	12.86
w/o SE block	0.4005	2.13	5.70
w/o spatial attention	0.3990	2.14	5.70
w/o DSPP	0.3885	0.51	3.17

The results indicate that the effect of each component is not homogeneous and varies with respect to the kind of information module it abstracts.

The most important influencing module for performance is the DSPP. We can see that its removal leads to a drop of 0.0151 in mIoU, therefore verifying that capturing global structure and the ability to handle crowd

occlusion are both necessary and multi-scale context aggregation is critical for this.

A moderate and consistent improvement is attained using both the SE block and spatial attention modules. For the third experiment, discarding the SE block leads to 0.0031 mIoU and for the fourth, removing spatial attention descends by 0.0046. The FRC and BCP modules used for feature refinement improve the representation of channel-wise, and reduce irrelevant activations to background.

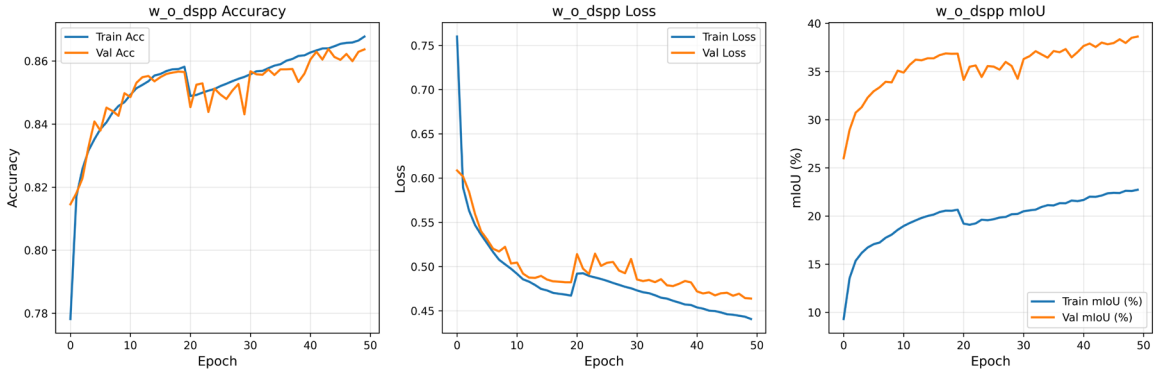
On CIHP, removing the grouping decoder does not decrease mIoU. Performance is not improved significantly either (from 5.70 GFLOPs to 12.86 GFLOPs), with the exception of a higher computation cost. This indicates that, in this evaluation setting, the grouping mechanism has a marginal impact on global metrics like mIoU.

The grouping decoder only acts locally—but it provides spatial consistency along object boundaries and enhances the smoothness of neighboring areas. These impacts are not always well represented in global metrics, but can be seen in qualitative outcomes, particularly when they engage in crowded or overlapping conditions.

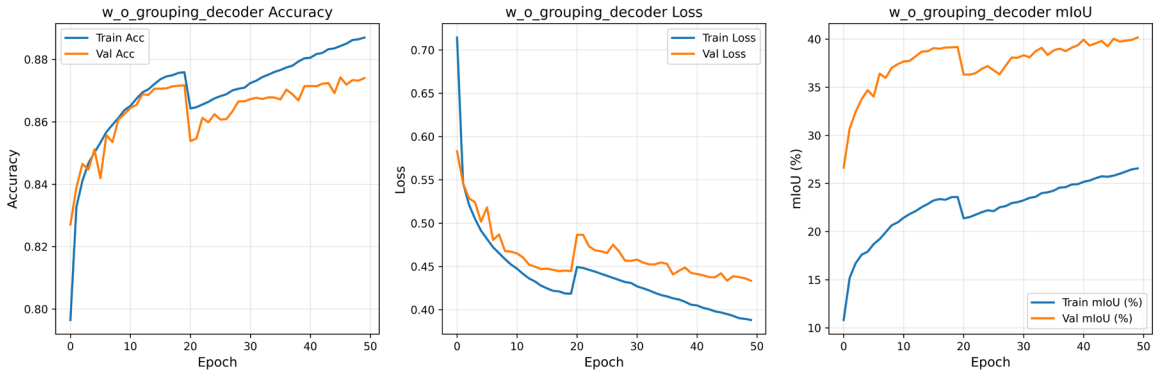
On the computational side, it leads to a more light-weight design of decoder due to already shown similar in accuracy but significantly reduced complexity with and without grouping.

In summary, these results suggest that the majority contribution to performance is from DSPP and attention and SE modules provide complementary refinements. Grouping largely adds to structural consistency, rather than global accuracy.

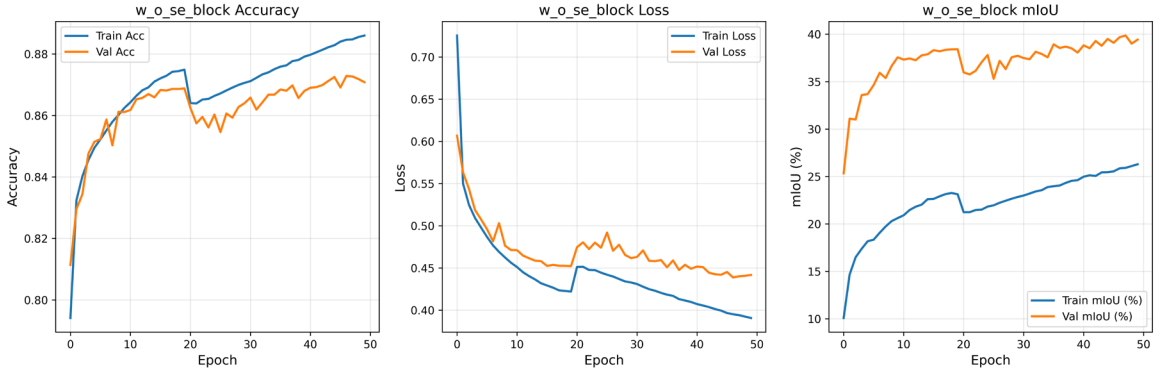
The training dynamics in Fig. 2 go beyond the static metrics, offering complementary evidence that sheds light on the role of each element. The DSPP-less variant (Fig. 2(a)) converges slower and achieves lower mIoU, which signals its inferior long-range contextual information gathering capacity. Such behavior indicates a direct dependence of both optimization efficiency and final performance on global context aggregation.



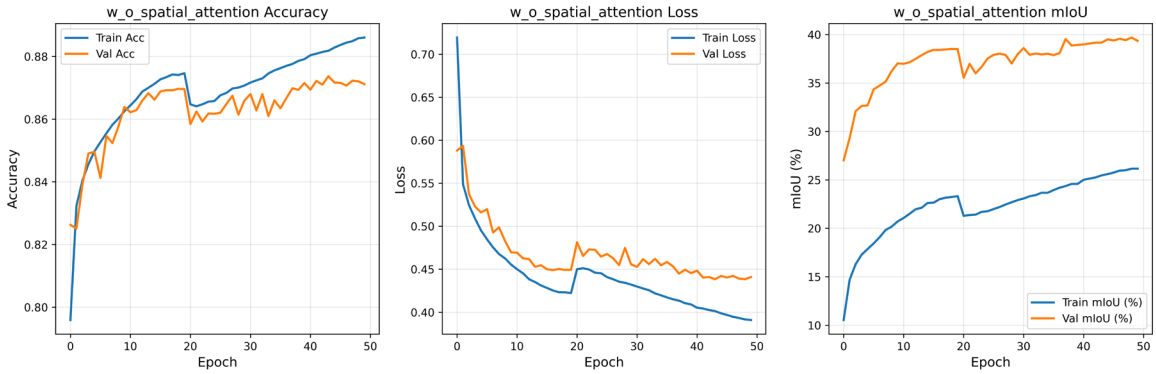
(a)



(b)



(c)



(d)

Fig. 2. Analysis on ablation variants in CIHP dataset and their training dynamics: (a) w/o DSPP, (b) w/o grouping decoder, (c) Block w/o SE, and (d) Block w/o spatial attention. In all plots, the evolution of training and validation accuracy, loss, and mIoU over epochs is reported. These five differences highlight differences in convergence speed, final performance and stability.

Spatial attention and SE modules are removed (Fig. 2(c) and (d)), causing more noise around validation curves, especially during later epochs. The presence of this instability suggests that these modules help stabilize the feature responses and improve generalization by cleaning up inter-mediate representations.

The behavior of the grouping decoder is shown in Fig. 2(b). The relative variant exhibits smoother convergence and a modest gain in training mIoU, suggesting less regularization during optimization. This indicates that the grouping mechanism provides a regularization effect, which limits learning to promote spatial coherence.

Here we see the difference between optimization behavior and final evaluation metrics. Our grouping decoder does not enhance global performance—mIoU, however, structural consistency at the local level. Scalar metrics do not capture such effects, and qualitative segmentation results reflect them too.

The training curves provide conclusive evidence that overall DSPP is mostly responsible for convergence and global representation, while attention and SE modules reinforce the consistency of results as well as refinement of features, and the grouping decoder imposes local structural constraints. This combination of components allows to achieve a balanced trade-off between accuracy, stability and efficiency.

E. Cross-Dataset Evaluation on LIP

We further perform experiments on the Look into Person (LIP) dataset to validate the cross-domain generalization performance of the proposed architecture, since LIP is more diverse than CIHP in terms of pose, scale and background complexity.

This is evaluated on 5000 held-out images under the same inference protocol. We summarize the results in Table IV.

The evaluation from CIHP to LIP establishes a solid understanding of the generalization behavior of our architecture over two datasets. As we discussed, performance variation is expected because of the distribution of data and complexity of scenes as seen on GeoQueries dataset, thus, degenerate annotations across the two datasets.

TABLE IV. CROSS-DATASET EVALUATION ON LIP

Model	mIoU	Params (M)	FPS
w/o SE block	**0.2640**	2.13	67.08
w/o grouping decoder	0.2574	2.53	66.59
w/o spatial attention	0.2559	2.14	65.03
Full model	0.2490	2.14	66.69
w/o DSPP	0.2391	0.51	67.39

It is a source of interest that this model in a sense does exhibit adaptability and its components have unique functions. Our full model still obtains competitive performance on LIP, indicating the generality of our design beyond the source domain. Furthermore, the slight performance improvements of variants without the SE block or grouping decoder indicate that the architecture is malleable and can be optimized based on data distributions.

Among the findings, one key result is that the overall contribution of DSPP module is stable over datasets. We show that removing it incurs the highest performance drop, which indicates that multi-scale context modeling is inherently domain-robust and other approaches are mostly used to capture global structure.

On the contrary, this group decoder and SE block showed a more specific behavior. These contents encode structural and statistical patterns learned from the source dataset, which can differ across domains. That sensitivity should not be seen as a shortcoming, but rather evidence that these modules are learning high-resolution dataset-specific feature information and must be refined to facilitate cross-dataset transfer.

These results offer a key insight into lightweight model design: global context modeling appears to have good dataset generalization, whereas refinement and structural consistency mechanisms can be domain-specific items that enhance performance. This indicates a productive avenue for future work, with an attention to domain-aware training and adaptive module assimilation.

In the end, the resulting architecture presents good and competitive performance under domain shift, with a deeper understanding of how each component contributes to what generalization is. This emphasizes that the design is very effective, but can be much more rewarding for cross-dataset environments.

F. Training Behavior

Next, let us examine how well the model generalizes when trained It is always a good idea to mention the normative behavior expected right before we plot the curves, which should be a stable gain in pixel-wise classification accuracy on both the training and validation splits with a consistent drop-off in loss until convergence, with no significant overfitting. In Fig. 3, we see the training dynamics depicted whereby the accuracy and loss improve for each split through its respective epochs.

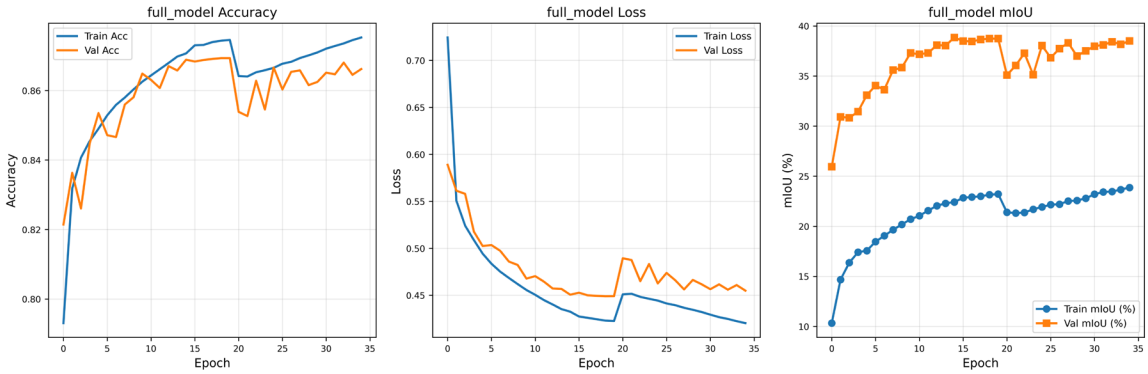


Fig. 3. CIHP training dynamics. Pixel accuracy (left), loss (middle) and mIoU (right) vs. evolution over the training epoch for both train/val datasets. These curves show stable convergence; validation trends closely follow training behavior, exhibiting both effective, well-controlled optimization and consistent generalization.

The training dynamics are depicted in Fig. 3. This gives us extra insight into how the model behaves. The training and validation accuracy increases steadily and converge around 30–35 epochs, whereas the loss is also decreasing in a smooth manner without oscillation. Good generalisation thus would entail a small gap between training and validation curves, which may be inferred in this figure proving there is no overfitting even as relative capacity of the model hardly seems optimal. Two reasons behind this behavior are: (i) the depthwise separable convolutions act as an implicit regulariser, (ii) which is simply a way to enforce some structural consistency (beyond pixel-wise supervision) via an auxiliary grouping loss. All these together mitigates the effect of optimisation causing them relatively less sensitive towards noise and local ambiguities providing a more robust convergence than conventional lightweight segmentation models.

G. Class-wise Performance Analysis and Error Assessment

Comprehensive metrics are, well, global: they will give you a summary of performance without telling you which categories are easy and which are hard. Inspects the distribution of errors across classes to better understand the strengths and weaknesses of the model. Overall, we anticipate that significant and common areas (background, torso, upper-clothes) will be segmented more consistently, while small or infrequent classes (e.g., scarf or bag) can still be difficult to aggregate, as shown in past CIHP works [25].

Fig. 4 shows the class-normalized confusion matrix over the CIHP categories and Fig. 5 lists the class-wise Intersection over Union (IoU) in sorted order. In tandem, these plots give you a close-up on where the model is succeeding as well as failing.

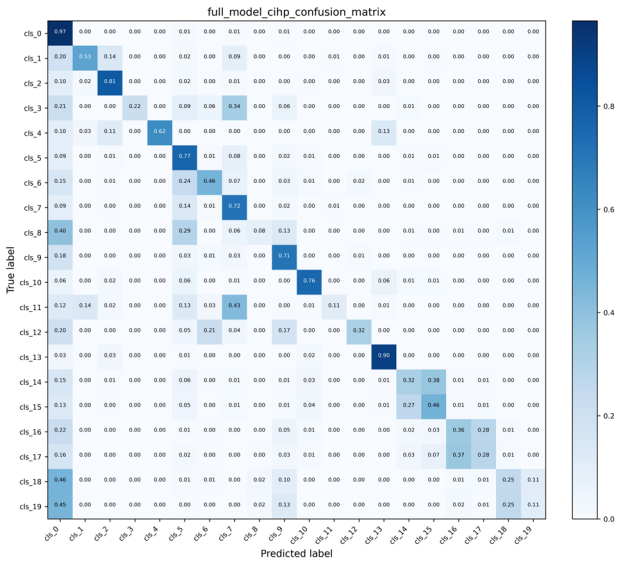


Fig. 4. Strong diagonal dominance is observed in the confusion matrix after normalization against model predictions over CIHP validation set. However, there remain residual ambiguities between visually and spatially correlated categories (especially upper clothes vs coats, shoes vs socks).

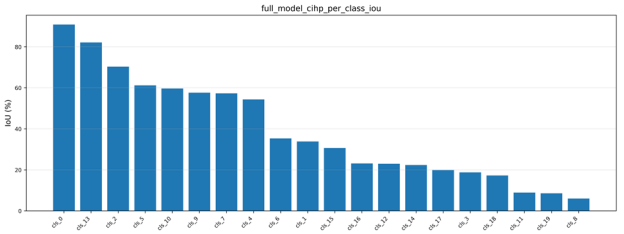


Fig. 5. As per-class IoU values on CIHP are ordered in descending order, we notice that large and more common regions (e.g. background, torso, upper-clothes) yield higher scores, and small or underrepresented categories manually pose bigger challenges for segmentation, such as scarf and bag.

The normalized confusion matrix in Fig. 4 is notably aligned with its diagonal, providing evidence captured by the model during training consistent semantics instead of global appearance-based features such as color and texture cues. The other confusion is mainly across semantically distinct correlated categories or classes exhibiting similar visual patterns, which are misclassifications among semantically similar categories, such as coats and upper-body garments, left and right limbs, and shoes versus socks. This pattern of behavior is consistent with what have been qualitatively reported for stronger but bulkier models [21], and it stems from the inherent difficulty of densely labelling crowded scenes.

We further visualize per-class IoU distribution to emphasize the advantages and disadvantages of the model. From this, we can see that high IoU scores are reached for larger and more frequent regions such as background, torso, and upper-clothes knowledge which suggests the model has learned to capture dominant structures.

The small or rare categories (scarf, bag, accessories) are still difficult-achieving IoU values are often below 25%. This is not only a result of class imbalance, but also the low spatial resolution and fewer number of features, which lightweight architectures are limited to.

They indicate that further gains are unlikely to result from simply increasing model size, but will likely come from more specific techniques (i.e., applying class-balanced losses, feature reweighting or high-resolution feature fusion). That is to say, the residual performance gap isn't something that architecture alone can fix; it's structural and data-driven.

H. Qualitative Evaluation and Contribution of DSPP, PGN, and Attention Modules

The quantitative results highlight the effectiveness improvement with respect to each proposed module, but do not elucidate how these two modules change the internal behavior of the network. To tackle this aspect, we perform a qualitative evaluation through visual inspection relying on three complementary analytical components:

- spatial attention heatmap representations, a way to evaluate whether the spatial features attention module attends to human parts;
- DSPP component multi-scale feature activations for a straw-man examination of the impact of varying dilation rates on contextual representation.
- qualitative comparisons to a lightweight baseline side-by-side, which we instrumentalize as the upper bound for final predictions.

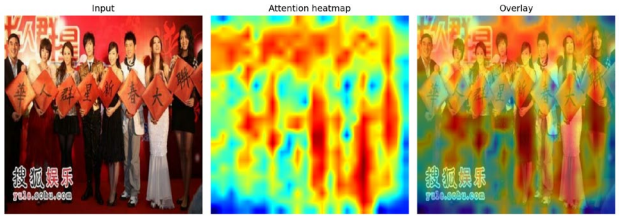


Fig. 6. Spatial attention visualization emphasizing human contours and fine-grained boundaries.

The impact of the spatial attention block is illustrated in Fig. 6. This image shows the input image, the attention map learned and regions overlap between them.

Then, the attention map as shown in Fig. 6 focuses energy around crowded human regions and fine-grained part boundaries, while attenuating activations in static back-ground structures, such as walls and floors, which are as-signed a very low weight. The behavior also accounts for the mIoU gain we observe by adding spatial attention mechanism built upon the DSPP representation: The net uses more capacity for ambiguous human parts and less for uninformative background, which follows the intuition behind attention mechanisms [5, 16].

We now turn to the DSPP module. Feature maps averaged over channels for various dilation rates (Fig. 7), all computed on the same scene. This also lets us visualize how the different receptive fields capture complementary structures.

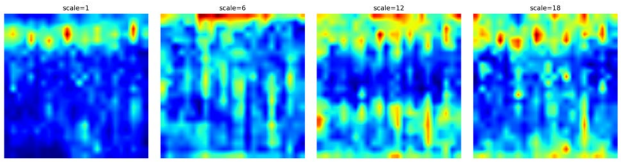


Fig. 7. Multi-scale feature responses under varying DSPP dilation rates (1, 6, 12, 18).

Conv. filters with small receptive fields (rate 1) that emphasized fine-grained patterns like fingers, hair and edges. With the increasing dilation rate (6, 12, 18) activations steadily move inwards towards larger areas of the upper-body structural patterns, limb configurations and spatial organization of the crowd. Concatenating these multi-scale representations are dynamically re-weighted to emphasize salient contextual information by attention, the decoder gets fine-grained boundary details information as well as low-frequency contextual representations. The multi-scale feature integration comes in handy, particularly in crowded environments, where local spatial information is usually ambiguous.

Finally, we evaluate how the proposed modules used during training affect the final segmentation masks. Qualitative comparisons between the input image, a Fast-DSPP+PGN+ATTN with Ground Truth (GT) and MobileNetV2+U-Net base-line (w/o DSPP, PGN, or ATTN) are shown in Fig. 8. For each case, we present the input visual sample, prediction from the baseline model, proposed method output and pixel-wise ground-truth segmentation map.

Fig. 8. Qualitative evaluation of segmentation results on CIHP.

Across these representative cases, the reference base-line approach is often joining spatially contiguous persons close together, oversmoothing part borders and mislabelling accessories or thin structures (i.e. arms, legs). The proposed model on the other hand, yields crispier silhouettes, reduced crosstalk under neighbouring distinct individuals and improved consistency in limb labeling and head. The Dilated Spatial Pyramid Pooling (DSPP) module protects fine-grained components from capturing background clutter, the PGN-inspired grouping module enforces spatial consistency at the local level across boundaries, and the SE-refined skip connection filter out low-level noise. To summarize these points, we conclude that the proposed architecture fundamentally reshapes the behavior of the model rather than just being a purely superficial change which accounts for the quantitatively improved results seen in Tables I and II.

In general, the visual analysis provides evidence that the model is not simply manipulating numerical metrics but rather modifying how the model processes spatial information. DSPP, grouping and attention work together to pro-duce more structured and meaningful predictions, often in regions where light-weight models fail.

I. Limitations and Discussion

Experimental results show a definitive tradeoff between efficiency, accuracy and generalization. The proposed model performs competitively in the presence of strict constraints but not evenly across datasets.

CIHP: By incorporating multi-scale context, attention mechanism, and grouping of scalars leads to improve convergence consistency, as well as final performance. By comparison, cross-dataset evaluation on LIP reveals domain shift sensitivity. Specifically, learned structural relationship derived components, such as grouping mechanisms, perform poorly in scenarios where testing domains are out-of-training domain data distributions. This suggests that while lightweight architectures need to find a trade-off between accuracy and efficiency, they also need to be robust against distribution shift.

While the proposed model offers a positive efficiency-accuracy trade-off, it also has multiple drawbacks. For starters, performance on rare or small categories is still very weak. This is primarily the result of class imbalance and loss of spatial resolution in fine structures representation. Dealing with the problem might rest upon datacentric solutions like class re-weighting, focal loss, or dedicated refinement modules.

Second, the model takes just one image as input and does not take advantage of temporal information. It also limits the amount of consistency one can impose at frames in dynamic scenes. Temporal cues can add robustness and stability to video-based applications as they allow them to carry context information, especially in surveillance and behavior analysis tasks.

Finally, while the model is ready for frontier deployment, additional optimizations are attainable. Using methods like structured pruning, quantization or knowledge distillation could help to decrease memory and latency whilst retaining most of the segmentation accuracy.

These limitations imply that future gains may not come primarily from architectural changes alone, but require consideration of factors like data distribution, temporal modeling, and deployment-specific constraints.

V. Conclusion

This paper introduces a new lightweight human parsing framework that is developed for real-time computation on edge-oriented devices. The proposed approach with a MobileNetV2 backbone, dilated contextual aggregation and boundary-aware pixel grouping, and lightweight attention mechanisms strikes a good balance between the pixel-wise segmentation accuracy and computation tested under the CIHP evaluation protocol benchmark. With a mere 2.14 M parameters, 5.70 GFLOPs, and an average inference speed of 51.9 FPS, experimental results show that accurate human parsing can be performed without the need for large-scale backbone networks or computationally expensive pipelines.

Moreover, the results suggest that organizing pixel-level context information in a structured way offers an effective manner to reason occlusions, crowded interactions and fine-grained boundary regions. These functionalities render the proposed pipeline suitable for embedded sensing applications such as real-time surveillance, sports analysis, and mobile augmented-reality systems.

In future work, we plan to extend the framework with lightweight temporal reasoning mechanisms, allowing inference from short-range motion cues and compression techniques like pruning and factorization of models. In addition, extensions for domain-adaptation to retrieval in case of varying imaging conditions and heterogeneous edge hardware platforms will be explored.

Data Availability Statement

The dataset used in this study is the Crowd Instance-level Human Parsing (CIHP) dataset, which contains over 38,000 images annotated with pixel-level labels across 20 semantic categories. The dataset was accessed and used in accordance with its license and terms of use. Requests regarding the data usage may be addressed to the corresponding author, Abderrahim Ouza (abderrahim.ouza@uit.ac.ma).

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

A.O.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing—original draft, Visualization. M.E.G.: Methodology, Validation, Writing—review & editing, Project administration. A.C.: Resources, Writing—review & editing, Project administration; all authors have read and approved the final version of the manuscript.

ACKNOWLEDGMENT

The work was conducted using institutional resources available at Ibn Tofail University and ENSAM, Mohammed V University in Rabat. The authors would like to express their grateful thanks towards Ibn Tofail University, and the École Nationale Supérieure d'Arts et Métiers (ENSAM), Mohammed V University in Rabat for providing an institutional support and for giving access to lab resources. The authors are grateful to the institutions where this work was performed for providing facilities and resources.

REFERENCES

[1] Q. Liu, Y. Dong, and X. Li, "Multi-stage context refinement network for semantic segmentation," *Neurocomputing*, vol. 535, pp. 53–63, 2023.

[2] T. Li *et al.*, "Enhanced multi-scale networks for semantic segmentation," *Complex & Intelligent Systems*, vol. 10, no. 2, pp. 2557–2568, 2024.

[3] L. Hu, X. Zhou, J. Ruan *et al.*, "Aspp+-lanet: A multi-scale context extraction network for semantic segmentation of high-resolution remote sensing images," *Remote Sensing*, vol. 16, no. 6, 1036, 2024.

[4] S. Shahed, S. M. Arman, S. H. Sami *et al.*, "A hybrid approach for facial parsing using transfer learning," *Scientific Reports*, vol. 16, 3405, 2026.

[5] L. Yang, Z. Liu, T. Zhou *et al.*, "Part decomposition and refinement network for human parsing," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 6, pp. 1111–1114, 2022.

[6] M. Lou, J. Meng, Y. Qi *et al.*, "Mcrnet: Multi-level context refinement network for semantic segmentation in breast ultrasound imaging," *Neurocomputing*, vol. 470, pp. 154–169, 2022.

[7] Q. Yi, G. Dai, M. Shi, Z. Huang, and A. Luo, "Elanet: Effective lightweight attention-guided network for real-time semantic segmentation," *Neural Processing Letters*, vol. 55, no. 5, pp. 6425–6442, 2023.

[8] Y. Zhou and P. Y. Mok, "Enhancing human parsing with region-level learning," *IET Computer Vision*, vol. 18, no. 1, pp. 60–71, 2024.

[9] X. Xiao, Y. Zhao, F. Zhang *et al.*, "BASeg: Boundary aware semantic segmentation for autonomous driving," *Neural Networks*, vol. 157, pp. 460–470, 2023.

[10] G. Xu, J. Li, G. Gao *et al.*, "Lightweight real-time semantic segmentation network with efficient transformer and cnn," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 15897–15906, 2023.

[11] K. Lu, J. Cheng, H. Li, and T. Ouyang, "MFAFNet: A lightweight and efficient network with multilevel feature adaptive fusion for real-time semantic segmentation," *Sensors*, vol. 23, no. 14, p. 6382, 2023.

[12] M. I. Hosen and T. Aydin, "Edge devices friendly multi-human parsing with lightweight encoding and multi-scale self-attention based decoding," *Multimedia Tools and Applications*, vol. 84, no. 22, pp. 25027–25047, 2024.

[13] X. Zhang, Y. Chen, M. Tang *et al.*, "Human parsing with part-aware relation modeling," *IEEE Transactions on Multimedia*, vol. 25, pp. 2601–2612, 2022.

[14] H. Zhao, J. Shi, X. Qi *et al.*, "Pyramid scene parsing network," in *Proc. IEEE Conference on Computer Version and Pattern Recognition*, 2017, pp. 2881–2890.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015, pp. 234–241.

[16] R. Wang, S. Chen, C. Ji *et al.*, "Boundary-aware context neural network for medical image segmentation," *Medical Image Analysis*, vol. 78, 102395, 2022.

[17] A. G. Howard, M. Zhu, B. Chen *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint, arXiv:1704.04861, 2017.

[18] M. Sandler, A. Howard, M. Zhu *et al.*, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[19] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.

[20] K. Gong, X. Liang, Y. Li *et al.*, "Instance-level human parsing via part grouping network," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 770–785.

[21] M. Knolle, G. Kaissis, F. Jungmann *et al.*, "Efficient, high-performance semantic segmentation using multi-scale feature extraction," *PLOS One*，vol. 16, no. 8, e0255397, 2021.

[22] P. Wang, L. Li, F. Pan, and L. Wang, "Lightweight bilateral network for real-time semantic segmentation," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 27, no. 4, pp. 673–682, 2023.

[23] K. I. Rashid, A. Hussain, C. Yang, and C. Huang, "Hybrid semantic segmentation with broad context and attention encoded network for urban street scenario," *Engineering Applications of Artificial Intelligence*, vol. 165, 113528, 2026.

[24] J. C. H. Ong, S. L. H. Lau, M. Z. P. Ismadi *et al.*, "Feature pyramid network with self-guided attention refinement module for crack segmentation," *Structural Health Monitoring*, vol. 22, no. 1, pp. 672–688, 2023.

[25] Z. Su, M. Chen, E. Haung *et al.*, "MVSN: A multi-view stack network for human parsing," *Neurocomputing*, vol. 465, pp. 437–450, 2021.

[26] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3260–3271, 2020.

[27] X. Zhang *et al.*, "Part-aware context network for human parsing," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8971–8980.

[28] F. Xia, P. Wang, X. Chen *et al.*, "Joint multi-person pose estimation and semantic part segmentation," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6769–6778.

[29] J. Wang, K. Sun, T. Cheng *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[30] K. Gong *et al.*, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 932–940.