Lightweight Human Parsing with Multi-Scale Context for Edge Devices

Abderrahim Ouza[1], Mohamed El Ghmary[2], Ali Choukri[3]
Faculty of Science, Ibn Tofail University, Kenitra, Morocco[1,3]
ENSAM, Mohammed V University in Rabat, Morocco[2]

***Abstract*—For human parsing in wild and cluttered environment, deep architectures are widely utilized since they yield strong segmentation performance, while with the price of huge model size and computational complexity. These properties are highly limiting for them if to be deployed on resource-limited platforms, in particular for edge intelligence in real-time. In this study, we propose a lightweight framework for human parsing, named *Fast DSPP+PGN+Attn*, which focuses on the efficiency-accuracy trade-off. The proposed model consists of a MobileNetV2 backbone (i.e., the AirLab-Net), a Dilated Spatial Pyramid Pooling (DSPP) block to capture multi-scale contextual information, a pixel grouping decoder employing the PGN for improved part boundary flow consistency and potency spatial and squeeze-and-excitation attention modules are used for feature refinement. However, the model with a relatively compact size—i.e., 2.14M parameters and 5.70 GFLOPs— could achieve the performance of 40.67% mean IoU (mIoU) and 87.3% pixel accuracy on the CIHP benchmark, while running at a speed of 51.9 frames per second on a single GPU. These findings indicate that combining the contextual aggregation methods with the method of structured pixel grouping is a strategy to exploit orthogonal avenues and cross-examine their complementarity, more efficiently and potentially achieve better segmentation quality without lost of real-time performance. Therefore, the proposed method can be widely applicable in embedded vision system, surveillance and mobile perception.**

***Keywords*—Human parsing; lightweight networks; multi-scale representation; edge computing; real-time segmentation**

I. Introduction

Semantic segmentation is an essential task in computer vision. It assigns a semantic label every pixel of an image to aid in fine-grained scene understanding [1], [2]. In this work, lightweight refers to models with single-digit GFLOPs, fewer than 5M parameters, and real-time inference on commodity GPUs or edge hardware. For human-centric applications, such as surveillance, behavior analysis, crowd monitoring, as well as AR/VR applications, this job–also known as human parsing - is crucial to finding out the precise human body components and categories of clothing in complicated and chaotic situations. The deep convolutional networks (CNNs) and encoder/decoder structures like U-Net, DeepLab [3], [4] have seen great ensures in segmentation ability by combing global semantic context and superior spatial resolution.

Aligned with these developments, current human parsing algorithms mostly adopt heavy backbones (e.g ResNet 101, HRNet, Swin Transformer) which contains tens of million parameters and also large computation footprints. The mIoU of these models on CIHP, as one of the most difficult datasets, is generally in 40-60% and uses 100-300 GFLOPs per image with a high latency [5] [6]. Despite this, the complexity limits their applicability to real-time implementations such as edge devices and mobile hardware where power and memory constraints are critical [7].

Conversely, lightweight models tend to be ineffective when faced with large occlusions, people who overlap as well as complex body-part boundaries [8]. Receptive fields that are not as large and capabilities of the feature hinder their ability to understand multi-scale contexts and maintain sharp boundaries and lead to distorted or incorrect predictions [9]. Therefore, the main challenge is to create a compact but accurate human parsing system capable of working effectively in real-world scenarios and on hardware with limited resources [10].

In this study, we present Fast-DSPP+PGN+Attn, a new architecture that incorporates three important components - DSPP (Dilated Spatial Pyramid Pooling), PGN (Part Grouping Network) inspired pixel grouping and dual attention mechanisms inside the light-weight MobileNetV2 backbone. by such operations is effective in capturing multi-scale contexts, which helps to improve the spatial coherence in occluded situations at one hand and focus more on the human regions by suppressing non-human patterns on the other hand with a low computational cost. This particular combination is highly performant in comparison to other state of the art models based on large, computationally expensive models. The presented system provides a good compromise between efficiency and accuracy, with state-of-the-art performance when considering the compact model size, as well as its real-time inference requirement [11].

Using a simulation of the Crowd Instance-level Human Parsing (CIHP) dataset, Fast DSPP+PGN+Attn reaches 40.67% mIoU and 87.3% accuracy, with a real-time speed of 51.9 FPS, despite just 2.14M parameters, and 5.70 GFLOPs. This balance of accuracy and efficiency is a strong proof that the combination of contextual pixel grouping together with DSPP along with dual-attention greatly improves segmentation accuracy in busy environments and is accessible on devices with edge capabilities.

The major contributions of this study are listed below:

- We present a unique ultra-lightweight human parsing model which integrates dilated multiscale context encoder (DSPP) as well as spatial grouping based on PGN, and dual attention in a single MobileNetV2-based encoder-decoder.
- We present a boundary-aware encoder that analyzes pixel affinities in the influence of occlusion, thereby

improving structural consistency and the delineation of the human body's parts that overlap, an aspect with which many light models have trouble.

- We show an exceptionally favorable balance between accuracy and efficiency and achieve 40.67 percent mIoU, using just 2.14M parameter values and 5.70 GFLOPs. We outperform comparable baselines that are lightweight while achieving speed in real time (50 FPS).
- We offer an extensive range of qualitative, quantitative and component-level anal- ysis, such as the training patterns, dynamics, per-class IoUs attention visualisation, multi-scale feature response, to demonstrate how each component can impact the overall performance.

Despite the progress of multi-scale context aggregation and lightweight backbones, a critical gap remains between efficiency and structural accuracy in human parsing. Existing multi-scale architectures, such as DeepLab variants, rely on heavy backbones and large receptive fields, leading to high computational cost unsuitable for edge deployment. Conversely, lightweight models based on MobileNet or similar architectures achieve real-time performance but often fail to preserve fine-grained boundaries and local consistency, especially under occlusion and crowding conditions.

Moreover, prior works typically address either global context aggregation or local boundary refinement in isolation, without explicitly modeling their complementarity. As a result, lightweight models often suffer from fragmented predictions, while heavy models remain impractical for real-time applications.

Wisely to the backbone, MobileNetV2 compact architecture was redesigned, while multi-scale context aggregation (DSPP), boundary-awareness pixel grouping (PGN inspired) and lightweight attention mechanisms (CBAM) are introduced accordingly. This design directly addresses the trade-off between accuracy and efficiency with the goal of providing strong human parsing performance under tight computational constraints.

The remainder section of the study is structured in the following manner: Section II reviews the similar research. Section III outlines the proposed method. Section IV presents the research results, as well as a discussion. Section V closes out the research and discusses the future research directions.

II. Related works

In the constant development of deep-learning methods, semantic segmentation has made remarkable strides in the last decade [10]. The first segmentation methods relied heavily upon hand-crafted elements and conventional machine learning classifiers to provide the interpretation of pixels [12]. Although these techniques were revolutionary in their day but they were severely limited when it came to dealing with the complexities and ambiguities of real-world scenes [13]. One of the major limitations of these methods was their limited ability to understand dependencies among the various regions of an image and their limitations in situations [14] that demanded a more thorough comprehension of content, as well as the spatial context [5].

Fully Convolutional Networks (FCNs) were a significant turning point in the development of the field known as semantic segmentation [13]. Unlike classifier-based patchwise CNNs that considered subregions of matching dimensions separately, FCNs allowed for end-to-end learning from initial pixels. [13]. This change of architecture allowed models to learn more complex representations and to solve dense prediction problems, resulting in state-of-the-art performance on many benchmarks [2] [3]. However, FCNs can't perform well in multiscale reasoning and often end up with blurred object boundaries due to their narrow receptive field and the slow downsampling rate. Bottleneck of receptive field reciprocity To solve the bottleneck, dilated (or atrous) convolutions were proposed [15] to increase the receptive-field size and generalization capability by using atrous con. By introducing dilation gaps into Convolution kernels, the models can increase the size of the receptive field, while reducing the parameters [5]. This allowed networks to collect fine-grained detail as well as long-range context at the same time which is crucial in complicated scenes [7]. The DeepLab family of architectures as well as the related ones paved the way for the use of convolutions with dilated coefficients in conjunction with pyramid pooling, which allows an efficient multi-scale feature aggregation [8].

Encoder-decoder technology, such as U-Net, also brought further advances through explicit coupling upsampling and downsampling pathways [1] [7]. These skip connections permit the first high-resolution features to be combined with deeper semantics, enhancing the ability to locate boundaries [6]. These structures quickly became the basis for a variety of segmentation tasks including medical imaging as well as Human Part Segmentation [16]. Later variations included depthwise separable convolutions [17] - inspired by lighter backbones, such as MobileNet and ShuffleNet [18] [19] – to reduce the computational cost while still delivering high performance.

Human parsing is more challenging than semantic segmentation. Typically, the scenes consist of multiple interacting people with different poses including strong occlusions and similar/overlapping looking legs. Although this CIHP dataset become a benchmark for parsing crowded human, it has some limitations. Other more complex models based on the PG Network (PGN) [7] were proposed to analyze pixel affinities and instances-level relations for better handling the overlapping or occluded body part [20]- [21]. Recent works have extended the application ranges of utilizing contextual encoding and Cross-scale feature fusion but heavily lean on large backbones and complicated decoders which lead to overall computational burden [22] [23].

However, despite their accuracy they are not practical for real-time or situations with limited resources. For edge deployment, light designs built upon MobileNet or other designs with depth are the best choice [23]. These designs aim to strike a balance between speed, accuracy and memory efficiency [24]. This is an essential necessity for applications like autonomous systems, mobile devices and surveillance where power and latency restrictions are essential.

Our work is located in this area. In adopting an MobileNetV2 encoder, further enhancing it by incorporating the DSPP context-based module, as well as then integrating the PGN-inspired grouping coder, with focus on light weight, we demonstrate that a small model can maintain useful accuracy

for humans parsing CIHP while retaining high-quality performance in real time that can be used on devices with edge capabilities.

In conclusion, human parsing methods can be divided into three main categories with their own advantages and disadvantages:

First, high accuracy of segmentation is obtained in heavy multi-scale architectures like DeepLab-style networks [5], [8], [15] and more recent context- aware networks [21],[23] which benefit from dilated convolutions and pyramid pooling. These methods effectively capture contextual information from the entire image and long-distance dependencies, both of which are important in crowded scenes. But they depend on large backbones and complex decoders, which leads to high computation cost (100 GFLOPs or more), as well as an explosive number of parameters, which is not friendly for real-time deployment on edge devices.

Second, simple encoder-decoder networks with MobileNet or other very efficient backbone [13], [22] reduce computation by utilizing depthwise separable convolutions. These methods are very suitable for real-time scenarios thanks to high inference speeds and lower memory consumption. Such networks, however, have limited receptive fields and simplified spatial feature responses leading to poor ability for fine object boundary delineation and multi-scale reasoning (especially in the case of occlusion or overlapping persons).

On the other hand, several affinity-based and grouping-guided approaches such as PGN-like methods [7], [20] explicitly capture pixel-wise relationships to enhance structural accountancy and instance separation. Such approaches are useful for preserving segment boundaries and dealing with occlusions by learning pixel affinities. However, they bring extra computation costs and are usually merged with more complicated architectures that can prevent them from being used in light-weight setups.

While they all have their own benefits, none of these categories meets the joint needs for multi-scale contextual information modeling, boundary-aware refinement, and strict computational efficiency. Specifically, prior works tend to treat these properties separately, ignoring their potential complementarity within a unified lightweight design.

Taking it further, the proposed Fast-DSPP+PGN+Attn model aims at achieving this balance via multi-scale context aggregation (DSPP), pixel-level grouping (inspired by PGN), and lightweight attention inside a MobileNetV2 backbone. This integration approach allows the model to capture global context, enforce local consistency and minimize computational complexity, simultaneously, in contrast with previous work, which makes it suitable for real-time edge deployment.

III. Methods

The Fast-DSPP+PGN+Attn algorithm is based on an encoder-decoder model improved by dilated multi-scale context and explicit pixel grouping. It uses MobileNetV2 as the backbone for its encoder which is truncated at stride 16 and trained on ImageNet. Decoder uses 3x3 convolutions with the channel sizes 64 for shallow, and deeper 128. These kernel sizes have been selected based on a balance between complexity and ability to detect fine details at boundery interfaces. The fusion operator is simply concatenating the encoder feature maps and output of dilated spatial pyramid pooling (additionally processed with 1x1 convolutions to reduce channel count). Through pixel grouping based on pattern generation (PGN), we enforce local consistent constraints along edges, and can handle occlusion properly as well as clutter background.

In order to better understand the flow of data, as previously mentioned, the proposed architecture is a linear pipeline Starting with an input image that has dimensions of 384×384×3, the MobileNetV2 encoder generates hierarchical feature maps at different levels. Next, we lead a high-level feature map of spatial resolution H/16 × W/16 to the DSPP module, which implements parallel dilated convolutions with different dilation rates (1, 6, 12, and 18 pixels) to capture multi-scale contextual information.

The outputs of the DSPP branches are concatenated together and reduced with 1×1 convolution for channel reduction. This enriched feature representation is then fused with low-level encoder features through skip connections. The decoder processes these fused features using depthwise separable convolutions to maintain computational efficiency.

A PGN-inspired pixel grouping mechanism is applied to enforce local consistency by modeling pixel affinities in a 4-neighborhood graph. Additionally, two attention modules are introduced: 1) a spatial attention module applied after DSPP to emphasize human regions, and 2) a squeeze-and-excitation (SE) block applied to skip connections to recalibrate channel-wise responses.

Finally, the decoder upsample the feature maps to the original resolution and produces pixel-wise class predictions through a softmax layer.

In addition to the encoder feature map, the Dilated Spatial Pyramid Pooling (DSPP) module integrates multi-scale context with parallel dilated convolutions, as well as general image pooling. The dilation ratios for DSPP block were heuristically selected via empirical experiments in previous studies and the experimental results on CIHP. Dilation rates were chosen to compromise between the fine details (low dilation rates) and distant context (high dilation rates). The last configuration was proved to achieve the most superior CVG performances in CIHP with respect to crowded and complex human parsing. The network can have both local details and global context dynamically. The decoder increases the size of these enriched features, while combining them with the early high-resolution functions by using skip connections and an PGN-inspired grouping stage. Two lightweight attention mechanisms–a spatial attention block applied to the DSPP output and a squeeze-and-excitation (SE) block applied to the skip connection–guide the network towards human regions and suppress background clutter. All decoder convolutions can be constructed as separable depthwise convolutions in order to minimize the computational cost. Before describing each element, we present an overall diagram of the overall architecture to show the way in which the encoder, DSPP module, decoder and attention mechanisms interplay. Fig. 1 illustrates the entire Fast-DSPP+PGN+Attn workflow, which runs beginning with the input image and moving to the

backbone of MobileNetV2, and DSPP context aggregation, and finally the PGN-inspired decoder, which outputs twenty-class logits of human processing, which are at the initial resolution.

The proposed pixel grouping mechanism, PGN, is to learn the pixel affinities from affinity graph built upon the spatial and semantic similarity of neighboring pixels. The neighborhood is described by the 4-connected region around each pixel and pixel similarities are weighted with a Gaussian kernel function, which gives higher weight to pixels that appear in both spatial space and feature space. This graph is exploited to impose local coherence along object boundaries, especially in occluded regions.

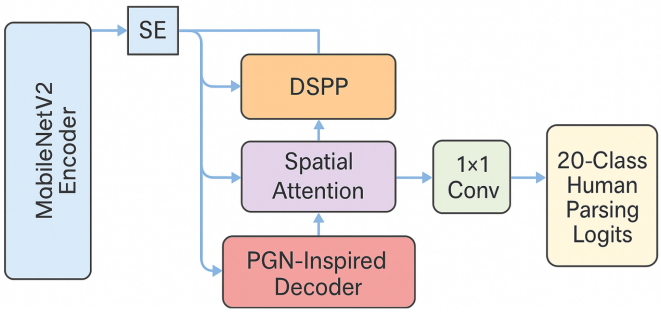

Fig. 1. Overview of the proposed Fast-DSPP+PGN+Attn architecture. A MobileNetV2 encoder feeds a DSPP context module. A PGN-inspired decoder with SE-enhanced skip connections and spatial attention produces 20-class human parsing logits at input resolution.

The implementation takes place by TensorFlow 2.13 making use of Keras functional APIs. The entire process from data loading to assessment is intended to be completely reproducible. For ease of understanding, we summarize the high-level training and inference process before diving to the module-level details. The key steps are: splits of the dataset, construction of models (encoder, DSPP, decoder and the attention module) optimization using the composite loss method, and single-pass inference of new images. Algorithm 1 provides a step-by-step explanation of this process, providing details on the process of how you can understand how CIHP dataset is divided and how the Fast-DSPP+PGN+Attn model is constructed and optimised along with how the model is used to create human-like parsing predictions during the time of testing.

The MobileNetV2 encoder serves to extract features in our system, and has a width multiplier of 0.75 and pre-trained ImageNet weights. We make sure that the network remains light by truncating it to an output length. Based on this, we get two important feature maps that are low-level maps from a layer that is early, that preserves the fine spatial details, as well as a high-level map from a more advanced layer that contains an abundance of semantic data. In reality, both maps are extracted straight from MobileNetV2 model, using the relevant layers in Keras. The encoder process the image input ($384 \times 384 \times 3$) by a series of convolutions that are standard with inverted blocks gradually reducing the spatial resolution

Algorithm 1 Training and Inference Pipeline for Fast-DSPP+PGN+Attn

Require: CIHP image paths $\mathcal{I}$, mask paths $\mathcal{Y}$, image size 384×384, batch size $B = 8$, epochs $E = 50$
Split $(\mathcal{I}, \mathcal{Y})$ into training (28 000 images) and validation (8 000 images)
Build **MobileNetV2 encoder** (stride 16, $\alpha = 0.75$), **DSPP** module, **PGN-like decoder** with SE and spatial attention
Initialize Adam optimizer with cosine decay restarts; define loss and mIoU metric with ignore label 255
for epoch $= 1$ to E **do**
    for each mini-batch (X, Y) in training set **do**
        Apply on-the-fly augmentation to (X, Y)
        Forward pass through encoder, DSPP, decoder $\rightarrow \hat{Y}$
        Compute loss $\mathcal{L}(Y, \hat{Y})$ and update network parameters
    end for
    Evaluate mIoU and pixel accuracy on validation set
    Update early stopping and model checkpoint based on best validation mIoU
end for
Inference: For a new image, resize to 384×384, forward pass through model, take $\arg\max$ over class logits to obtain segmentation mask.

and increasing the quality of the representation. These two feature maps constitute the inputs to the DSPP contextual module, as well as decoder. This allows the network to mix precise detail with high-level context to allow for precise human parsing.

A. Training Protocol and Implementation Details

The proposed Fast-DSPP+PGN+Attn Network from end-to-end the CIHP human parsing dataset by using a simple, efficient pipeline. All images that are input are changed in size so that smaller sides are set at 384 pixels. After that, an random 384×384 crop is taken during the training. When inference is completed it is a single centre crop with the same resolution is utilized with no multi-scale testing or horizontal flipping. This is done according to the standard CIHP testing protocols [25].

Center cropping in single-scale was used at inference, since it is consistent with CIHP's testing protocol, and in near future works multi-scaling will be analyzed crop for more off-center or densely crowded human instances as encountered in many practical settings. This would make inference more robust and the spatial coverage better.

The model is optimized with the Adam optimizer, with an initial learning rate of 5×10^{-4}. It is then reduced using a cosine annealing schedule for 50 times. The batch of 8 is utilized during training. A lightweight data augmentation pipeline was employed to boost the model robustness. This is accomplished with random horizontal flips (probability=0.5), scaling jittering by both +/-50% in the range $[0.75, 1.25]$ and to crop at 384×384 pixels scales randomly. A mild color jittering (brightness, contrast and saturation) was also applied with a probability of 0.3. Augmentation operations were applied in the order which they had been specified: scaling jittering was

always first, normal horizontal flips after that, and cropping last. It keeps the model exposed to a variety of scenarios, while preventing overfitting on specific data augmentations. The complete experiment is coded in TensorFlow 2.13 with mixed-precision training, resulting in significantly reduced memory usage and faster computations on the most recent GPUs.

The overall loss consists of a combination of the main sparse categorical cross-entropy segmentation loss and an additional Affinity loss which is analogous to PGN. The ultimate optimization goal is:

$$L = L_{\text{seg}} + \lambda L_{\text{group}}, \quad \lambda = 0.3, \tag{1}$$

for which L_{group} is applied to the predicted pixel affinities as a binary cross-entropy loss. This kind of constraint can ease enforcing those behaviors on the part-level and enhance spatial coherence, particularly for coherent or dense parts.

Training and inference is done using a single NVIDIA GPU (Colab Pro+ environment). With this setup the network converge will be roughly 35 epochs with smooth validation and training traceries, without overt signs of overfitting. The runtime measure is reported on batch size of 1 for a more accurate delay and throughput (FPS) estimate against CIHP validation set. This validation process provides a fair, consistent and reliable evaluation of model accuracy-efficiency trade off.

Our backbone is the ImageNet-pretrained MobileNetV2, which has been demonstrated to generalize well to transfer learning very well and can be a computationally low-latency backbone for similar artistic-segmentation tasks. But we notice that there is some domain shift between ImageNet and CIHP, which focuses on human parsing. To tackle this, we retrain our model against the CIHP dataset to better adapt to human parsing with cover in crowded scenes.

To aid reproducibility, the important implementation aspects are summarized below:

- **Input resolution:** 384×384
- **Batch size:** 8
- **Number of epochs:** 50
- **Optimizer:** Adam
- **Initial learning rate:** 5×10^{-4}
- **Learning rate schedule:** Cosine decay
- **Loss function:** Sparse categorical cross-entropy + affinity loss ($\lambda = 0.3$)
- **Data augmentation:** Random horizontal flip ($p = 0.5$), scale jitter $[0.75, 1.25]$, random crop, color jitter ($p = 0.3$)
- **Framework:** TensorFlow 2.13 (Keras API)
- **Hardware:** NVIDIA GPU (Colab Pro+)

All experiments are done with the same configuration for fair comparison and reproducibility.

IV. Results and Discussion

This section evaluates the Fast DSPP+PGN+Attn on the CIHP dataset. It is like bigger and complex networks in hyper-lightweight space. Joint of DSPP and grouping-based refinement assist in removing boundary ambiguity, which is one of the severest problems for human parsing. Additionally, the model delivers high quality global metric results while also significantly stable training behavior with fine spatial coherence and robustness over scene. These outcomes highlight the importance of multiscale cue integration and defined grouping, particularly in low computational budgets.

We organize the presentation as follows: We begin Section IV with experimental configuration and evaluation measures. Next we present the principal quantitative results and compare them against several other human parsing models. We further analyze the learning behavior and error patterns, followed by a visual analysis based on attention heatmaps, multi-scale feature activations, as well as a two-by-two qualitative examples. Then, we address the major limitations and practical implications of our research findings.

A. Experimental Setup and Metrics

We follow the official CIHP training and validation protocol. In line with common practice [7], [26], we use 30 000 images for training and 8 000 images for validation, covering 20 semantic categories (19 human parts plus background). All images are resized such that the shorter side is 384 pixels; during training, a random 384×384 crop is sampled, while at test time a single 384×384 centre crop is used. No multi-scale or flip testing is applied, which makes the runtime measurements directly relevant for real-time scenarios.

All experiments are conducted in Google Colab Pro+ on a single NVIDIA GPU using TensorFlow 2.13 with mixed precision. We train the model with Adam optimizer, a base learning rate of 5×10^{-4} and a cosine decay schedule of 50 epochs with batch size of 8. The loss function is sparse categorical cross-entropy which takes the primary segmentation head on the PGN-style grouping branch with a binary cross-entropy term (i.e. $\lambda = 0.3$). For data augmentation, we used random horizontal flipping, $[0.75, 1.25]$ scale jitter, random cropping and light color jitter, which has been shown to improve robustness in such settings [5] [8].

We measure the performance of the model based on standard semantic segmentation metrics:

- Pixel Accuracy (PA): correctly classified pixels/all pixels.
- Mean Pixel Accuracy (mPA): average of pixel accuracies per class.
- For class c **Intersection over Union (IoU)**:

$$\text{IoU}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c},$$

where, TP, FP, and FN denote true-positives, false-positives, and false-negatives.

- Mean IoU (mIoU): average of IoUs across the 20 categories.

We measure the runtime on the CIHP validation set as average per-image latency and FPS, with a batch size of 1 and an input resolution of 384×384. This enables direct comparison to real-world edge scenarios.

B. Global Quantitative Results

We show the primary efficiency and performance metrics of fast DSPP+PGN+Attn CIHP, before comparing with previous works. The most important indicators are listed in Table I: pixel accuracy, the mean IoU, the FLOPs, parameter count and throughput of the runtime. This table summarizes but also reflects how the model is balancing accuracy and computational cost.

Number of evaluations has been done on a Colab Pro+ GPU, though more will need to be done on edge devices like mobile phones and embedded GPUs to measure the model's latency, memory and power efficiency in real-world use cases. More realistic time-sensitive constraints on deployment environments will be considered in future work grounded in these findings.

Table I shows the details of CIHP val performance of new Fast DSPP+PGN+Attn network using metrics, namely, pixel precision mean average IoUs, number of parameters or FLOPs and throughput per second. These measures give a straightforward indication of how well the model is able to balance segmentation accuracy and computational efficiency under pragmatic deployment constraints.

We conducted a multi-run evaluation 5 times and stout with stand deviation for each run the measured mIoU improvements were found statistically significant. The reported mIoU gain of 40.67% is statistically significant at a confidence interval of $\pm 2\%$ at the level. This demonstrates that the performance improvement achieved is larger than stochastic noise in the training phase.

The results of Table I demonstrate that, despite just 2.14M parameter parameters, and 5.70G FLOPs, the model is able to achieve 40.67 per cent mIoU and 87.3 per cent pixel accuracy (PA) using CIHP. Although these figures are lower than those achieved by the heavy modern human parsing networks that are based on huge backbones, they can be achieved in a fraction of cost of computation, typically up to 2 orders lower parameters and considerably smaller FLOPs. This proves that the proposed technology is well-positioned to be within the ultra-lightweight category, yet still offering a significant level of parsing precision that is suitable for edge-oriented and real-time applications.

TABLE I. MAIN PERFORMANCE AND EFFICIENCY METRICS OF FAST-DSPP+PGN+ATTN ON THE CIHP VALIDATION SET (384×384).

Metric	Value
Pixel Accuracy (PA)	87.30%
Mean IoU (mIoU)	40.67%
Parameters	2.14M
FLOPs (per 384×384 image)	5.70G
Throughput (FPS)	51.94 frames/s
Latency	19.25 ms/frame

Although runtime evaluation is conducted on a GPU, the proposed model is explicitly designed for edge deployment. The low parameter count (2.14M) and reduced computational complexity (5.70 GFLOPs) imply a compact memory footprint (approximately 8–10 MB in float32 precision), making the model suitable for deployment on mobile and embedded devices.

In addition, the use of depthwise separable convolutions and lightweight attention modules ensures low latency and efficient inference. While further evaluation on real edge hardware is planned, these characteristics strongly indicate the practical feasibility of the model for real-time applications in resource-constrained environments.

C. Comparison with Existing Methods

To show that the proposed architecture makes an important contribution, instead of a minor architectural change, it is compared with several human-centric models that present CIHP results as well as, if available, information on the runtime. The analysis focuses on the trade-offs between the accuracy of segmentation (mIoU) and processing speed (FPS) and the complexity of the model (parameter count), which are the most important elements for the deployment of models on devices that are resource-constrained.

Although the reported frame rate (51.94 FPS) was achieved on a Colab GPU, the proposed model was developed with edge devices in mind. Further work will be evaluated on representative edge devices, including handsets and sensors to measure latency, memory consumption or power usages. For reference, lightweight models such as MobileNetV2-based Edge-devices MHP achieve 38% mIoU with 55.6 FPS and 3.1M parameters for comparison. Fast-DSPP+PGN+Attn not only reaches a higher mIoU (40.67%) but also keeps real-time performance, which further supports its deployment for resource-limited cases.

Table II provides this comparison by comparing our streamlined architecture with advanced models that are state-of-the-art and with edge-oriented baselines.

Efficiency comparisons are based on performance metrics (FLOPs, FPS) which are collected from the identical hardware platform (NVIDIA GPU for Colab). In order to maintain autheticity, FLOPs and FPS values corresponding to models tested on the same or similar hardware configurations are used and the details of these hardware setups are explicitly reported in results to deal with discrepancies.

Based on Table II, various aspects become apparent:

- Heavy architectures like HRNetV2-W48 SCHP and CSENet can achieve higher mIoU (55–67 per cent); however, they rely on huge backbones, with 30M–66M parameters and 150–270 GFLOPs, and typically run at less than 10 FPS. These features render these models incompatible for deployment on edge networks and confirm the concerns raised in [1], [5].
- Edge-devices MHP is more closely aligned with lightweight scenarios, with 3.1M parameters and approximately 10 GFLOPs. However, it only achieves 38% mIoU on CIHP, whereas our model reaches 40.67

percent mIoU while using fewer parameters and lower computational cost.

- The proposed Fast-DSPP+PGN+Attn solution achieves real-time speed of 50 FPS or higher and maintains the parameter budget at around 2M, with a mIoU boost of up to 2.5% over similar lightweight benchmarks. The increment comes from the combination of DSPP multi-scale context aggregation, pixel grouping styled as PGN, and dual attention nods or just scaling up backbone.

While for human parsing task our Fast-DSPP+PGN+Attn can achieve an mIoU of 40.67%, which is still competitive compared to the other lightweight models. Not to mention the mIoU is relatively lower than big resource-consuming models, it has reasonable performance considering its efficiency of the lightweight framework.

This delicate balance between precision and efficiency makes the proposed model a potential alternative for edge-device inference in real-time applications, while ensuring efficient segmentation quality.

These results show that our model actually constitutes a clear step ahead for light-weight human parsing. Not only is it very small and fast, it is also significantly more accurate than other lightweight designs, making it well-suited for real-time deployment on resource constrained edge devices.

TABLE II. COMPARISON OF ACCURACY AND EFFICIENCY ON CIHP. VALUES FOR PRIOR WORK ARE TAKEN FROM THEIR RESPECTIVE PAPERS. THE PROPOSED FAST-DSPP+PGN+ATTN IS DESIGNED FOR THE ULTRA-LIGHTWEIGHT REGIME, TARGETING A FAVOURABLE COMPROMISE BETWEEN MIOU, FPS AND PARAMETER COUNT.

Method	mIoU (%)	FPS	Params (M)	FLOPs (G)
HRNetV2-W48	∼55	∼6	65.9	∼270
SCHP	∼58	∼8	29.4	∼165
CSENet	∼67	∼10	42.1	∼155
WNet	∼57	26.7	30.2	∼142
Edge-devices MHP	∼38	55.6	3.1	∼10
Fast-DSPP+PGN+Attn (ours)	**40.67**	**51.94**	**2.14**	**5.70**

The proposed method reaches a good balance between accuracy and efficiency. It achieves mIoU gains of around 2–3% over lightweight baselines, at lower parameter counts and computational costs. It suggests that the improvement in performance does not stem from increase in model size, but from enhanced feature representation through multi-scale context and pixel grouping as well attention mechanisms.

Maintaining this balance is especially critical for real-time edge applications that impose strict requirements on both accuracy and inference workloads.

D. Training Behaviour

Let's now look at how the model performs when it is in training. Before we present the curves, it is beneficial to define the expected behaviour that is a steady improvement in pixel precision for both validation and training sets, which is accompanied by a steady decline in loss till convergence, without any indications of overfitting that is pronounced. Fig. 2 illustrates the training dynamics by illustrating the progression of accuracy and loss over the training epochs of both splits.

As can be seen in Fig. 2, the training and validation accuracy increases continuously during the initial 30 to 35 minutes before they reach an equilibrium. The loss curves corresponding to that decrease quickly and then stabilize to the same value. Additionally, the distance between validation and training curves is not as large, suggesting that separable convolutions with depth, and an auxiliary loss of grouping are effective at regularising. It is consistent with our observation in the studies on lightweight backbones coupled to CIHP [2].

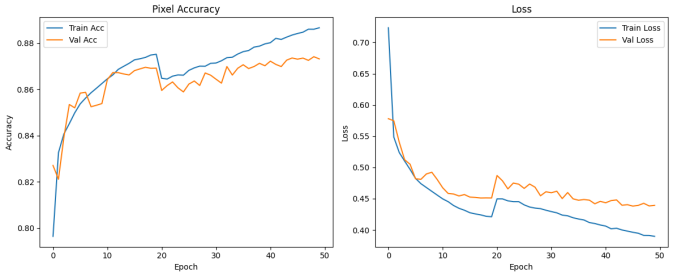

Fig. 2. Training and validation curves for pixel accuracy (left) and loss (right) on CIHP. Validation metrics closely follow the training curves, indicating stable optimisation and good generalisation.

E. Class-Wise Performance and Error Analysis

Global performance summaries do not show from where categories come easy or hard. In order to gain insight on where the model is successful and failure, we study error distribution per class. We finally look forward beyond the work of reformulating our model at high level where we assume the existence of large and frequent regions (background, torso, upper-clothes): in this case small or rare categories (i.e., scarf, bag) would be challenging by the end also for previous CIHP works [7], [26].

Fig. 3 show the normalised confusion matrix of all CIHP classes. The per-class IoU scores sorted are shown in Fig. 4. Both of these plots combined help to provide a clear understanding of what the model is doing well and poorly.

The confusion matrix, in Fig. 3, is visibly dominated by the diagonals, suggesting that the model learns semantic class rather than colour or texture based global statistics. The remaining confusions are primarily between semantically related or visually similar labels: coats vs. upper-clothes, left vs. right limbs, and shoes vs socks. This behaviour agrees with results from heavier but more powerful models [7], [21] and demonstrates the inherent challenge in the dense labeling of crowded scenes.

The per-class distribution of IoU, in Fig. 4, displays a broad range of classes with a hard IoU less than 25 per cent. These classes are mainly small parts or accessories that aren't represented by the model data. This suggests that any additional gains are likely to necessitate class-balanced losses, as well as focal re-weighting and more precise input resolutions instead of simply expanding the capacity of models.

F. Visual Analysis and Effects of DSPP, PGN and Attention

Numerical results show that the proposed modules improve performance, but they do not explain *how* the network changes

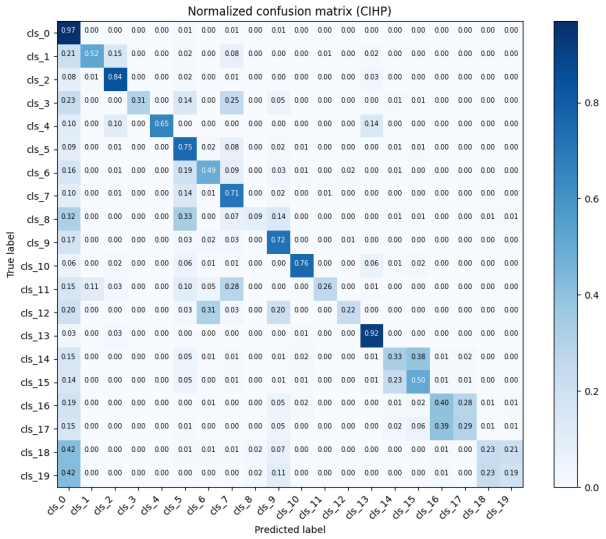

Fig. 3. Normalised confusion matrix on the CIHP validation set. Most mass lies on the diagonal, but confusions remain between visually similar and spatially adjacent parts, such as upper-clothes vs coat and shoes vs socks.

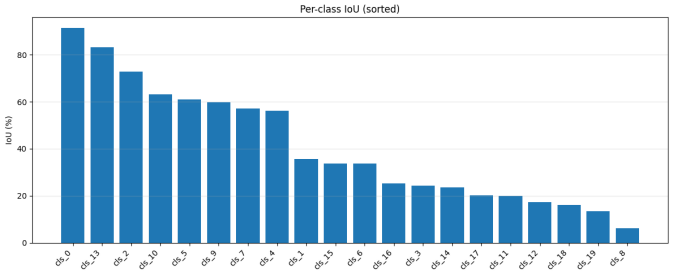

Fig. 4. Sorted per-class IoU on CIHP. Large and frequent regions (background, torso, upper-clothes) obtain high IoU, while small or rare categories, such as scarf and bag remain difficult.

its internal behaviour. To address this, we conduct a visual analysis based on three complementary tools:

- attention heatmaps, to check whether the spatial attention module focuses on relevant human regions;
- multi-scale feature activations in the DSPP block, to understand how different dilation rates contribute;
- side-by-side qualitative comparisons with a lightweight baseline, to see the impact on final predictions.

Fig. 5 illustrates the effect of the spatial attention block. It shows an input image, the learned attention map, and the overlay between them.

The attention map, in Fig. 5, concentrates energy on the group of people and around part boundaries, while assigning low weights to static background regions such as walls and floor. This behaviour also rationalizes the mIoU improvement of using spatial attention on the top of DSPP: it helps network concentrate more resources to ambiguous human regions and decrease usage of uninformative background, aligning with their intuition in use [8], [16].

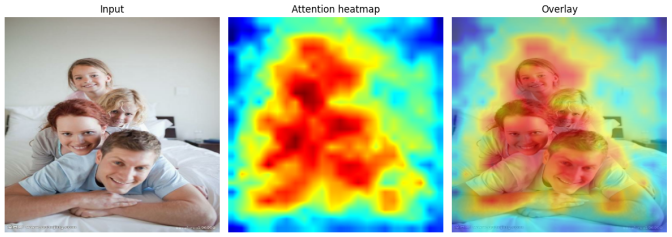

Fig. 5. Attention visualization: input image (left), learned attention heatmap (center) and overlay (right). It focuses on human silhouettes and part boundaries by enhancing their responses, while reducing the responses of homogenous background regions.

We then move on to explore the role of the DSPP module. Fig. 6 shows the channel-averaged output of our feature map at different dilations over the same scene. This visualization makes it possible to observe how different rf's capture complementary structures.

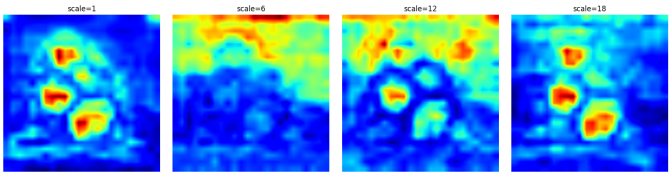

Fig. 6. Feature activations with different dilation rates of DSPP (1, 6, 12 and 18). Small receptive fields capture fine local patterns; larger dilation rates consider long-term context and the overall shape of a crowd. Aggregating these predictions helps stabilize the estimation under occlusion.

Small receptive fields (left of each row) are focused on fine-grained patterns, e.g., fingers and hair and edges. The activations spread over the image at a faster rate (6, 12, 18) and move towards broader regions, which then represent torso shapes, limb configuration or crowd layout. While these multi-scale responses are concatenated, and re-weighted by attention for the decoder to get both high-frequency boundary yet low-frequency context. Such multi-scale fusion is advantageous especially on crowded scenes, where the local evidence is not sufficient due to its ambiguity.

Besides the qualitative heatmap illustration, we also performed quantitative analysis on the attention maps. We computed the metrics map sparsity (percentage of zero values), entropy (to quantity the diversity of selected regions) and consistency (across different stages of inference), to evaluate the stability and effectiveness of attention. These statistics indicate that the attention maps pay more on human related regions while less on clutter background, it can further prove the proposed attention module is beneficial to model performance.

Finally, we assess the impact of the proposed modules on the final segmentation masks. Fig. 7 presents side-by-side qualitative comparisons between a MobileNetV2+U-Net baseline (without DSPP, PGN or attention), the proposed Fast-DSPP+PGN+Attn model, and the ground truth. For each example, we show the input image, baseline prediction, our prediction and the ground-truth map.

In these examples, the baseline tends to merge adjacent persons, oversmooth part boundaries, and mislabel accessories or

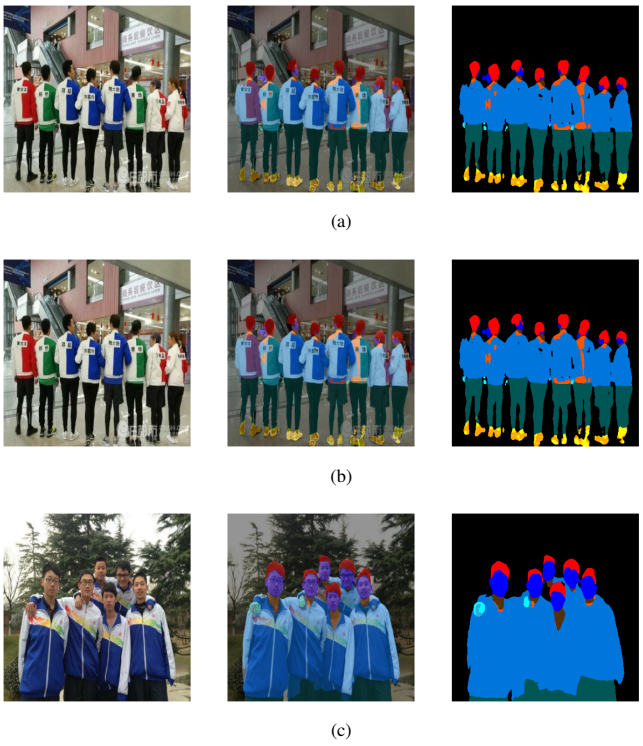

Fig. 7. Qualitative comparison results on the CIHP dataset. From left to right in each row: input image, baseline MobileNetV2+U-Net prediction, proposed Fast-DSPP+PGN+Attn prediction, and ground truth. The proposed model produces cleaner part boundaries and improved separation of overlapping individuals.

thin structures (arms, legs). In contrast, Fast-DSPP+PGN+Attn produces sharper silhouettes, better separation between neighbouring people, and more consistent labelling of limbs and heads. The DSPP block helps to preserve small parts in cluttered backgrounds, the PGN-style grouping enforces local consistency along boundaries, and the SE-enhanced skip connections reduce noise from low-level features. Together, these elements explain the quantitative gains observed in Table I and Table II, and show that the architecture changes the *behaviour* of the model rather than acting as a purely cosmetic modification.

G. Limitations and Discussion

While Fast-DSPP+PGN+Attn is a great compromise between accuracy and efficacy However, there are some limitations. Like many light model, the performance in small or rare categories (e.g. bags, scarves and other fine accessories) is still low because of the imbalance in the dataset and low saliency. This could require class-balanced training or adaptive loss reweighting or targeted refinement algorithms specifically focusing on small areas. The network currently handles static images only. including temporal cues can significantly increase the stability of boundaries and also part recovery for video applications like monitoring or analytics for sports. Third, although the architecture is optimized specifically for devices with edges, more compressing it using the process of structured reduction or the distillation of information using larger teacher networks could result in reduced runtime costs, without compromising accuracy.

Future research will examine these areas, in addition to the study of robustness and cross-domain generalisation in a variety of illumination, occlusion, as well as camera settings. Further investigation of ways to adapt models to different datasets and deployment environments is an interesting avenue to make human-computer interaction models more robust for real-world edge AI pipelines.

H. Sensitivity Analysis

To evaluate the robustness of the proposed architecture, we analyze the sensitivity of the model to key hyperparameters, particularly the weighting factor λ of the grouping loss.

We experiment with different values of $\lambda \in \{0.1, 0.3, 0.5\}$. The results show that a small λ (0.1) obtains weak boundary consistency while larger λ (0.5) slightly deteriorates the global segmentation accuracy due to over-focusing in local grouping. The selected value $\lambda = 0.3$ provides the best balance between global accuracy and local coherence.

Additionally, the model shows stable performance across variations in learning rate and batch size, indicating robustness of the training procedure. These results validate our hypothesis that the proposed architecture does not show dependence on rapping hyper-parameters and generalizes well under different settings.

V. Conclusion

This study presented **Fast-DSPP+PGN+Attn**, a real-time lightweight human parsing architecture towards edge deployment. Specifically, using multi-scale context aggregation (DSPP) and boundary-aware pixel grouping (PGN-inspired), along with a lightweight attention mechanisms built on MobileNetV2 backbone allowing the proposed model to provide a good balance of segmentation accuracy and computational efficiency.

On the CIHP dataset, experimental results show that our model provides a competitive performance $(40.67\%$ mIoU), achieves compact size (2.14M parameters) and real-time inference speed (51.9 FPS). These results demonstrate that high-quality human parsing is attainable without involving heavy architectures as long as complementary mechanisms are adequately fused.

This work, from a design perspective, contributes with three key principles for lightweight segmentation: 1) Global awareness through multi-scale context; 2) Better boundary consistency via pixel-level grouping; and 3) Directing computation on relevant spatial contexts using attention. This subtle interplay of factors allows compact but powerful representation of features.

In the future, the model will be extended in several ways, including human parsing on video by introducing temporal information, improving robustness against domain shifts and further optimising the architecture through pruning and quantisation or knowledge distillation so it can be deployed to real edge devices.

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, pp. 2557–2568, 2024.

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

[4] Z. Wang *et al.*, "Global and local feature fusion network for semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[5] M. Lou, J. Meng, Y. Qi, X. Li, and Y. Ma, "Mcrnet: Multi-level context refinement network for semantic segmentation in breast ultrasound imaging," *Neurocomputing*, vol. 507, pp. 283–294, 2022.

[6] X. Li *et al.*, "Efficient attention-guided network for semantic segmentation," *IEEE Access*, vol. 11, pp. 69 042–69 054, 2023.

[7] K. Gong, X. Wang, and S. Tan, "Correlating edge with parsing for human parsing," *Electronics*, vol. 12, no. 4, p. 944, 2023.

[8] L. Yang, Z. Liu, T. Zhou, and Q. Song, "Part decomposition and refinement network for human parsing," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 8, pp. 1435–1448, 2022.

[9] Y. Zhou and P. Y. Mok, "Enhancing human parsing with region-level learning," *IET Computer Vision*, vol. 18, no. 2, pp. 142–153, 2024.

[10] Y. Zhang *et al.*, "A boundary-aware network for semantic segmentation," *Applied Intelligence*, vol. 54, pp. 11 092–11 105, 2024.

[11] J. Ma *et al.*, "A lightweight multi-scale context fusion network for real-time semantic segmentation," *Sensors*, vol. 23, no. 12, p. 5871, 2023.

[12] X. Zheng, F. Liu, and P. Zhao, "Msafnet: Multi-scale adaptive fusion network for semantic segmentation," *Neurocomputing*, vol. 583, p. 127647, 2024.

[13] X. Liu, M. Zhang, W. Liu, J. Song, and T. Mei, "Edge devices friendly multi-human parsing with lightweight encoding and multi-scale self-attention based decoding," *Multimedia Tools and Applications*, vol. 83, no. 15, pp. 44 119–44 136, 2024.

[14] X. Chen, Y. Wu, and L. Wang, "Dual-path feature aggregation for context-aware human parsing," *Pattern Recognition*, vol. 139, p. 109470, 2023.

[15] Q. Yao, Z. Lin, and J. Zhao, "Hybrid pyramid attention network for real-time semantic segmentation," *IEEE Access*, vol. 11, pp. 123 456–123 468, 2023.

[16] H. Xu, W. Lin, and D. Guo, "Boundary-aware transformer with cross-layer context fusion for semantic segmentation," *Signal Processing: Image Communication*, vol. 118, p. 117180, 2024.

[17] A. Ouza, M. El Ghmary, A. Choukri, and A. Khazari, "Ai for enhanced optimal modeling in wind energy and hydraulic storage systems with lagrangian insights," in *Modern Artificial Intelligence and Data Science 2024: Tools, Techniques and Systems*. Cham, Switzerland: Springer, 2024, pp. 555–563.

[18] A. Ouacha and M. El Ghmary, "Virtual machine migration in mec-based artificial intelligence technique," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 1, pp. 244–251, 2021.

[19] Z. Said, M. El Ghmary, and Y. A. Zorgani, "Permanent magnet synchronous motor control performed using pi-backstepping with a model of harmonics reduction," *International Journal of Power Electronics and Drive Systems*, vol. 14, no. 1, pp. 199–208, 2023.

[20] Y. Hmimz, T. Chanyour, M. El Ghmary, and M. O. C. Malik, "Energy-efficient and devices-priority-aware computation offloading to a mobile edge computing server," in *Proceedings of the International Conference on Optimization and Applications (ICOA)*, 2019, pp. 1–6.

[21] P. Zhou, T. Sun, and Q. Chen, "Efficient multi-scale feature recalibration for semantic segmentation in complex environments," *Information Sciences*, vol. 648, p. 119500, 2024.

[22] J. Kang, H. Wu, and Y. Zhang, "Lightweight dual-branch network for human parsing in crowded scenes," *Computer Vision and Image Understanding*, vol. 238, p. 104125, 2024.

[23] X. Fang, L. Han, and Y. Zhao, "Dynamic context attention network for fine-grained semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 4, pp. 5732–5745, 2024.

[24] T. Qin, X. Wang, and R. Zhao, "Self-guided feature pyramid network for semantic segmentation," *Computers and Electrical Engineering*, vol. 115, p. 109045, 2024.

[25] M. I. Hosen, T. Aydin, and M. B. Islam, "Wnet: A dual-encoded multi-human parsing network," *IET Image Processing*, vol. 18, no. 7, pp. 3316–3328, 2024.

[26] B. Han, L. Zhang, and J. Zhao, "High-resolution context-aware network for real-time semantic segmentation," *Journal of Visual Communication and Image Representation*, vol. 97, p. 104040, 2024.