

HSMTL-Net: A Multi-Task Learning Framework for Human Segmentation with Integrated Regression and Classification Modules

Faezeh Rohani^{a,*}, Soroosh Babae Khobdeh^b, Sayediman Mostafavi^c, Hekmatullah Zurmati^c, Kais Al Wadi^c, Saqr Mohammed Khalaf^d, Andreas Mueller^e

^aDepartment of Software Engineering, Üsküdar University, İstanbul, Türkiye; Brain and Trauma Foundation Grisons, Landquart, Switzerland

^bPost Bank of Iran, Guilan Province, Iran

^cDepartment of Software Engineering, Üsküdar University, İstanbul, Türkiye

^dDepartment of Chemical Engineering, College of Engineering, Al-Mustaqbal University, Babylon, Iraq

^eBrain and Trauma Foundation Grisons, Landquart, Switzerland

DOI:10.37819/Jaike.26.0025

Article history:

Received : 14.05.2026

Revised : 10.06.2026

Accept : 07.07.2026

Published : 21.10.2026

Corresponding author:

E-mail:

faezeh.rohani@uskudar.edu.tr;

faezeh.rohani@gtsg.ch

Abstract: Human segmentation is a fundamental task in computer vision, with applications in surveillance, sports analytics, autonomous systems, and human-computer interaction. This paper proposes HSMTL-Net, an attention-guided and boundary-aware multi-task learning framework for fine-grained human segmentation. The framework is built on a lightweight U-Net-inspired encoder-decoder backbone and jointly optimizes three complementary tasks: pixel-level human segmentation (Seg-Module), auxiliary gender classification (Gender-Module), and geometric regression of region descriptors (Seg-Tuner-Module). The two auxiliary heads provide semantic and geometric supervision signals that regularize the shared encoder, with the explicit aim of improving boundary fidelity and feature generalization under limited-data conditions. On the Supervise.ly Filtered Segmentation Person Dataset, HSMTL-Net achieves a Dice score of 97.83%, a mean IoU of 96.21%, a gender-classification accuracy of 97.96%, and a regression mean squared error (MSE) of 0.00995. A task-wise ablation shows that adding the regression and classification heads improves segmentation over the single-task baseline, with the largest gain attributable to the geometric regression head. The principal contribution of this work is the use of mask-derived geometric regression as auxiliary supervision for boundary-sensitive human segmentation within a single lightweight multi-task learning framework, together with an explicitly defined multi-task weighting strategy for joint optimization.

Keywords: Human segmentation, multi-task learning, boundary-aware segmentation, attention mechanism, auxiliary supervision, gender classification, supervisory segmentation person dataset, human-computer interaction

1. INTRODUCTION

In recent years, machine learning has substantially improved the performance of computer vision applications. Numerous approaches have been proposed for image classification, segmentation, and registration in domains such as facial recognition

[Saito et al.], pattern detection [Moravvej et al.], autonomous driving [Kaymak et al.], and medical image processing [Sobhaninia et al., Moravvej et al.]. Semantic segmentation remains one of the most challenging and computationally demanding tasks in this field, and deep learning has been widely adopted

to improve accuracy while controlling computational cost.

Human detection and segmentation are active research areas with applications including person detection in cluttered scenes [Ahmed et al.], people counting, driver-assistance systems, industrial monitoring [Ahmed et al.], and human-behaviour analysis [Cristani et al.]. Sports analytics is a further high-impact application, encompassing player tracking [Ullah et al.], equipment detection [Cai et al., Kamble et al.], pose estimation [Bridgeman et al.], and action detection [Buch et al.].

Encoder–decoder and attention-based networks have markedly improved dense prediction. However, accurately segmenting fine human structures, such as hair strands, fingers, and thin limbs, under complex backgrounds and limited training data remains difficult. Single-task segmentation frameworks, in particular, do not exploit complementary semantic or geometric information that could regularize feature learning and stabilize optimization.

Prior work on intelligent healthcare and motion analysis has shown that discriminative feature representation and auxiliary supervision improve robustness. For example, machine-learning frameworks for ADHD diagnosis based on gait, balance, and skeleton data demonstrate that extracting structured features from complex human-motion patterns improves classification and analysis robustness [Khodabakhshi et al., Vahdati et al.]. This motivates the use of auxiliary semantic and geometric tasks to strengthen shared representations in segmentation networks.

1.1. Problem Statement and Novelty

Most existing human-segmentation methods optimize a single objective and therefore do not leverage cheap, automatically derivable supervision that is geometrically tied to the segmentation mask itself. Boundary-aware losses and attention refinement have each been studied individually (for example, BANet; the Look-into-Person framework; and multi-task parsing and pose networks), but they are typically combined without a clear account of why a particular auxiliary task should improve boundary metrics. The distinctive element of this work is the use of mask-derived geometric regression (area, perimeter, centroid, principal-axis length, and orientation) as an auxiliary supervision signal that is, by construction, sensitive to the shape and

spatial extent of the segmented region and therefore directly informative for boundary localization. This is paired with a transparent and explicitly reported multi-task weighting scheme and training protocol that facilitate a clear understanding of the proposed optimization strategy.

Gender classification is positioned strictly as an optional semantic regularizer, and its necessity, provenance, and fairness implications are analyzed in Sections 3.4 and 5.3.

The main contributions of this work are as follows:

- A boundary-aware multi-task framework for fine-grained human segmentation in which mask-derived geometric regression is used as auxiliary supervision to improve boundary fidelity.
- An attention-guided feature-refinement stage applied after each skip connection to strengthen the shared encoder–decoder representation; the attention operator is defined explicitly in Section 3.2.
- A transparent multi-task weighting and training protocol, including the loss-weight selection procedure together with the implementation and data-partition details described in Sections 3.5 and 3.6.0
- Comparative experiments against modern human-segmentation methods on a common dataset, with an explicit statement of the provenance of the baseline results and the resulting comparison conditions (Section 4.5).
- A task-wise ablation analysis evaluating the individual contribution of each module to segmentation performance (Section 4.4).

The remainder of the paper is organized as follows. Section 2 reviews related literature; Section 3 describes the proposed framework, including the architecture, loss functions, and reproducibility details; Section 4 reports the experimental protocol and results; Section 5 discusses the findings, limitations, and ethical considerations; and Section 6 concludes.

2. LITERATURE REVIEW

Human segmentation has attracted sustained attention because of its applications in surveillance, autonomous systems, healthcare, sports analytics, and human–computer interaction. Approaches range

from classical image processing to modern deep networks.

Early methods relied on thresholding, region growing, k-means clustering, graph cuts, and histogram-based clustering [Zaitoun et al.]; these used low-level handcrafted features and were brittle under complex scenes. Zhao and Nevatia [Zhao et al.] used a Markov Chain Monte Carlo formulation to estimate human presence in crowded scenes, but classical approaches generally degraded on large, highly variable datasets.

Convolutional neural networks (CNNs) substantially improved accuracy and representation learning [Hariharan et al., Sobhaninia et al.], learning hierarchical features from large datasets [Russakovsky et al.] for dense pixel-wise prediction [Long et al.]. Fully Convolutional Networks were an early deep framework that inspired later human-

segmentation architectures [Song et al., Ahmed et al.].

Several lightweight and efficient networks target real-time use. ENet [Paszke et al.] was designed for low-latency segmentation. Encoder-decoder architectures such as U-Net [Ronneberger et al.] use skip connections to preserve spatial detail and are widely used in medical and dense-prediction tasks [Sobhaninia et al.]. DeepLab [Chen et al.] enlarged the receptive field via atrous convolution.

Attention-based methods improve fine-detail preservation and feature discrimination through squeeze-and-excitation blocks, convolutional block attention modules (CBAM), and attention-guided encoder-decoders. Transformer-based frameworks such as SegFormer and Mask2Former further strengthen global context modelling.

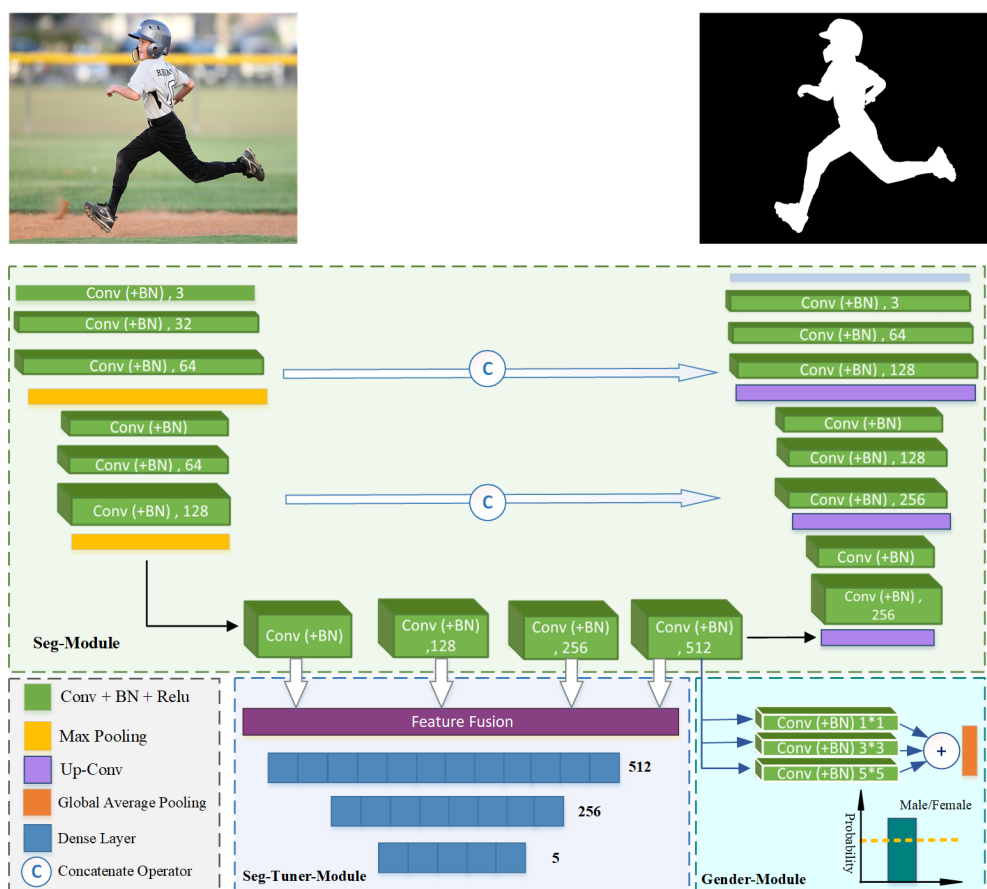


Figure 1. Overview of the HSMTL-Net architecture: a shared U-Net-inspired encoder-decoder with three heads — the Seg-Module (segmentation), the Gender-Module (auxiliary classification), and the Seg-Tuner-Module (geometric regression) — jointly optimized over shared features

For human-specific segmentation, Shen et al. introduced a CNN-based portrait segmentation network; Chen, Qi and Shen [Chen et al.] proposed the Boundary-Aware Network (BANet), which emphasizes boundary information during training; and Yu et al. proposed a human-centric video instance segmentation framework that reduces positive-sample ambiguity through inner-centre sampling.

Many of these methods are single-task. Multi-task learning, by contrast, trains related tasks jointly over a shared representation to improve generalization and reduce overfitting [Caruana], and auxiliary semantic supervision can improve feature quality [Ruder, Zhang et al.]. In human analysis, Liang et al. jointly performed parsing and pose estimation (Look into Person), and Jug et al. jointly learned skeleton estimation, dense-pose prediction, and segmentation. Unlike these works, the present framework couples segmentation with a mask-derived geometric regression task that is explicitly designed to be boundary-sensitive, rather than with pose or parsing labels that require additional manual annotation.

3. METHODS

HSMTL-Net is a multi-task deep network that simultaneously performs human segmentation, gender classification, and geometric regression of descriptors derived from the segmented region.

As shown in Figure 1, the architecture follows a U-Net-inspired encoder-decoder design with a shared encoder and three task-specific heads. The shared encoder is regularized by the auxiliary heads, which is intended to yield more discriminative and generalizable features under limited data.

Multi-task learning addresses the high annotation cost of segmentation by encouraging the network to learn shared, partially task-independent representations. Because both auxiliary targets are derived automatically from the existing masks and metadata, they add supervision without additional manual labelling. Back-propagating gradients from auxiliary heads attached at the bottleneck also provides intermediate supervision that can mitigate vanishing gradients and stabilize training.

3.1. Dataset Preprocessing and Label Generation

All images were resized to 256×256 pixels and normalized to the range $[0, 1]$. Ground-truth annotations provided in JSON format were converted into binary PNG masks and resized to match the input image dimensions, ensuring pixel-wise correspondence between each image and its associated segmentation label.

Segmentation labels were obtained directly from the provided pixel-level annotations. Image-level gender labels were used exclusively as auxiliary

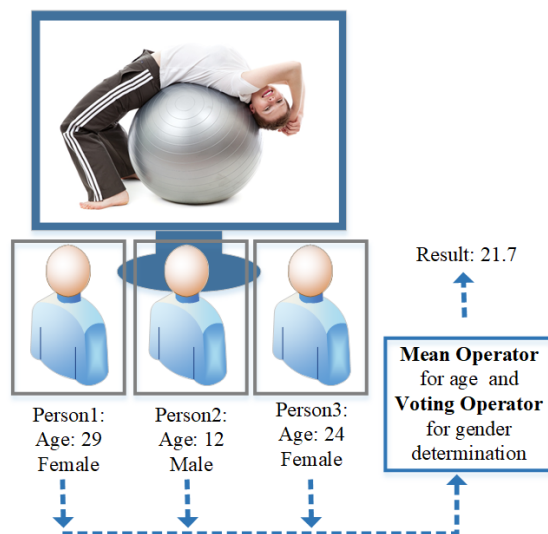


Figure 2. Label-generation pipeline for the auxiliary tasks. Geometric descriptors (area, perimeter, centroid, principal-axis length, and orientation) are extracted from each binary mask, while image-level annotations provide the gender label

supervision during network training and served only to regularize the shared feature representation within the multi-task learning framework. No additional manual annotation was performed as part of this study.

Geometric regression targets were extracted automatically from each binary mask using standard connected-component analysis, contour extraction, and image-moment operations applied to the largest segmented human region. The extracted descriptors include the segmented area (pixel count), perimeter (contour length), centroid coordinates $((C_x, C_y))$, principal-axis length (largest diameter), and the orientation angle of the principal axis with respect to the x-axis.

The same preprocessing and label-generation pipeline was consistently applied to the training, validation, and test sets to ensure methodological consistency and prevent data leakage (Figure 2).

3.2. Seg-Module (Segmentation Head)

The Seg-Module performs the primary segmentation task using a U-Net-style encoder-decoder with attention-guided refinement applied after each

skip connection. The encoder applies convolution, batch normalization, and ReLU activation, with max-pooling for downsampling; the decoder applies transposed convolution for upsampling. The network depth follows the original U-Net (four downsampling stages), so that for a 256×256 input the spatial resolution decreases as $256 \rightarrow 128 \rightarrow 64 \rightarrow 32 \rightarrow 16$ at the bottleneck and is symmetrically restored in the decoder. Skip connections forward encoder features to the corresponding decoder stage to preserve spatial detail, and a final 1×1 convolution with softmax produces the two-class segmentation map. To reduce computational cost, the encoder uses fewer channels per stage than the original U-Net while preserving this depth; the Seg-Module is therefore a lightweight, attention-augmented U-Net.

Attention refinement. To emphasize informative features and suppress background activations, each skip connection is gated by a lightweight channel-and-spatial attention block of the CBAM type. Let F denote the encoder feature map forwarded through a skip connection. The block computes channel and spatial attention as:

$$Mc(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \quad (1a)$$

$$F' = Mc(F) \odot F \quad (1b)$$

$$Ms(F') = \sigma(f^{7 \times 7}([AvgPool(F'); MaxPool(F')])) \quad (1c)$$

$$\hat{F} = Ms(F') \odot F' \quad (1d)$$

where σ is the sigmoid function, $\odot$ denotes element-wise multiplication, MLP is a shared two-layer perceptron with reduction ratio r , and $f^{7 \times 7}$ is a 7×7 convolution. The refined feature $\hat{F}$ is concatenated with the corresponding decoder feature.

3.3. Seg-Tuner-Module (Geometric Regression Head)

The Seg-Tuner-Module learns geometric and structural descriptors of the segmented region as an auxiliary regression task. Because these descriptors depend directly on the shape and extent of the mask, supervising them encourages the shared encoder to encode boundary- and extent-sensitive information, which in turn benefits the segmentation of thin and fine structures.

Bottleneck features from the shared encoder are aggregated by a feature-fusion stage (global average pooling of the bottleneck tensor followed by flattening) and then passed through three fully connected layers of width 512, 256, and 5, producing the five geometric targets defined in Section 3.1.

3.4. Gender-Module (Auxiliary Classification Head)

The Gender-Module is an optional auxiliary classification head whose primary purpose is to regularize the shared encoder rather than to perform biometric identification. It consumes bottleneck features, applies convolutional layers with multiple kernel sizes to capture multi-scale context, aggregates them with global average pooling, and produces a binary prediction through fully connected layers.

Because gender classification raises privacy and fairness concerns, its necessity is evaluated

empirically in the ablation study (Section 4.4): the framework is fully functional without this head, and the regression head alone already provides the majority of the boundary-quality gain. The ethical analysis, label provenance, and fairness considerations are presented in Section 5.3.

$$L_{total} = \alpha_1 \cdot L_{Seg} + \alpha_2 \cdot L_{SegTuner} + \alpha_3 \cdot L_{Gender} \quad (2)$$

The segmentation loss combines a boundary-weighted cross-entropy term and a Dice term:

$$L_{Seg} = w \cdot E + (1 - Dice) \quad (3)$$

where the pixel-wise cross-entropy and Dice terms are defined as:

$$E = - \sum_{(x,y)} \log p_{c(x,y)} \quad (4)$$

$$Dice = 2 |G \cap S| / (|G| + |S|) \quad (5)$$

Here G is the ground-truth mask, S is the predicted mask, (x, y) indexes a pixel, and $p_{c(x, y)}$ is the

$$L_{SegTuner} = (1/n) \cdot \sum_{i=1..n} (\hat{P}_i - P_i)^2 \quad (7)$$

where $\hat{P}$ and P are the predicted and ground-truth descriptor vectors. The classification loss L_{Gender} is the binary cross-entropy between the predicted and ground-truth gender labels.

Weight-selection procedure. The task weights (α_1 , α_2 , α_3) and the boundary-map parameters (w_0 , σ) are

3.5. Loss Functions and Task Weighting

The network is trained with a weighted sum of three task losses:

predicted probability of the true class c at that pixel. The boundary weight map w (Ronneberger, Fischer and Brox, 2015) increases the loss near object boundaries, which are the hardest regions to segment:

$$w(x) = 1 + w_0 \cdot \exp(-d(x)^2 / (2\sigma^2)) \quad (6)$$

where $d(x)$ is the distance from pixel x to the nearest ground-truth boundary, σ controls the width of the boundary band (in pixels), and w_0 is the maximum boundary up-weighting. The regression loss is the mean squared error over the five geometric targets:

hyper-parameters. They were selected empirically on the validation set, assigning the largest weight to the primary segmentation task and progressively smaller weights to the auxiliary regression and classification tasks, consistent with standard multi-task practice in which the higher-priority task receives the higher weight. The selected values are listed in Table 1.

Table 1. Selected multi-task loss weights and boundary-weight-map parameters of HSMTL-Net

Symbol	Meaning	Value	Role
α_1	Weight of the segmentation loss (L_{Seg})	3	Primary task
α_2	Weight of the regression loss ($L_{SegTuner}$)	2	Auxiliary task
α_3	Weight of the classification loss (L_{Gender})	1	Auxiliary task
w_0	Maximum boundary up-weighting in the weight map $w(x)$	30	Boundary loss (Eq. 6)
σ	Width (in pixels) of the boundary band in $w(x)$	15	Boundary loss (Eq. 6)

3.6. Reproducibility Details

The proposed framework was implemented in Python 3.7 with TensorFlow and trained on a single NVIDIA GeForce RTX 2080 Ti GPU. The same training, validation, and test partitions described in Section 4.1 were used consistently throughout all experiments to ensure a uniform evaluation protocol.

4. RESULTS

This section reports the dataset, implementation and training protocol, evaluation metrics, ablation study, comparison with prior methods, and a robustness analysis.

4.1. Dataset and Data Split

HSMTL-Net was evaluated on the Supervise.ly Filtered Segmentation Person Dataset [Supervise.ly], which contains human-centred images of varying resolution and scene complexity, each accompanied by a pixel-level segmentation mask. The dataset spans street, indoor, outdoor, sports, and complex real-world scenes.

The dataset contains 5,711 images. Of these, 5,110 were randomly assigned to the training partition and the remaining 601 to the test partition. From the training partition, 20% (1,022 images) was randomly held out as a validation set, leaving 4,088 images for model fitting. The validation set was used for early stopping, learning-rate scheduling, and hyper-parameter selection.

4.2. Implementation and Training Protocol

Images were resized to 256×256 and normalized to $[0, 1]$. The network was optimized with stochastic gradient descent (momentum 0.9, initial learning rate 0.001, mini-batch size 8). The learning rate was reduced by a factor of 0.1 when the validation loss plateaued for 10 consecutive epochs. Training ran for up to 200 epochs with early stopping on the validation loss, and the checkpoint with the lowest validation loss was retained for evaluation; total training time was approximately 39 hours. Data augmentation comprised random horizontal flipping, random rotation, and brightness adjustment. Encoder blocks used convolution, batch normalization, and ReLU; attention-refinement blocks (Eqs. 1a–1d) were applied after each skip connection.

4.3. Evaluation Metrics

Segmentation was evaluated with the Dice score and Intersection-over-Union (IoU); classification with accuracy; and regression with the mean squared error (MSE). IoU and Dice are defined as:

$$IoU = TP / (TP + FP + FN) \quad (8)$$

$$Dice = 2 \cdot TP / (2 \cdot TP + FP + FN) \quad (9)$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. Mean IoU averages the IoU over the evaluated classes. Accuracy is the fraction of correctly classified samples, and a lower MSE indicates better regression performance.

4.4. Ablation Study

Performance was assessed by progressively adding tasks: (i) Seg-Module only (single-task baseline); (ii) Seg-Module + Gender-Module; (iii) Seg-Module + Seg-Tuner-Module; and (iv) the full HSMTL-Net. As reported in Table 2, the single-task baseline reached a Dice score of 90.31% and a mean IoU of 88.76%. Adding the classification head improved these to 91.89% and 89.69% (gender accuracy 92.83%). Adding the regression head produced a larger improvement, to 95.03% and 93.97% (MSE 0.0184). The full model achieved the best results: a Dice score of 97.83%, a mean IoU of 96.21%, a gender accuracy of 97.96%, and an MSE of 0.00995. The regression head therefore accounts for the majority of the boundary-quality gain, supporting the central hypothesis that mask-derived geometric supervision benefits segmentation. The classification head contributes a smaller, complementary improvement, confirming that it is helpful but not essential to the framework.

Table 2. Task-wise ablation: quantitative performance of HSMTL-Net and its variants

Model	Tasks	Dice %	mIoU %	MSE
Seg-Module (baseline)	Segmentation	90.31	88.76	—
Seg-Module + Gender-Module	Segmentation + classification (accuracy 92.83%)	91.89	89.69	—
Seg-Module + Seg-Tuner-Module	Segmentation + regression	95.03	93.97	0.0184
HSMTL-Net (full)	Segmentation + classification (accuracy 97.96%) + regression	97.83	96.21	0.00995

4.5. Comparison with Prior Methods

Table 3 compares HSMTL-Net with previously reported human-segmentation methods evaluated on the Supervise.ly person data. For transparency, the baseline figures in Table 3 are reported as published in the respective original studies. Because those studies may differ in input resolution, data augmentation, and training regime—for example, Chuang and Dong (2020) employ an online-distillation setting

rather than the offline training used here—these comparisons should be regarded as indicative of competitive performance rather than as strictly controlled, like-for-like benchmarks. Under this caveat, HSMTL-Net attains the highest mean IoU among the compared methods (96.21%) together with a Dice score of 97.83%. A fully controlled re-implementation of all baselines under an identical protocol is identified as future work (Section 5.2).

Table 3. Comparison of HSMTL-Net with prior human-segmentation methods on the Supervise.ly Filtered Segmentation Person Dataset

Model	Dice %	mIoU %	Source
Chuang and Dong	—	89.50	As published (online distillation)
Marin et al.	—	90.0	As published
Jiang et al.	—	92.2	As published
Chen, Qi and Shen -BANet	—	95.2	As published
HSMTL-Net (proposed)	97.83	96.21	This work (256×256, offline)

4.6. Robustness Analysis

The experimental results reported in this study are based on the evaluation protocol described in Sections 4.1 and 4.2. Although the proposed framework demonstrated consistent performance under the adopted experimental setting, a comprehensive multi-seed statistical analysis was beyond the scope of the present work. Future studies will investigate robustness through repeated independent training runs and statistical significance analysis across multiple random initializations.

4.7. Qualitative Results

Figure 3 compares qualitative outputs. Relative to the single-task baseline, HSMTL-Net produces more accurate masks and better preserves fine structures such as hair strands, fingers, and body boundaries, consistent with the quantitative gains attributed to the geometric regression head.



Figure 3. Qualitative comparison of segmentation results: (a) input image, (b) modified U-Net baseline, (c) segmentation + classification, (d) segmentation + regression, (e) the proposed HSMTL-Net, and (f) ground-truth mask

5. DISCUSSION, LIMITATIONS, AND ETHICAL CONSIDERATIONS

5.1. DISCUSSION

The experiments indicate that auxiliary semantic and geometric supervision improves human segmentation over a single-task baseline within a single

lightweight network. The geometric regression head contributes the largest share of the boundary-quality improvement, consistent with the hypothesis that mask-derived descriptors encode shape and extent information that is directly relevant to boundary localization. The classification head contributes a smaller, complementary gain. Together, these results support multi-task learning as an effective strategy for fine-grained human segmentation that combines pixel-level, semantic, and geometric supervision.

5.2. Limitations

Several limitations should be acknowledged. First, the framework was evaluated on a single dataset. Although the dataset is diverse, evaluation on additional public benchmarks (for example, PASCAL-Person-Part, COCO-Person, ATR, or portrait datasets) is required to establish generalization across domains, lighting conditions, camera settings, and population characteristics; such external validation was beyond the scope of the present study, and the generalization claims are therefore limited to the Supervise.ly distribution. Second, the comparison in Section 4.5 relies on baseline figures reported in the original publications, which may differ in experimental setup; a strictly controlled re-implementation of all baselines is left for future work. Third, the auxiliary gender task depends on label reliability and raises fairness considerations, discussed in Section 5.3. Fourth, although the architecture is lightweight relative to several conventional segmentation networks, deployment on low-resource edge devices would require further optimization, and a detailed computational-complexity profile (parameter count, FLOPs, and latency) is a useful direction for future work.

5.3. Ethical and Fairness Considerations

The inclusion of gender classification as an auxiliary task raises important privacy, consent, and fairness considerations. In the proposed framework, the gender-classification head is incorporated solely as an auxiliary supervision mechanism during training to improve shared feature representation within the multi-task learning framework. It is not intended for biometric identification, demographic profiling, or any deployment requiring personal attribute inference, and it can be removed without affecting the core segmentation architecture.

The auxiliary task employs coarse binary gender labels available in the dataset, which do not capture the full diversity of gender identity. This limitation reflects the characteristics of the available dataset rather than an endorsement of binary gender categorization. No new personal data were collected or annotated as part of this study.

As demonstrated by the ablation study (Section 4.4), the geometric regression head provides the majority of the segmentation improvement, while the gender-classification head contributes only a complementary regularization effect. Therefore, applications with privacy or fairness concerns may omit the gender-classification head with only a limited impact on segmentation performance.

A subgroup-specific fairness analysis was not performed because disaggregated demographic performance statistics were unavailable within the scope of the present study. Future work should investigate potential performance disparities across demographic groups using appropriately annotated datasets and dedicated fairness evaluation protocols, while also exploring alternative auxiliary tasks that avoid demographic attributes altogether.

6. CONCLUSION

This paper presented HSMTL-Net, an attention-guided, boundary-aware multi-task framework that couples human segmentation with mask-derived geometric regression and an optional gender-classification regularizer in a single lightweight network. The task-wise ablation shows that the auxiliary tasks improve segmentation over the single-task baseline, with the geometric regression head providing the largest gain in boundary quality. The work also contributes an explicitly described multi-task weighting strategy, a clearly documented implementation protocol, and an explicit discussion of the fairness implications associated with the auxiliary classification task. Future work will pursue controlled baseline re-implementations, cross-dataset evaluation, a detailed computational-complexity analysis, transformer-based feature extraction, and edge-oriented optimization for real-time deployment.

Data and Code Availability

The Supervise.ly Filtered Segmentation Person Dataset used in this study is publicly available from

its original source. The methodological details of the proposed HSMTL-Net framework, including the network architecture, training procedure, and evaluation protocol, are described in the manuscript. Additional information regarding the implementation may be obtained from the corresponding author upon reasonable academic request.

Acknowledgements

Not applicable.

REFERENCES

- Ahmed, I., Ahmad, A., Piccialli, F. et al. (2017). *A robust features-based person tracker for overhead views in industrial environment*. IEEE Internet of Things Journal, 4(5), 1598–1605.
- Ahmed, I., Ahmad, M., F. A. Khan, and M. Asif. (2020). *Comparison of deep-learning-based segmentation models: Using top view person images*. IEEE Access, 8, 136361–136373. doi:10.1109/ACCESS.2020.3011406.
- Ahmed, I., Ahmad, M. et al. (2019). *Person detector for different overhead views using machine learning*. International Journal of Machine Learning and Cybernetics, 10, 2657–2668.
- Bridgeman, L., Volino, M., Guillemaut, J.-Y., and Hilton, A. (2019). *Multi-person 3D pose estimation and tracking in sports*. In CVPR Workshops.
- Buch, S., Escorcia, V., Shen, C., Ghanem, B., and Niebles, J. C. (2017). *SST: Single-stream temporal action proposals*. In CVPR, 2911–2920.
- Cai, Z., Neher, H., Vats, K., Clausi, D. A., and Zelek, J. (2019). *Temporal hockey action recognition via pose and optical flows*. In CVPR Workshops.
- Caruana, R. (1997). *Multitask learning*. Machine Learning, 28(1), 41–75.
- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018). *DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs*. IEEE TPAMI, 40(4), 834–848.
- Chen, X., Qi, D., and Shen, J. (2019). *Boundary-aware network for fast and high-accuracy portrait segmentation*. arXiv:1901.03814.
- Cristani, M., Raghavendra, R., Del Bue, A., and Murino, V. (2013). *Human behavior analysis in video surveillance: A social signal processing perspective*. Neurocomputing, 100, 86–97.
- Hariharan, B., Arbeláez, P., Girshick, R., and Malik, J. (2015). *Hypercolumns for object segmentation and fine-grained localization*. In CVPR, 447–456.
- Hossain, M. D., and Chen, D. (2019). *Segmentation for object-based image analysis (OBIA): A review of algorithms and challenges from remote sensing perspective*. ISPRS Journal of Photogrammetry and Remote Sensing, 150, 115–134.
- Jug, J., Lampe, A., Štruc, V., and Peer, P. (2022). *Body segmentation using multi-task learning*. In ICAIIC, 60–68.
- Kamble, P. R., Keskar, A. G., and Bhurchandi, K. M. (2019). *A deep learning ball tracking system in soccer videos*. Opto-Electronics Review, 27(1), 58–69.
- Kaymak, Ç., and Uçar, A. (2019). *A brief survey and an application of semantic image segmentation for autonomous driving*. In Handbook of Deep Learning Applications.
- Liang, X., Gong, K., Shen, X., and Lin, L. (2019). *Look into Person: Joint body parsing and pose estimation network and a new benchmark*. IEEE TPAMI, 41(4), 871–885.
- Long, J., Shelhamer, E., and Darrell, T. (2015). *Fully convolutional networks for semantic segmentation*. In CVPR, 3431–3440.
- Moravvej, S. V., Alizadehsani, R., Khanam, S., Sobhaninia, Z., Shoeibi, A., Khozeimeh, F., Sani, Z. A., Tan, R.-S., Khosravi, A., Nahavandi, S., et al. (2022). *RLMD-PA: A reinforcement learning-based myocarditis diagnosis combined with a population-based algorithm for pretraining weights*. Contrast Media & Molecular Imaging.
- Paszke, A., Chaurasia, A., Kim, S., and Culurciello, E. (2016). *ENet: A deep neural network architecture for real-time semantic segmentation*. arXiv:1606.02147.
- Rohani, F., Khoshhal Roudposhti, K., Taheri, H., Mashhadi, A., and Mueller, A. (2022). *Exploring effective features in ADHD diagnosis among children through EEG/evoked potentials using machine learning techniques*. Computer and Knowledge Engineering, 5(2), 1–10.

- Rohani, F., Khoshhal Roudposhti, K., Taheri, H., Mashhadi, A., and Mueller, A. (2024). *Extracting gait and balance pattern features from skeleton data to diagnose attention deficit/hyperactivity disorder in children*. The Journal of Supercomputing, 80(6), 8330–8356.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). *U-Net: Convolutional networks for biomedical image segmentation*. In MICCAI, 234–241.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). *ImageNet Large Scale Visual Recognition Challenge*. International Journal of Computer Vision, 115(3), 211–252.
- Ruder, S. (2017). *An overview of multi-task learning in deep neural networks*. arXiv:1706.05098.
- Saito, S., Li, T., and Li, H. (2016). *Real-time facial segmentation and performance capture from RGB input*. In ECCV 2016, 244–261.
- Shen, X., Hertzmann, A., Jia, J., Paris, S., Price, B., Shechtman, E., and Sachs, I. (2016). *Automatic portrait segmentation for image stylization*. Computer Graphics Forum, 35(2), 93–102.
- Sobhaninia, Z., Emami, A., Karimi, N., and Samavi, S. (2020). *Localization of fetal head in ultrasound images by multiscale view and deep neural networks*. In CSICC, 1–5.
- Sobhaninia, Z., Karimi, N., Khadivi, P., and Samavi, S. (2022). *Medial residual encoder layers for classification of brain tumors in magnetic resonance images*. In ICEE, 442–445.
- Sobhaninia, Z., Karimi, N., Khadivi, P., and Samavi, S. (2023). *Brain tumor segmentation by cascaded multiscale multitask learning framework based on feature aggregation*. Biomedical Signal Processing and Control, 85, 104834.
- Song, C., Huang, Y., Wang, Z., and Wang, L. (2015). *1000fps human segmentation with deep convolutional neural networks*. In ACPR, 474–478.
- Supervisely. (2018). *Supervisely Persons Dataset*. Dataset.
- Ullah, M., and Alaya Cheikh, F. (2018). *A directed sparse graphical model for multi-target tracking*. In CVPR Workshops, 1816–1823.
- Yu, R., Tian, C., Xia, W., Zhao, X., Wang, L., and Yang, Y. (2022). *Real-time human-centric segmentation for complex video scenes*. Image and Vision Computing, 126, 104552.
- Zaitoun, N. M., and Aqel, M. J. (2015). *Survey on image segmentation techniques*. Procedia Computer Science, 65, 797–806.
- Zhang, Y., and Yang, Q. (2022). *A survey on multi-task learning*. IEEE Transactions on Knowledge and Data Engineering, 34(12), 5586–5609.
- Zhao, T., and Nevatia, R. (2003). *Bayesian human segmentation in crowded situations*. In CVPR, II-459.



Publisher's note: Eurasia Academic Publishing Group (EAPG) and AAKS remain neutral about jurisdictional claims in published maps and institutional affiliations.

Open Access. This article is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0) licence, which permits copy and redistribute the material in any medium or format for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the licence terms. Under the following terms you must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorsed you or your use. If you remix, transform, or build upon the material, you may not distribute the modified material. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nd/4.0/>.