

FingerSeg: Highly-Efficient Dual-Resolution Architecture for Precise Finger-Level Semantic Segmentation

Gibran Benitez-Garcia¹ ^a and Hiroki Takahashi^{1,2,3}

¹Graduate School of Informatics and Engineering, The University of Electro-Communications, Japan

²Artificial Intelligence eXploration Research Center (AIX), The University of Electro-Communications, Japan

³Meta-Networking Research Center (MEET), The University of Electro-Communications, Japan

Keywords: Semantic Segmentation, Finger Segmentation, DDRNet, Real-Time CNN, IPN-Hand Dataset.

Abstract: Semantic segmentation at the finger level poses unique challenges, including the limited pixel representation of some classes and the complex interdependency of the hand anatomy. In this paper, we propose FingerSeg, a novel architecture inspired by Deep Dual-Resolution Networks, specifically adapted to address the nuances of finger-level hand semantic segmentation. To this end, we introduce three modules: Enhanced Bilateral Fusion (EBF), which refines low- and high-resolution feature fusion via attention mechanisms; Multi-Attention Module (MAM), designed to augment high-level features with a composite of channel, spatial, orientational, and categorical attention; and Asymmetric Dilated Up-sampling (ADU), which combines standard and asymmetric atrous convolutions to capture rich contextual information for pixel-level classification. To properly evaluate our proposal, we introduce IPN-finger, a subset of the IPN-Hand dataset, manually annotated pixel-wise for 13 finger-related classes. Our extensive empirical analysis, including evaluations of the synthetic RHD dataset against current state-of-the-art methods, demonstrates that our proposal achieves top results. FingerSeg reaches 73.8 and 71.1 mIoU on the IPN-Finger and RHD datasets, respectively, while maintaining an efficient computational cost of about 7 GFLOPs and 6 million parameters at VGA resolution. The dataset, source code, and a demo of FingerSeg will be available upon the publication of this paper.

1 INTRODUCTION

Hand segmentation represents a dense prediction problem, focused on identifying each pixel associated with a hand in binary segmentation frameworks (Urooj and Borji, 2018), and in more advanced applications, distinguishing between left and right hands (Bandini and Zariffa, 2020). This segmentation is often a preliminary step in diverse tasks, ranging from hand gesture recognition (HGR) to human behavior analysis (Dadashzadeh et al., 2019; Likitlersuang et al., 2019; Benitez-Garcia et al., 2021b). While effective for broad categorization, conventional left and right hand detection falls short in applications requiring finer granularity. Specifically, finger-level segmentation may provide a clearer distinction in gestures, particularly those involving ambiguity between the number of fingers involved. This level of detail is crucial for accurately distinguishing between nuanced gestures, as illustrated in Figure 1, where the standard

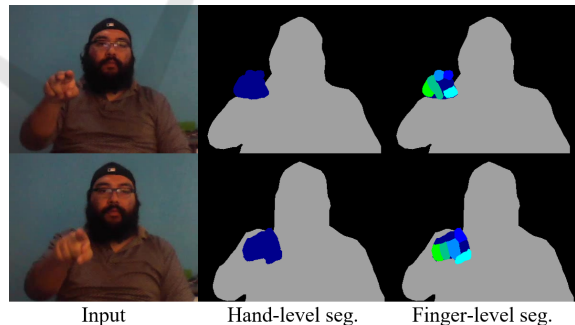



Figure 1: Comparison between gestures based on one and two fingers. The finger segmentation granularity helps to determine the number of fingers involved.

hand segmentation can help localize the hand but fails to determine the number of fingers shown in the gesture. Therefore, in this paper, we propose FingerSeg, a novel segmentation architecture explicitly designed for finger-level hand semantic segmentation.

FingerSeg is inspired by the principles of Deep Dual-Resolution Networks (DDRNet (Pan et al.,

^a  <https://orcid.org/0000-0003-4945-8314>

2022)) and tailored to overcome the unique challenges posed by finger segmentation, such as the limited pixel area of certain fingers and the complex spatial relationships within hand anatomy. FingerSeg introduces advanced feature fusion and attention mechanisms to capture the subtle distinctions between fingers. This enhanced granularity in segmentation is not only crucial for applications in HGR and sign language recognition but also holds significant promise in medical fields, where precise hand and finger movements are essential (Rangesh and Trivedi, 2018; Likitlersuang et al., 2019).

Efficiency in computation is essential when integrating finger segmentation into real-time applications. The demand for lightweight models to ensure prompt and responsive performance is paramount. In this way, DDRNet offers an efficient architecture optimized for speed and accuracy. However, it does not inherently address the mentioned challenges of finger segmentation. To bridge this gap, we propose three novel modules that do not impose excessive computational costs. Our proposal includes the Enhanced Bilateral Fusion (EBF), which improves feature merging precision; the Multi-Attention Module (MAM), which provides nuanced processing of features; and the Asymmetric Dilated Up-sampling (ADU), which enhances feature resolution effectively. Each module is carefully designed to contribute to the overall efficacy, ensuring that FingerSeg remains highly efficient while setting new standards in segmentation detail and accuracy.

To thoroughly evaluate the capabilities of FingerSeg, we introduce IPN-finger, a curated subset of the IPN-Hand dataset for HGR (Benitez-Garcia et al., 2021a). Specifically, we defined 13 finger-related classes and manually annotated 1000 frames at pixel level. Moreover, our evaluation extends to the Rendered Hand Pose Dataset (RHD) (Zimmermann and Brox, 2017), which, to the best of our knowledge, is the only publicly available dataset offering finger-level pixel-wise annotations. Thus, including real-world samples of the IPN-finger dataset serves as a critical benchmark to validate the precision and effectiveness of finger segmentation approaches.

In our experimental analysis, FingerSeg not only demonstrates an exceptional balance between accuracy and computational efficiency but also presents a significant improvement over our baseline model. Specifically, FingerSeg overcomes DDRNet by approximately 3%, achieving a mIoU of 73.8 on the IPN-Finger dataset. FingerSeg demands only about 7 GFLOPs and 6 million parameters for a 640x480 image resolution, which allows faster than real-time performance across different platforms. Consequently,

FingerSeg presents a valuable solution for finger segmentation, serving as an efficient preprocessing step and a robust framework for real-time applications.

The main contributions are summarized as follows:

- Introduction of FingerSeg, an architecture specifically designed for efficient and accurate finger-level hand segmentation.
- Development and integration of three novel modules: Enhanced Bilateral Fusion (EBF), Multi-Attention Module (MAM), and Asymmetric Dilated Up-sampling (ADU), to achieve state-of-the-art results with minimal computational cost increase.
- Compilation of the IPN-Finger dataset, comprising 1000 frames with pixel-wise annotations across 13 classes, including the palm, all fingers of each hand, and the overall shape of the person.
- Superior performance of FingerSeg with results of approximately 73.8 and 71.1 mIoU on the IPN-Finger and RHD datasets, respectively, surpassing real-time semantic segmentation approaches, including notable methods like DDRNet (Pan et al., 2022) and PIDNet (Xu et al., 2023).

2 RELATED WORK

2.1 Hand Segmentation

Hand segmentation has been an active research topic in recent years, with significant implications for diverse applications. Notable among these are hand gesture recognition (HGR), RGB-based hand pose estimation, and the analysis of egocentric interactions (Bandini and Zariffa, 2020).

Segmentation serves as a preprocessing step for HGR, enhancing subsequent processes like classification. Studies such as (Dadashzadeh et al., 2019) and (Benitez-Garcia et al., 2021b) have illustrated how effective segmentation can boost the accuracy of HGR systems. Binary segmentation, in particular, has been a staple in preprocessing for hand pose estimation. For instance, HandSegNet (Zimmermann and Brox, 2017) underpins the 3D hand pose estimation from RGB frames. Likewise, the end-to-end trainable framework proposed in (Baek et al., 2019) utilizes segmentation masks to facilitate 3D hand pose reconstructions from 2D joint estimations.

In the context of egocentric vision, robust hand segmentation has proven critical for action and activity recognition involving hands. This has led to the development of methods that rely on accurate

hand segmentation as a precursor to activity recognition (Li et al., 2019). A significant advancement in binary hand segmentation is presented in (Cai et al., 2020), which showcases a Bayesian CNN framework enhancing generalizability across diverse domains.

However, a noticeable gap in the existing literature is the lack of focus on hand segmentation with the granularity necessary to distinguish individual fingers. This paper seeks to fill that gap by introducing an approach specifically designed for this purpose, extending the scope and applicability of hand segmentation in computational vision.

2.2 Real-Time Semantic Segmentation

Advanced semantic segmentation techniques typically depend on preserving high-resolution features while implementing convolutions with extensive dilation rates to broaden receptive fields, as seen in methods like PSPNet (Zhao et al., 2017) and DeepLabV3+ (Chen et al., 2018). Despite their high accuracy, the computational intensity and complex pooling mechanisms of these methods often prohibit real-time performance.

In contrast, real-time segmentation algorithms consider more efficient architectures, such as lightweight encoder-decoder or bilateral pathway designs. These often incorporate compact pyramidal pooling and depth-wise convolutions. DABNet (Li and Kim, 2019), for example, leverages Depth-wise Asymmetric Bottleneck modules, which combine factorized depth-wise convolutions in a bottleneck structure to extract local and contextual information jointly, obviating the need for extensive pooling modules. Other approaches like HardNet (Chao et al., 2019) and FASSDNet (Rosas-Arias et al., 2021) use classic encoder-decoder architectures relying on Harmonic Dense Blocks. These blocks are engineered to reduce memory usage and computational density, addressing the challenges of the dense blocks proposed by DenseNet. On the other hand, a significant development in real-time semantic segmentation is DDRNet (Pan et al., 2022), which introduces a dual-resolution backbone comprising low- and high-resolution branches with a one-to-one correlation between paths. This includes bilateral connections to foster efficient information exchange between context (low-resolution) and detail (high-resolution) branches. More recently, PIDNet (Xu et al., 2023) advances the field with a three-branch network architecture. This design parses detailed, contextual, and boundary information through separate branches, utilizing boundary attention to guide the fusion of detailed and contextual information.

Our choice of DDRNet as a baseline derives from its efficient dual-resolution approach and the potential for enhancements in finger-level segmentation. This framework allows for an optimal balance between detail capture and computational efficiency, making it an ideal foundation for our FingerSeg model.

3 FINGERSEG NETWORK

The proposed architecture for finger-level hand segmentation is designed to address the specific challenges inherent in this task. Our methodology is inspired by the efficient architecture of DDRNet, which we have significantly adapted and enhanced to cater to the nuanced requirements of segmenting individual fingers. At the core of FingerSeg are three key modules: Enhanced Bilateral Fusion (EBF), Multi-Attention Module (MAM), and Asymmetric Dilated Up-sampling (ADU). Each of these modules plays a pivotal role in refining the segmentation process, ensuring both high accuracy and computational efficiency. In the following sections, we delve into the intricacies of these modules, explaining how they collectively contribute to the superior performance of FingerSeg in finger-level segmentation tasks.

3.1 Network Overview

FingerSeg’s architecture, as illustrated in Figure 2, is built upon a dual-resolution backbone that bifurcates from a single trunk into two parallel branches, each operating at a distinct resolution. The high-resolution branch aims to generate detailed feature maps at $1/8$ the resolution of the input image. Notably, this branch excludes any downsampling operations to preserve high-resolution information, maintaining a one-to-one correspondence with the low-resolution branch to form deep, detailed representations. Conversely, the low-resolution branch, akin to DDRNet’s design, employs multiple downsampling operations within its Residual Blocks (RB) to produce feature maps at a reduced $1/64$ resolution. This structure not only captures rich contextual information but also contributes to the network’s overall efficiency.

EBF blocks are employed to integrate detailed and contextual features at two critical points within the architecture. Additionally, the end of the low-resolution pathway incorporates the Deep Aggregation Pyramid Pooling Module (DAPPM), as utilized in DDRNet (Pan et al., 2022). The DAPPM enriches semantic information without compromising inference speed by processing lower-resolution feature maps. It operates on feature maps at a $1/64$ resolution, utiliz-

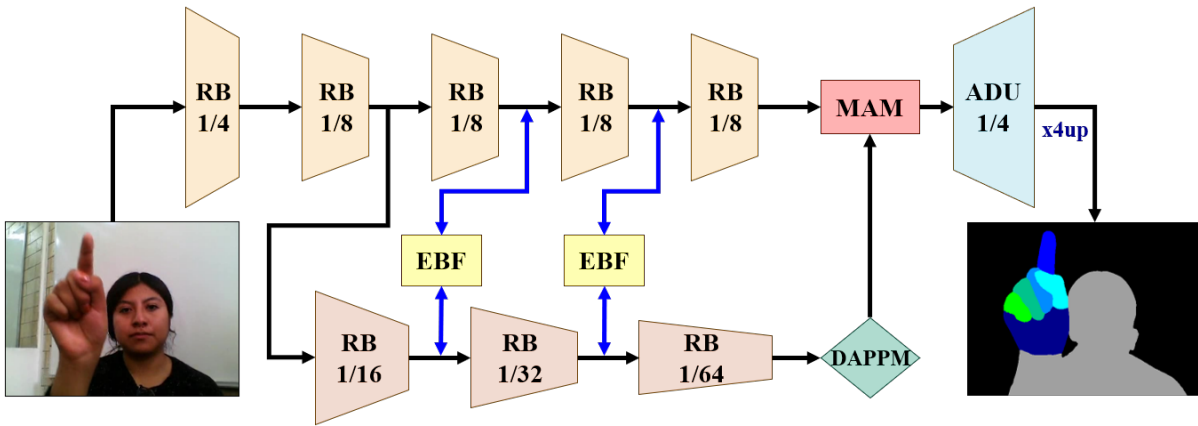


Figure 2: Overview of the FingerSeg architecture. The dual-resolution network starts as a single trunk and divides into two branches: the high-resolution branch (top) and the low-resolution branch (bottom). Enhanced Bilateral Fusion (EBF) blocks facilitate the integration of detailed and contextual features. The Deep Aggregation Pyramid Pooling Module (DAPPM) at the end of the low-resolution branch enriches semantic information. Feature maps are then refined by the Multi-Attention Module (MAM) and upsampled by the Asymmetric Dilated Up-sampling (ADU) module for final segmentation.

ing large pooling kernels to create multi-scale feature maps, and incorporates global average pooling for additional image-level detail. A cascading fusion strategy involving successive upsampling and 3x3 convolutions is applied to integrate varying scales of features, which are then unified and compacted using a 1x1 convolution.

Subsequently, the refined outputs from the DAPPM and the high-resolution branch converge within our MAM block before being upsampled by the ADU module, leading to the final segmentation prediction. The segmentation scores are subsequently upscaled using bilinear interpolation, aligning with the supervision of a standard cross-entropy loss function. The forthcoming subsections will delve into the specifics of the proposed modules.

3.2 EBF: Enhanced Bilateral Fusion

The EBF block, depicted in Figure 3, is designed to effectively merge high-resolution detail with low-resolution context features. High-to-low resolution features undergo a transformation involving channel expansion and spatial reduction through a 3x3 convolution with a stride of 2. Concurrently, low-to-high resolution features are channel-wise compacted using a 1x1 convolution and subsequently spatially enlarged via bilinear interpolation (x2up).

Once the feature maps from both resolutions are matched in spatial and channel dimensions, they are combined through element-wise addition. Building upon the concept introduced by CBAM (Woo et al., 2018), we further refine the fused features by applying channel and spatial attention mechanisms to each respective branch. This dual attention schema ensures

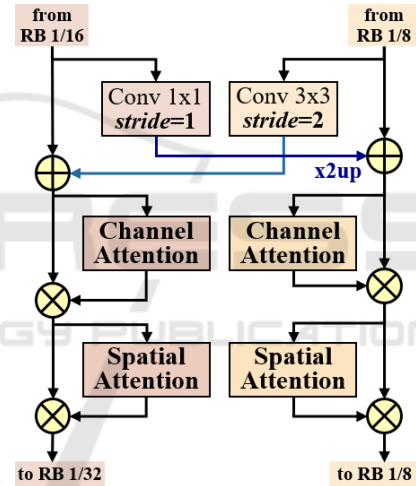


Figure 3: The structure of the Enhanced Bilateral Fusion (EBF) block.

that the most relevant features are emphasized, enhancing the quality of the subsequent feature representations for the precise segmentation tasks at hand.

3.3 MAM: Multi-Attention Module

The Multi-Attention Module (MAM), as shown in Figure 4, is crucial in refining the feature fusion process within FingerSeg. It starts by upscaling the output feature maps from the DAPPM using bilinear interpolation to match the spatial dimensions of the high-resolution feature maps. These aligned feature maps are then combined via element-wise summation. This fusion sets the stage for a series of specialized attention mechanisms, extending beyond the scope of CBAM to address the unique challenges of

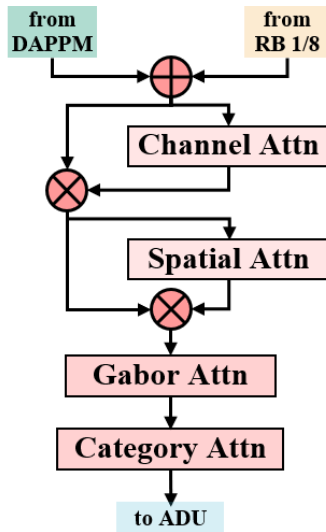


Figure 4: Configuration of the Multi-Attention Module (MAM).

finger segmentation.

We employ a Gabor Attention mechanism (Richards et al., 2022) after channel and spatial attention, specifically designed to be robust to the orientation of features. This mechanism utilizes Gabor-modulated convolutions, where convolutional weights are multiplied by Gabor filters across different rotation parameters to generate orientation-sensitive feature representations. The Gabor Attention then calculates correlations along the orientation axis, which are crucial for interpreting the diverse positioning of fingers.

Further, the Category Attention Block (CAB) (He et al., 2020) is integrated to address the challenges associated with the imbalanced data characteristic of finger segmentation, where pixel representation of different fingers varies significantly. The Category Attention of CAB operates in a class-specific manner, allocating an equal number of feature channels to each class. Thus, it mitigates channel bias and amplifies inter-class feature distinction, ensuring each category receives equal treatment.

MAM’s strategy, which includes channel, spatial, orientational, and categorical attention, fortifies the feature richness. This robustness is essential for FingerSeg’s capability to accurately segment fingers in varying poses and alignments, which is critical for the following ADU block and final segmentation process.

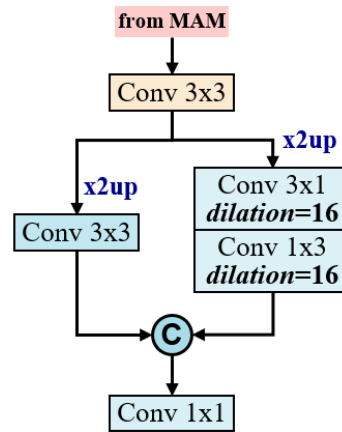


Figure 5: Schematic representation of the Asymmetric Dilated Up-sampling (ADU) block.

3.4 ADU: Asymmetric Dilated Up-Sampling

As shown in Figure 5, the process of ADU begins with a 3×3 convolution that further refines the feature maps received from MAM. After this initial refinement, two pathways are used to process the features in parallel. One branch utilizes asymmetric convolutions with dilation to effectively capture contextual information from the feature maps, which have been previously upsampled using bilinear interpolation. This branch’s dilated convolutions enable a broader receptive field, facilitating the assimilation of context without loss of resolution. Simultaneously, the second branch, employing a direct upsampling approach without dilation, concentrates on learning fine-grained details to enhance the spatial resolution of the features. The outputs of both branches are then concatenated, providing a composite feature map that embodies contextual and detailed attributes.

The concatenated feature maps undergo a final 1×1 convolution for the pixel-level classification. This fusion of asymmetric dilation and direct detail learning in the ADU block optimizes the balance between contextual understanding and detail preservation, a critical aspect for accurately segmenting fingers.

4 IPN-FINGER DATASET

The scarcity of hand datasets offering annotations beyond binary pixel labels presents a significant challenge in advancing hand segmentation research. For instance, EgoHands (Bambach et al., 2015) offers pixel-wise annotations for activity recognition across four classes, limited to distinguishing only between the user’s and others’ left and right hands. Simi-



Figure 6: Class labels in the IPN-Finger dataset.

larly, the dataset from (Benitez-Garcia et al., 2021b) provides annotations for 500 frames of left and right hands, targeted at touchless screen interactions. The WorkingHands dataset (Kim et al., 2020) is another substantial contribution, featuring over 400 thousand frames of thermally and RGB-D captured "hands using tools." However, these datasets do not address finger-level granularity. On the other hand, to the best of our knowledge, the Rendered Hand Pose Dataset (RHD) (Zimmermann and Brox, 2017) is the only publicly available dataset with finger-level pixel-wise annotations, comprising 43,986 synthetically generated images.

To fill the void of real-world, finger-specific annotations, we have extended the IPN-Hand dataset with finger-level semantic annotations. The IPN-Hand dataset, known for capturing genuine interactions with touchless screens, was the ideal candidate for this subset. We selected 1000 frames showcasing a range of finger positions, spanning the 13 static and dynamic gestures and the variety of the 28 scenes and backgrounds present in the dataset. This subset, dubbed IPN-Finger, represents samples from 50 different subjects and defines 13 classes that comprise the palms, all fingers on each hand, and the person's shape. Figure 6 provides a visual guide to these classes. The annotation was performed manually using the LabelMe toolbox (Russell et al., 2008), ensuring precise and comprehensive pixel-level labeling.

4.1 Dataset Statistics

The IPN-Finger dataset exhibits a considerable class imbalance in terms of the frequency of class appearances across images and the pixel area occupied by each class, as shown in Table 1. The table also highlights the pixel area covered by each class, given a standard image size of 640x480 pixels, where the area percentage reflects the proportion of the image that a class occupies. From this table, we can see that obviously, the 'person' class is the most prevalent, appearing in all 1000 images and covering 26.45% of the

Table 1: Distribution of classes across the dataset.

Class	Images	Area (%)	Area (pix.)
person	1000	26.45	285 ²
l_thumb	61	0.34	32 ²
l_palm	107	1.30	63 ²
l_index	77	0.34	32 ²
l_medium	90	0.39	35 ²
l_ring	97	0.37	34 ²
l_pinky	96	0.28	29 ²
r_thumb	729	0.55	41 ²
r_palm	853	1.59	70 ²
r_index	840	0.60	43 ²
r_medium	851	0.66	45 ²

image area on average. Conversely, the classes representing individual fingers occupy significantly less space, with most covering less than 1% of the image area. Notably, the left pinky class ('l.pinky') is the least represented in terms of area, averaging an area of approximately 29x29 pixels.

This imbalance extends to the visibility of hands within the images: the left hand is notably less present, appearing in fewer than 110 images. Such disparities underscore the challenges of finger segmentation, specifically in training robust models capable of accurately segmenting classes of varying pixel representations and frequencies. The stark contrast between the dataset's most and least represented classes accentuates the necessity of a model like FingerSeg, which is adept at handling the intricacies of finger-level segmentation within this uneven scenario.

5 EXPERIMENTAL RESULTS

5.1 Datasets

To assess the performance of our FingerSeg architecture and compare it with state-of-the-art methods, we evaluated two datasets: our IPN-Finger and the publicly available RHD dataset (Zimmermann and Brox, 2017).

For the IPN-Finger dataset, we randomly chose a fixed subset of 850 images for training and 150 for testing. To address the challenge of class representation imbalance across images, we expanded the dataset offline by mirroring each image, effectively doubling the number of images for training and testing to 1700 and 300, respectively. The RHD dataset's standard data split was employed, comprising 41,258 images for training and 2,728 for testing. It is important to note that the RHD dataset's original annotations span 31 classes, reflecting individual finger pha-

lages. We consolidated these annotations into the same 13 classes defined for our evaluation.

5.2 Implementation Details

All experiments were conducted using Python 3.7.16 and PyTorch 1.10.2, with CUDA 12.0 acceleration, on an Intel Core i7-9700K desktop paired with an Nvidia RTX 2080Ti GPU. To ensure a fair comparison among all evaluated models, we standardized the training settings across the board. The Stochastic Gradient Descent (SGD) algorithm was employed as the optimizer for all models, in conjunction with a cross-entropy loss function as suggested by the online bootstrapping strategy (Wu et al., 2016).

Data augmentation techniques were uniformly applied to each dataset, including random horizontal flips, random scaling, and random cropping to 480x480 size. Before augmentation, images from both datasets were upscaled to a uniform resolution of 640x640 pixels. However, for testing, images were evaluated at their native 640x480 resolution without cropping. Each model was trained with a batch size of 32 images. Specifically for the RHD dataset, models were trained from scratch over 60,000 iterations. In contrast, for the IPN-Finger dataset, we fine-tuned the models pre-trained on RHD for an additional 35,000 iterations.

5.3 Ablation Study

The efficacy of segmentation models is traditionally measured by the mean intersection-over-union accuracy (mIoU). Alongside mIoU, we also provide insights into the model complexity by reporting the number of parameters and computational cost measured in GFLOPs.

This ablation study dissects the incremental contributions of the proposed modules integrated into the DDRNet architecture. Our baseline is the DDRNet23 slim variant (Pan et al., 2022), which comprises 5.73 million parameters with a computational cost of 5.55 GFLOPs. We systematically enhance this baseline by sequentially incorporating our proposed modules. The results of this study are shown in Table 2.

The addition of each module demonstrates a significant improvement in mIoU, as illustrated in Table 2. The full implementation of our FingerSeg model, which includes all three modules, shows a mIoU of 73.79, outperforming the baseline by a significant margin. It is noteworthy that FingerSeg achieves these results with only a moderate increase in parameters and GFLOPs.

We also compare FingerSeg against the more

Table 2: Ablation study showing the enhancements of the proposed modules compared with the baseline (DDRNet23_slim), including the complex DDRNet23_full for reference. Results on the IPN-Finger dataset.

Method	Params	GFLOPs	mIoU
DDRNet23_slim	5.73M	5.55	70.70
+EBF	5.84M	5.56	71.21
+EBF+MAM	6.15M	6.31	72.52
+EBF+MAM+ADU (FingerSeg)	6.20M	7.30	73.79
DDRNet23_full	20.30M	21.79	74.37

complex DDRNet23 full model for a broader perspective. Although the full model has nearly triple the parameters and computational complexity of FingerSeg, the improvement in mIoU is marginal. Specifically, the DDRNet23 full model registers a mIoU of 74.37, a modest increment over FingerSeg’s 73.79 mIoU. FingerSeg’s design strategy demonstrates that strategic module enhancements can yield near-comparable accuracy while significantly reducing the computational burden of more complex models. This underscores the effectiveness of each integrated module in FingerSeg and emphasizes the model’s capacity to achieve high-level accuracy with a more efficient use of resources.

5.4 Per-Class Analysis

The task of finger segmentation poses varying degrees of difficulty across different classes, mainly due to the size and frequency of occurrence of each finger within the dataset. Our analysis, detailed in Table 3, suggests that the smaller fingers, particularly the left pinky—the least represented class—pose the most significant challenge.

When examining the performance metrics, it is evident that our FingerSeg model outperforms the baseline DDRNet23 slim, especially in the classes where size and representation pose a challenge. Notably, FingerSeg achieves a significant improvement in the segmentation of the left palm, with an increase in accuracy from 68.6 to 73.1 mIoU, indicating a prominent enhancement in distinguishing this particular region compared to the baseline. On the other hand, the left pinky, being the smallest and least represented class, shows no critical increase from the baseline performance. Another point of interest is the anomalously low accuracy for the right palm in DDRNet’s results, which FingerSeg effectively addresses, improving accuracy by over 20 percentage points.

Figure 7 offers visual insights into the performance improvements by illustrating qualitative comparisons between FingerSeg and the baseline DDR-

Table 3: FingerSeg per-class mIoU comparison with the baseline on the IPN-Finger dataset.

Method	l_thu	l_pal	l_ind	l_med	l_rin	l_pin	r_thu	r_pal	r_ind	r_med	r_rin	r_pin
DDRNet	69.8	68.6	67.6	71.2	68.3	67.4	69.4	53.7	68.2	70.4	67.3	67.9
FingerSeg	69.9	73.1	71.2	69.4	67.8	67.5	69.8	75.6	70.8	71.4	67.7	68.3

Table 4: Comparative analysis of FingerSeg and state-of-the-art semantic segmentation methods on the IPN and RHD datasets.

Method	Params (M)	GFLOPs	mIoU (IPN)	mIoU (RHD)
FastSCNN (Poudel et al., 2019)	1.134	1.03	64.66	55.05
DABNet (Li and Kim, 2019)	0.755	6.12	64.26	61.22
FC-HardNet (Chao et al., 2019)	4.119	5.19	68.93	68.59
FASSDNet (Rosas-Arias et al., 2021)	2.845	6.60	69.58	68.50
DDRNet23_slim (Pan et al., 2022)	5.734	5.55	70.70	69.03
PIDNet (Xu et al., 2023)	7.625	7.20	71.98	70.91
FingerSeg	6.196	7.30	73.79	71.15

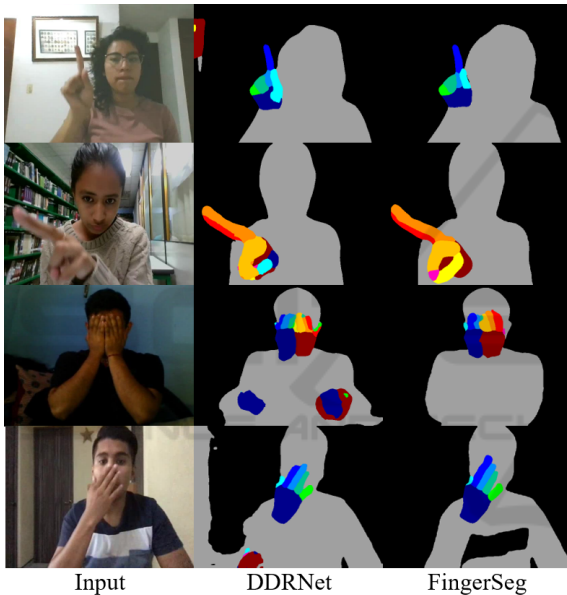


Figure 7: Qualitative comparison of DDRNet (baseline) vs. FingerSeg.

Net on test images from the IPN-finger dataset. The third row particularly tells that DDRNet erroneously classifies skin regions as part of the right and left palms, a significant error that underscores difficulties in discerning hand anatomy. Moreover, the baseline occasionally misclassifies objects and background elements with hands, suggesting limitations in its dual-resolution architecture when learning fine anatomical structures. In stark contrast, FingerSeg exhibits no such classification errors, reinforcing the notion that our integrated modules considerably enhance the model’s ability to segment and differentiate hand and finger regions accurately.

In general, FingerSeg consistently maintains or improves upon the baseline accuracies, affirming the

effectiveness of the integrated modules tailored to address the intricacies of finger segmentation. This per-class improvement demonstrates that FingerSeg is well-suited for the detailed task at hand, capable of precisely discerning between closely situated and similarly sized classes.

5.5 Comparisons with State-of-the-Art Approaches

The comparative analysis shown in Table 4 showcases the performance of FingerSeg against several leading semantic segmentation models: FastSCNN (Poudel et al., 2019), DABNet (Li and Kim, 2019), FC-HardNet (Chao et al., 2019), FASSDNet (Rosas-Arias et al., 2021), DDRNet (Pan et al., 2022), and PIDNet (Xu et al., 2023). Notably, FingerSeg achieves the highest mIoU scores on both the IPN and RHD datasets, reaching 73.79 and 71.15, respectively. This marks a substantial improvement over other approaches while maintaining competitive computational efficiency.

FastSCNN, while being the most computationally lightweight model with only 1.03 GFLOPs, falls short in mIoU performance. DABNet offers the lowest parameter count at 0.755 million, yet its mIoU scores do not compete with FingerSeg, emphasizing the latter’s superior balance of model complexity and segmentation capability. FC-HardNet and FASSDNet present themselves as intermediate options in terms of parameters and GFLOPs. While these models offer competitive mIoU scores, particularly FC-HardNet’s performance on the RHD dataset, they still do not reach the benchmark set by FingerSeg.

Our baseline model, DDRNet, demonstrates robust performance with a mIoU of 70.70 on IPN and 69.03 on RHD. However, PIDNet, one of the

most recent models, stands out with high mIoU scores of 71.98 and 70.91 on the IPN and RHD datasets, respectively. Yet, FingerSeg overcomes PIDNet in accuracy while requiring fewer parameters. The GFLOPs of FingerSeg and PIDNet are closely matched, underscoring FingerSeg’s architectural optimizations that allow for high accuracy without a substantial increase in computational demand.

In summary, FingerSeg sets new standards in segmentation accuracy and exhibits a notorious balance between computational requirements and model complexity. This performance is particularly important given the fine-grained nature of the finger segmentation task, proving the worth of FingerSeg’s design.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we introduced FingerSeg as an advanced solution for finger-level hand segmentation. Through meticulous design and the integration of specialized modules (EBF, MAM, and ADU), FingerSeg has demonstrated a significant leap forward in the accuracy and efficiency of semantic segmentation for nuanced hand gestures. The empirical results, bolstered by thorough ablation studies and comparisons with state-of-the-art methods, affirm FingerSeg’s standing as a leading solution to the presented task. Moreover, the creation and annotation of the IPN-Finger dataset have not only facilitated the development of FingerSeg but also enriched the resources available to the research community. By offering this dataset publicly, alongside the FingerSeg model, we anticipate stimulating further innovation and exploration in the detailed segmentation of hands and fingers.

Looking ahead, the integration of FingerSeg into multimodal hand gesture recognition (HGR) systems presents promising future work. Its application as an additional modality can potentially enrich the interpretative capabilities of HGR, particularly in complex or nuanced scenarios. Exploring the synergy between FingerSeg’s detailed segmentation and other modalities will be instrumental in developing more intuitive and natural user interfaces, contributing significantly to advancements in human-computer interaction.

ACKNOWLEDGEMENTS

This work is supported by a Research Grant (S) at Tateisi Science and Technology Foundation.

REFERENCES

- Baek, S., Kim, K. I., and Kim, T.-K. (2019). Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1067–1076.
- Bambach, S., Lee, S., Crandall, D. J., and Yu, C. (2015). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1949–1957.
- Bandini, A. and Zariffa, J. (2020). Analysis of the hands in egocentric vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*.
- Benitez-Garcia, G., Olivares-Mercado, J., Sanchez-Perez, G., and Yanai, K. (2021a). Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In *2020 25th international conference on pattern recognition (ICPR)*, pages 4340–4347. IEEE.
- Benitez-Garcia, G., Prudente-Tixteco, L., Castro-Madrid, L. C., Toscano-Medina, R., Olivares-Mercado, J., Sanchez-Perez, G., and Villalba, L. J. G. (2021b). Improving real-time hand gesture recognition with semantic segmentation. *Sensors*, 21(2):356.
- Cai, M., Lu, F., and Sato, Y. (2020). Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14392–14401.
- Chao, P., Kao, C.-Y., Ruan, Y.-S., Huang, C.-H., and Lin, Y.-L. (2019). HarDNet: A Low Memory Traffic Network. In *The IEEE International Conference on Computer Vision (ICCV)*.
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *The European Conference on Computer Vision (ECCV)*, pages 801–818.
- Dadashzadeh, A., Targhi, A. T., Tahmasbi, M., and Mirme-hdi, M. (2019). Hgr-net: a fusion network for hand gesture segmentation and recognition. *IET Computer Vision*, 13(8):700–707.
- He, A., Li, T., Li, N., Wang, K., and Fu, H. (2020). Cabnet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 40(1):143–153.
- Kim, S., Chi, H.-g., Hu, X., Vegesana, A., and Ramani, K. (2020). First-person view hand segmentation of multimodal hand activity video dataset. In *British Machine Vision Conference (BMVC)*.
- Li, G. and Kim, J. (2019). DABNet: Depth-wise Asymmetric Bottleneck for Real-time Semantic Segmentation. In *British Machine Vision Conference (BMVC)*.
- Li, M., Sun, L., and Huo, Q. (2019). Flow-guided feature propagation with occlusion aware detail enhancement for hand segmentation in egocentric videos. *Computer Vision and Image Understanding*, 187:102785.
- Likitlersuang, J., Sumitro, E. R., Cao, T., Visée, R. J., Kalsi-Ryan, S., and Zariffa, J. (2019). Egocentric video: a

- new tool for capturing hand use of individuals with spinal cord injury at home. *Journal of neuroengineering and rehabilitation*, 16(1):1–11.
- Pan, H., Hong, Y., Sun, W., and Jia, Y. (2022). Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes. *IEEE Transactions on Intelligent Transportation Systems*, 24(3):3448–3460.
- Poudel, R. P., Liwicki, S., and Cipolla, R. (2019). Fast-scnn: fast semantic segmentation network. In *British Machine Vision Conference (BMVC)*.
- Rangesh, A. and Trivedi, M. M. (2018). Handynet: A one-stop solution to detect, segment, localize & analyze driver hands. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1103–1110.
- Richards, F., Xie, X., Paiement, A., Sola, E., and Duc, P.-A. (2022). Multi-scale gridded gabor attention for cirrus segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3733–3737. IEEE.
- Rosas-Arias, L., Benitez-Garcia, G., Portillo-Portillo, J., Olivares-Mercado, J., Sanchez-Perez, G., and Yanai, K. (2021). Fassed-net: Fast and accurate real-time semantic segmentation for embedded systems. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):14349–14360.
- Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2008). Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173.
- Urooj, A. and Borji, A. (2018). Analysis of hand segmentation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4710–4719.
- Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.
- Wu, Z., Shen, C., and Hengel, A. v. d. (2016). High-performance semantic segmentation using very deep fully convolutional networks. *arXiv preprint arXiv:1604.04339*.
- Xu, J., Xiong, Z., and Bhattacharyya, S. P. (2023). Pidnet: A real-time semantic segmentation network inspired by pid controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19529–19539.
- Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid Scene Parsing Network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zimmermann, C. and Brox, T. (2017). Learning to estimate 3d hand pose from single rgb images. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 4903–4911.