# A Comparison of Modern Deep Neural Networks Architectures for Cross-section Segmentation in Images of Log Ends

Felipe Alfredo Nack, Marcelo Ricardo Stemmer, and Maurício Edgar Stivanello

***Abstract*—The semantic segmentation of log faces constitutes the initial step towards subsequent quality analyses of timber, such as quantifying properties like mechanical strength, durability, and the aesthetic attributes of growth rings. In the literature, works based on both classical and machine learning approaches for this purpose can be found. However, more recent architectures and techniques, such as ViTs or even the latest CNNs, have not yet been thoroughly evaluated. This study presents a comparison of modern deep neural network architectures for cross-section segmentation in images of log ends. The results obtained indicate that the networks using the ViTs considered in this work outperformed those previously evaluated in terms of both accuracy and processing time.**

**Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/8585**

***Index Terms*—CNNs, Deep neural networks, Segmentation Transformers, Wood log ends.**

## I. Introduction

The processing of wooden log faces allows for the extraction of features from processed trunks within the timber industry. These features can be employed to estimate the quality of the logs. As elucidated in [1], the quality of a wooden log is delineated by several attributes, including mechanical strength, aesthetic appeal, resistance to fungi and insects, among others. Apart from quality assessment, another pivotal endeavor intrinsic to the process routinely embraced by the timber industry pertains to the determination of the volume or dimension of the logs. This information is indispensable in ascertaining the yield of derivatives that each log is capable of yielding.

Among the important stages in this context, lies the measurement of cross-sectional profiles of the timber logs. Frequently, this stage is executed manually. In this case, the process becomes labor-intensive, time-consuming, and the derived measurements are subject to imprecision [2]. For this rationale, in recent years, systems have been proposed to mechanize the associated activities.

Computer vision has emerged as a pivotal tool in the development of automated systems for this purpose. In this approach, cameras are employed to capture images of the wooden log faces, which are subsequently subject to analysis by a processing system. The most crucial phase of a processing system in this context is the semantic segmentation of the log faces, which constitutes the preliminary step toward subsequent quality analyses of the timber. The outcome of the segmentation stage is an image with an indication of which pixels are part of the surface of a given log. Thus, its outcome directly influences subsequent analysis processes such as the quantification of properties like mechanical strength, durability, and aesthetic attributes of the growth rings, which are examples of factors that determine the quality of a given log.

In the literature, works based on both classical [3]–[5] and machine learning approaches [1], [2] for this purpose are found. Among these, it is possible to observe that the methods that employ machine learning approaches outperform those based on classical processing techniques in terms of results. It is observed, however, that despite the good results already achieved, more current architectures or techniques, such as vision transformers (ViTs) or even more current convolutional neural networks (CNNs), have not yet been evaluated in this application domain.

Since the increase of precision in the segmentation process can improve the performance of automatic analyzes of wood logs, then it is important to evaluate the performance of modern artificial neural networks in this scope of application. In this sense, the present work presents a comparison of the most current segmentation techniques, confronting the results obtained by these with those previously found in the literature. In general, this work aims to deliver the following contributions:

- Employing modern deep neural networks in the segmentation of wood log faces;
- Compare the results obtained with previous works found in the literature;
- Make the implementation and image dataset available to the academic community.

In section II, approaches found in the literature for wood segmentation using computer vision are described. Section III presents modern architectures of artificial neural networks used in the developed work. Section IV describes the evaluation method used. Section V presents the experimental results obtained. Section VI presents conclusions and suggestions for future work.

Felipe Alfredo Nack and Marcelo Ricardo Stemmer are with the Federal University of Santa Catarina, Florianópolis, Brazil (e-mails: felipe.nack@posgrad.ufsc.br and marcelo.stemmer@ufsc.br).

Maurício Edgar Stivanello is with the Federal Institute of Santa Catarina, Florianópolis, Brazil (e-mail: mauricio.stivanello@ifsc.edu.br).

## II. Literature Approaches to Wood Segmentation

As the main focus of this work is deep learning architectures, we focused on analyzing publications from the last five years, making few exceptions for some classic techniques. The survey of related works was carried out generally following the systematic methodology proposed in [6]. The defined research questions were:

- Which articles involve some research in the topic of semantic segmentation of log ends?
- Which deep neural networks are mostly used in the task of semantic segmentation?

The searches were conducted mainly on the Google Scholar and IEEE Xplore platforms. We selected a few keywords for this search: "Semantic Segmentation", "Log Ends", "Review", "CNN" or "Convolutional Neural Networks", "Transformers" and "Classic". Initially we searched for the terms "Semantic Segmentation of Log Ends" and "Semantic Segmentation review" to understand how progress is being made in this task. Then we searched for the terms "Semantic Segmentation with CNN", "Semantic Segmentation with Transformer" and "Classic Semantic Segmentation" to evaluate the options available in the field of semantic segmentation. Similar variations of these search strings were also used to broaden the results. Following the directions stablished by [6], several works were selected for reading based on the response of searching platforms and, finally, the works were filtered by the relevance of their abstract to the theme proposed here. Mainly, the inclusion criteria were: (1) works written in English, Spanish or Portuguese, (2) deep neural networks proposed between 2015 and 2023, (3) papers using CNN, Transformers or classical techniques for segmentation and (4) papers presenting deep neural networks widely adopted in the semantic segmentation task.

Initially, 33 papers were selected through the initial steps of the systematic review. Then, using the exclusion and inclusion criteria, 17 papers remained, were [1] was found to be the most correlated one. Among the remaining papers, neural networks were chosen as follows:

1) Best performing network on [1] based on our proposed evaluation metrics;
2) Widely adopted CNN for semantic segmentation;
3) Best performing Transformer reviewed;
4) Second best performing Transformer reviewed;
5) Fastest Transformer reviewed.

Some works found in the literature already contemplate attempts to carry out the segmentation process for the described application. In general, we can divide the approaches found into two broad categories: those based on classical computer vision techniques and those based on the use of artificial neural networks.

### A. Approaches Based on Classical Computer Vision

In [3] we find an example of an approach based on classical image analysis techniques to perform wood segmentation. The images of wood trunks are submitted to a selection in the HSV color space, whose main objective is to isolate only pixels with tones similar to that of the wood. The result of this selection is used as input for some morphological erosion and dilation operations, which eliminate noise and close possible holes in the wood. Finally, the result of the morphological operations is used as input for an algorithm that calculates the smallest possible convex envelope of the remaining pixels, resulting in the final segmentation mask. Despite obtaining good results in very controlled environments, this technique is not accurate for other environments, whether outdoors or environments where there are different species of wood. Another classic technique found in the literature is the one used by [4], where we verify the use of the circular Hough transform together with several other manipulations in color spaces. Finally, a third approach can be found in [5], this one using a cluster growth technique, which presents excellent results within what is expected among classical techniques.

### B. Approaches Based on the Use of Artificial Neural Networks

Semantic segmentation approaches involving artificial neural networks are generally new. In [2] it is possible to verify the use of a convolutional neural network called VGG-16 [7] to perform the segmentation of wood trunks. In this first work, the accuracy found by the authors was 0.97. [1] presents an extensive comparison between different techniques for segmentation, mainly using artificial neural networks. In this work we can find a modified version of the U-Net architecture [8], an implementation of the Mask R-CNN architecture [9], an implementation of the RefineNet architecture [10], an implementation of the SegNet architecture [11], an implementation of the K-means method [12] and finally an implementation of the active contour technique (snake) [13]. The work carried out by [1] corresponds to an extensive comparison between artificial intelligence techniques in a series of databases and was able to achieve excellent results. These results will be used as a basis for comparison in this work.

## III. Architectures

Five deep neural networks were selected for evaluation based on the method exposed in Section II. Clearly, FastFCN [14] and UNet [8] were selected for being the most popular CNNs in semantic segmentation task, UNet was also used by [1]. SegFormer [15], Swin [16] and Twins [17] were selected as representatives of Transformers-based networks. Transformers-based networks are here considered what we call modern networks, due to their application being relatively new to the task of semantic segmentation in general and specifically in the task of segmentation of log ends. All five selected networks showed, during the literature review, great relevance and wide adoption in the topic of semantic segmentation.

### A. FastFCN Architecture

The first evaluated deep neural network was proposed by [14]. The original FCN (Fully Convolutional Network) typically performs the downsampling process through convolutional layers and pooling layers, resulting in a very small feature map when compared to the original image. This small feature map ends up not being able to represent information

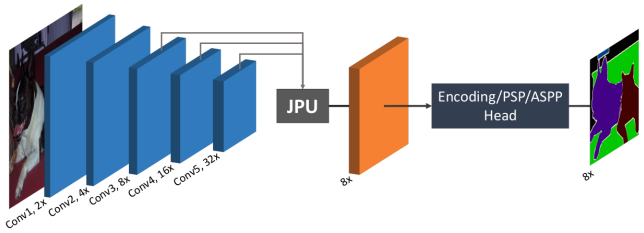


Fig. 1. Simplified FastFCN architecture [14].

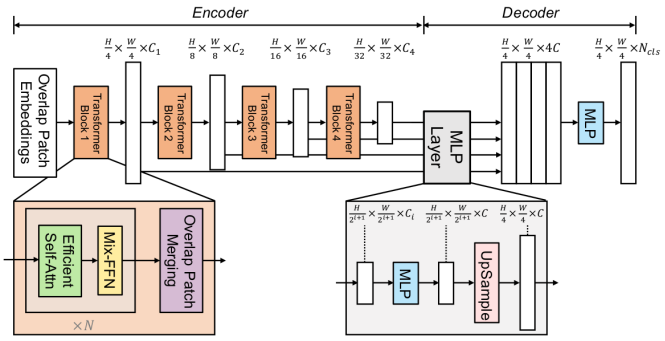


Fig. 2. Simplified SegFormer architecture [15].

related to the edges of the segmented objects. Alternatively, networks such as DeepLab [18] propose to replace the last downsampling layers of the original FCN with dilated convolution layers. This tradeoff allows the final feature map to have higher resolution. However, the computational cost of this new network becomes high, preventing its use in situations that require real-time processing.

In this sense, the FastFCN architecture has as main objective to achieve performance similar to that found in similar networks using only a fraction of the computational cost. Therefore, in their work, the authors propose to reduce the cost of dilated convolution operations through a new upsampling module. The module in question is called JPU (Joint Pyramid Upsampling). In short, the method uses the original FCN as backbone and sequentially applies the JPU module to augment the final feature map. In Fig. 1 the architecture of FastFCN is presented in a simplified way. This network performs well on benchmark datasets such as Pascal Context (53.13% mIoU). The network implemented in this work is a variation of the one proposed by the original authors. We used a ResNet-50 network as a backbone, we kept the JPU module and finally we used an Encode Head.

### B. Simplified U-Net Architecture

The second deep neural network used is known as U-Net and was initially proposed by [8] to perform the semantic segmentation of images focused on medical problems. U-Net is a network well known for learning quickly, providing good results and being able to work with smaller datasets. This network is composed of a contraction path that is responsible for capturing the context and an expansion path, symmetrical to the contraction path, which allows the precise location of the class of each pixel. The U-Net implementation used in this work is very similar to the one found in the original article. We use a U-Net network with 5 stages, however, there is a relevant change which is the size of the input images, now with 2048x1024 pixels.

### C. SegFormer Architecture

The third deep neural network used was proposed by [15] and is called SegFormer. According to the authors, SegFormer is a network designed to perform semantic segmentation in images through the unification of Transformers with Multi Layer Perceptrons (MLPs) of low computational cost for decoding. Fig. 2 presents a simplified proposal for the SegFormer network. The first functionality that the SegFormer network implements is a Transformer-type encoder layer that delivers multi-scale characteristics. This encoder layer does not need positional coding, avoiding the problem known as "interpolation of positional codes" which, in short, affects the network performance when the inputs in the training and testing phases have different sizes. The second feature that this network presents is the use of simple decoders, or low-cost MLPs, which according to the authors is a key feature for the good results achieved.

The SegFormer architecture, at the time of publication, achieved state-of-the-art performance on the ADE20k benchmark dataset (51% mIoU). The authors of this network end up developing 5 variants, with SegFormer-B0 being the lightest version, with fewer parameters, and SegFormer-B5 being the most robust version, with many more parameters. In this article we will be performing the implementation of the SegFormer-B0 network.

### D. Swin Architecture

The fourth deep neural network used in this work was proposed by [16] and is called by the authors Swin Transformer. This architecture is proposed to function as a general purpose backbone for tasks involving computer vision. Swin Transformer's main objective is to overcome the problems encountered in conventional Transformers, such as the ViT proposed by [19], which are: handling high resolution images, capturing global and local contexts, computational efficiency and preservation of spatial information. The solution to these problems is through a hierarchical Transformer that uses the concept of displaced windows. As seen in Fig. 3, each layer adopts shifted windows to perform the selt-attention process, which allows the windows of layer $l$ to get connections to each other in later layers ( $l+1$, $l+2$, ..., $l+n$).

The implementation of this work uses the Swin Transformer as a backbone and the UperNet [20] as a segmentation method. As the Swin Transformer managed to achieve, at the time of publication, results similar to the state of the art, it was chosen as a modern network to have its performance tested in this application domain.

### E. Twins Architecture

The fifth deep neural network used was proposed by [17] and is called Twins by the authors. The Twins architecture tries

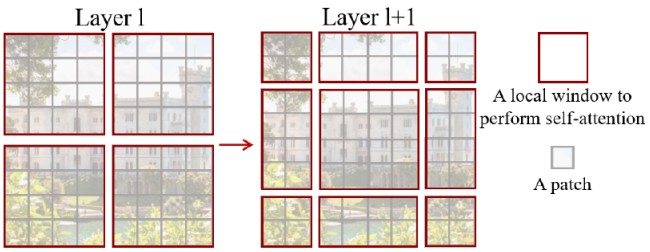


Fig. 3. Swin Transformer Window Shift Mechanism [16].

to tackle the same problems posed by the Swin Transformer [16] authors mentioned earlier. However, in their article, the authors indicate that the Swin Transformer architecture still has a field of perception that is limited by windows, in addition to the fact that its unequal-sized windows make implementations difficult in ONNX and TensorRT.

Taking these problems into account, the authors conclude that the solution lies in the design of the spatial attention mechanism. As a proposal, the authors propose the use of two attention mechanisms- (i) locally-grouped self-attention (LSA) and (ii) global sub-sampled attention (GSA), where the LSA is responsible for capturing specific information and short distance and the GSA is responsible for capturing global and long distance information.

In this sense, the Twins backbone was also chosen as a representative of modern networks to be evaluated in this application domain. Specifically, this work uses the Twins-SVT-L backbone implementation together with the UperNet method to perform dense prediction.

## IV. Evaluation Method

This section describes the method and tools used to evaluate experimental results. All implementation files and image datasets are available in the following GitHub repository [21].

### A. Image Dataset Description

The image dataset used to obtain the experimental results corresponds to a subset of the base provided by [1]. In this image base there are samples of Norwegian spruce and Douglas fir. There are a total of 2342 images of straight sections of wood logs, and for each of them there is an image with masks indicating which of the pixels correspond to the straight section of the wood. The base is divided into 5 smaller sets, grouped according to acquisition characteristics: Sbg_TS3, Sbg_TS12, Lumix, Huawei and Ane.

In Fig. 4 there are some sample images of each *subset*. From the original base, the set called Sawmill, mentioned by the authors of the referred work but not available, was not used. In Table I more details are presented for each set, including information on the equipment used in the acquisition. Each image has a manually created *ground truth* equivalent.

### B. Evaluation Metrics

Two metrics were used to evaluate the performance of deep neural networks: *Mean Pixel Accuracy* and *Mean Intersection over Union*. These metrics are widely used to evaluate the performance of neural networks in the semantic segmentation task [1] [15] [16] [17] [18] [22] [23] and therefore were selected.

***Mean Pixel Accuracy*** (mAcc) is the percentage measure of how many *pixels* were classified correctly. Let $TP_i$ be the number of *pixels* classified correctly for the class *i* and $P_i$ the total number of *pixels* in the class *i*, the mAcc (*Mean Pixel Accuracy*) metric is calculated as:

$$mAcc = \frac{\sum_{n=1}^{i} \frac{TPi}{P_i}}{i} \tag{1}$$

***Mean Intersection over Union*** (mIoU) is a measure that relates the intersection between the network prediction and the *ground truth* with the area of union between the network prediction and the *ground truth*. For any prediction with $i$ classes, take $OA_i$ as the intersection area of the $i$ class and take $U_i$ as the union area of the $i$ class, then calculate the metric mIoU as:

$$mIoU = \frac{\sum_{n=1}^{i} \frac{OA_i}{U_i}}{i} \tag{2}$$

### C. Implementation and Training of the Nets

The necessary developments for surveying the experimental results were made using the MMSegmentation toolbox [24]. MMSegmentation is an open source framework implemented on top of PyTorch aimed at image segmentation. It provides a variety of pre-trained models and tools for training new image segmentation models.

This framework provides implementations of several image segmentation models, from classic models to more recent and advanced architectures. Users can even train custom segmentation models with their own datasets using available tools. The framework offers performance evaluation metrics to measure the effectiveness of trained models in segmentation tasks. After training, the models can be used to perform inference on new images for object segmentation. It was designed to be flexible and extensible, allowing users to customize and adapt settings as needed for their own applications. Using this tool, the networks described in Section III were implemented and configured.

Two training strategies were adopted to evaluate the performance of the networks. In the first one, each network was trained for each subset of images individually. In the second, all subsets were grouped into a single image database, and then each of the networks was trained using all images at once.

The image base used underwent a *data augmentation* process and, therefore, it was necessary to carefully divide the images into training and validation groups to avoid network bias. As described by [1], the augmentation techniques employed were: scaling, rotation, vertical and horizontal shift, zooming and shearing. Each image generated as augmentation was a random combination of these techniques. On average, each original image wields another 10 augmented versions of itself. In this sense, the training and validation sets were divided manually, with a proportion of about 70% of the

Sbg_TS3 Sbg_TS12 Lumix Huawei Ane

Fig. 4. Samples from the image dataset.

TABLE I
DATASET INFORMATION

| Subset Name | Camera | Number of Images | Image Size | Wood type |
|---|---|---|---|---|
| Sbg_TS3 | Canon EOS 70D | 1504 | 1368 x 912 | Norwegian spruce |
| Sbg_TS12 | Canon EOS 5D Mark II | 768 | 2048 x 1365 | Douglas fir |
| Lumix | Panasonic DMC-FZ45 | 37 | 4320 x 3240 | Douglas fir |
| Huawei | Huawei PRA-LX1 | 22 | 3968 x 2976 | Norwegian spruce |
| Ane | Huawei ANE-LX1 | 11 | 4608 x 3456 | Douglas fir |

images being used for training and the rest for validation. For training, inference and testing purposes, all images were resized to 2048x1024.

In training, the U-Net and FastFCN networks used the SGD optimizer while the others used the Adam optimizer. All training was carried out until it was not possible to detect advances in minimizing the losses of each network.

Training was conducted on an AWS EC2 model g4dn.2xlarge virtual machine with the following configuration:

- GPU: 1x Nvidia Tesla T4
- CPU: 8x vCPU (Intel(R) Xeon(R) Platinum 8259CL CPU @ 2.50GHz)
- RAM: 32GB
- Storage: 256GB SSD

## V. EXPERIMENTAL RESULTS

The Table II presents the inference results of all networks for each subset of images and also for the complete set. In Fig. 5 the segmentation results are presented for an image of each subset compared with the respective ground truth. White pixels correspond to correct predictions of the wood, black pixels correspond to correct predictions of the background and red pixels correspond to incorrect predictions of any of the classes.

The FastFCN network showed the best performance in subset ane and excellent result in subset Sbg_TS12. As can be seen in Fig. 5, the network presents good accuracy both in terms of pixel estimation and in format and size, although it is possible to verify errors located at the edge of the wood for some of the sets (Lumix, Sbg_TS3, All), in addition to distant wood noises for other cases (Huawei).

The U-Net network presented, in general, the worst performance among all networks. When trained on the Ane dataset, U-Net was able to surpass the SegFormer by 0.49 points on the mIoU metric and was also able to surpass the SegFormer and Twins by 0.24 and 1.32 on the mAcc metric, respectively. When trained on the Sbg_TS12 dataset, U-Net was able to only surpass FastFCN by 2.61 points on mIoU and 3.17 points on mAcc. The metrics also indicate that in these subsets U-Net was able to have good pixel accuracy and good accuracy in predicting the shape of the wood, despite presenting small flaws in the background. On the other hand, in the other subsets it is verified that both the prediction of pixels and format was greatly impacted. The U-Net results, in general, show several flaws in the inner region of the wood log (Huawei, Lumix, All) and also classify many lateral noises as wood (Huawei, Lumix). In the subset Sbg_TS3, U-Net only presented the problem of underdimensioning the size of the wood, maintaining a format similar to that expected.

The SegFormer network achieved results considered good and consistent. In all training cases his pixel accuracy was above the 90% mark, while his mIoU measure remained above 85%. The inference images clearly demonstrate that this network does not present major problems with noise or even with the shape of the segmented object, that is, it only presents some slight deformation at the edges of the wood. In general, their metrics are very promising and indicate that modern networks provide greater stability in their inference.

The Swin network presented results with the highest overall quality in this work. All pixel precision metrics were above 95%, while the mIoU metric remained above 90%. Considering the size of the image sets, these results are very promising and highlight again the advances of modern architectures using the transformer mechanism. The inference results produced by Swin are very similar to those produced by SegFormer, however, they are more accurate.

The Twins network generated the worst results among the networks that use the transformer mechanism. On Ane dataset, Twins was able to outperform SegFormer by 0.51 points at mIoU metric, but lost all other comparisons to Swim and SegFormer. Regarding its performance against CNN-based models, it outperformed both U-Net and FastFCN, at mIou and mAcc, on Sbg_TS12, Sbg_TS3 and All. For Ane, it lost

TABLE II
RESULTS OF ALL PROPOSED ARCHITECTURES

| Source | Ane | | Lumix | | Huawei | | Sbg_TS3 | | Sbg_TS12 | | All | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metrics | mAcc | mIoU | mAcc | mIoU | mAcc | mIoU | mAcc | mIoU | mAcc | mIoU | mAcc | mIoU |
| FastFCN | **97.58** | **95.10** | 89.72 | 78.46 | 94.20 | 88.76 | 83.50 | 67.79 | 92.30 | 87.04 | 83.23 | 64.06 |
| U-Net | 96.31 | 93.09 | 84.04 | 70.91 | 83.76 | 74.00 | 70.06 | 49.33 | 95.57 | 89.65 | 73.37 | 53.50 |
| SegFormer | 96.07 | 92.60 | 92.78 | 85.62 | 96.32 | 90.87 | 94.05 | 86.52 | 99.12 | 98.13 | 95.86 | 90.71 |
| Swin | 97.17 | 94.83 | **95.33** | **90.98** | **96.92** | **93.20** | **96.75** | **93.42** | **99.26** | **98.43** | **97.49** | **93.82** |
| Twins | 94.99 | 93.12 | 89.43 | 81.45 | 88.94 | 84.53 | 89.98 | 78.99 | 98.10 | 95.74 | 94.21 | 87.37 |

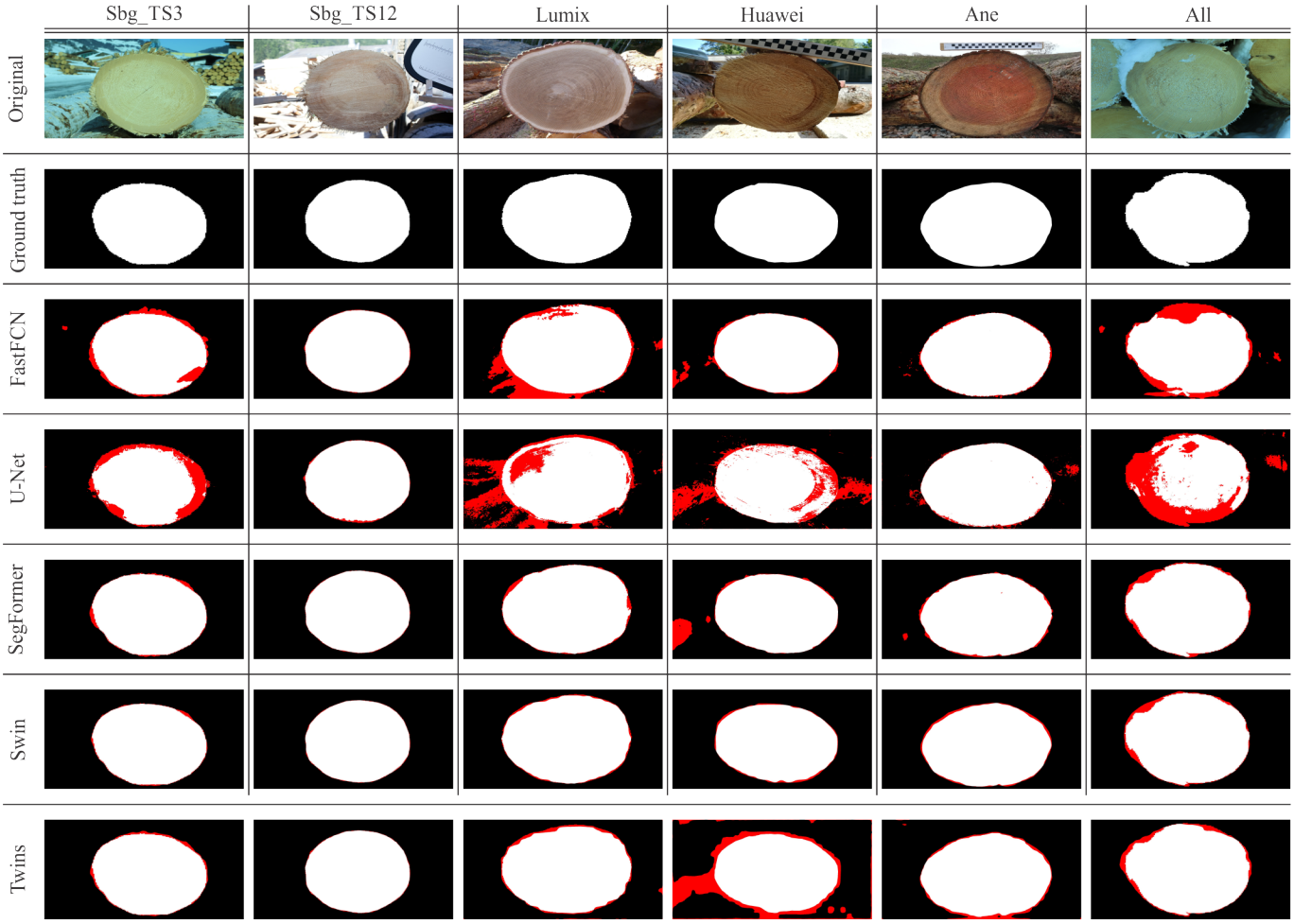


Fig. 5. Result of the inference of networks on an image from each set of images base.

on mAcc to U-Net (1.32 points) and FastFCN (2.59 points) and on mIoU it lost to FastFCN (1.98 points). For Lumix, it lost for FastFCN at mAcc (0.29 points) and won other comparisons. For Huawei, it lost both comparisons to FastFCN and won both comparison from U-Net. Considering these results, it is possible to argue that Twins outperforms FastFCN and U-Net in most cases. Similar to Swin, Twins results are also quite similar to those produced by SegFormer, both in format and pixel precision. The only exception here is the result seen on the Huawei subset which had major glitches related to confusion with the background and edges of the image.

In general, we can separate the results into two broader categories: networks based on CNNs (U-Net and FastFCN) and networks based on Transformers (SegFormer, Swim and Twins). The metrics results make it clear that Transformer-based architectures achieve better results the proposed task. This difference is due to the attention mechanism in networks based on Transformers, which allows capturing global and local contexts simultaneously, enabling both good pixel-level accuracy and good shape-level accuracy (IoU). CNNs, on the other hand, have limited attention-windows due to the size of their kernels and network depth, which do not efficiently capture these more complex contexts. In this specific case, we are not only attempting to segment wood but specifically the central wood log in the image. The results show that CNNs faced difficulties mainly in these cases.

Among the modern Transformer-based networks, the Swim architecture seems to perform better in this task due to its com-

plexity, primarily because of its Spatially Variant Attention. On the other hand, the Twins architecture demonstrates good processing efficiency owing to its innovative design, which suggests processing data through different pathways. Lastly, SegFormer appears to be a good tradeoff, offering moderate speed and delivering good results.

The average inference time of each of the proposed networks can be seen in Table III. Unlike the training, the inferences were performed on a personal computer, without the use of graphics cards. For that, a Ryzen(R) 7-4800H processor together with 16GB of RAM memory @ 3200MHz was used. We adopted this approach to assess the speed of networks in production, where GPUs are often not available.

TABLE III
AVERAGE INFERENCE TIME FOR EACH ARCHITECTURE IN SECONDS (SEC)

| | FastFCN | U-Net | SegFormer | Swin | Twins |
|---|---|---|---|---|---|
| Inference (s) | 6.7 | 9.1 | 5.1 | 7.2 | 2.7 |

Evaluating the results of Table III together with the results present in Table II it is evident that the computational complexity of networks does not imply their performance in a given task. In this work, it is possible to verify that the network with the worst segmentation results (U-Net) also presents the worst results related to speed. Furthermore, it is important to point out that networks that use the transformer mechanism will not necessarily be slow, given the example given by the Twins network in this work, reaching by a large margin the highest speed among all for inference on a CPU.

When the results of this paper and the results presented by [1] are compared, it is possible to notice a slightly increase of the $mAcc$ and $mIoU$ metrics. Nonetheless, the biggest improvement noticed was in the consistency of the networks across the different datasets. As shown in Table II, the Swin architecture was able to produce consistent results across all datasets, unlike RefineNet in [1]. Finally, the results of this paper indicates that new deep neural networks, specially Transformer based, are preferred over old ones in this specific task.

## VI. CONCLUSIONS

In this work, we evaluated different deep neural network architectures tuned for segmentation of images of log ends. All networks were able to successfully perform the task of segmenting the straight section of wood logs. Networks that use convolution as a basis (U-Net and FastFCN) and that had already been evaluated in previous works for the proposed application presented the worst general results in this work. On the other hand, the networks that use the transformer block considered in this work presented the best results overall.

Even with the small size of some subsets, it was possible to achieve high performance in semantic segmentation. It is important to point out that the computational cost of networks that use the transformer grows rapidly according to the size of the images and, therefore, in this work we chose to reduce the images before training. Anyway, it is concluded here that the Swin transformer is the most suitable network when we are looking for precision in this task, while the Twins transformer is the most suitable network when we are looking for speed.

In order to enhance the study focused on segmenting wood logs, there are several potential strategies to consider. Firstly, expanding the dataset size beyond the existing images could be beneficial, as larger datasets often lead to improved model generalization. Likewise, an interesting evaluation could also be the assessment of these networks in the task without the use of data augmentation. Additionally, addressing any imbalance or unequal distribution among the different wood log classes within the dataset may help in achieving more accurate and unbiased model predictions. Furthermore, exploring a broader spectrum of evaluation metrics beyond the ones previously employed, such as the Dice coefficient, precision, and recall, may provide a more comprehensive assessment of model effectiveness. Lastly, experimenting with alternative architectures or novel approaches in addition to the present architectures might uncover better-suited models for this specific segmentation task. Implementing these measures should bolster result reliability and contribute to a more nuanced understanding of the performance exhibited by each model in the context of segmentation of log ends.

## ACKNOWLEDGMENTS

This work has been supported by the Brazilian research agencies CAPES and CNPq. We would also like to thank Remí Decelle and Ehsaneddin Jalilian for providing the original database with wood images.

## REFERENCES

[1] R. Decelle and E. Jalilian, "Neural networks for cross-section segmentation in raw images of log ends," in *2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS)*, 2020, pp. 131–137.

[2] N. Samdangdech and S. Phiphobmongkol, "Log-end cut-area detection in images taken from rear end of eucalyptus timber trucks," in *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, 2018, pp. 1–6.

[3] F. A. Nack, "Sistema de medição do volume de toras de madeira utilizando visão computacional." Blumenau, SC, 2021. [Online]. Available: https://repositorio.ufsc.br/handle/123456789/228449

[4] B. Galsgaard, D. H. Lundtoft, I. Nikolov, K. Nasrollahi, and T. B. Moeslund, "Circular hough transform and local circularity measure for weight estimation of a graph-cut based wood stack measurement," in *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2015, pp. 686–693.

[5] R. Schraml and A. Uhl, "Similarity based cross-section segmentation in rough log end images," in *Artificial Intelligence Applications and Innovations: 10th IFIP WG 12.5 International Conference, AIAI 2014, Rhodes, Greece, September 19-21, 2014. Proceedings 10*. Springer, 2014, pp. 614–623.

[6] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, no. 2004, pp. 1–26, 2004. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=29890a936639862f45cb9a987dd599dce9759bf5

[7] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2018.

[10] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5168–5177.
[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
[12] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '07. USA: Society for Industrial and Applied Mathematics, 2007, p. 1027–1035.
[13] T. Chan and L. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.
[14] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation," *arXiv preprint arXiv:1903.11816*, 2019.
[15] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," 2021.
[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2021, pp. 9992–10 002.
[17] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," 2021.
[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.
[20] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," 2018.
[21] F. A. Nack, M. E. Stivanello, and M. R. Stemmer, "Modern wood segmentation," https://github.com/NackFelipe/ModernWoodSegmentation, 2024.
[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
[24] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.

**Felipe Alfredo Nack** received his BSc degree in Control and Automation Engineering from the Federal University of Santa Catarina (UFSC), Brazil in 2021, and is finishing his MSc degree in Automation and Systems Engineering from the Federal University of Santa Catarina (UFSC). Presently he works with computer vision systems at Bunge. His interests include mechatronics, computer vision systems and linear algebra.

**Marcelo Ricardo Stemmer** received his BSc degree in Electrical Engineering from the Federal University of Santa Catarina (UFSC), Brazil in 1982, MSc degree in Electrical Engineering from the Federal University of Santa Catarina (UFSC), Brazil in 1985 and Doctors degree in the RWTH Aachen University, Germany in 1991. He held his Post-Doctoral Internship at the Laboratoire dInformatique de Paris VI (LIP6, Pierre et Marie Curie University, Paris, France) in 2004. Presently, he is Titular Professor at the Department of Automation and Systems of the Federal University of Santa Catarina, in Florianopolis, Brazil.

**Maurício Edgar Stivanello** received his BSc degree in Computer Science from the Regional University of Blumenau (FURB), Brazil in 2005, MSc degree in Electrical Engineering in 2008 and Doctor's degree in Automation and Systems Engineering in 2013 from the Federal University of Santa Catarina (UFSC), Brazil. He held his Post-Doctoral Internship at the Department of Automation and Systems (UFSC) in 2020. Since 2010, he is a Full-Time Professor at the Federal Institute of Santa Catarina, Brazil. His interests include mechatronics and computer vision systems.