Lightweight Semantic Segmentation Model for Disaster Area based on TransUNet

Hyunsoo Kim
Dept. of Computer Science and Engineering
Chungnam National University
Daejeon, Republic of Korea
hyunsoo99kim@gmail.com

Ilmo Kang
The Division of Computer Convergence
Chungnam National University
Daejeon, Republic of Korea
rkddlfah0319@cnu.ac.kr

Donghun Baek
The Division of Computer Convergence
Chungnam National University
Daejeon, Republic of Korea
202102643@o.cnu.ac.kr

Haechan Jung
Dept. of Linguistics
Chungnam National University
Daejeon, Republic of Korea
junghc0304@o.cnu.ac.kr

Catherine Morales
Computer and Information Technology
Purdue University
West Lafayette, Indiana, USA
catherinemorales02@gmail.com

***Abstract*—In this paper, we propose a lightweight and real-time capable variant of the TransUNet model for disaster area semantic segmentation tasks, optimized for deployment on edge devices. While traditional TransUNet models have demonstrated performance in medical image segmentation, their size and complexity hinder real-time inference on resource-constrained platforms. To address this limitation, we used pre-trained R-ViT-Ti_16 model for the encoder part which maintains competitive performance despite a smaller model size. Additionally, we added an extra layer to the decoder and utilized only one skip connection to further reduce complexity. Several experiments were conducted to refine the model, including comparisons of different Vision Transformer (ViT) models, optimizers, loss functions, and activation functions. Our results demonstrated that the combination of R-ViT-Ti_16 with SGD optimizer, Log-Cosh Dice loss, and TanhExp activation function yielded a compact model with 87.5% of the performance of the baseline TransUNet while using only 13% of the model size. The final model was compressed using TensorRT; however, this step introduced significant performance degradation, indicating limitations in achieving real-time efficiency with this particular setup. This research highlights both the potential and the challenges of deploying highly efficient segmentation models on edge devices, suggesting the need for further optimization to balance performance and resource efficiency in real-time applications.**

***Index Terms*—Unmanned aerial vehicle(UAV), Semantic Segmentation, Vision Transformer(ViT), Model Compression, Real-Time Performance**

I. Introduction

Recent advancements in technology have significantly enhanced the role of unmanned aerial vehicles (UAVs) across various industries, with their utilization in disaster relief becoming increasingly prominent. UAVs are highly useful for real-time data collection and supporting rescue operations in disaster scenarios. UAVs provide real-time data collection and damage assessment during disaster situations, enabling rapid and efficient responses to large-scale emergencies. UAVs play a crucial role in quickly and accurately assessing the state of affected areas and providing essential information to rescue teams, which is particularly evident in major natural disasters.

UAVs have become essential tools for rapidly and efficiently assessing damage and providing necessary information to rescue teams in large-scale disaster situations. UAVs play a critical role in monitoring disaster-stricken areas and supporting rescue operations, facilitating effective disaster response through timely and accurate information. The technological advantages and practical applications of semantic segmentation to UAVs significantly contribute to improving the efficiency of disaster management.

However, semantic segmentation on UAV still faces several technical challenges, including power supply issues, processing capability limitations, and unreliable communication channels. These technical constraints remain a significant challenge, particularly for real-time response in disaster area segmentation. Therefore, there is a need to develop models that improve UAVs' real-time processing capabilities and operate effectively in resource-constrained environments.

Recent developments in Transformer-based models have shown excellent performance in image segmentation tasks, with various attempts to implement these models on small devices. TransUNet [1], originally used for medical image segmentation, combines the U-Net architecture with Transformers to effectively segment complex anatomical structures. TransUNet achieved an average Dice Similarity Coefficient (DSC) of 77.48% on the Harvard "Synapse multi-organ segmentation" dataset, surpassing the performance of traditional U-Net and U-Net++ models [1]. This performance highlights TransUNet's usefulness in medical image analysis. In this study, we apply TransUNet to image segmentation tasks in disaster areas and aim to make the model suitable for use in resource-constrained environments by reducing its size.

We propose a modified TransUNet model that uses a R-ViT-Ti_16 as the encoder, reduces the number of skip connections, and incorporates the TanhExp activation function along with

979-8-3503-6463-7/24/$31.00 ©2024 IEEE

Log-Cosh Dice Loss. This model retains 87.5% of the baseline TransUNet's performance while reducing the model size to 13%. However, real-time inference on the Jetson Orin Nano using TensorRT did not meet expectations. These results demonstrate both the potential and challenges of applying complex models in resource-constrained environments and indicate areas for further improvement.

II. Related works

A. TransUnet

Chen et al., 2021, proposed a novel model [1] that integrates both Transformers and U-Net [2] architectures for medical image segmentation. Traditional CNN/FCN-based medical image segmentation models have limitations in global context modeling. To address this issue, TransUNet was developed as a segmentation model leveraging transformers, which were initially designed for machine translation and have achieved state-of-the-art performance in numerous NLP tasks. This model builds upon the established U-Net structure.

While transformers are effective for global context modeling in image and video segmentation, when used alone, they often lack sufficient capability for capturing local context, which can restrict their performance. Conversely, CNN-based architectures such as U-Net excel in local context modeling but have limitations in global context representation. TransUNet is proposed as a hybrid architecture that capitalizes on the strengths of both U-Net and transformers, effectively combining their advantages in local and global context modeling.

TransUNet has demonstrated success in medical image segmentation by integrating convolutional layers with transformer architectures. Specifically, on the Synapse multi-organ segmentation dataset, TransUNet achieved a Dice Similarity Coefficient (DSC) of 77.48%, significantly outperforming traditional U-Net models. Additionally, TransUNet reported a Hausdorff Distance (HD) of 31.69 mm, indicating superior boundary delineation compared to other models. The model also excelled in segmenting specific organs, achieving a DSC of 87.23% for the aorta and 94.08% for the liver, underscoring its capability to accurately capture both global and local contextual information.

However, one limitation is its reliance on large models like ViT-B, which makes it resource-intensive and less suitable for edge devices with limited computational capacity. This highlights the need for a more lightweight and efficient architecture that can run on edge devices while maintaining strong performance in semantic segmentation.

B. Vision Transformer(ViT)

Transformer is a model used for sequence modeling in natural language processing tasks. Based on the attention mechanism, the global dependency of the sequence can be modeled. ViT recently expanded the Transformer architecture to the different computer vision tasks. In general, ViT produces similar or lower performance than ResNet-based models that previously achieved state-of-the-art in computer vision task. However when the models are trained on larger datasets Transformers can obtain better inductive bias. The best model reached accuracy of 88.55% on ImageNet, 90.72% on ImageNet-Real, and 94.55% on CIFAR-100 [3]. Despite this remarkable performance, their typical use of large model sizes (e.g., ViT-B, ViT-L) can be a significant drawback when deployed in resource-constrained environments. Recent efforts towards developing smaller versions of ViT, such as R-ViT-Ti_16, present a promising direction for creating lightweight, high-performance models suitable for edge computing.

Transformer architectures have primarily been used with in natural language processing tasks, but have recently been applied effectively in vision tasks such as image segmentation. This transformer-based image segmentation study enables the capture of global information and effective feature learning.

C. TensorRT

TensorRT [4] is a model optimization engine for high-performance deep learning inference on NVIDIA GPUs using techniques such as quantization, layer and tensor fusion, kernel tuning. TensorRT supports 32-bit, 16-bit floating point and 8-bit quantized floating point precision. By using lower precision, memory usage and computation cost can be reduced, but the accuracy can be decreased. To solve this problem, TensorRT provides calibration. For these reasons, TensorRT is used to make a low precision with high accuracy model for edge devices such as Jetson Orin Nano.

D. Semantic Segmentation in Disaster Area

Image segmentation plays an important role in disaster-related fields, and various studies have been conducted for this purpose. Mainly CNN-based methods or traditional image segmentation techniques have been applied to segmentation of disaster area images. Semantic segmentation is effective in recognizing disaster areas and establishing countermeasures in disaster areas. According to the Nur Atirah Muhadi et al [5], accurate water levels are measured by real-time monitoring of water level fluctuations in flooded areas through semantic segmentation based on CNN. When using water level gauges and ultrasonic sensors, which are existing water level measurement methods, installation and maintenance Maintenance costs increase. However, since surveillance cameras are already installed in many places, economic benefits can be pursued through semantic segmentation using cameras. This study promotes efficient resource distribution and disaster area recognition in disaster areas by performing semantic segmentation of real-time disaster area images using UAVs, rather than surveillance cameras.

In disaster scenarios, instant and real-time inspection is crucial for effective response and decision-making. Therefore, achieving real-time semantic segmentation is of paramount importance. While TransUNet has proven effective in medical imaging, it is not well-suited for disaster area inspection due to its relatively large parameter count of 105 million. This makes the model less optimal for deployment on edge devices where computational resources are constrained. To address the demands of real-time disaster area segmentation there is a need

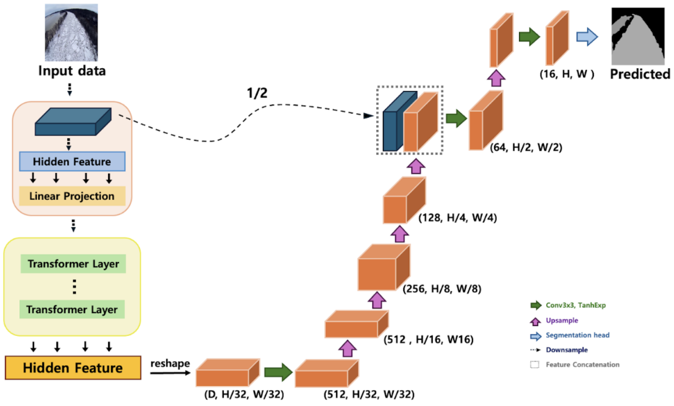

Fig. 1. Model Structure

for more compact models that can still deliver high accuracy, making them viable for use on edge devices.

III. Methodology

The architecture of the model to solve this problem is illustrated in Fig. 1. It utilizes R-ViT-Ti_16 as an encoder part, Log-Cosh Dice Loss and Cross Entropy as a loss function, TanExp as an activation function in convolution layer, SGD as an optimizer.

A. Network Architecture

In our modified TransUNet we used the R-ViT-Ti_16 model, chosen for its compact size and effective residual modules. This choice aligns with our goal of developing a lightweight model that retains competitive performance. The R-ViT-Ti_16's smaller footprint, combined with its efficient design, enables significant reduction in model size while preserving essential features necessary for accurate segmentation.

B. Skip Connection

The basic TransUNet utilizes three skip connections to integrate features from different levels of the network, enhancing the detail and accuracy of the segmentation output. However, to mitigate the increase in model size and complexity, we used a single skip connection. This strategic reduction in the number of skip connections helps maintain a balance between model complexity and computational efficiency, ensuring that the model remains practical for real-time applications on resource-constrained devices.

C. Decoder

For the decoder part, we added an additional upsampling layer to further enhance the resolution of the segmentation output. This additional step is important for refining the details in the final segmentation maps. By increasing the resolution of the decoder's output, we can achieve more precise and detailed segmentations, which is particularly important for high-quality image segmentation in practical applications.

D. Activation Function

Tanh Exponential(TanhExp) [6] is an activation function which can improve the performance for lightweight neural networks. It is defined as the following equation.

$$f(x) = x \tanh(e^x) \tag{1}$$

TanhExp has a similar figure to the other activation functions such as Mish [7] and Smish, but it requires less calculation. It also shows a steeper gradient near zero that can accelerate convergence of the network than others. We replaced the ReLU [8] activation function with TanhExp in our model. This substitution is driven by the need for improved non-linearity modeling and better stability during training. TanhExp offers a smoother gradient and reduces the likelihood of issues related to activation saturation, which can be beneficial in lightweight networks where maintaining stable and effective learning is crucial. This choice enhances the model's overall learning efficiency and performance.

E. Loss Function

1) Dice loss: The Dice score coefficient(DSC) [9] is used to assess segmentation performance when a ground truth is available by measuring how much the ground truth and a prediction are overlapped.

Using DSC, Dice loss has been devised as a loss function and it is defined as the following equation. R is a ground truth with voxel values r_n and P is a prediction with image elements p_n. The term is used here to avoid the numerical issue of dividing by 0.

$$DL(P,R) = 1 - \frac{\sum_{n=1}^{N} p_n r_n + \epsilon}{\sum_{n=1}^{N} p_n + r_n + \epsilon} - \frac{\sum_{n=1}^{N} (1-p_n)(1-r_n) + \epsilon}{\sum_{n=1}^{N} 2 - p_n - r_n + \epsilon} \tag{2}$$

2) Log-Cosh Dice Loss: The Log-Cosh Dice loss is a loss function using the Log-Cosh approach with Dice Loss [9]. It was proposed for its tractable nature while containing the features of the dice coefficient. Log-Cosh Dice Loss is defined as the following equation.

$$L_{lc-dce} = log(cosh(DiceLoss)) \tag{3}$$

This function remains continuous and finite after its first derivative, ensuring smoother gradients and more stable optimization during training. It encapsulates the key features of the Dice coefficient, making it well-suited for tasks like segmentation where class imbalance is a concern. By introducing the Log-Cosh function, the loss is less sensitive to large variations in the input, which can help prevent gradient explosions or vanishing gradients. This leads to more robust training, especially in scenarios where rapid changes in the predicted outputs might otherwise destabilize the learning process.

We used Log-Cosh Dice Loss due to its advantages in segmentation tasks, especially for lightweight models. This loss function combines the benefits of the Dice coefficient with a log-cosh penalty, which helps in handling outliers and stabilizing the training process. By smoothing the differences between predicted and true values, Log-Cosh Dice Loss ensures that the model converges effectively, providing robust segmentation results while accommodating the constraints of a compact model. Also we utilized Cross Entropy Loss for balancing the segmentation task between the classes.

F. Optimizer

Among the various optimization techniques explored, Stochastic Gradient Descent(SGD) proved to be the most effective. SGD is favored for its simplicity, efficiency, and robustness. It provides stable convergence and performs well in scenarios involving lightweight models. The choice of SGD aligns with our objectives of optimizing performance while managing computational resources, ensuring that the model meets the requirements for real-time inference on edge devices.

IV. Implementation

The implementation of our model involved several stages, including training setup, model compression and deployment.

A. Training Configuration

The training process was configured with specific parameters to ensure effective model training and convergence. For the training configuration, we set the learning rate at 0.01. The batch size was set to 24 to balance memory usage and training stability. The number of epochs was set to 200 to provide sufficient training time for the model to learn from the data. While we initially intended to experiment with different hyperparameter settings, constraints on time and resources necessitated the use of fixed values for these parameters. Instead, the focus was directed towards evaluating different activation functions, loss functions, and ViT sizes.

B. Pre-trained Model Utilization

We utilized the pre-trained R-ViT-Ti_16 model as the encoder component of our modified TransUNet architecture. This pre-trained model was selected for it's efficient feature extraction capabilities and compact size. The integration process involved incorporating the pre-trained weights into the encoder, allowing the model to leverage previously learned features.

C. Compression and Deployment

To optimize the model for real-time inference on the Jetson Orin Nano, we used fp16 method from TensorRT for model compression. The model was converted into a TensorRT-optimized format, which involved reducing precision, merging layers, and optimizing operations to enhance inference speed and efficiency. After compression, the model was deployed on the Jetson Orin Nano. We addressed various challenges during deployment to ensure that the model performed efficiently within the constraints of the edge device.

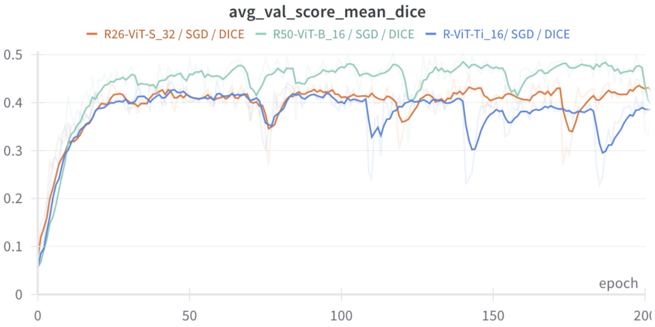

Fig. 2. Performance test by ViT models

V. Experiment

A. Dataset

The input to our method is an UAV-view disaster image I associated with 14 semantic categories based on a dataset provided from Low-Power Computer Vision Challenge(LPCV) 2023 [10]. The study used these images to train our own model.

B. Experiment Setup

All experiments were conducted using the LPCV dataset, which consists of UAV-captured imagery of disaster-affected areas. Our objective was to develop a model that could perform accurate segmentation while maintaining real-time processing capabilities on the Jetson Orin Nano. We structured our experiments into four aspects, progressively testing different configurations to achieve optimal performance. The setup were evaluated based on average mean dice score function from LPCV 2023 [10].

1) Model Size vs. Performance (ViT Variants): We first explored the impact of different Vision Transformer (ViT) models on segmentation performance. For these tests, we kept the optimizer (SGD) and the loss function (Dice Loss + Cross Entropy) consistent, changing only the ViT model in the encoder. The small R-ViT-Ti_16 model, despite its reduced size, showed competitive performance, making it a viable candidate for our lightweight model which is shown in the Fig. 2.

2) Optimizer Comparison: We fixed the ViT model (R-ViT-Ti_16) and Dice Loss while experimenting with various optimizers, including ADAMW, ADAM, SGD, and LION. Among these, SGD provided the best overall performance, striking a balance between training stability and final accuracy, and was selected for further experiments. The results can be checked in the Fig. 3.

3) Loss Function Experimentation: With the R-ViT-Ti_16 and SGD fixed, we compared the performance of two loss functions: Dice Loss and Log-Cosh Dice Loss. While the results were similar in terms of segmentation accuracy, the Log-Cosh Dice Loss demonstrated potential advantages in terms of model stability and gradient smoothness, leading us to select it as the primary loss function due to its suitability for lightweight models.

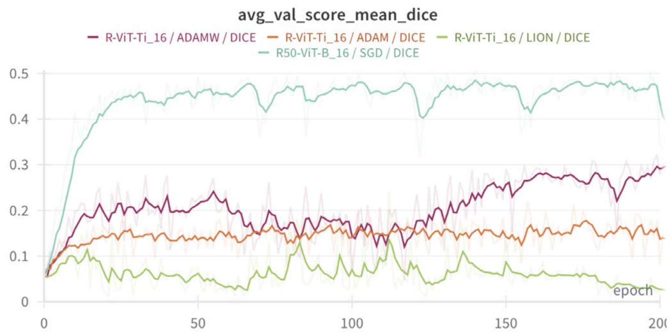

Fig. 3. Performance test by Optimizers

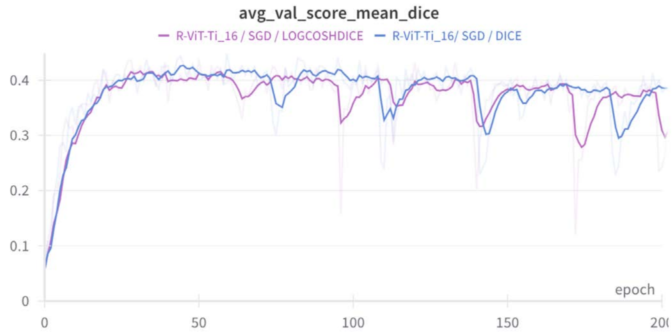

Fig. 4. Performance test by Loss Functions

4) Activation Function Substitution: Finally, we replaced the traditional ReLU activation function with the TanhExp function in an effort to further reduce the model size and improve computational efficiency. Our experiments showed no significant loss in performance with this substitution, making TanhExp a favorable alternative for resource-constrained environments.

C. Model Performance

After selecting the final configuration (R-ViT-Ti_16, SGD, Log-Cosh Dice Loss, TanhExp), we compared the performance of our modified TransUNet with the baseline TransUNet (R50-ViT-B_16, SGD, Dice Loss). Our model achieved 87.5% of the baseline performance while reducing the model size to just 13% of the original, illustrating its effectiveness for UAV-captured disaster area segmentation tasks.

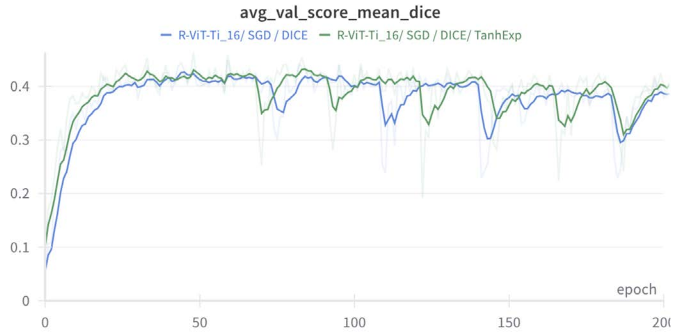

Fig. 5. Performance test by Activation Function

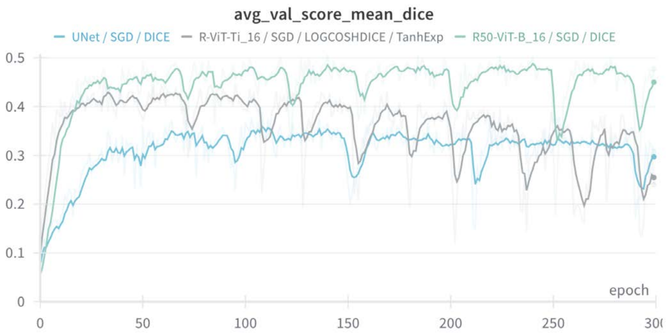

Fig. 6. Our Model VS TransUNet VS UNet

D. Real-Time Deployment

Despite achieving satisfactory segmentation performance during offline testing, our model encountered significant challenges when deployed on the Jetson Orin Nano. After compressing the model using TensorRT, we observed that the model was unable to perform real-time inference, resulting in a black screen output. This highlighted the limitations of current model compression techniques for edge devices, suggesting the need for further optimization and refinement.

VI. Conclusion

In this work, we presented a modified version of TransUNet designed for UAV-captured disaster area segmentation with a focus on real-time deployment on edge devices such as Jetson Orin Nano. Our approach utilized a lightweight R-ViT-Ti_16 model as the encoder, a simplified decoder with fewer skip connections, and TanhExp as the activation function. We also employed Log-Cosh Dice Loss to optimize performance while maintaining model efficiency.

Through a series of experiments, we demonstrated that our modified model achieved 87.5% of the baseline performance while reducing the model size to only 13% of the original TransUNet. This significant reduction in size highlights the potential for deploying efficient segmentation models in resource-constrained environments.

However, despite these promising results, our efforts to deploy the model in real-time on the Jetson Orin Nano revealed substantial challenges. After compressing the model using TensorRT, the model failed to perform effectively during inference, resulting in a black screen. This outcome underscores the limitations of current model compression techniques and the need for further research into optimization strategies that can maintain both performance and functionality on edge devices.

In future work, we aim to explore advanced compression and optimization techniques, refine the model architecture further, and improve real-time inference performance on embedded devices. Our findings offer valuable insights into the trade-offs between model size, performance, and deployment feasibility, paving the way for more robust real-time segmentation solutions in UAV-based disaster management applications.

Acknowledgement

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW) supervised by the IITP(Institute of Information & communications Technology Planning & Evaluation) in 2024(2021-0-01435).

References

[1] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[4] NVIDIA Corporation. *Developer Guide: NVIDIA Deep Learning TensorRT Documentation*, 2023. Accessed: 2024-08-21.

[5] Nur Atirah Muhadi, Ahmad Fikri Abdullah, Siti Khairunniza Bejo, Muhammad Razif Mahadi, and Ana Mijic. Deep learning semantic segmentation for water level estimation using surveillance camera. *Applied Sciences*, 11(20), 2021.

[6] Xinyu Liu and Xiaoguang di. Tanhexp: A smooth activation function with high convergence speed for lightweight neural networks. *IET Computer Vision*, 15:136–150, 02 2021.

[7] Diganta Misra. Mish: A self regularized non-monotonic activation function. In *British Machine Vision Conference*, 2020.

[8] Abien Fred Agarap. Deep learning using rectified linear units (relu). *ArXiv*, abs/1803.08375, 2018.

[9] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7, 2020.

[10] Look into Person Challenge. Lpcv dataset. https://lpcv.ai/2023LPCVC/introduction, 2023. Accessed: 2024-08-21.