Semantic Segmentation with FCN, UNET and Transformer-based Models

Atharva Vinay Sapre **Soham Shinde**

Overview

The project aims to improve the accuracy of semantic segmentation, which involves assigning labels to each pixel in an image. This task is crucial for various applications, including medical image analysis, autonomous driving, and video surveillance. To achieve this goal, the project explores different deep learning models such as Unet, Unet with attention, Unet with residual connections, RA-Unet, Unet-ASPP, and Swin-Unet, and evaluates their effectiveness in improving semantic segmentation accuracy. The project's outcome will deepen the understanding of various semantic segmentation approaches and identify the most promising models for future research and development.

The Unet model, which uses an encoder-decoder network with skip connections, is the baseline model for comparison. The other models improve on this architecture by adding attention mechanisms, residual connections, recurrent attention modules, and atrous spatial pyramid pooling modules. Swin-Unet utilizes the Swin Transformer architecture to further enhance the semantic segmentation performance. The project's ultimate goal is to improve the accuracy and efficiency of semantic segmentation models, making them more practical for real-world applications.

1 Related Work

1.1 Fully Convolutional Networks for Semantic Segmentation

This paper proposes a novel approach for Semantic Segmentation task by making the use of fully convolutional networks (FCNs), allowing end-to-end learning of pixel-wise labels. FCNs are an extension of convolutional neural networks (CNNs) that preserve spatial information throughout the network. The authors modify the VGG-16 architecture by replacing the fully connected layers with convolutional layers. Skip connections are introduced along with transposed convolutional layers to upsample the obtained feature maps to the original image size.

The proposed FCN architecture achieves state-of-the-art performance on benchmark datasets and is a standard technique in the field of computer vision. Since the publication of this paper, FCNs have been widely used in different encoder architectures, paving the way for future research in computer vision.

1.2 U-Net: Convolutional Networks for Biomedical Image Segmentation

The paper proposes a convolutional network architecture called U-Net which achieves state-of-the-art performance on biomedical image segmentation tasks. The Unet architecture consists of an encoder-decoder structure with skip connections. These skip connections allow low-level and high-level features to be combined in a more effective manner. The Unet architecture consists of an encoder-decoder structure with skip connections. The encoder progressively reduces the spatial resolution of the input image using convolutional layers and max-pooling layers, while the decoder increases the spatial resolution of the output using upsampling layers and convolutional layers. The skip connections connect corresponding encoder and decoder layers at the same spatial resolution, enabling the model to capture both local and global features. Prior to the Unet architecture, semantic segmentation models typically suffered from information loss due to the downsampling process, resulting in reduced segmentation accuracy.

The authors trained and evaluated their model on the ISBI 2012 EM Segmentation Challenge dataset and achieved a Dice coefficient of 0.804, outperforming all other methods at the time. They also demonstrated that the skip connections were critical to the success of the model, as removing them led to a significant decrease in segmentation accuracy. Overall, the U-Net architecture has proven to be a versatile and effective tool for various image segmentation tasks, particularly in the biomedical domain.

1.3 RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans

The authors observed that the original Unet architecture was suffering from information loss and gradient vanishing problems, which lead to suboptimal segmentation results. Therefore, attention mechanisms were proposed to selectively focus on informative regions of the input image. The attention Unet architectures use attention modules to highlight relevant features in the encoder and decoder networks, improving the quality of the segmentation. However, attention mechanisms can introduce additional computational overhead, making the network slower and more difficult to train.

To address this issue, the authors propose a revised attention Unet (RA-Unet) architecture that combines attention and residual connections to achieve better segmentation accuracy. The RA-Unet architecture uses residual connections to improve gradient flow and handle the vanishing gradient problem. In addition, it uses attention mechanisms to selectively focus on informative regions of the input image, while minimising the computational overhead. The RA-Unet architecture consists of an encoder, a decoder, and a final convolutional layer. Attention modules are added to the encoder and decoder networks, and residual connections are added to both the encoder and decoder networks.

This RA-Unet architecture was evaluated on several benchmark datasets, achieving a state of art performance with a mean IoU of 85.5% on the PASCAL VOC 2012 dataset. This result demonstrated the effectiveness of the RA-Unet architecture in semantic segmentation tasks and highlighted the importance of combining attention and residual connections for improved segmentation accuracy.

1.4 DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs

The paper proposes an architecture for semantic image segmentation which combines deep convolutional neural networks (CNNs), atrous convolution, and fully connected conditional random fields (CRFs). The traditional approaches based on CNNs suffer from a tradeoff between spatial resolution and field-of-view, which limit their ability to capture fine-grained details. To address this, the authors propose the use of atrous convolution, which allows the network to increase the receptive field without reducing the spatial resolution. This is achieved by inserting holes in the filters used in the convolutional layers.

The authors also introduce multi-scale prediction, which involves applying the network at different scales to capture both fine and coarse details. This is done by concatenating the output of multiple networks trained at different scales. To refine the segmentation results, the authors use a fully connected CRF, which models the dependencies between neighbouring pixels. This allows the network to incorporate higher-level contextual information, such as object boundaries and shape. The proposed approach achieved a state-of-the-art performance on benchmark datasets such as PASCAL VOC 2012 and Cityscapes.

1.5 Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

This paper proposes a new transformer-based architecture called Swin Transformer. It overcomes the limitations of traditional transformer architecture by using a shifted-window strategy to capture both the local and global contextual information. The authors mention that transformer-based models are limited in their ability to capture local information because the window size is fixed.

The Swin Transformer is evaluated on several image classification benchmarks, and the results show that it outperforms the state-of-the-art transformer-based models. The authors highlight the importance of capturing local information for achieving high accuracy in image classification tasks. In addition, the Swin Transformer's computational efficiency makes it a promising tool for real-world applications, where computational resources are limited. The proposed architecture can also be applied to vision tasks like object detection and semantic segmentation.

1.6 Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation

The authors propose a new architecture called Swin-Unet for semantic segmentation tasks by combining the Swin Transformer and Unet decoder. The Swin Transformer uses a hierarchical structure of non-overlapping local windows to capture spatial dependencies and a shifted-window mechanism to increase the receptive field. The Unet decoder improves segmentation performance by incorporating skip connections and Swin Transformer blocks. The Swin-Unet approach achieved state-of-the-art results on several benchmark datasets with better accuracy, faster training and inference time, and fewer parameters. The authors conducted ablation studies to demonstrate the effectiveness of the Swin-Unet architecture.

The Swin-Unet architecture addresses the limitations of existing segmentation models by capturing long-range dependencies and multi-scale features while maintaining a simple and efficient architecture. The authors suggest that the Swin-Unet architecture has the potential to improve the performance of other computer vision tasks. Therefore, converting Unet into Swin-Unet is a promising approach for semantic segmentation tasks, which can lead to better accuracy and faster processing times, making it a suitable architecture for segmentation tasks.

2 Method and Study

2.1 Preprocessing the data

In our image segmentation project, we preprocess input images using the preprocess() function. This function takes an image file path as input, opens the image, crops two regions of size (256, 256) from the input image, and resizes them to (128, 128). The first cropped image is normalized to values between 0 and 1, while the second cropped image is converted to a numpy array. We also create a black mask of the same size as the second cropped image to store the pixel-wise segmentation labels. Next, we perform a color to class mapping on each pixel of the second cropped image by finding the class that the pixel's color is closest to using Euclidean distance. The corresponding class label is assigned to the corresponding pixel in the mask array. Finally, we reshape the mask array to a three-dimensional array with a single channel. We then preprocess all images in the training and validation directories. This function takes the file paths of the directories as input and uses the preprocess() function to preprocess each image. The preprocessed images and masks are then stored in four separate lists. These lists can be used to train and evaluate a machine learning model that performs image segmentation.

2.2 FCN based Baseline Model

In our baseline approach, we removed skip connections because we aimed to evaluate the performance of the model without any additional shortcuts or optimizations [1]. Skip connections are commonly used in neural networks to improve the flow of information and avoid the problem of vanishing gradients, which can occur in deep networks. However, they also increase the number of trainable parameters in the model and can potentially introduce overfitting if not used properly. By removing skip connections, we wanted to establish a simple and fair comparison between different architectures and assess the effectiveness of each approach without any confounding factors.

We developed a U-Net architecture for road segmentation on the Cityscapes dataset, where the model takes an input of size 128 x 128 x 3 and is designed to classify 31 different classes. The architecture consists of five downsampling layers, each followed by a dropout layer, and three upsampling layers, also followed by dropout layers. Each downsampling layer consists of two convolutional layers with a kernel size of 3 x 3, padding of "same," and strides of 1. Each convolutional layer is followed by a batch normalization layer and a ReLU activation function. The first convolutional layer in each downsampling layer is also used for the residual connection. Max-pooling layers with a pool size of 2 x 2 and stride of 2 are applied after each convolutional layer.

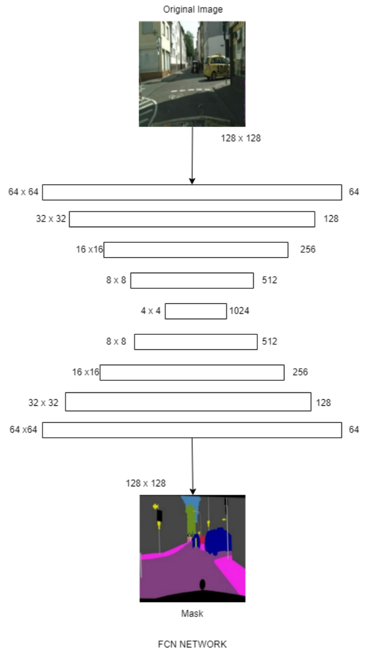


Figure 1: FCN

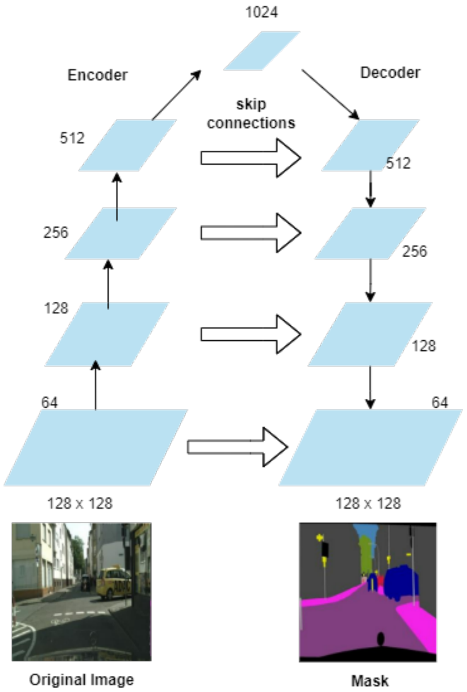


Figure 2: UNET

The upsampling layers are each composed of an up-sampling layer with a size of 2 x 2, followed by a dropout layer, and two convolutional layers with the same specifications as the convolutional layers in the downsampling layers. The last convolutional layer in the model has a filter size of 1024, and the final output is of the same size as the input.

Our baseline approach was motivated by the Fully Convolutional Networks (FCN) for semantic segmentation research paper[1], which proposed a U-Net architecture for semantic segmentation. We aimed to design a simple model that could serve as a benchmark for evaluating more complex architectures. By removing skip connections, we were able to assess the effectiveness of each approach without any confounding factors.

2.3 Unet Model and its variations

We have implemented the Unet architecture in Python using TensorFlow for a segmentation task. Our model consists of several convolutional layers, batch normalization layers, max-pooling layers, and upsampling layers to learn the features of the input images and predict the segmentation masks. The input images have a size of 128x128x3, where 3 represents the number of color channels (RGB).

We have used "same" padding in the first 2D convolutional layer with 64 filters and a kernel size of 3x3, and ReLU activation function. The output of the first layer is then passed through batch normalization and another 2D convolutional layer with similar settings. After that, we have used a max-pooling layer with a pool size of 2x2 and a stride of 2, and dropout regularization with a rate of 0.2.

The second downsample follows a similar pattern to the first downsample, but with 128 filters in the convolutional layers. The third, fourth, and fifth downsample follow the same pattern, but with 256, 512, and 1024 filters in the convolutional layers, respectively. Then, we have used a series of upsampling layers to gradually increase the size of the feature maps and output the final segmentation masks.

The U-Net model is widely used in medical image segmentation tasks due to its effectiveness in handling small datasets and its ability to capture both local and global features of the input images. We modified the Unet code to build several new models including Attention Unet, Residual Unet,

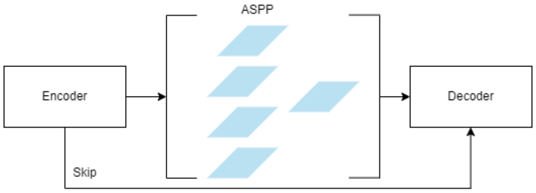


Figure 3: ASPP

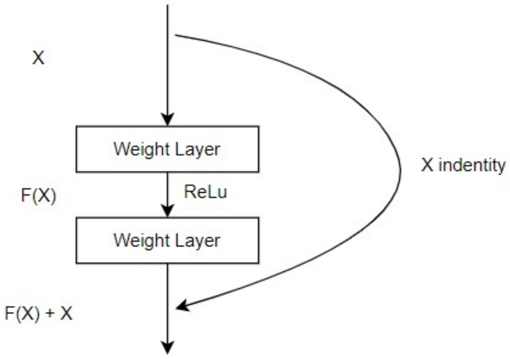


Figure 4: FCN

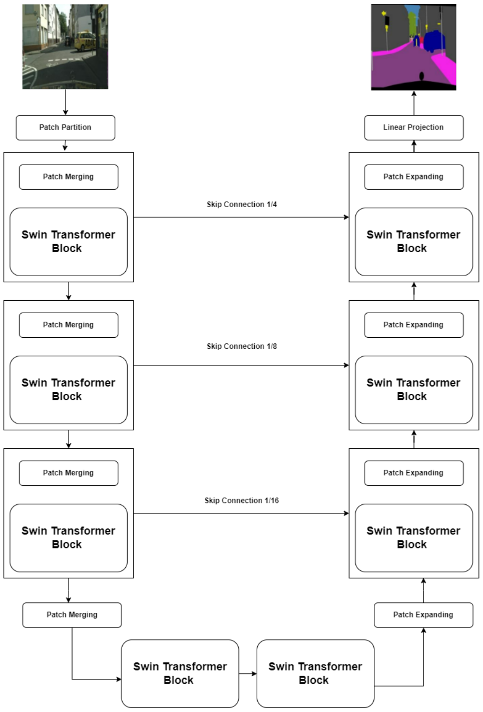


Figure 5: SWIN-UNET

RA-Unet, and ASPP-Unet, based on research papers that we studied, which highlighted specific improvements that could be made to the original Unet model.

The Attention Unet model incorporates attention gates to selectively emphasize features in the image that are most relevant to the task at hand. This helps to improve the accuracy of the segmentation output. Similarly, the Residual Unet model adds residual connections to the Unet architecture, which can help to address the issue of vanishing gradients during training, leading to more stable training and faster convergence.

The RA-Unet model includes residual connections and an attention mechanism to improve the flow of information through the network and focus on important regions in the image, respectively. RA-Unet is particularly effective for segmenting complex scenes.

The ASPP-Unet model incorporates ASPP (Atrous Spatial Pyramid Pooling) to extract features at multiple scales and better capture both global and local context. This modification is useful for segmenting objects of varying sizes within an image. By incorporating these modifications, we can create several new models with improved performance.

2.4 Swin - Unet (Vision Transformer based Unet architecture)

We also experimented with the Swin-Unet model after exploring various Unet variations. The Swin-Unet model is a combination of the Swin Transformer and the Unet architecture[5].

The Swin Transformer is a new type of transformer that is designed specifically for image processing tasks. It uses a hierarchical structure that breaks down the image into smaller patches and applies self-attention within each patch. The Swin Transformer also uses a shifted window mechanism to reduce the number of computations required to process large images.

Swin-Unet is a combination of the Swin Transformer and the U-Net architecture. The Swin Transformer is a recent architecture that has shown excellent results in image classification tasks. It is based on the self-attention mechanism used in transformers, but it introduces a hierarchical structure to improve efficiency. The Swin Transformer is built using several blocks, including Patch Merging, Shifted Window Attention, and Swin Block. We built the Swin UNet model by modifying the Swin Transformer codebase. Specifically, we added skip connections and upsampling layers along with MLP heads to enable the network to perform semantic segmentation. (References: https://github.com/berniwal/swin-transformer-pytorch)

The Swin-Unet architecture builds on top of the Swin Transformer by replacing the final classification layer with a decoder, which allows for semantic segmentation. The decoder is based on the U-Net architecture and consists of an upsampling path and a series of convolutional blocks.

The Swin Block is the fundamental building block of the Swin Transformer and consists of several sub-layers, including a 1x1 convolution, a shifted window attention layer, and a 3x3 convolution. The Patch Merging layer combines adjacent patches in the feature map to reduce the spatial resolution, while the Shifted Window Attention layer performs self-attention within a local region to capture spatial relationships.

The upsampling path in the Swin-Unet decoder consists of several blocks, including a transposed convolution layer, a concatenation layer that combines the output of the corresponding downsampling path, and several convolution layers. The final output of the Swin-Unet model is a segmentation mask with the same resolution as the input image.

Comparing Vision Transformers with Unets and CNNs, Vision Transformers have shown to be highly effective for image classification tasks. They are able to capture long-range dependencies in the image and have shown excellent results on large-scale datasets. However, U-Net and CNNs are still the most popular choices for image segmentation tasks due to their ability to capture spatial relationships and preserve spatial resolution. Vision Transformers can be adapted for segmentation tasks by incorporating a decoder architecture, similar to the Swin-Unet model.

3 Experiments

The task of Semantic Segmentation was performed on the CityScapes Image Pairs Dataset, consisting of urban street images with resolution of 256x256 pixels. To achieve this, the images were preprocessed to generate masks, which assigned each pixel in the image its corresponding category label. The dataset was split into training, testing, and validation, respectively, with 2975, 300, and 200 images in each set. We trained three different models, each with their own variations, for epochs with early stopping conditions for varying hyperparameters using sparse categorical loss for backpropagation. The models were evaluated using the mean IOU and Pixel Accuracy metrics.

3.1 Experiments

3.1.1 Performance Comparison to the Baseline Model:

We implemented the baseline model of FCN from scratch, which gave an initial test mean IOU of 0.281. We saw an improvement in performance with the UNET model, achieving a test mean IOU of 0.341. Variations of UNET with Attention, Residual, and ASPP logic were also implemented and trained. The ASPP variant of UNET gave the test performance of mean IOU of 0.354 with other and these models giving same performance as the original UNET. However, the use of Transformers in the form of SWIN-UNET resulted in slightly improved performance with a test mean IOU of 0.393. This suggests that the architecture of the transformer model may be well-suited for this task.

3.1.2 Performance Comparison to State-of-the-Art Methods:

The state-of-the-art methods for Semantic Segmentation on the CityScapes Dataset have reported mean IOU values ranging from 0.55 to 0.65. In contrast, the best-performing model in this project achieved a test mean IOU of 0.393, which is lower than the state-of-the-art methods. This difference may be due to various factors, such as the quality and quantity of data, hyperparameter tuning, preprocessing, and pretraining.

3.1.3 Hyperparameter Tuning:

The choice of hyperparameters, such as learning rate, optimizer, loss function, and batch size, can significantly affect the model's performance. In this project, the models were trained with a variable learning rate ranging from 1e-5 to 1e-4, Adam optimizer, and Sparse Categorical Cross-Entropy loss. Batch sizes of 4, 8, 16, and 32 were used. Hyperparameter tuning resulted in finding the most suitable values for achieving the best result from each specific model. Additionally, increasing the resolution of images from 128 to 256 improved the test mean IOU by 0.03-0.04 for most of the models. Transfer learning using pre-trained ResNET weights was also applied to the FCN and UNET model, but it did not improve the model's performance.

3.1.4 Analysis of Transformer Model Performance:

We observed that the transformer model, SWIN-UNET, performed slightly better than the CNN models of UNET and FCN, achieving a test mean IOU of 0.393. One possible reason for the similar performance may be that transformer models are able to capture the spatial information needed for semantic segmentation tasks. However, UNET ASPP will be suitable for semantic segmentation tasks where objects of different sizes need to be accurately segmented, allowing for multi-scale feature extraction.

3.2 Metrics

3.2.1 Pixel Accuracy

Pixel accuracy is a metric used for evaluating the performance of semantic segmentation models. It measures the percentage of correctly classified pixels in the predicted segmentation mask when compared to the ground truth segmentation mask. The equation for pixel accuracy is:

Pixel Accuracy = (Number of Correctly Classified Pixels) / (Total Number of Pixels)

This metric provides a simple and intuitive way of evaluating the overall accuracy of a model's segmentation predictions. However, it does not take into account the spatial coherence of the segmentation and may not be sensitive to class imbalance in the dataset. Therefore, it is often used in conjunction with other metrics such as Intersection over Union (IoU) to provide a more comprehensive evaluation of the model's performance.

3.2.2 mIOU

mIOU or mean Intersection over Union computes the average intersection over union (IOU) for all classes. IOU is calculated by dividing the intersection of predicted and ground truth pixels by their union. For a given class, the IOU is calculated as: IOU = (True Positive) / (True Positive + False Positive + False Negative) where True Positive is the number of correctly classified pixels for the given class, False Positive is the number of pixels predicted as the given class but actually belong to another class, and False Negative is the number of pixels that actually belong to the given class but are misclassified as another class. mIOU is calculated as the average IOU over all classes:

mIOU = (1/n) * $\sum IOU(i)$

where n is the number of classes and IOU(i) is the IOU for class i. A higher mIOU indicates better segmentation performance.

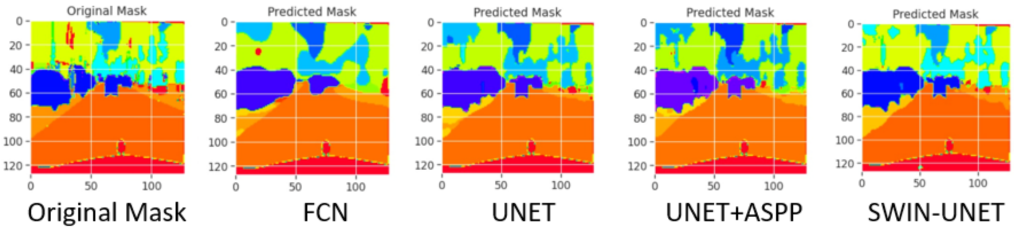
3.3 Results

Results for 128 x 128 pixels resolution:

	Valid Accuracy	Valid mIOU	Test Accuracy	Test mIOU
FCN	0.810	0.264	0.810	0.266
UNET	0.846	0.339	0.841	0.322
UNET + RES	0.843	0.344	0.841	0.326
UNET + ATT	0.842	0.343	0.841	0.330
UNET + RES + ATT	0.838	0.337	0.838	0.325
UNET + ASPP	0.846	0.342	0.838	0.356
SWIN UNET	0.834	0.370	0.821	0.363

Results for 256 x 256 pixels resolution:

	Valid Accuracy	Valid mIOU	Test Accuracy	Test mIOU
FCN	0.824	0.288	0.812	0.281
UNET	0.847	0.343	0.838	0.341
UNET + ASPP	0.854	0.361	0.846	0.354
SWIN UNET	0.861	0.395	0.859	0.393



Original Mask FCN UNET UNET+ASPP SWIN-UNET

Figure 6: Inference using the models

4 Conclusion

In this project, we explored the task of Semantic Segmentation on the CityScapes Image Pairs Dataset using three different architectures, each with their own variations. We started with a baseline FCN model and gradually improved its performance using UNET and its ASPP variant. We also experimented with a transformer based SWIN UNET, which outperformed other models, achieving a test mean IOU of 0.393. However, its performance is significantly lower than the state-of-the-art methods, indicating the need for further improvements.

Our experiments also highlighted the importance of hyperparameter tuning, model architecture, and preprocessing in achieving accurate semantic segmentation on complex datasets such as CityScapes. We found that increasing the resolution of images and transfer learning can potentially improve the performance of the models. Furthermore, we observed that transformer models perform slightly better than CNN models and can be used for semantic segmentation. This suggests that future work can explore transformer-based models further for better performance.

Future work can focus on developing more efficient and effective models for semantic segmentation on CityScapes and other similar datasets. One promising direction is the use of multi-feature selection and ensembling techniques to combine the strengths of different models and improve their performance. Additionally, exploring different types of data augmentation techniques can improve the robustness of the models and enhance their generalisation ability. Finally, applying semantic segmentation to other domains such as medical imaging and autonomous driving can further demonstrate the potential of this technique for solving real-world problems.

References

[1] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” arXiv.org, 2014. https://arxiv.org/abs/1411.4038

[2] Ronneberger, O., Fischer, P., &; Brox, T. (2015, May 18). U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv.org. Retrieved April 27, 2023, from https://arxiv.org/abs/1505.04597

[3] Jin, Q., Meng, Z., Sun, C., Cui, H., &; Su, R. (2020, December 1). Ra-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans. Frontiers. Retrieved April 27, 2023, from https://www.frontiersin.org/articles/10.3389/fbioe.2020.605132/full

[4] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., &; Yuille, A. L. (2017, May 12). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv.org. Retrieved April 27, 2023, from https://arxiv.org/abs/1606.00915

[5] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., &; Guo, B. (2021, August 17). Swin Transformer: Hierarchical vision transformer using shifted windows. arXiv.org. Retrieved April 27, 2023, from https://arxiv.org/abs/2103.14030

[6] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., &; Wang, M. (2021, May 12). Swin-Unet: Unet-like pure Transformer for Medical Image segmentation. arXiv.org. Retrieved April 27, 2023, from https://arxiv.org/abs/2105.05537

[7] He, K., Zhang, X., Ren, S., &; Sun, J. (2015, December 10). Deep residual learning for image recognition. arXiv.org. Retrieved April 27, 2023, from https://arxiv.org/abs/1512.03385

[8] M. Cordts et al., "The Cityscapes Dataset for Semantic Urban Scene Understanding," arXiv:1604.01685 [cs], Apr. 2016, Available: https://arxiv.org/abs/1604.01685