CS-433: Machine Learning Project 2

Peh Jin Yang, Celest Angela Tjong, Lee Ee Cheer
Department of Computer Science, EPFL Lausanne, Switzerland

***Abstract*—In this study, the critical task of road segmentation in satellite images was addressed using a novel approach. A comprehensive literature review explores the different semantic segmentation models, leading to the selection of U-Net and CBAM for this study. The model leverages the strengths of a weighted loss function, with a U-Net integrated with CBAM (Convolutional Block Attention Module). A series of ablation studies and hyperparameter tuning demonstrates the effectiveness of the proposed approach. Data augmentation coupled with custom-weighted loss functions are utilised to address class imbalance in the dataset. In addition, an analysis of the different models was conducted to evaluate which model is best for this task.**

I. Introduction

Road segmentation predicts and categorises road pixels in an image to assist in the extraction of accurate road networks [1]. This report delves into the application of semantic segmentation to road detection in satellite images, a task critical for extracting meaningful information from the vast amount of geospatial data at our disposal.

A. Background

Image segmentation is classified into three main groups, semantic segmentation, instance segmentation and object detection [2], [3]. Semantic segmentation is a technique that divides an image into distinct semantic areas and classifies these regions according to specified categories, in our case, background and road. Our focus is to delineate road areas from the surrounding environment in satellite images through the classification of pixels into semantic classes.

B. Motivation

Traditional image processing techniques often fall short in handling the complexity of real-world scenes, such as extracting road information from satellite imagery. Recently, researchers have taken to deep and machine learning model approaches [4]. In particular, Convolutional Neural Networks (CNN) and Vision Transformers (ViT) have shown remarkable success in semantic segmentation tasks.

II. Literature Review

A. Studying past models

Over the years, developments in machine learning and deep learning-based methods brought about notably improved performances in semantic segmentation [5]. A timeline of methods throughout the years is visualised in Fig. 1. One of the first neural networks used to undergo semantic image segmentation emerged in 2015 is Fully Convolutional Networks (FCN) [6]. FCN takes every fully connected layer in existing CNN architectures and replaces them with fully-convolutional ones, enabling end-to-end pixel-wise predictions. It transforms an image of any size into a spatial map where each pixel is assigned to a predefined label [7]. Nevertheless, there are some drawbacks to this approach such as the struggle to capture fine-grained details, maintain spatial information and speed suitable for real-time inference.

To address poor localisation, convolutional encoder-decoder-based models integrate skip connections to enable extraction of both high-level and low-level features. First proposed in 2015 by Noh et al. [8], this method consists of an encoder and a decoder. The encoder employs convolutional layers to capture and extract features from the input image, while the decoder processes the feature vector to generate a map of pixel-wise class probabilities. An example is U-Net, introduced by Ronneberger et al. in 2015 [9] for biomedical image analysis. As indicated in its name, U-Net is a symmetric U-shaped network that recursively connects outputs of the contracting path to the corresponding expansive path. This allows the network to capture both high-level semantic information and fine-grained details. U-Net is often benchmarked against traditional CNN architectures and is a state-of-the-art FCN, allowing it to handle input images of varying sizes, making it suitable for segmentation tasks on images of different dimensions. Other examples of such models include SegNet in 2017 [10] and HRNet in 2019 [11].

However, a drawback of U-Net is its focus on capturing local details through skip connections, which causes it to fall short of incorporating global context information. To counter this, Attention U-Net incorporates attention gates (AG) that selectively focus on important features and less on extraneous regions, allowing the model to capture both local and global context information. Attention U-Net was proposed by Oktay et al. [12] in 2018, for image segmentation.

Regardless, a more lightweight design is the Convolutional Block Attention Module (CBAM), made by Woo et al. in 2018 [13], which can be integrated into various other CNN architectures, including U-Net. The foundational idea behind CBAM was initially presented in 2016, by Chen et al., "Spatial and Channel-wise Attention in Convolutional Networks (SCA-CNN)" [14], which introduced the notion of merging both attentions into multi-layered attention mechanisms. Subsequently, the CBAM paper demonstrated the module's broad utility [15]. The CBAM's channel attention module focuses on emphasising meaningful features, deciding what part of the input image is more important, as illustrated in Fig. 2. Meanwhile, the spatial attention module locates important areas within a single channel by generating a map based on inter-spatial relationships, involving operations such as average-pooling and max-pooling along the channel axis, as shown in

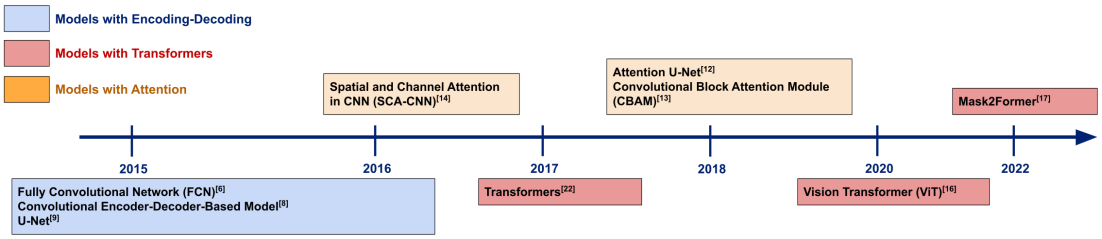

Fig. 1: Timeline of Deep Learning Models for Semantic Segmentation [6], [8], [14], [22], [12], [13], [16], [17]

Fig. 3. The combination of both modules creates more robust feature maps and enhances the model's performance [13].

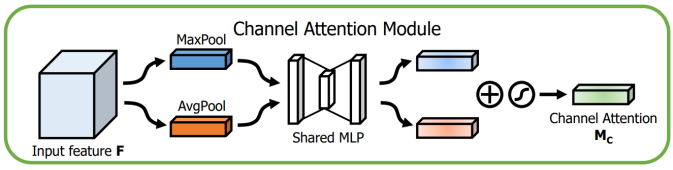

Fig. 2: Diagram of Channel Attention, *adapted from [13]*

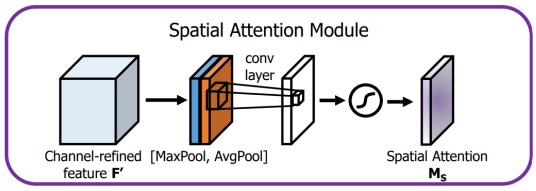

Fig. 3: Diagram of Spatial Attention, *adapted from [13]*

However, these attention mechanisms are typically more localised and may fail to capture global context at times. Thus, in recent years, self-attention networks emerged, specifically Transformers. ViTs were developed in 2020 by Dosovitskiy et al. [16] for computer vision and are applicable in image processing and classification tasks. Self-attention mechanisms are employed to capture long-range dependencies within an input image. An example is the Masked-attention Mask Transformer (Mask2Former) [17], which comprises a multi-head self-attention layer, a key component in ViT, a multi-layer perceptron and a layer normalisation. Nonetheless, semantic segmentation using ViTs results in low resolution due to the number of convolution and pooling layers, and regular ViT models are simply not suitable for segmentation tasks as they lack dedicated segmentation heads [18].

B. Deduction

The U-Net architecture is central to semantic segmentation tasks and more recent models used for semantic segmentation has been built on it. Unlike U-Nets, CNNs use fully connected layers as the final layers and are optimised for image classification problems. As for ViTs, it has seen favourable results in image classification. However, they are not directly applicable to dense prediction tasks like image segmentation, due to their patch partitioning scheme. In any case, the addition of CBAM's modules facilitates a better understanding of both global and local contexts in the input image. This is crucial for road segmentation, where the model needs to recognise long-range dependencies (global context) and fine details (local context) simultaneously.

III. Methodology

A. Models

The U-Net and CBAM U-Net will be utilised in this project. To introduce attention mechanism into the CBAM U-Net model, a CBAM module is inserted at each of the skip connections in the original U-Net architecture. Hence, attention is applied to the feature maps from the contracting path before it is cropped and concatenated with the feature maps from the expansive path.

B. Metrics for Segmentation Models [19]

Pixel Accuracy: Ratio of properly classified pixels divided by total number of pixels.

$$PA = \frac{\sum_{i=0}^{K} p_{ii}}{\sum_{i=0}^{K} \sum_{j=0}^{K} p_{ij}}$$

Intersection Over Union: Ratio of the intersection area between the predicted segmentation map and the ground-truth to the union area. A smoothing term, ϵ, is added to prevent division by zero.

$$IoU = J(A, B) = \frac{|A \cap B| + \epsilon}{|A \cup B| + \epsilon}$$

Dice Coefficient: Ratio of twice the intersection area between the predicted and ground-truth maps, relative to the total number of pixels in both images. In the context of binary semantic segmentation, Dice is equivalent to F1.

$$\text{F1/Dice} = \frac{2|A \cap B|}{|A| + |B|}$$

C. Loss Functions

BCE loss [20] measures the dissimilarity between predicted pixel-wise probabilities and corresponding ground-truth labels. Minimising BCE loss during training encourages the segmentation model to generate pixel-wise probability maps that closely match the ground-truth.

$$\text{BCE}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)]$$

The IoU loss is derived from the IoU metric. It optimises the overlap area between predicted and ground-truth by penalising deviations from ground-truth boundaries. This is highly

important in image segmentation tasks where it is crucial to precisely capture the object boundaries and shapes.

$$\text{IoU Loss}(y, \hat{y}) = 1 - \text{IOU}$$

Due to class imbalances, weighted variations of the loss functions were implemented, namely Weighted BCE loss (W.BCE) and Weighted IOU loss (W.IOU) Using class weights, a higher weightage hence importance was given to the minority road class.

$$\text{W.BCE} = -\frac{1}{N}\sum_{i=1}^{N} w_r y_i \log(\hat{y}_i) + w_b(1 - y_i)\log(1 - \hat{y}_i)$$

$$\text{W.IOU} = C - (w_b IoULoss(y, \hat{y}) + w_r IoULoss(y, \hat{y}))$$

To calculate the weighted IoU loss, the IoU for each class was first calculated and then multiplied by its respective class weights. Finally, the loss is derived as a constant C with the weighted sum of the classes' IoU subtracted.

D. Data Augmentation

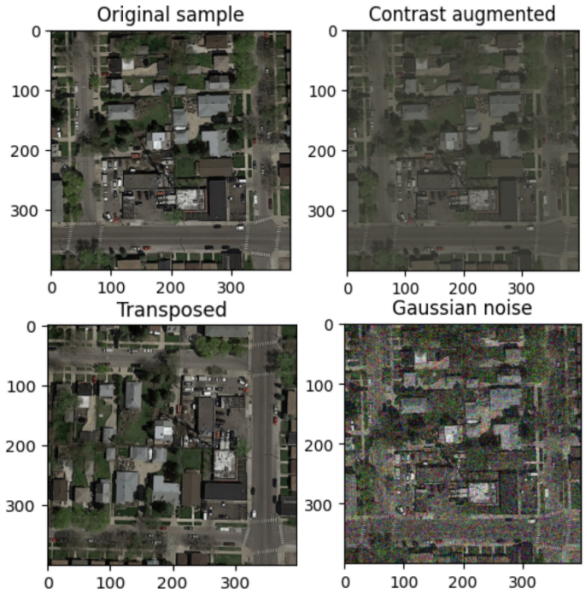

Fig. 4: Example of augmented samples

Augmented samples were generated to ensure that they are realistic while adding enough perturbation to the original samples. Augmented samples were generated using the following methods:

- Colour Augmentation: Brightness, Contrast, Saturation
- Rotation Augmentation: Flip, Rotation, Transpose
- Noise Augmentation: Gaussian, Cloud, Black Patch

IV. Experiments

Attention was first employed by integrating CBAM modules in the vanilla U-Net model to help the model focus on specific parts of the image due to the abundance of noise in the images. By analysing the gradient activations of trials 1 and 2, a decrease in gradient activations along thinner roads, with a corresponding increase in gradient activations along the main road is observed.

TABLE I: Experiments

Trial	Model	Loss	Augmented	Patched
1	Vanilla U-Net	W. BCE	No	No
2	CBAM U-Net	W. BCE	No	No
3	CBAM U-Net	W. BCE	Yes	No
4	CBAM U-Net	W. IoU	Yes	No
5	CBAM U-Net	W. IoU	Yes	Yes

Data augmentation was then performed to prevent the model from overfitting. Overfitting was diagnosed with noticeable disparity between train and validation losses. Moreover, the gradient activations from trial 2 seem to suggest a high correlation with the presence of cars. As shown in Fig. 6. (learning curve figure), data augmentation makes the model more robust to new unseen predictions.

While data augmentation helped to prevent overfitting, it also meant that the model's prediction were more sensitive to different entities. This resulted in predictions that were less smooth where predictions on long stretches of road may be broken apart. Here, using a weighted IoU loss resulted in smoother predictions by directly optimising over the overlap area between the predicted and ground-truth.

Finally, patching of the images was performed to allow the model to attend to multiple regions of interest in the image. As seen in the gradient activation of Trial 5, less significant regions of interest are now attended to as well which leads to an overall improvement in the model's performance. Patching also allows the model to handle varying input sizes.

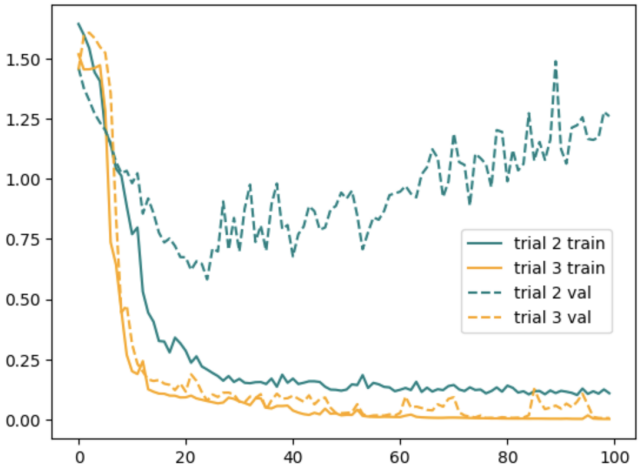

Fig. 6: Learning curves for trial 2 and 3. The small difference in trial 3's train and validation loss shows that overfitting has been mitigated.

V. Results

In each experiment, 10 pairs of images and ground-truths from the original data set were used as the validation data set.

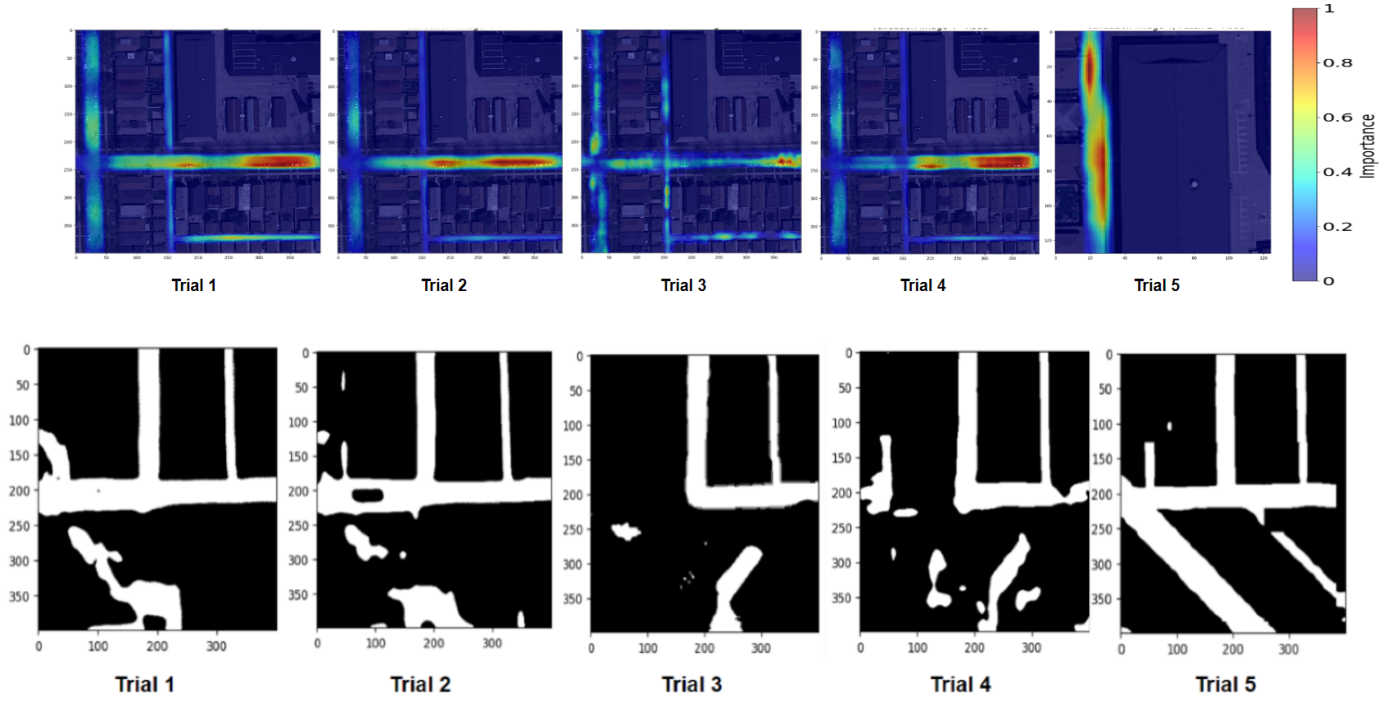

Fig. 5: Model's Gradient Activation and Prediction. For trial 5, only the top middle patch's gradient activation is shown. *Code referenced from [21]*

TABLE II: Experiment Results on Validation set

Trial	Pixel Accuracy	IoU	Dice
1	0.925 ± 0.086	0.586 ± 0.457	0.707 ± 0.227
2	0.918 ± 0.098	0.588 ± 0.511	0.699 ± 0.530
3	0.921 ± 0.101	0.547 ± 0.509	0.665 ± 0.522
4	0.928 ± 0.092	0.762 ± 0.277	0.835 ± 0.271
5	0.863 ± 0.126	0.908 ± 0.147	0.946 ± 0.110

TABLE III: Hyper-parameter tuning

Parameter	Value
Learning Rate	0.00001
Epochs	100
Batch size	16
Optimiser	Adam
MaxPooling Dropout	0
L2 Regularisation	0

Observing both the metrics and prediction on the validation data set, we concluded that Trial 5 exhibited the best performance. Indeed, the better performance is also exhibited in the gradient activations where multiple regions of interest over the entire image are attended to.

Finally, hyper-parameter tuning was performed using KerasTuner. 25 trials were performed using a Bayesian Optimiser with an EarlyStopping of patience 3. The configurations of the tuner and tuned hyper-parameters are shown below in Table II. While the tuning results suggested an optimal learning rate of 0.01, 0.00001 was used instead to prevent the model from converging to a local optimum too quickly. Moreover, regularisation like adding Dropout layers was not included to prevent excessive addition of Bias. The final model was then used to generate a test F1 score of 0.826 and pixel accuracy of 0.907 on AICrowd.

VI. Discussion

Semantic segmentation requires a precise delineation of classes for accurate labeling. At the same time, there is also a need for global context integration to understand the relationships between distant pixels. To meet these requirements, the attention mechanism was integrated with a vanilla U-Net.

Moreover, there was an issue of class imbalance with almost 80% pixels being background pixels. To address this, realistic data augmentation methods and class weights were employed.

Together, these strategies enabled the training of a classifier with good performance. For future improvements, more sophisticated attention mechanisms (e.g. triplet attention) and model architectures (e.g. ViT) could be used instead.

VII. Summary

In this project, a U-Net integrated with lightweight attention module CBAM has been applied to satellite image segmentation. Moreover, to overcome the imbalance in classes, custom-weighted IoU loss and data augmentation techniques were applied. We hope that this project will serve as a basis for future improvements in the field of semantic segmentation for satellite images.

VIII. APPENDIX

A. Ethical risks

The individuals and communities impacted by this risk primarily encompass those whose private properties and activities may be unintentionally captured in satellite images. The adverse effect revolves around the potential violation of individuals' privacy, exposing their locations, behaviors, and living spaces without their awareness or consent. The seriousness of this risk is substantial, given the sensitive nature of personal privacy. The likelihood of its occurrence depends on the specific use case and the effectiveness of measures in place to alleviate such concerns.

To address these security concerns, a possible approach would be to "censor" parts of the images used for training that contain private information. For example, black patches could be placed on these areas. Fig. 7. shows an example of applying a black patch to censor private information.

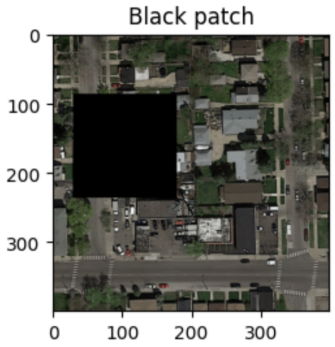

Fig. 7: Train image with black patch "censoring"

However, adding such modifications to the images may introduce more bias in the model and this may negatively affect the model's performance.

The paper by Kiya et al.[23] suggests the use of encrypting images and models to alleviate privacy concerns. Rather than adding perturbations to the train images, access to the train images and models is restricted instead. By designing models that can be trained with encrypted images, personal privacy concerns can be mitigated while ensuring that the model's performance is maintained. While the team was not able to implement it in this project, it is definitely an approach worth considering for future works.

REFERENCES

[1] O. Ozturk, M. S. Isik, M. Kada, and D. Z. Seker, "Improving Road Segmentation by Combining Satellite Images and LiDAR Data with a Feature-Wise Fusion Strategy," Applied Sciences, vol. 13, no. 10, p. 6161, May 2023, doi: 10.3390/app13106161.

[2] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images", ISPRS J. Photogramm. Remote Sens., vol. 117, pp. 11-28, Jul. 2016.

[3] W. H. M. W. Mohtar, A. M. Muad, M. Porhemmat, H. Ab. Hamid and S. S. Whayab, "Measuring scour level based on spatial and temporal image analyses", Struct. Control Health Monitor., vol. 28, no. 1, pp. e2645, Jan. 2021.

[4] A. Abdollahi, B. Pradhan, N. Shukla, S. Chakraborty and A. Alamri, "Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review", Remote Sens., vol. 12, no. 9, pp. 1444, May 2020.

[5] B. Baheti, S. Innani, S. Gajre, and S. Talbar, "Semantic scene segmentation in unstructured environment with modified DeepLabV3+," Pattern Recognit. Lett., vol. 138, pp. 223–229, 2020, doi: 10.1016/j.patrec.2020.07.029.

[6] 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, pp. 3431– 3440, 2015.

[7] Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M., and Stilla, U., "Semantic segmentation of aerial images with an ensemble of CNNs", in ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., III-3, 473–480, https://doi.org/10.5194/isprs-annals-III-3-473-2016, 2016.

[8] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1520–1528.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[10] 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.

[11] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," arXiv preprint arXiv:1909.11065, 2019.

[12] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999, 2018.

[13] S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon, "CBAM: convolutional block attention module", in Proc. European Conference on Computer Vision, ECCV. 2018, pp. 3-19.

[14] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.S. Chua, "SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning," in Proc.

of Computer Vision and Pattern Recognition (CVPR), 2017.
[15] D. Misra, “Attention Mechanisms in Computer Vision: CBAM,” Paperspace Blog, https://blog.paperspace.com/attention-mechanisms-in-computer-vision-cbam/, 2020.
[16] 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,” arXiv preprint arXiv:2010.11929, 2020.
[17] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, ”Masked-attention Mask Transformer for Universal Image Segmentation,” arXiv.org, https://arxiv.org/abs/2112.01527, 2021.
[18] Z. Xu, W. Zhang, T. Zhang, Z. Yang, and J. Li, ”Efficient transformer for remote sensing image segmentation”, in Remote Sensing 13 (18), 3585, 2021.
[19] S. Minaee, Y. Boykob, F. Porikli, A. Plaza, N. Kehtarnavaz, D. Terzopoulos, ”Image Segmentation Using Deep Learning: A Survey,” arXiv.org, https://arxiv.org/abs/2001.05566, 2020.
[20] K. Yadav, “An in-depth exploration of loss functions in deep learning,” LinkedIn, https://www.linkedin.com/pulse/in-depth-exploration-loss-functions-deep-learning-kiran-dev-yadav, 2023.
[21] K. Vinogradova, A. Dibrov, and G. Myers, ”Towards Interpretable Semantic Segmentation via Gradient-Weighted Class Activation Mapping (Student Abstract),” Proceedings of the AAAI Conference on Artificial Intelligence, 34(10), 13943-13944. https://doi.org/10.1609/aaai.v34i10.7244, 2020.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ”Attention Is All You Need”, arXiv preprint arXiv:1706.03762.
[23] H. Kiya, T. Nagamori, S. Imaizumi, S. Shiota, ”Privacy-Preserving Semantic Segmentation Using Vision Transformer,” Journal of Imaging, 8(9):233. https://doi.org/10.3390/jimaging8090233, 2022.