AUTOMATED VISUAL DEFECT DETECTION USING DEEP LEARNING

BY

LOH XIAO

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONOURS)

Faculty of Information and Communication Technology

(Kampar Campus)

MAY 2022

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: Automated Visual Defect Detection Using Deep Learning

Academic Session: MAY 2022

I LOH XIAO

(CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,

(Author's signature)

(Supervisor's signature)

Address:
11-10-3, Lintang Macallum 2,
10300 Georgetown,
Pulau Pinang

Dr Ng Hui Fuang
Supervisor's name

Date: 8 September 2022

Date: 9 September 2022

FACULTY/INSTITUTE* OF INFORMATION AND COMMUNICATION

UNIVERSITI TUNKU ABDUL RAHMAN

Date: 8 September 2022

SUBMISSION OF FINAL YEAR PROJECT/DISSERTATION/THESIS

It is hereby certified that ***Loh Xiao*** (ID No: 19ACB06100) has completed this final year project/ dissertation/ thesis* entitled "Automated Visual Defect Detection Using Deep Learning" under the supervision of Dr Ng Hui Fuang (Supervisor) from the Department of Computer Science Faculty/Institute* of Information and Communication.

I understand that University will upload softcopy of my final year project / dissertation/ thesis* in pdf format into UTAR Institutional Repository, which may be made accessible to UTAR community and public.

Yours truly,

Loh Xiao

DECLARATION OF ORIGINALITY

I declare that this report entitled "**Automated Visual Defect Detection Using Deep Learning**" is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature : 

Name : LOH XIAO

Date : 8 September 2022

ACKNOWLEDGEMENTS

I would like to express thanks and appreciation to my supervisor, Dr Ng Hui Fuang who has given me a bright opportunity to engage in this “Automated Visual Defect Detection Using Deep Learning” project. This is a valuable opportunity for me to take the first step towards deep learning research. Throughout the project, I learned a lot of leading and excellent concepts and knowledge regarding semantic image segmentation. A million thanks and appreciation to you.

Finally, I must say thanks to my parents and my family for their love, support, and continuous encouragement throughout the course.

ABSTRACT

A manufacturing defect is a flaw that causes a product to deviate from its intended design, thereby losing its quality and no longer having its due value. There are two kinds of methods employed by the manufacturing industries to ensure that the manufactured products are well-conditioned and free of any defects, namely human quality inspection and artificial intelligence visual inspection. However, the former faces many limitations and problems, resulting in low quality control efficiency, while the latter is relatively reliable and effective. Artificial intelligence visual inspection is a technique which utilises computer vision and deep learning technology to mechanically "see" a product and determine whether it has defects, without any human involvement. The main goal of this project is to study and develop various automated defect detection models by utilizing state-of-the-art deep learning segmentation algorithms, including U-Net, Double U-Net, SETR, TransU-Net, TransDAU-Net, CAM and SEAM to perform semantic segmentation in fully supervised and weakly supervised learning manners. Model analysis and evaluation are performed to compare the performance of all the algorithms in a variety of aspects. In this project, a magnetic tile defect dataset and a production item surface defect dataset are employed to train and evaluate deep learning segmentation models. By applying the models, the detected defects will be segmented and classified, and the output results will be displayed to the user through a coloured segmentation mask for each defect. Manufactures will be able to automate the industrial inspection process by implementing the deep learning models proposed in this project. Quality control procedures of the industries will be sublimated to another level by the automation of product defect detection.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

C	Feature channels
c	Number of possible classes to detect
H	Height of image/feature map
P	Height/width of image/feature patch
W	Width of image/feature map

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AQL	Acceptable Quality Levels
CAM	Class Activation Map / CAM model
CNN	Convolutional Neural Network
CPA	Malaysian Consumer Protection Act 1999
DDI-Net	Densely Dilated Inception Network
EM	Electron Microscopy
ECR	Equivariant Cross Regularization
ER	Equivariant Regularization
FCN	Fully Convolutional Network
FN	False Negative
FP	False Positive
IDE	Integrated Development Environment
IoU	Intersection Over Union
MDI	Multi-scale Dilated Inception
mIoU	Mean Intersection Over Union
PR	Precision Rate
PA	Pixel Accuracy
PCM	Pixel Correlation Module
ReLU	Rectified Linear Activation Function
SEAM	Self-supervised Equivariant Attention Mechanism
SETR	SEgmentation TRansformer
SGD	Stochastic Gradient Descent Algorithm
TN	True Negative
TP	True Positive
VGG	Visual Geometry Group
ViT	Visual Transformer
WSSS	Weakly Supervised Semantic Segmentation

CHAPTER 1 PROJECT BACKGROUND

In this chapter, we present the background and motivation of our research, our contributions to the field, and the outline of the thesis.

1.1 Introduction

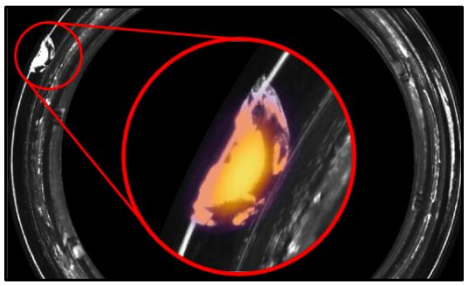


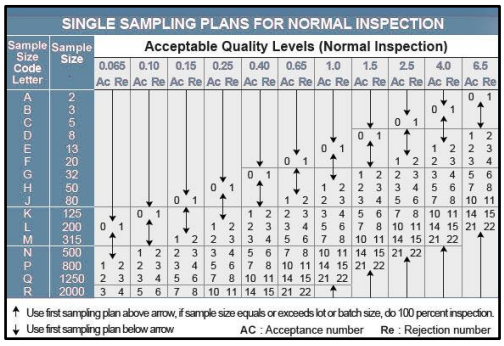
Figure 1.1 Manufacturing defect.

Figure 1.2 Semantic segmentation.

Defects in products before and during the manufacturing process are inevitable. This problem may be due to poor product design or human errors and mechanical failures in the manufacturing process, causing the product to deviate from the expected design. By all means, a defective product loses its original value, because it has no quality assurance, and more seriously, it will bring danger to users. According to s. 68.1 of the Malaysian Consumer Protection Act 1999 (CPA) [1], if any damage is caused in whole or in part by a product defect, the producer of the product will be liable for the damage. Therefore, manufacturer parties always try to detect and destroy all product defects to leave a good impression on customers, maintain reputation, and avoid unnecessary troubles.

Deep learning is one of the most notorious technologies in recent years which helps the computer to perform what humans can do naturally — understand the contents of the image. In traditional object detection algorithms, features of an object have to be handcrafted, and alternatively, deep learning algorithms grant computers the ability to learn features on its own. Semantic image segmentation is one of the deep learning concepts that predict and label each pixel in a digital image in accordance with its corresponding class, thereby partitioning the image into multiple segments, making it easier to observe and analyze objects of interest. The output of segmentation model, a pixel-level labelled image, is referred as a mask. Deep learning algorithms allow the computers to extract the unique features from a set of labelled training data, and thus be able to classify the input image accordingly at a pixel-level in a real-time scenario.

1.2 Problem Statement



SINGLE SAMPLING PLANS FOR NORMAL INSPECTION

Sample Size Code Letter	Sample Size	Acceptable Quality Levels (Normal Inspection)																							
		0.065		0.10		0.15		0.25		0.40		0.65		1.0		1.5		2.5		4.0		6.5			
		Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re		
A	2																						↓	0	1
B	3																			↓	0	1	↑		
C	5																	↓	0	1	↑				
D	8															↓	0	1	↑						
E	13													↓	0	1	↑			↓	1	2	↓	1	2
F	20											↓	0	1	↑			↓	1	2	↑			2	3
G	32									↓	0	1	↑			↓	1	2	↓	2	3	↓	3	4	3	4
H	50							↓	0	1	↑			↓	1	2	↓	2	3	↓	3	4	↓	5	6	5	6
J	80					↓	0	1	↑			↓	1	2	↓	2	3	↓	3	4	↓	5	6	↓	7	8	7	8
K	125			↓	0	1	↑			↓	1	2	↓	2	3	↓	3	4	↓	5	6	↓	7	8	↓	10	11	10	11
L	200	↓	0	1	↑			↓	1	2	↓	2	3	↓	3	4	↓	5	6	↓	7	8	↓	10	11	14	15	14	15
M	315					↓	1	2	↓	2	3	↓	3	4	↓	5	6	↓	7	8	↓	10	11	↓	14	15	21	22	21	22
N	500	↓			1	2	↓	2	3	↓	3	4	↓	5	6	↓	7	8	↓	10	11	↓	14	15	↓	21	22	↑		
P	800	↓	1	2	↓	2	3	↓	3	4	↓	5	6	↓	7	8	↓	10	11	↓	14	15	↓	21	22	↑				
Q	1250	↓	2	3	↓	3	4	↓	5	6	↓	7	8	↓	10	11	↓	14	15	↓	21	22	↑							
R	2000	↓	3	4	↓	5	6	↓	7	8	↓	10	11	↓	14	15	↓	21	22	↑										

↑ Use first sampling plan above arrow, if sample size equals or exceeds lot or batch size, do 100 percent inspection.
↓ Use first sampling plan below arrow AC : Acceptance number Re : Rejection number

Figure 1.3 Acceptable quality level sampling table.

As mentioned earlier, manufactured products are suffered from defects which hamper the intended use of the products. To be specific, a surface defect may demolish the appearance of an artwork or adornment, making it no longer exquisite and losing its due ornamental value. Defects may also affect the functionality and practicality of an accessory equipment or component part. Most industrial companies will manage the defective products based on the acceptable quality levels (AQL), which measures the company's tolerance for each defect. Nonetheless, the various appearances and characteristics of defects often lead to misclassification of their type and severity, and ultimately lead to mishandling issue. In addition, some defects tend to be miniature and difficult to identify. In the worst case, a defective component is not noticed and applied to a life-critical system, it may cause the entire system to malfunction, resulting in serious injury or even death.

In order to ensure that the manufactured products are free of defects, most small-scale industries that do not acquire advanced technologies will hire work forces to inspect the quality of the products. The advantage of human inspection lies in the extraordinary interpretative ability of human beings, which enables us to perceive even previously unknown objects. Moreover, with the combination of human eyes, we are able to analyze and recognize objects of interest in a wide field of vision. However, human visual perception has limitations that may influence the performance of visual inspection. As an emotional creature, human perceptions are easily distinctive due to subjectivity. Human eyes are also sensitive to lighting and color. If they overwork beyond their limits, problems such as overlook and eye fatigue may arise. In essence, human quality inspection may lose accurate judgements and measurements by cause of physical and emotional limitations.

1.3 Motivation

Figure 1.4 Artificial intelligence visual inspection.

In this day and age, due to the low efficiency of human inspection, artificial intelligence (AI) has been introduced to take over the position. As a subset of AI, deep learning technology can perform image processing and defect detection without human assistance after learning the features of specific defects from the supplied training dataset. With this, product inspection and analysis work can be fully automated. Advanced deep learning models may have better interpretative ability than the human brain, indicating that they are more tolerable to variations. Furthermore, AI inspection has relatively better performance and sustainability by virtue of the robustness of a computerized system. For instance, it is not restrained by physical capacity and capable to operate at a higher speed while maintaining marvelous detection accuracy. In short, in most scenarios, machine can easily surpass human in terms of visual defect detection with the help of deep learning technology.

The aim of the thesis is to propose deep learning semantic segmentation models for automatic defect detection, which can be applied in the industrial quality inspection applications, despite of the type of product it is allocated for inspection. The model shall be able to segment and classify the defects in the input image captured by industrial inspection camera, and display the segmentation results to the user. The detection accuracy of the model shall be as good as possible to ensure that all defects can be correctly identified and labelled, while its computation speed shall fulfil the basic requirement of industrial inspection. After all, it shall separate the detected defects from the irrelevant background and display segmentation masks on the screen in the corresponding color of each defect category.

1.4 Objectives

The main objective of the project is to research and develop automatic defect detection models by using well-known existing deep learning algorithms. The foremost algorithm will be reviewed, modified and extended to the built model architectures as the key deliverables of the project. Since this project mainly focuses on semantic segmentation, several state-of-the-art deep learning semantic segmentation-based algorithms plus two improved models proposed in this project are selected for "seed matching", namely U-Net, Double U-Net, SETR, TransU-Net, TransDAU-Net, CAM and SEAM algorithms. In fact, all the mentioned existing algorithms have achieved marvelous performance in semantic segmentation problem.

Many researchers have claimed that Transformer-based models outperform CNN-based models in visual recognition tasks due to their robust working principles and less computational resource requirements. Therefore, efforts are being made to compare CNN-based models and Transformer-based models in various aspects, such as design patterns of model architecture, segmentation performance, computational cost, etc. Additionally, research will be conducted to determine if blending these two algorithms in a single model architecture, that is, a hybrid CNN-Transformer architecture will yield better results. To meet the performance requirements of industrial visual inspection, attention is aimed at the cost and accuracy of the models so as to genuinely inspect hundreds or thousands of objects on a production line with insignificant error rates.

The traditional semantic segmentation task that applies fully supervised setting is time-consuming and expensive because of the need to prepare pixel-level class-specific annotations for the entire dataset images. Weakly supervised semantic segmentation is therefore proposed by the community to exploit weak supervisions that are relatively easily obtainable to achieve equivalent segmentation performance. In this project, advanced weakly supervised semantic segmentation methods that utilize image-level classification labels to localize objects of interest, i.e., surface defects, without severely impacting model performance, will be explored and implemented to lift the burden of industrial labour in collecting pixel-level annotations.

1.5 Project Scope and Direction

As the title suggests, this project focuses on the construction of deep learning models that can detect, segment and classify defects that appear on the surface of defective item. Developed models will be highly evaluated based on the fundamental requirements of industrial inspection services, and it is expected that the models will be widely used in product development and manufacturing stages. With the help of these defect semantic segmentation models, manpower is no longer needed throughout the inspection process. Thus, human errors and ill-constructed products are anticipated to be greatly minimized by the segmentation models, thereby advancing and maintaining the quality of the manufactured products.

The basic architecture of segmentation models involved in this project can be mainly divided into two sections, namely encoder and decoder. The encoder part acts as a contextual feature extractor, where the patterns or representations of defects in the image, such as corners, edges, etc., will be learned by the models through the implementation of convolution filters. The extracted features are then passed to the decoder part to restore the dimension of the feature map to that of the input image, thereby recovering the spatial information neglected by the encoder part. In the case of fully supervised setting, the ultimate convolutional layer in the decoder part is responsible for making pixel-level classification and outputting segmentation maps/masks. Alternatively, in weakly supervised setting, the last fully connected layer serves for making image-level classification and producing image-level classification labels. Later, the predicted masks or class labels are compared with the annotated ground-truth masks or class labels, and filter weights are adjusted during backpropagation to better learn to recognize the features, thereby reducing the loss, also interchangeably referred to as prediction error.

The deep learning models built in this project will be trained and evaluated on two large benchmarks consisting of images of defective or unblemished production items and magnetic tiles, respectively. As such, the final deliverable of this project will be trained deep learning models that can detect surface defects in common production items or magnetic tiles, thereby increasing the efficiency, productivity and accountability of inspection processes in various manufacturing industries.

1.6 Contributions

Deep learning algorithms are becoming mature as the technology develops. There are many studies which utilize state-of-the-art deep learning techniques for defect detection. However, most researchers have focused on developing semantic segmentation models that are tuned to target only specific types of defects. A model proficient in detecting specific defects may not achieve good detection performance for other types of defects. Moreover, there are many existing deep learning algorithms that have not been explored and utilized in the field of defect detection research.

In this project, a variety of well-known fully supervised and weakly supervised deep learning algorithms are studied, modified and applied to construct respective defect semantic segmentation models. These models will be evaluated and compared in many aspects to determine the most superior model that suits for defect detection. Additionally, the defect semantic segmentation models are trained to be more robust and possess higher generalization performance. Most notably, two improved models with outstanding segmentation performance were developed throughout the research, namely Double U-Net and TransDAU-Net. Double U-Net complements the U-Net model and achieves better detection results, while TransDAU-Net alleviates the high-cost issue of the TransU-Net model while achieving similar or better performance. By that, some fresh concepts on visual defect detection using deep learning techniques are devoted for extensive studies in the future work. The proposed deep learning models can be extended and applied to the embedded system of industrial inspection cameras for quality control automation. Manufacturing industries no longer have to worry about product quality degradation due to troublesome defects.

1.7 Report Organization

The details of this study can be found in the following sections. Chapter 2 reviews some related work on semantic segmentation problems. In addition, the limitations of the models proposed in these studies will be analyzed. In Chapter 3, the deliverable requirements for this project will be mentioned. Most importantly, the entire flow of the built deep learning system will be explained in detail, including the architectural design, setup, training, and validation steps of each model. This chapter will also describe the project timeline. Then, Chapter 4 describes the performance of each model by visualizing prediction results of the models for each defect category and analyzing evaluation results of various metrics. At last, Chapter 5 reports the conclusion of the project.

CHAPTER 2 LITERATURE REVIEW

2.1 Previous Works on Defect Detection

2.1.1 U-Net: Convolutional Networks for Biomedical Image Segmentation

Convolutional networks are often used for image classification tasks, where the input image is classified as a whole and does not include object localization. However, for biomedical tasks, the network is expected to locate objects of interest by classifying each image pixel accordingly. Training the network in a sliding window setting is able to generate the desired output, but this approach is slow and requires a trade-off between contextual and spatial information. Therefore, O. Ronneberger, et al. [2] proposed a network that complements the common contracting convolutional network with successive layers consisting of upsampling operators to remedy all the mentioned prior art problems. The proposed network, also known as U-net, achieves the ability to generate more accurate segmentations with relatively few training images.

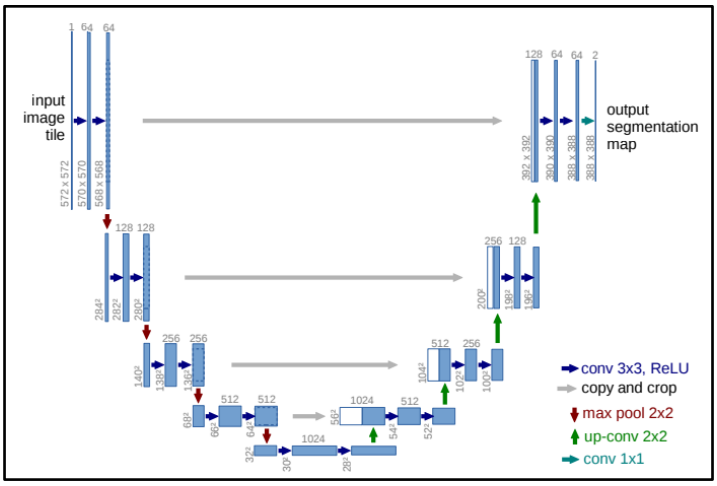


Figure 2.1 Architecture of the proposed U-net model [2].

As shown in Figure 2.1, the architecture of U-net consists of a contracting path that performs downsampling and an expanding path that performs upsampling. For each downsampling block, it consists of two unpadded convolutions with kernel size 3 × 3, each followed by a ReLU activation function. A 2 × 2 max-pooling layer with stride 2 is appended at the end of each block for downsampling. Note that compared to the previous downsampling block, the number of feature channels per convolution is doubled to obtain more semantic knowledge while reducing spatial knowledge. Conversely, each upsampling block starts with the concatenation of the output supplied from the previous block with the corresponding

cropped feature maps in the contracting path, followed by two 3 × 3 unpadded convolutions with ReLU activation and a 2 × 2 deconvolution that reduces the number of feature channels by a factor of two.

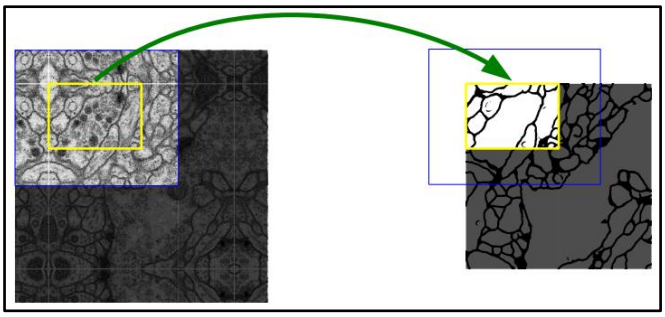


Figure 2.2 Tiling strategy for large image segmentation [2].

Skip connections are introduced in the U-net architecture, where the layers in the contracting path are concatenated with the upsampling layers in the expanding path with a vast range of feature channels. This allows fine-grained features extracted in the downsampling path of the network to be assembled and embraced in successive convolutional layers. Instead of a typical fully connected layers at the end of the network, 1 × 1 convolution with *N* feature channels is used to map the feature vector produced by the previous layer, where *N* denotes the number of possible classes. More importantly, only the effective part of each convolution will be used to allow the tiling strategy to be applied, as shown in Figure 2.2. Segmentation predictions for boundary regions (i.e., the yellow regions in Figure 2.2) are accomplished by extrapolating the missing context using mirroring techniques. To compensate for the lack of data, data augmentation techniques including elastic deformation are applied to force the network to learn invariance to simulated yet realistic deformations.

Stochastic gradient descent (SGD) is implemented to train the network. Since the entire network uses unpadded convolutions, the resolution of the output image is relatively small compared to the input image. Magnitude of momentum is set to 0.99 so that larger proportion of previously seen training images get to decide the update in current optimization step. To enhance the model's ability to split touching objects of the same class and detect objects with lower pixel frequency, weight map is pre-computed and introduced in the cross-entropy loss function to force the network to pay more attention to certain pixels, such as the separation boundaries between touching objects. Moreover, initial weights for each convolution come from a Gaussian distribution with a standard deviation of $\sqrt{2/N}$, where *N* represents the total number of incoming nodes responsible for computing a neuron.

Table 2.1 Segmentation results on EM segmentation challenge (left) and ISBI cell tracking challenge (right) [2]

Rank	Group name	Warping Error	Rand Error	Pixel Error
	** human values **	0.000005	0.0021	0.0010
1.	u-net	**0.000353**	0.0382	0.0611
2.	DIVE-SCI	0.000355	0.0305	0.0584
3.	IDSIA [1]	0.000420	0.0504	0.0613
4.	DIVE	0.000430	0.0545	**0.0582**
⋮				
10.	IDSIA-SCI	0.000653	**0.0189**	0.1027

Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	**0.9203**	**0.7756**

O. Ronneberger, et al. [2] have applied U-net to perform segmentation task on the dataset provided by the EM segmentation challenge, which contains 30 training images of neuronal structures in electron microscopic (EM) recordings. U-Net has achieved a warping error of 0.0003529, a rand-error of 0.0382 and a pixel error of 0.0611 without any further pre- or postprocessing. U-Net outperformed state-of-the-art algorithms in terms of warping error and ranked second concerning rand error. They also applied U-Net to the cell segmentation task proposed by the ISBI Cell Tracking Challenge. The task consists of two datasets, where the first dataset contains 35 partially annotated cell images recorded by light microscopy, while the second dataset contains 20 partially annotated HeLa cell images recorded by differential interference contrast microscopy. The average IoU achieved by U-Net in the two datasets were 92.03% and 77.56%, respectively, which were the best results among other algorithms.

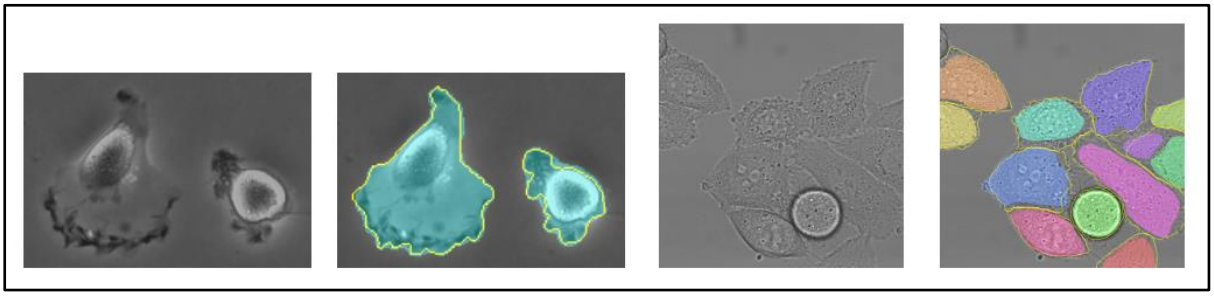


Figure 2.3 Input images and segmentation outputs of U-Net on ISBI cell tracking challenge [2].

In conclusion, U-Net architecture proposed by O. Ronneberger, et al. [2] successfully performs accurate segmentation on a variety of biomedical tasks without the need for extensive annotated images by leveraging data augmentation with elastic deformations. To this day, the U-Net architecture is still considered the de facto standard and has contributed immeasurably to the field of medical image segmentation.

2.1.2 Dense Dilated Inception Network for Medical Image Segmentation

U-Net model with encoder-decoder architecture achieves splendid performance in medical image segmentation task. The architecture is well designed, where the encoder path focuses on feature extraction, and the decoder path performs segmentation based on these extracted features. However, the typical U-Net architecture suffers from some shortcomings, making it difficult to learn multi-scale features and generalize to other tasks. S. A. Bala and S. Kant [3] therefore pointed out that widening and deepening the parameter space would resolve aforementioned issues. To overcome the vanishing gradient effect, and to increase network parameters as well as computational cost that accompany network widening and deepening, they propose an improved U-Net model, also known as Densely Dilated Inception Network (DDI-Net).

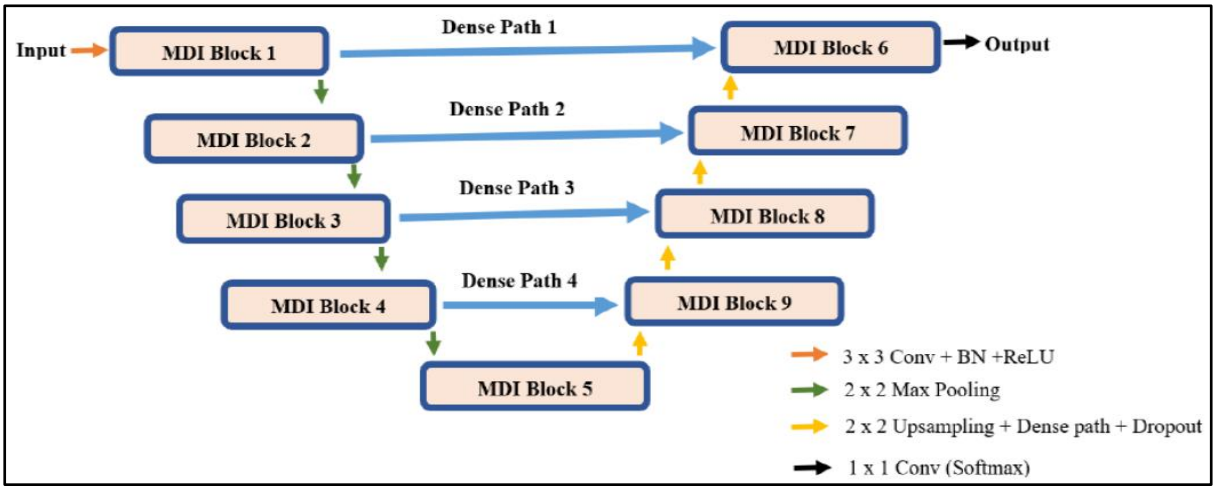


Figure 2.4 Model architecture of DDI-Net [3].

Architecture design of DDI-Net is inspired by U-Net, Dense-Net, Inception module, and Dilation convolution by aggregating dense path and multi-scale dilated inception blocks (MDI) into U-Net. Skip connections introduced in U-Net model cause redundant features to be passed from the encoder part to the decoder part, which in turn causes the segmentation accuracy to suffer. Also, feature maps of encoder and decoder have a large semantic gap, leading to discrepancies when fused together during training. Thus, S. A. Bala and S. Kant [3] replace the skip connections with the dense paths, where each path comprises densely connected 3 × 3 convolution layers and a bottleneck layer. This allows the model to perform in-depth supervision by alleviating low-level features extraction in the encoder and spatial knowledge recovery in the decoder. Dense paths advocate feature reuse and prevent model overfitting with its regularization effect.

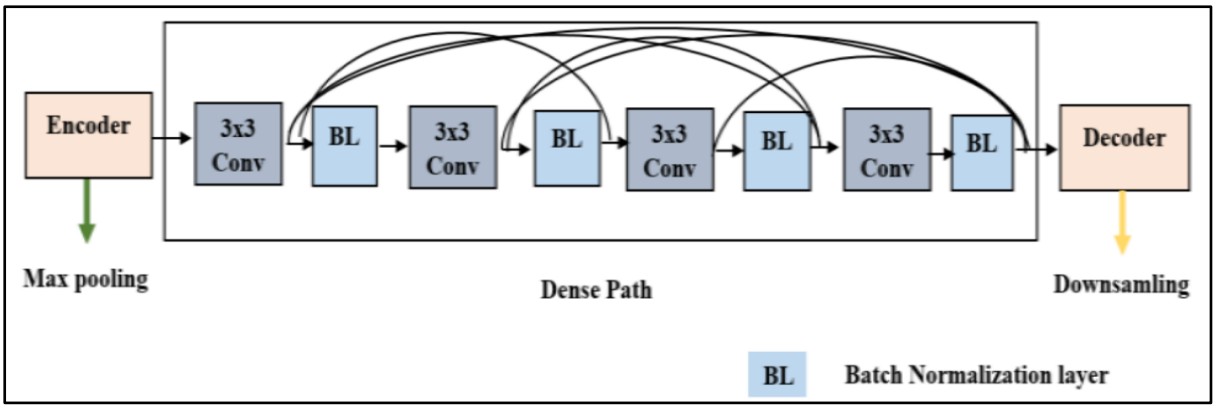


Figure 2.5 Dense path architecture [3].

Objects of interest in medical image segmentation often have multiple scales. A network that can learn and capture multi-scale features is thus necessary. However, the typical inception module designed for multi-scale feature learning utilizes convolutions with large kernel size to capture large scale features, which in turn escalate the number of parameters and computational cost. Therefore, the MDI module containing dilated convolutions is proposed. Similar to large convolution kernel, dilated convolution can capture large-scale features by expanding the receptive field, but it will not increase the parameters and computational cost. MDI module is used in both encoder and decoder paths of DDI-Net to extract and aggregate multi-scale feature maps.

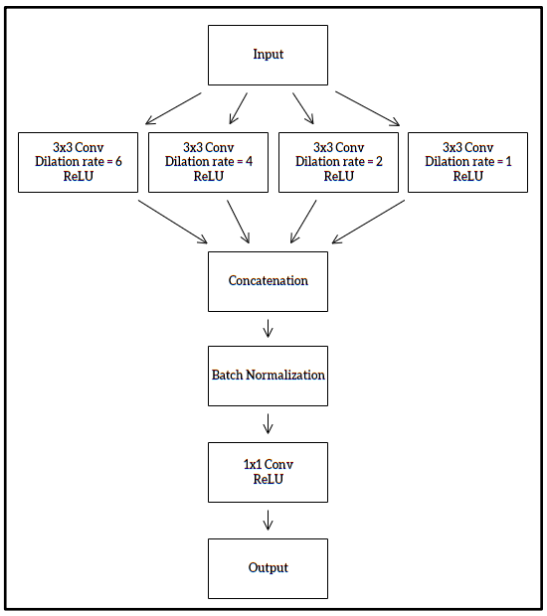


Figure 2.6 Multi-scale dilated inception block architecture [3].

The MDI block starts with four 3 × 3 convolutional layers, each with a different dilation rate (1, 2, 4, and 6). The receptive fields of these convolutions are 3 × 3, 5 × 5, 9 × 9, and 13 ×

13 respectively. Outputs of these convolutions will be concatenated and passed through a batch normalization layer to speed up training and enhance model stability. At the end of MDI block, a 1 × 1 convolution is applied to reduce the feature channels. Each convolutional layer implements ReLU as the activation function. Experiments conducted by S. A. Bala and S. Kant [3] verified that more multi-scale features will be learnt with a relatively low computational complexity and fewer parameters by replacing the original convolution block in U-Net model with proposed MDI block.

$$Dice\ (GT, SR) = \frac{2|GT \wedge \mathrm{SR}|}{(|GT| + |SR|)}, where\ GT = Ground\ truth\ results, SR = Model\ segmentation\ results$$

DDI-Net was tested on brain tumor, hippocampus, and heart segmentation tasks. The brain dataset contains 484 multiparametric MRI scans from patients diagnosed with glioblastoma or low-grade glioma; the hippocampus dataset includes scans from 260 stable adults and adults with non-affective psychosis Images; the cardiac dataset consists of 20 left atrial MRI images. Dice-score is implemented as the evaluation metric for these three tasks and the results are shown in the Table 2.2 below.

Table 2.2 Segmentation results on brain tumor, hippocampus and heart datasets [3]

Brain Tumor Segmentation

Networks	Edema	Non-enhancing	Enhancing
NDN	0.71	0.60	0.72
nnU-Net	0.68	0.48	0.68
DDI-Net (ours)	**0.82**	**0.68**	**0.79**

Hippocampus Segmentation

Networks	Anterior	Posterior
NDN	0.88	0.89
nnU-Net	0.90	0.89
DDI-Net (ours)	**0.92**	**0.90**

Heart Segmentation

Networks	Left Atrial
NDN	0.85
nnU-Net	0.93
DDI-Net (ours)	**0.95**

In conclusion, the results show that the proposed DDI-Net outperformed state-of-arts approaches. The two features — dense path and MDI block allows the model to go deeper and wider without experiencing vanishing gradient problem but with smaller parameter space. By that, it is easier for the encoder to extract low-level features, which relieves the decoder from recovering lost spatial knowledge. The generalization performance of the model in various segmentation tasks will also be improved.

2.1.3 TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation

U-Net architecture which employs a symmetric encoder-decoder network design with skip connections has achieved great success in medical image segmentation. However, as one of the CNN-based approaches, U-Net performs poorly in long-range dependency information learning due to the locality aspect of convolution operations. Thenceforth, Transformers, which acquire strong global interactions modelling capabilities and exhibit superior transferability to downstream tasks, have become favored for alternative architectures. In spite of that, J. Chen, et al. [4] found that directly using a Transformer-alone encoder to develop fine-grained feature representations followed by an upsampling path to restore them to full resolution, did not yield a satisfactory result. This is because Transformers only focus on modelling global context at all layers, resulting in the production of poor-quality features representations which lack detailed localization information. To overcome the preceding limitation, J. Chen, et al. [4] proposed TransUNet which adopts a hybrid CNN-Transformer architecture to spawn features containing fine spatial information and rigid global context for all pixels.

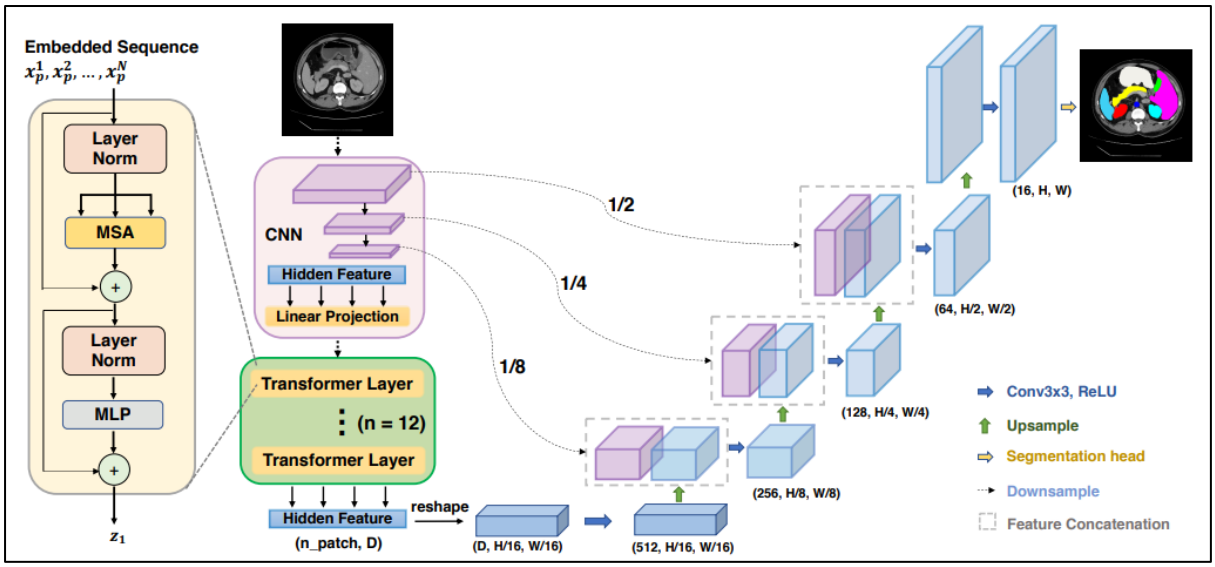


Figure 2.7 Overview of the TransUNet framework [4].

Architectural design of TransUNet is inspired by U-Net, where the features produced by the hybrid CNN-Transformer encoder are upsampled to the input image resolution, with skip connections imported at the intermediate stages to leverage fine-grained features extracted in the downsampling path for enabling precise localization. Different from U-Net, TransUNet aggregates Transformer into the encoder section to enrich global interactions of all pixels in

the feature maps generated by CNN, which acts as an initial feature extractor. This inventive design allows the utilization of intermediate high-resolution CNN feature maps in the decoder section, and thus resulting in better segmentation performance compared to naïve usage of pure Transformer as the encoder.

Prior to the Transformer section, hidden features extracted by the CNN will be reshaped into a sequence of flattened 2D patches, resulting in a total of N image patches of size $P \times P$, where $N = \frac{\text{Resolution of Feature Maps, } \mathrm{H} \times \mathrm{W}}{\mathrm{P}^2}$. Later, the vectorized image patches are mapped into a latent D-dimensional embedding space using a trainable linear projection. In order to preserve the positional information of the patches, the corresponding position embeddings are encoded into the patch embeddings. The Transformer section is made up of L layers, each consisting of a Multihead Self-Attention (MSA) block and a Multi-Layer Perceptron (MLP) block, serving the purpose of enriching each patch in the sequence with other patches and extracting features from the self-enriched features respectively. The ultimate encoded feature representation generated by the Lth layer will be reshape from $\frac{\mathrm{H} \times \mathrm{W}}{\mathrm{P}^2} \times D$ to $\frac{\mathrm{H}}{\mathrm{P}} \times \frac{\mathrm{W}}{\mathrm{P}} \times D$.

There are two approaches to upsample the reshaped feature map to the full resolution H × W, the first of which is to directly bilinearly upsample it after applying a 1 × 1 convolution to reduce the feature channels D to the number of possible classes. However, this naive upsampling strategy leads to loss of low-level contextual information due to the large gap between feature resolution and original image resolution. The second approach is consequently introduced to compensate for such information loss. J. Chen, et al. [4] referred to the second strategy as a cascaded upsampler, where multiple stacked upsampling blocks are cascaded to help escalate the resolution of feature map gradually from $\frac{\mathrm{H}}{\mathrm{P}} \times \frac{\mathrm{W}}{\mathrm{P}}$ to $H \times W$. Each upsampling block comprises an upsampling layer with a factor of 2, a 3 × 3 convolution layer, and a ReLU layer. This method enables the use of skip connections, thereby facilitating feature aggregation at different resolution levels.

J. Chen, et al. [4] evaluated the proposed TransUNet framework under various settings and eventually came up with several optimal configurations as to number of skip connections, input resolution, sequence length and patch size, and model scaling. Previous experimental studies have shown that segmentation performance generally fluctuates with the number of skip connections. Also, extending the input resolution which in turn increasing the sequence length N shows robust improvements, but comes with a trade-off in computational cost. Since

the sequence length *N* is inversely proportional to the square of patch size, a smaller patch size implies a better segmentation performance in general. An explanation has been raised where longer input sequences encourages the modelling of more complex dependencies between each element. Lastly, larger model size with a significant number of parameters guarantees better performance at the expense of larger computational cost.

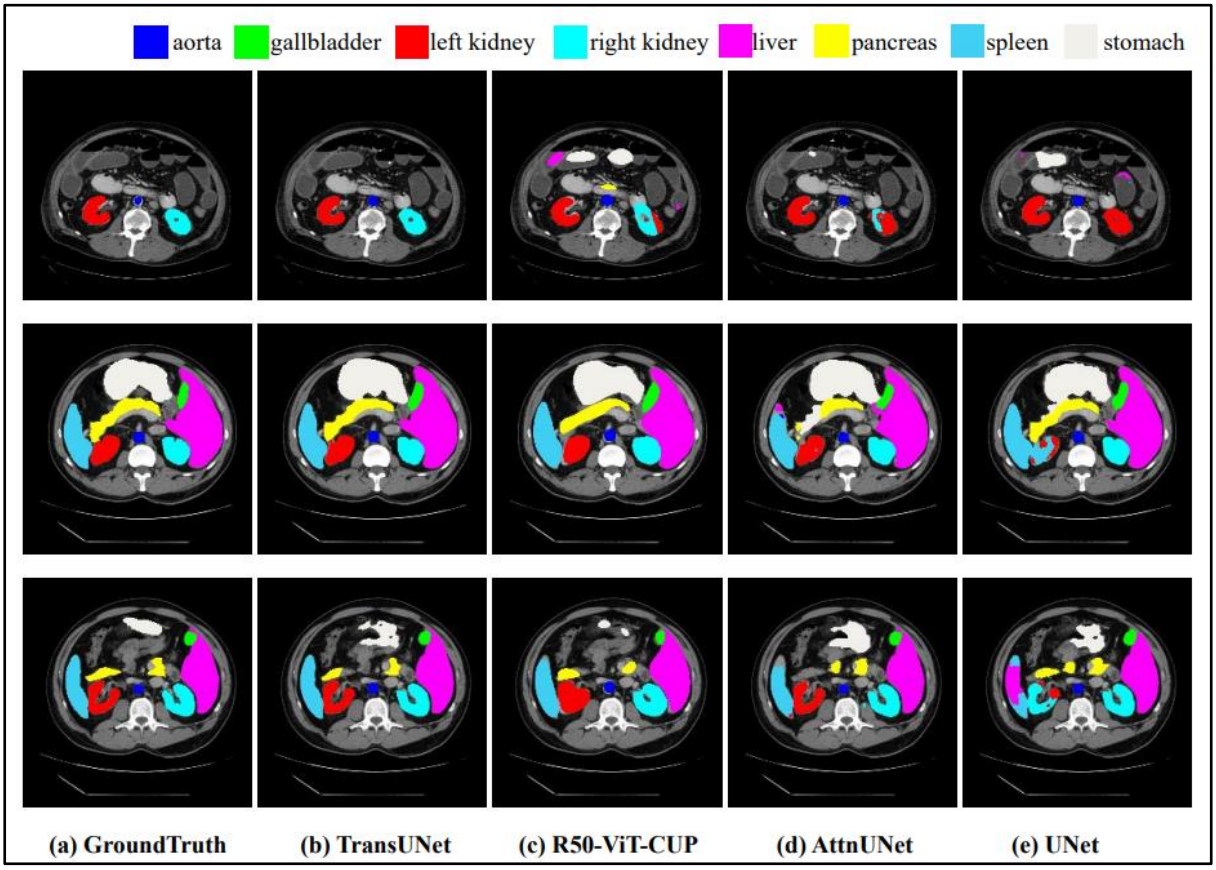


Figure 2.8 Qualitative comparison of TransUNet and other frameworks on the Synapse dataset by visualization [4].

From the results shown in Figure 2.8, it is observed that TransUNet possesses a stronger ability to suppress under-activated and over-activated regions on the final segmentation mask, indicating that it outperforms other frameworks in global context modelling and semantic distinguishment. Additionally, chaste prediction of both left and right kidneys by TransUNet demonstrates its ability to generate finer segmentations and retain structural information. In a nutshell, TransUNet enjoys the advantages of both strong global context and fine-grained CNN features via a hybrid CNN-Transformer architecture together with skip connections, thereby achieving performance superior to other well-known frameworks.

2.1.4 Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation

Recently, weakly supervised semantic segmentation that utilizes weak supervisions has been highly generalized compared to the traditional fully supervised setting as it is difficult to generate accurate pixel-level class-specific annotations for the entire dataset images, especially when the dataset is large-scale or the objects of interest are miniature. Class activation mapping concept proposed by B. Zhou, et al. [5], which projects the weights of the final dense layer on to the convolutional feature maps, became the standard method for locating objects by using image classification labels, and has been continuously refined and improved by researchers. Nonetheless, the end product, class activation maps (CAMs) generated by this concept has two serious limitations, in which CAMs are susceptible to under-activation and over-activation, and are inconsistent when derived from same input images at different scales due to the supervision gap between fully and weakly supervised semantic segmentation. Therefore, Y. Wang, et al. [6] proposed a self-supervised equivariant attention mechanism (SEAM), in association with the implementation of Siamese network to cope with aforementioned limitations.

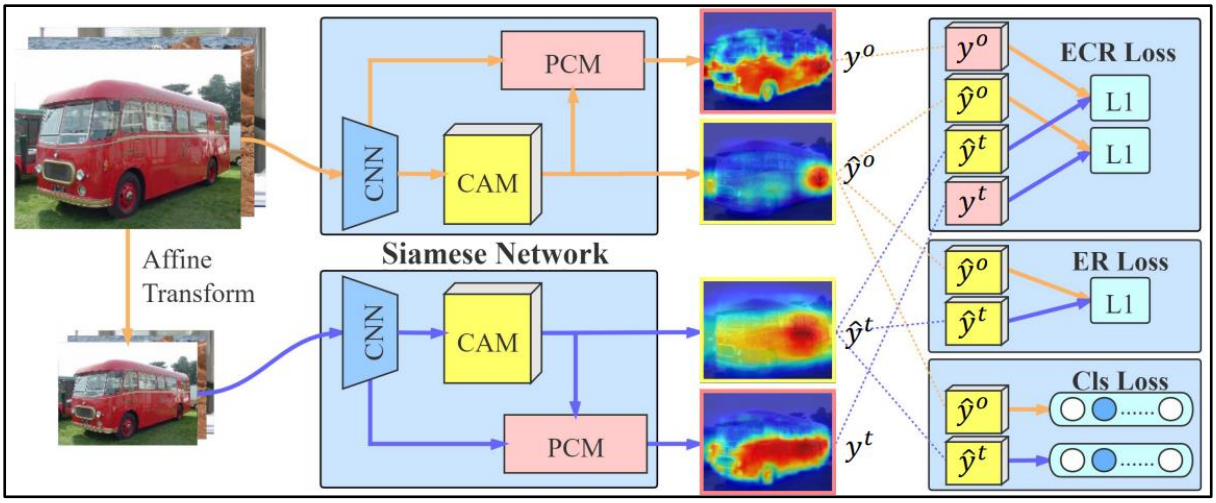


Figure 2.9 The Siamese network architecture of SEAM method [6].

The significant inconsistency issue with CAMs is that implicit equivariance constraint cannot be introduced into the network since image-level classification labels are not affected by affine transformations including rescaling. A shared-weight Siamese structure is therefore proposed to integrate equivariant regularization on the network. One branch applies the desired transformation to augment the input images before feeding them to the network, while the other branch applies the same transformation on the output activation maps. In the end, the output activation maps from both branches are regularized as follows [6]:

$$R_{ER} = |F(A(I)) - A(F(I))|, where\ F(\cdot)\ denotes\ the\ network, A(\cdot)\ denotes\ the\ affine\ transformation \quad (1)$$

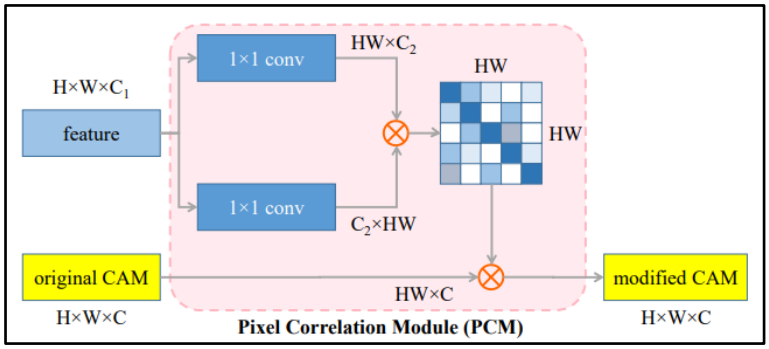


Figure 2.10 The structure of PCM [6].

To improve the contextual information of original CAMs, a self-attention module alike, Pixel Correlation Module (PCM) is introduced at the end of the network to pose global interactions of all pixels, resulting in finer pixel-wise prediction results. Unlike the classic self-attention module, PCM adopts cosine distance to compute inter-pixel similarity matrix, along with ReLU activation function and L1 normalization to suppress over-activated background regions. Furthermore, PCM removes residual connections and several embedding functions to respectively preserve the activation strength of the original CAMs and avoid overfitting on imprecise pixel-level supervision. Overall operation of PCM can be described by the following equation [6]:

$$y_i = \frac{1}{c(x_i)} \sum_{\forall j} ReLU\left(\frac{\theta(x_i)^T \theta(x_j)}{|\theta(x_i)| \cdot |\theta(x_j)|}\right) \hat{y}_j, where\ y_i\ denotes\ the\ original\ CAM, \hat{y}_j\ denotes\ the\ revised\ CAM \quad (2)$$

Each branch of the network will have a global average pooling layer appended at the end to vectorize the CAM for image-level classification purposes. The classification loss will be the average loss of the two branches, where the network is trained with a multi-label soft-margin loss on each branch. Apart from that, an equivariant regularization (ER) loss is applied to ensure the consistency of CAMs from two branches. Note that the original CAMs are utilized for the aforementioned loss calculations instead of revised CAMs that have went through PCM module as the revised CAMs fall into a local minimum so quickly that all pixels in the image are predicted to be of the same class. In order to improve the quality of CAMs while avoiding CAM degradation during PCM processing, an equivariant cross regularization (ECR) loss that utilizes the original CAM from another branch to regularize the PCM output maps is proposed. The final loss of SEAM is obtained by summing the classification loss, ER loss and ECR loss, where classification loss supervises the generation of grainy CAMs, ER loss ensures consistent predictions and ECR loss aggregates PCM module with the backbone of the network.

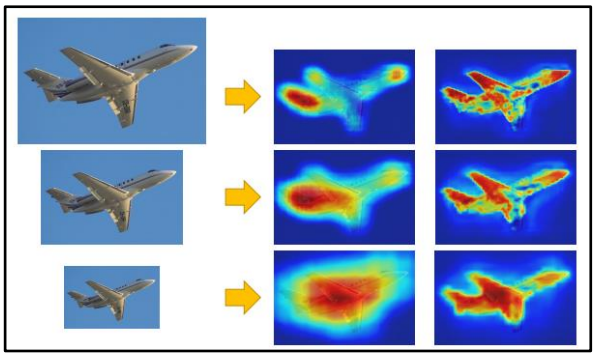


Figure 2.11 Consistent CAMs prediction by SEAM over scaling transformation [6].

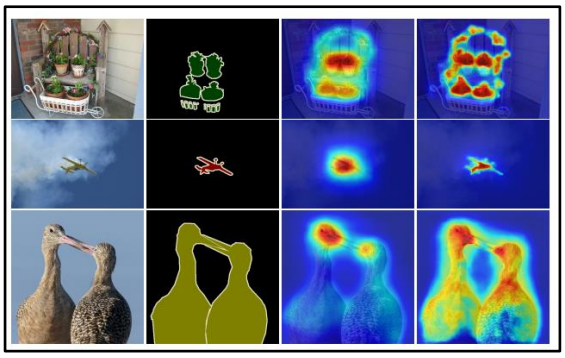


Figure 2.12 Fine CAMs prediction by SEAM with full object activation coverage [6].

Figure 2.11 visualizes the effect of input image rescaling on conventional CAMs (col. 2) and CAMs generated by the SEAM approach (col. 3). Traditional CAMs suffers from severe inconsistency issue, whereas CAMs produced by SEAM method that emphasize equivariant regularization and narrow the supervision gap between fully supervised and weakly supervised settings do not. Besides, Figure 2.12 depicts that CAMs generated by SEAM (col. 4) have fewer extraneous background regions that are falsely activated compared to conventional CAMs (col. 3), all thanks to the installation of PCM module that enables the learning of semantic boundary information from self-supervision.

Table 2.3 Performance comparisons on PASCAL VOC 2012 dataset [6]

Methods	Backbone	Saliency	*val*	*test*
CCNN [25]	VGG16		35.3	35.6
EM-Adapt [24]	VGG16		38.2	39.6
MIL+seg [27]	OverFeat		42.0	43.2
SEC [19]	VGG16		50.7	51.1
STC [33]	VGG16	√	49.8	51.2
AdvErasing [32]	VGG16	√	55.0	55.7
MDC [34]	VGG16	√	60.4	60.8
MCOF [36]	ResNet101	√	60.3	61.2
DCSP [4]	ResNet101	√	60.8	61.9
SeeNet [15]	ResNet101	√	63.1	62.8
DSRG [16]	ResNet101	√	61.4	63.2
AffinityNet [2]	ResNet38		61.7	63.7
CIAN [10]	ResNet101	√	64.1	64.7
IRNet [1]	ResNet50		63.5	64.8
FickleNet [21]	ResNet101	√	64.9	65.3
Our baseline	ResNet38		59.7	61.9
Our SEAM	ResNet38		64.5	65.7

Y. Wang, et al. [6] evaluated proposed SEAM model on PASCAL VOC 2012 dataset, which achieved state-of-the-art performance in terms of mean intersection over union, without possessing a relatively large network structure or an improved saliency detector. Hence, it is verified that the CAMs generated by SEAM are of better quality, not to mention the consistency enforcement, which allows the generated CAMs to be consistent across different transformed input images, while more tending towards ground truth segmentation masks.

2.2 Limitation of the Previous Studies

By referring to the model evaluation results of the previous state-of-arts that have been mentioned, all of the proposed models were capable to complete the segmentation tasks with a good performance. However, for the first three studies, the authors only constructed, trained and validated the proposed models based on the objects of interest in specific field. U-Net proposed by O. Ronneberger, et al. [2] only applied on biomedical segmentation applications; DDI-Net proposed by S. A. Bala and S. Kant [3] was limited to MRI image segmentation tasks; TransUNet proposed by J. Chen, et al. [4] served for general medical image segmentation only. Most importantly, none of the models proposed in previous studies have been evaluated in the defect segmentation task, indicating that these models may not perform as well as the authors have claimed when extended to work on surface defect segmentation tasks. For instance, organs are often larger in size and more pronounced than general surface defects. Consequently, the network architecture that used to perform multi-organ segmentation not be liable in detecting relatively small-scale defects, that is, surface defects in this case. Thus, most of these proposed models are reproduced, modified and improved in this study to verify their effectiveness.

Besides, each proposed model has some drawbacks. According to S. A. Bala and S. Kant [3], U-Net architecture encounters difficulties in learning multi-scale features due to the restrictions to the constant 3×3 convolution filters. Furthermore, the network is not robust enough, resulting in poor generalization performance when dealing with other segmentation tasks. On the other hand, convolutions with high dilation rate (4 and 6) are incorporated in the MDI block of DDI-Net, which is unfavorable to miniature object detection. R. Hamaguchi, et al. [7] highlighted that aggressively increasing dilation factors may fail to aggregate local features of small objects, resulting in poor model performance. Each Transformer layer usually requires a relatively large number of parameters, not to mention TransUNet has a total of 12 Transformer layers plus a CNN feature extractor at the beginning stage. It is foreseeable that TransUNet has the disadvantage of being computationally expensive due to a large number of parameters need to be trained. This exacerbates slow model training and inference problems. Lastly, SEAM is a complex framework with multiple regularization operations incorporated in the network. The final CAMs generated by SEAM during inference phase is expected to be mediocre than most segmentation masks predicted by fully supervised models, since SEAM only uses image-level labels as learning inputs during training stage.

CHAPTER 3 PROPOSED METHOD / APPROACH

The development process of the project can be divided into different stages, which were project pre-development, data pre-processing, model architecture construction, model training and model evaluation.

3.1 System Requirement

3.1.1 Tools to Use

The hardware(s) that has been for developing this system is:

1. Laptop

Table 3.1 Specifications of laptop

Description	Specifications
Model	Dell Inspiron 14 5480
Processor	Intel Core i5-8265U CPU @ 1.60GHz 1.80 GHz
Operating System	Windows 11
Graphic	NVIDIA GeForce MX250
Memory	12GB DDR4 RAM
Storage	1TB SATA HDD

The software(s) that has been used for developing this system is:

1. Anaconda

   Anaconda is a Python and R languages distribution which mainly serves for scientific computing. It provides common data science and deep learning tools such as scikit-learn and SciPy libraries to the end user. In this project, Jupyter Notebook, an available web-based integrated development environment (IDE) in Anaconda, is used for live coding, data visualization, model training, etc.

2. TensorFlow

   TensorFlow is an open-source machine learning framework developed by the Google Brain team. It has accumulated a series of machine learning and deep learning models, tools and algorithms that can help developers build, train and deploy machine learning models. Moreover, it allows developers to build and design their own neural networks by using dataflow graphs.

3. Keras

   Keras is an open-source Python library which runs on top of the TensorFlow platform and is designed for enabling fast-paced development and evaluation of deep neural networks by fast prototyping.

4. Google Colab

   Google Colab allows the developers to write and run Python in an interactive environment known as Colab notebook. By leveraging the platform, developers have access to free GPUs provided by Google to speed up code execution without any complex configuration compared to local CPU.

3.1.2 User Requirements

- User should be able to see the segmentation map generated by the models.
- User should be able to recognize the classes of the segmented defects.
- User should be able to identify the coverage area of the segmented defects in the input image.
- User should be able to obtain correct and accurate segmentation results from the models.

3.1.3 Verification Plan

The constructed defect segmentation models are anticipated to detect, segment and classify defects with high accuracy. Nonetheless, faults such as false segmentation and classification error may arise due to the changing circumstances. The suspected cases which may lead to such failures are listed as follows:

1. Defects occur on defective item with cluttered surface as displayed in Figure 3.1. The models potentially could not recognize the defect from the occluded background properly.
2. Similar appearance for different defect classes as displayed in Figure 3.2. The models could be confused and fail to correctly classify the defects.
3. Small-scale defect classes with an extremely small coverage area in the image as displayed in Figure 3.3. The models could neglect the defect or go to extremes and incorrectly predict larger coverage areas.

Therefore, a few verification steps are conducted to simulate the situations listed above, and ensure the capability of the models to counteract these possible failures. All verification test cases are documented in the following tables, with each representing the corresponding verification plan.

a) Cluttered surface

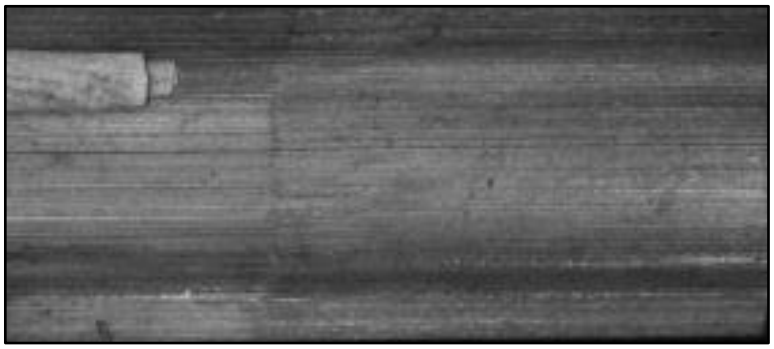


Figure 3.1 Defective magnetic tile with a cluttered surface.

Table 3.2 Verification P1

Procedure Number	P1
Method	Testing
Applicable Requirements	Recognize the defects in the image without distracted by the occluded background
Purpose/ Scope	To recognize only the actual defects on a cluttered surface
Items Under Test	Magnetic tile defect and production item surface defect
Precautions	The size of the defect in the image should be adequate
Special Conditions / Limitations	If the size of the defect is too small, the result presented may not be consistent
Acceptance Criteria	The models successfully segment and classify the defects in the image accordingly
Procedures	1. Feed images with defect(s) as well as occluded background into the models 2. Identify whether defect(s) is successfully segmented by the model through visualization of prediction results
Troubleshooting	Repeat the procedure

b) Different defect categories with similar appearance

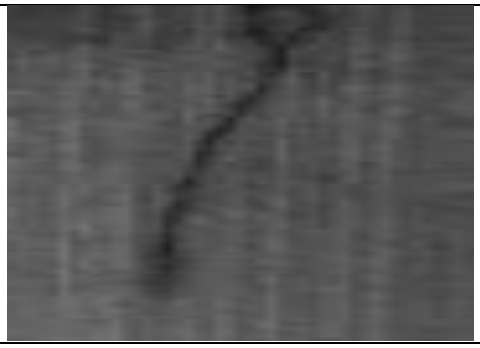
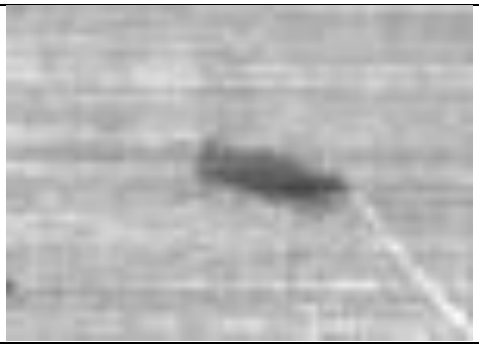
Crack	Blowhole
	

Figure 3.2 Similar-looking magnetic tile defects.

Table 3.3 Verification P2

Procedure Number	P2
Method	Testing
Applicable Requirements	Classify different defect types with similar appearance in the image
Purpose/ Scope	To classify all the defects appeared in the image accordingly
Items Under Test	Magnetic tile defect and production item surface defect
Precautions	The brightness of the image should be adequate
Special Conditions / Limitations	If the brightness of the image is too low or too high, the result presented may not be consistent
Acceptance Criteria	The models successfully segment and classify the defects in the image accordingly
Procedures	1. Feed images with similar looking defects into the models 2. Identify whether the defects are correctly segmented and classified by the model through visualization of prediction results
Troubleshooting	Repeat the procedure

c) Miniature defects

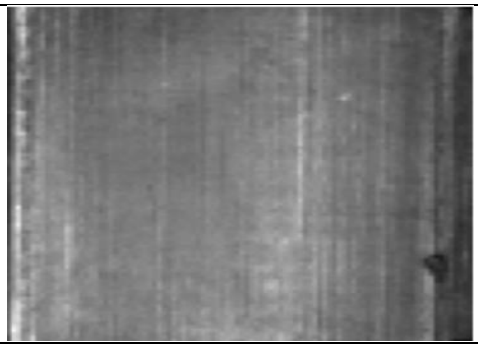
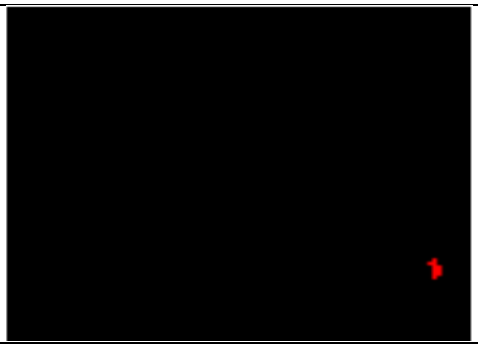
Image	Ground Truth Mask
	

Figure 3.3 Small-scale defect in image.

Table 3.4 Verification P3

Procedure Number	P3
Method	Testing
Applicable Requirements	Recognize the defects in the image regardless of their size or coverage area
Purpose/ Scope	To segment and classify all the defects appeared in the image accordingly
Items Under Test	Magnetic tile defect and production item surface defect
Precautions	The brightness of the image should be adequate
Special Conditions / Limitations	If the brightness of the image is too low or too high, the result presented may not be consistent
Acceptance Criteria	The models successfully segment and classify the defects in the image accordingly
Procedures	1. Feed images with extremely small defects into the models 2. Identify whether the defects are correctly segmented and classified by the model through visualization of prediction results
Troubleshooting	Repeat the procedure

3.2 System Design / Overview

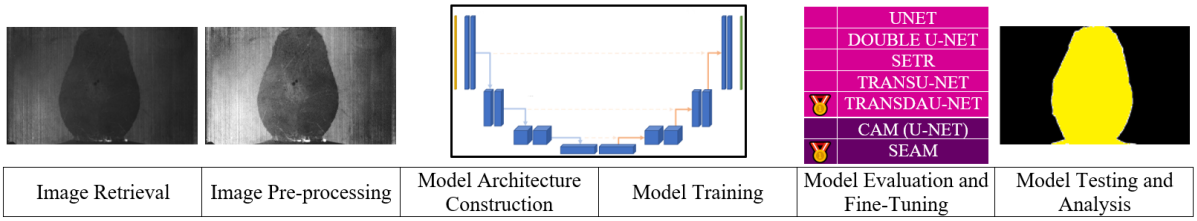


Figure 3.4 Block diagram of system design.

3.2.1 Image Retrieval

In this study, a magnetic tile defect dataset and a comprehensive production item defect dataset (KolektorSDD2), are referenced and used in the model training process, where the former dataset is originated from a journal authored by Y. Huang, et al. [8], and the latter dataset was made public by J. Božič, et al. [9]. The defects in these datasets are being labelled in accordance with their pixel level ground truth. In addition, the classes in the magnetic tile dataset are categorized into 6 categories, namely blowhole, crack, break, fray, uneven and free, each with different visual characteristics and properties, while the classes in the production item dataset are only divided into defective and non-defective. The magnetic tile dataset contains 1344 grayscale images, of which 952 images are defect-free; the production item dataset consists of 3335 color images, of which 2979 are defect-free. Note that both datasets will be adopted as data inputs to the models under fully supervised setting, while only production item dataset will be employed for weakly supervised learning models.

Table 3.5 Details of the magnetic tile defect dataset and production item defect dataset used

Feature \ Dataset	Magnetic Tile	Production Item
Total number of images	1344	3335
Anomaly images	392	356
Defect-free images	952	2979
Defect category	6	2
Image property	Grayscale	Color
Image size (pixels)	Vary	≈ 230 x 630

3.2.2 Image Pre-processing

Image pre-processing is a critical step to improve the image data before it is used as a model input. Segmentation results will vary depending on the implementation of different pre-

processing methods. In this project, all collected images from two datasets will be undergoing several pre-processing stages, including image resizing, pixel normalization and augmentation.

Since the magnetic tile dataset comprises images of vary resolution, and the production project dataset contains images of different heights and widths, image resizing is necessary to universalize the size of the images in both datasets. Furthermore, resolution of the images in training dataset is very decisive during model training. Training on images with small resolution may cause the inability of model to capture and learn context features, and the network of the model would be improbable to go very deep, while training on large resolution images will result in immense parameter space, leading to slow model training and high computational cost. After considering all the above factors, 96 × 96 pixels is chosen as the new image size for all images in both datasets. Annotations for the ground truth magnetic tile defects and production item surface defects will also be adjusted based on the resizing amplitude. In later process, a Siamese network will be constructed that accepts two input images of different sizes during the training phase. Hence, the production item dataset containing images resized to 48 × 48 pixels was replicated to feed forward in a later stage of model training.

Application of pixel normalization on dataset images is meant for modifying the pixel intensities to have similar distributions, thereby making the model easier to converge during training. Since the minimum and maximum pixel values of an image are 0 and 255 respectively, dividing all pixel values by 255 will scale the pixel intensities to a smaller range of 0 to 1. Therefore, the division step is carried out across all of the images in the magnetic tile and production item datasets as well as the ground truth masks.

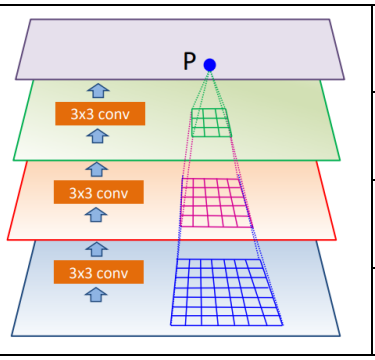
According to O. Ronneberger, et al. [2], data augmentation technique enhances the learning of invariance and robustness properties by the model when training samples are puny. Additionally, the number of defect-free images is much higher compared to the defect images in magnetic tile and production item datasets, which may cause models trained in subsequent steps to be biased towards predictions of defect-free pixels. For these reasons, in the case of magnetic tile images and masks, augmentation techniques such as flipping, cropping, padding are being applied to expand the dataset and ensure at least 1500 images per defect class. On the other hand, the production item training dataset is extended to 4791 images, of which 2085 are defect-free, by applying data augmentation techniques including translation, blurring, inversion, channel shuffling, brightness adjustment, etc. This critical step will compensate for insufficient training data and balance the number of each class, resulting in better model performance.

3.2.3 Model Architecture Construction

At this stage, we refer to and draw inspiration from the architecture design of state-of-arts mentioned in Chapter 2 to construct seven models, namely U-Net, Double U-Net, SETR, TransU-Net. TransDAU-Net, CAM and SEAM. The first five models are built in a fully supervised setting, in contrast, the remaining two models are meant for the weakly supervised semantic segmentation tasks. In fact, architecture of these models is an adaptation of a typical fully convolutional network.

Unlike common convolutional neural networks, fully convolutional network adopts an encoder-decoder architecture and replaces fully connected layers at the end of neural network with 1 × 1 convolution to classify each unit without losing the dimensionality of the input image. Generally, semantic segmentation models have similar widths and heights for input images and output maps. For instance, a 96 × 96 × 3 input image results in a 96 × 96 × c output map, where c represents the number of possible classes. This setup allows the model to learn feature representations and perform per-pixel classification based on local spatial inputs. However, fully convolutional network only supports fully supervised setting, where ground-truth segmentation map will be provided for comparison with the predicted output map at the end of the network. It is not applicable for weakly supervised learning approaches where only image-level annotations are provided as ground truth labels. Thus, in weakly supervised setting, 1 × 1 convolution will be removed, a global average pooling layer and a fully connected layer will be appended at the end of the fully convolutional network to reduce the rank of the output maps for spatial dimensions from 96 × 96 × C to c and perform image-level classification.

U-Net model is built with reference to the standard U-Net architecture proposed by O. Ronneberger, et al. [2]. For the sake of convenience, padding is implemented for each convolution in both downsampling blocks and upsampling blocks. This indicates that feature map cropping for encoder-decoder layer concatenation is no longer required, as boundary pixels will remain in every convolution. The output segmentation map of the model will have the same height and weight as the input image. Most importantly, the trademark of U-Net, skip connections connecting intermediate fine-grained feature maps with upsampling layers in the decoder section, are introduced and designed according to the original work. The integration of high-resolution feature maps and corresponding decoder layers maximizes the low-level information flow in the network, resulting in a more compact model with better segmentation performance.



Activation *p* at the 3rd activation map
Receptive field of *p* in the 2nd activation map = 3 × 3
Receptive field of *p* in the 2nd activation map = 5 × 5
Receptive field of *p* in the 1st activation map = 7 × 7

Figure 3.5 Receptive field of stacked 3 × 3 convolutional layers.

Receptive field is defined as the region of the input space that affects each element in the output activation map. For convolutional layer, the filter size coincides the receptive field of the layer. For instance, a 3 × 3 convolutional layer has a 3 × 3 receptive field. As a consequence, the receptive field increases linearly with the number of parameters in a convolutional neural network, implying greater computational costs if large filters are needed to expand receptive field to capture large-scale features.

Dilated convolution is therefore introduced to mitigate aforementioned problem. Nodes in the receptive field of a dilated convolutional filter are separated by r positions, where r is the dilation rate. Therefore, a larger input area will be covered by the same kernel size compared to convolution filter. For instance, a 3 × 3 dilated kernel with $r = 2$ has a 5 × 5 receptive field. This allows the receptive field to grow exponentially with the number of parameters, saving considerable computational cost. By replacing the convolutions in inception block invented by C. Szegedy, et al. [10] with dilated convolutions, multi-scale dilated inception (MDI) block proposed by S. A. Bala and S. Kant [3] debuts, promoting multi-scale features learning and receptive field expansion.

However, there are some drawbacks with the original MDI block. Due to the fact that S. A. Bala and S. Kant [3] incorporated convolutions with dilation rate $r = 6$ into the MDI block, inputs for the kernels are restricted to have a resolution of $\geq 13 \times 13$. Besides, although dilated convolution successfully expands receptive field, there are a lot of skipped pixels when a highly dilated convolution is used. This is detrimental to small objects segmentation as the model may fail to capture small-scale features, resulting in high false positive rates for these small-scale object classes. Therefore, modifications are made to the architecture of MDI blocks to embrace both sizable and miniature object detection by utilising convolutions with smaller dilation rate.

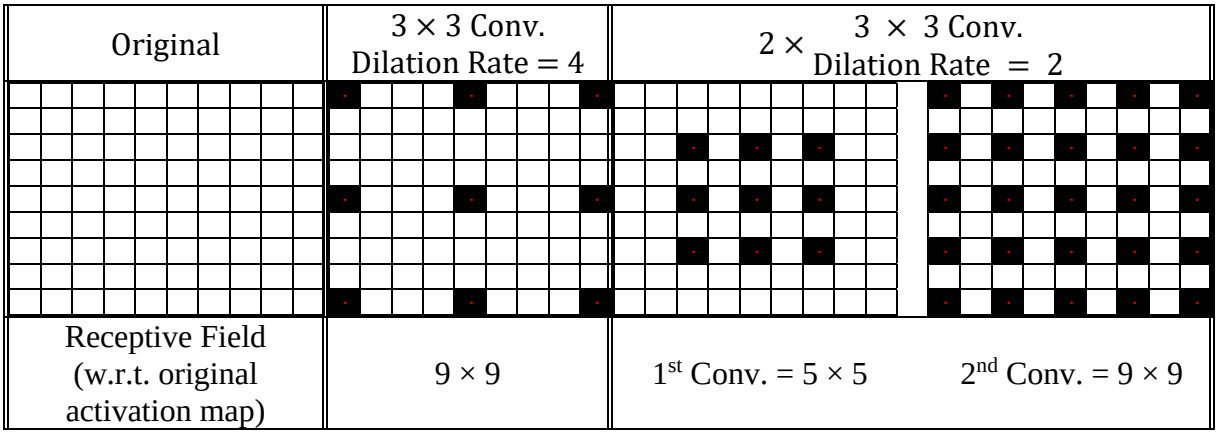


Figure 3.6 Receptive field of 3 × 3 kernel(s) with dilation rate of 2 and 4.

Instead of using convolutions with high dilation rates in MDI block, we stack multiple convolutions with small dilation rates. This approach reduces the number of skipped pixels while maintaining the benefit of using dilated convolution — enlarging the receptive field without incurring higher computational cost. Moreover, kernel inputs have a relatively small restrictions in terms of resolution, which is 5 × 5. Figure 3.6 illustrates that two stacked 3 × 3 convolution with $r = 2$ have the same receptive field as a 3 × 3 convolution with $r = 4$. In addition, the number of channels of dilated convolutions included in the modified MDI block is not fixed. As the network goes deeper, the number of channels per convolution increases. This is because the encoder gradually enhances the semantic knowledge with the depth of the network while neglecting the spatial knowledge. Since the modified MDI block has different depths in each path, it is more appropriate to call it a multi-depth dilated inception block.

Double U-Net model is a stack of two U-Nets, where the first U-Net is the pre-trained U-Net model mentioned earlier, while the second U-Net integrates multi-depth dilated inception blocks into the contracting path. By removing the last two layers, the first constructed U-Net is immediately attached to the top layer of second U-Net. Consequently, the model performs a two-stage encoding-decoding action, improving the process of learning contextual and spatial knowledge from additional low and high-level multi-scale features. Robustness of the network can also be guaranteed by deeper and wider architectural design. Apart from that, dilated convolutions are used extensively in the second U-Net, including layers in the bottom layer and in the expansion path. The receptive field of the network is hence dramatically expanded without placing a burden on the computational cost.

The SETR model architecture constructed in this project is a replica of SEgmentation TRansformer model proposed by S. Zheng, et al. [11]. SETR model breaks the traditional

nature of fully convolutional network by substituting consistent downsampling convolutional-based layers in encoder section with pure Transformer layers. This addresses the limitation of CNN in modelling long-range pixel-to-pixel dependency information due to its intrinsic locality nature. Different from the processing flow in the encoder part of the fully convolutional network, in SETR, input image is first partitioned into fixed-size patches, which are then linearly embedded by a convolutional layer and added with respective positional embeddings. Later, the resulting feature embedding vectors are feedforwarded to the Transformer-alone encoder. Each layer in the transformer encoder models the global interactions of all image patches while retaining the spatial information. As usual, to restore the resolution to the original input image, a progressive upsampling operation similar to the decoder in U-Net is applied to the fully attentive feature representation outputted from the encoder.

Although SETR overcomes the limitation of ordinary fully convolutional network with CNN-based encoder in modelling long-range dependencies at all layers, impulsive projection of input image to the Transformer encoder and lack of mechanisms to enrich the localization information in generated features leads to suboptimal segmentation performance, as observed by J. Chen, et al. [4], who have therefore proposed TransU-Net which combines the strengths of both CNN and Transformer via a hybrid CNN-Transformer architecture. The architectural design of TransU-Net starts with a common CNN encoder to extract low-level contextual information, followed by a Transformer encoder to model strong global context, and finally a decoder to recover the spatial information. To further improve the model performance, skip connections are also introduced in the intermediate layers, thereby enjoying the advantages of the typical U-Net model. For research purposes, TransU-Net framework is also reproduced in this study.

Most of the Transformer-based models, including TransU-Net, acquire relatively large parameter sizes, which are significantly more computationally expensive compared to basic CNN-based approaches. Thus, in this project, an improved model, TransDAU-Net is developed, with lower requirements in parameter size while achieving similar or better detection results. To improve the parameter efficiency of Transformer, the parameter sharing technique, CYCLE (REV) proposed by T. Sho and K. Shun [12] is adopted in TransDAU-Net, which promotes the reuse of parameters in the Transformer layers in reverse cyclic order. It is claimed that the parameter-sharing Transformer will achieve better performance than the vanilla Transformer when they have a similar parameter size. Since the aim of applying this technique is to narrow

the parameter size, the number of Transformer layers in Transformer encoder is reduced by 1/4 in TransDAU-Net compared to TransU-Net. To compensate for the possible performance loss due to the reduced number of Transformer layers, the classical skip connection mechanism is further enhanced. First, multi-depth dilated inception block is added in each skip connection to enlarge the receptive field, which in turn supports large-scale feature extraction of the network. Considering that the aggressive use of dilated convolutions may lead to difficulties in micro-object detection, each skip connection is split into two parallel paths, only one of which passes through the multi-depth dilated inception block. Moreover, at the end of each path, an attention module is introduced to actively suppress under-activation and over-activation in inappropriate regions, thereby alleviating the problem of poor representation of features transmitted in the network. The design of attention module refers to the attention gate incorporated in Attention U-Net proposed by O. Oktay, et al. [13].

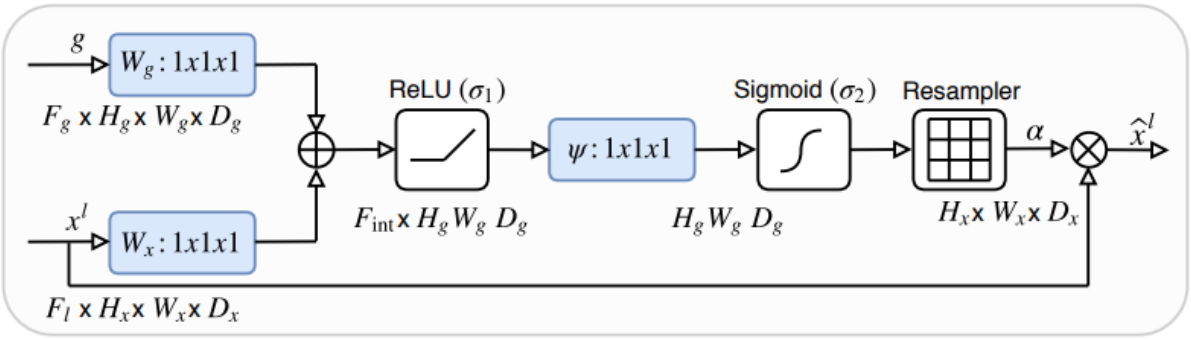


Figure 3.7 Attention gate (AG) schematic [13].

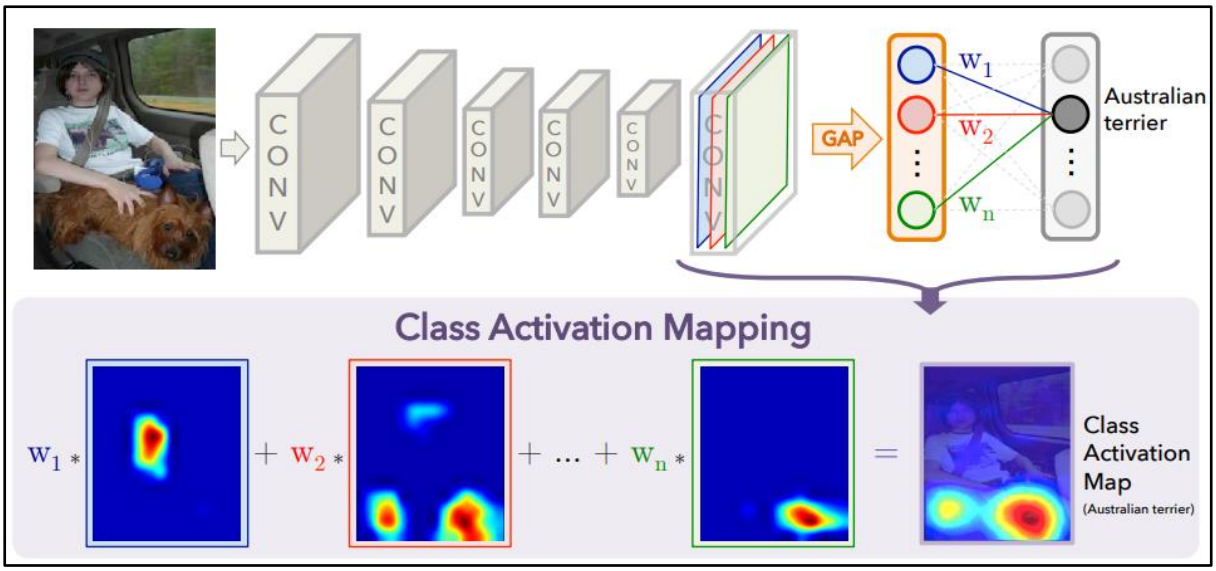


Figure 3.8 Class activation mapping technique [5].

This study builds two weakly supervised semantic segmentation models, CAM and SEAM, which use only image classification labels to segment objects of interest from irrelevant backgrounds. The first weakly supervised model, CAM, adopts U-Net as the backbone and uses the class activation mapping technique proposed by B. Zhou et al. [5] to generate class activation maps (CAMs) that highlight class-specific discriminative regions. Class activation mapping technique harnesses the ability of convolutional units to localize objects without any supervision. By projecting the weight of the final fully connected layer, which represents the importance of each feature map channel to class *c*, to last convolutional layer and compute a weighted sum of the feature maps, class-specific CAMs are generated.

The SEAM framework proposed by Y. Wang, et al. [6] that has already been introduced in Chapter 2, is reproduced in this study. To recap, SEAM has a Siamese network architecture combined with a PCM module, which resolves the inconsistency problem raised by the class activation mapping technique due to the supervision gap between full supervision and weak supervision, and suppresses the under-activation and over-activation of irrelevant regions, respectively. Both paths of Siamese network adopt U-Net as the backbone network. In addition, since U-Net is adopted as the backbone for both CAM and SEAM models, upsampling operation to restore the resolution of CAMs to the original input image size is no longer needed. By performing a min-max normalization followed by a thresholding operation, CAMs are transformed into normal segmentation maps.

Except SETR, all built models use pre-trained VGG-16 model as the encoder/backbone of the network (only the first U-Net for Double U-Net). By that means the weights of the convolution filters in these encoders are initialized from the VGG-16 model, which has been pre-trained on ImageNet benchmark with 1000 classes over 1.28 million training images. This approach, also known as transfer learning, prevents overfitting and improves the generalization performance of the trained model. Furthermore, those initialized weights are frozen and not updated by the optimizer to reduce the number of trainable parameters and thus scale down the computational cost.

3.2.4 Model Training

After applying augmentation techniques, the magnetic tile dataset has been expanded to contain 9632 images. Before feeding the images and labels to the models, the binary ground truth masks of magnetic tile dataset are transformed in such a way that each pixel is encoded according to the class it belongs to, with numeric value ranging between 0 and $c - 1$, where c represents the number of possible classes. Moreover, data in magnetic tile dataset is partitioned into training, validation and test set, in which the training set accounts for 81% of the data, the validation set accounts for 9%, and the remaining 20% is allocated to the test set. The production project dataset does not need to do the above operations, as it has only two possible classes and the dataset has been pre-divided into training and test sets by J. Božič, et al. [9]. The batch size of training, validation and test datasets is uniformly set to 5.

In this study, ReLU activation is applied to most intermediate layers to introduce nonlinearity. For the fully supervised models trained with the magnetic tile dataset and the weakly supervised models trained with the production item dataset, the softmax activation is used in the output layer to convert the class scores into probability distributions, while the sigmoid activation is used for the fully supervised models trained with the production item dataset to squash the output value between 0 and 1. On the other hand, multi-class cross-entropy loss or binary cross-entropy loss is adopted as the cost function for models with softmax activation or sigmoid activation in output layer, respectively, to measure the deviation of predicted labels from the actual labels. For SEAM model, additional losses such as ER loss and ECR loss are computed. “Adam” optimizer with a learning rate ranging from 1e−4 to 2e−5 is employed to update the weights of networks in the direction of the steepest descent in the cost surface. The number of epochs fluctuates with the learning rate, ranging from 20 to 50.

3.2.5 Model Evaluation and Fine-Tuning

Performance of the trained models is assessed and compared by inference on validation set. Various evaluation metrics are implemented to visualize performance of the models in various aspects. The confusion matrix shows the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) for each defect class. By utilising these indices, precision rate (PR), recall rate, pixel accuracy (PA), intersection over union (IoU) and dice score are computed. Moreover, the mean in intersection over union is also calculated for each model.

$$Precision\ Rate\ (PR) = \frac{TP}{TP + FP} \tag{3}$$

$$Recall\ Rate = \frac{TP}{TP + FN} \tag{4}$$

$$Pixel\ Accuracy\ (PA) = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Intersection\ over\ Union\ (IoU) = \frac{TP}{TP+FP+FN} \tag{6}$$

$$Dice\ Score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{7}$$

$$mean\ Intersection\ over\ Union\ (mIoU) = \frac{1}{c}\sum_{c} IoU, where\ c = number\ of\ classes \tag{8}$$

Table 3.6 Class weight assigned to each defect class in loss function

Magnetic Tile Defect Dataset						
Classes	Blowhole	Break	Crack	Fray	Uneven	Free
Sample Mask						
Class Weight	10.0	5.0	7.5	2.0	2.0	1.0

Production Item Surface Defect Dataset		
Classes	Non-defective	Defective
Sample Mask		
Class Weight	1.0	5.0

It is normal for the models to perform unsatisfactorily without adjustments as the classes in the magnetic tile dataset and production item dataset are highly imbalanced. Pixels of a particular class can appear inside an image more often than other classes. For instance, fray defect class in magnetic tile dataset occupies a larger area in the image than the blowhole defect class. This is calamitous as the model may be biased towards certain classes over others. Hence, assignment of class weights according to Table 3.6 is introduced to account for the imbalance issue by weighting and averaging the losses for each class. For example, every instance of blowhole defect class will be treated as 50 instances of free defect class in the loss function of a model trained with magnetic tile dataset based on the weight assignments shown in Table 3.6.

3.2.6 Model Testing and Analysis

The final step will be to evaluate all fine-tuned models on the test set. A similar assessment as in model evaluation phase will be carried out. The segmentation output of each model on random test images will be displayed and compared to each other. Thenceforth, strengths and weaknesses of each model will be analysed to prepare for future improvements.

3.3 Model Architecture

3.3.1 VGG-16 Network

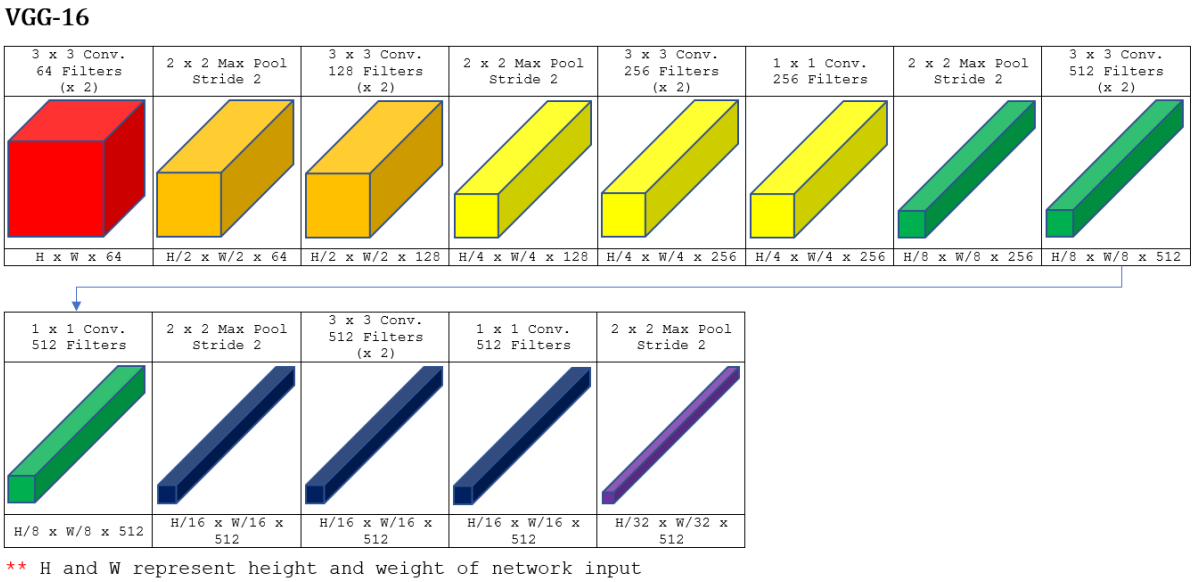


Figure 3.9 Architecture of VGG-16 network (without fully connected layers).

VGG-16 is a widely used convolutional neural network for image recognition. As the name suggests, there are 16 layers, including 13 convolutional layers and 3 fully connected layers, embodied inside the VGG-16 network. All layers are organized into homogeneous blocks, where the layers in each block have the same resolution. This is achieved by employing padded convolutions of stride 1. Downsampling operation is performed only by max-pooling layers of 2 × 2 kernel size with a stride 2. As the block progresses, the number of channels doubles and resolution decreases. This is because VGG-16 places more emphasis on contextual feature rather than spatial information to perform precise image classification. In this study, the last three fully connected layers of the VGG-16 network are removed to allow VGG-16 to serve as a feature extractor/backbone network for other models.

For the first two blocks of the VGG-16 network, each block contains two 3 × 3 convolutional layers, while for the next three blocks, each block contains two 3 × 3 convolutional layers, followed by a 1 × 1 convolution. A max-pooling layer is appended at the end of each block, reducing the resolution of activation map by a factor of two. The number of output channels per block is 64, 128, 256, 512 and 512 respectively. Given a 96 × 96 RGB image as input, the network will generate an output of shape 3 × 3 × 512.

3.3.2 U-Net

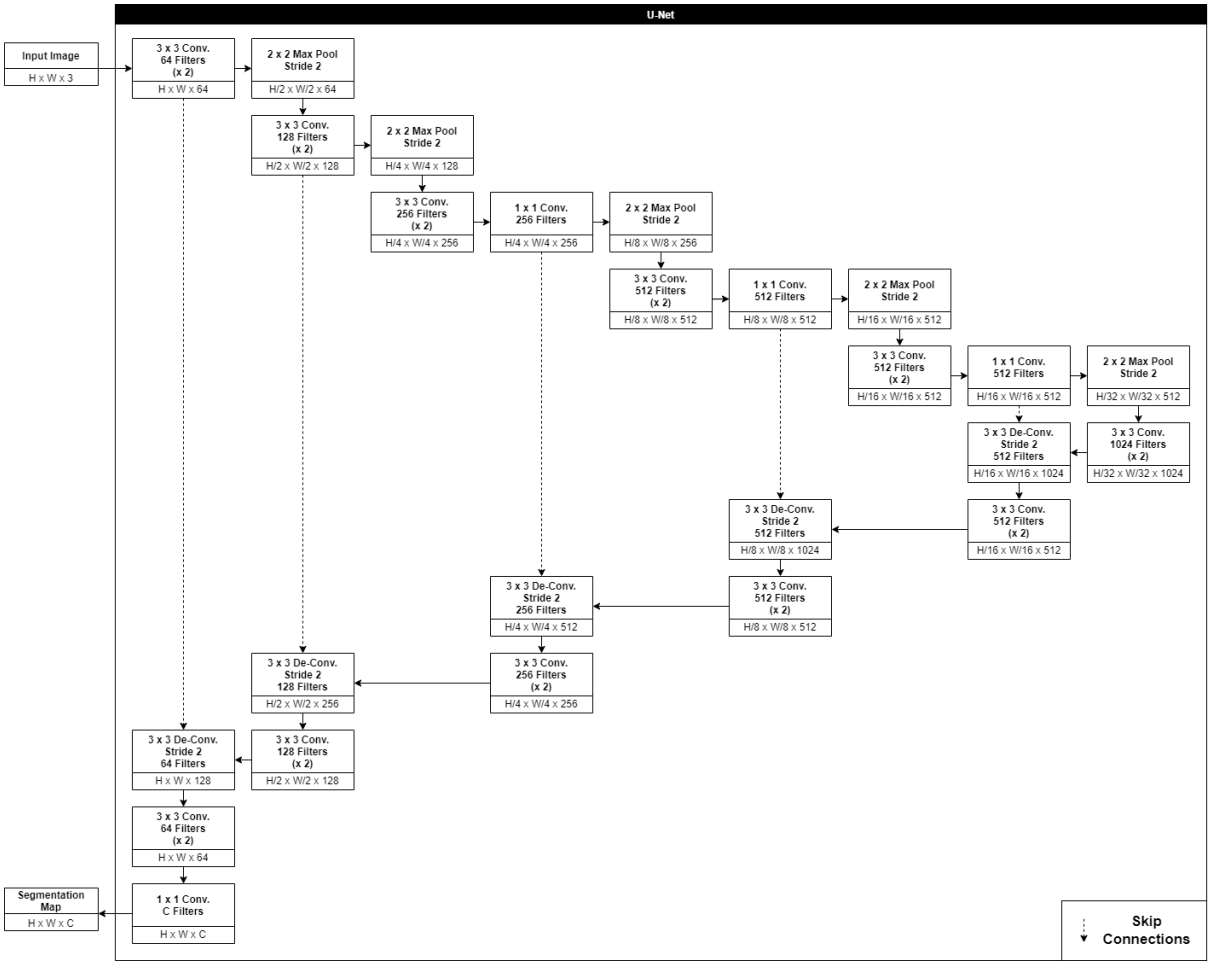


Figure 3.10 Detailed architecture of U-Net.

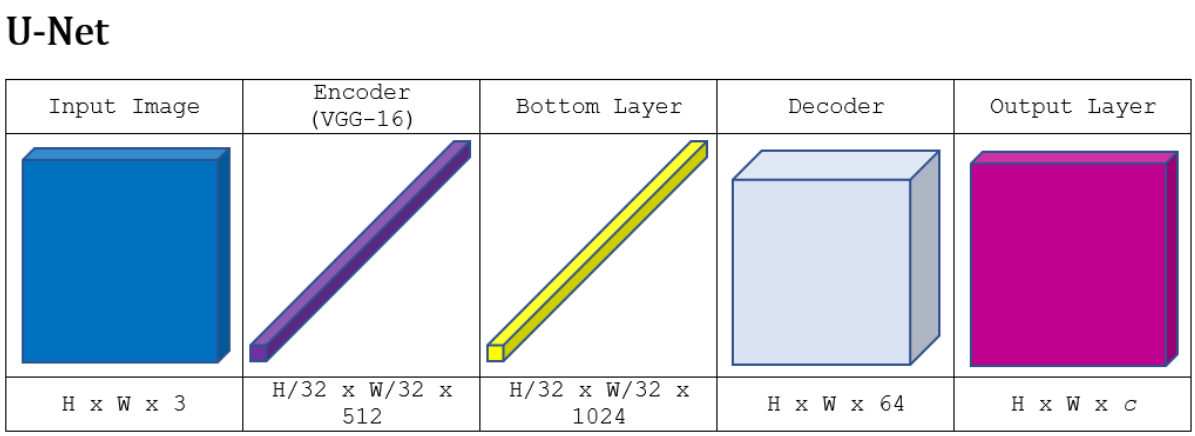


Figure 3.11 Architecture of U-Net.

Instead of building the network from scratch, the encoder in the U-Net model utilizes pre-trained VGG-16 model. Resolution of the activation map will be downsized to H/32 × H/32 × 512 after traversing contracting path. A bottom layer comprises two 3 × 3 convolutions with 1024 filters is then imported. After that, the process officially enters the expanding path. In contrast to the encoder that performs five downsampling operations with 2 × 2 max-pooling layers, the decoder in the U-Net model performs five upsampling operations by using 3 × 3 transposed convolutions. Skip connections are established in the U-Net model, where the feature maps generated from the encoder will be passed to the decoder part to be concatenated with the corresponding upsampled activation maps with the same resolution. Each concatenation layer is followed by two 3 × 3 convolutions, halving the number of output channels. Last but not least, a 1 × 1 convolution with *c* filters is applied as the output layer, where *c* is equal to the number of defect classes. The number of output channels of each upsampling block is 512, 512, 256, 128, and 64 respectively, which is the opposite of that of the downsampling block.

3.3.3 Multi-Depth Dilated Inception Block

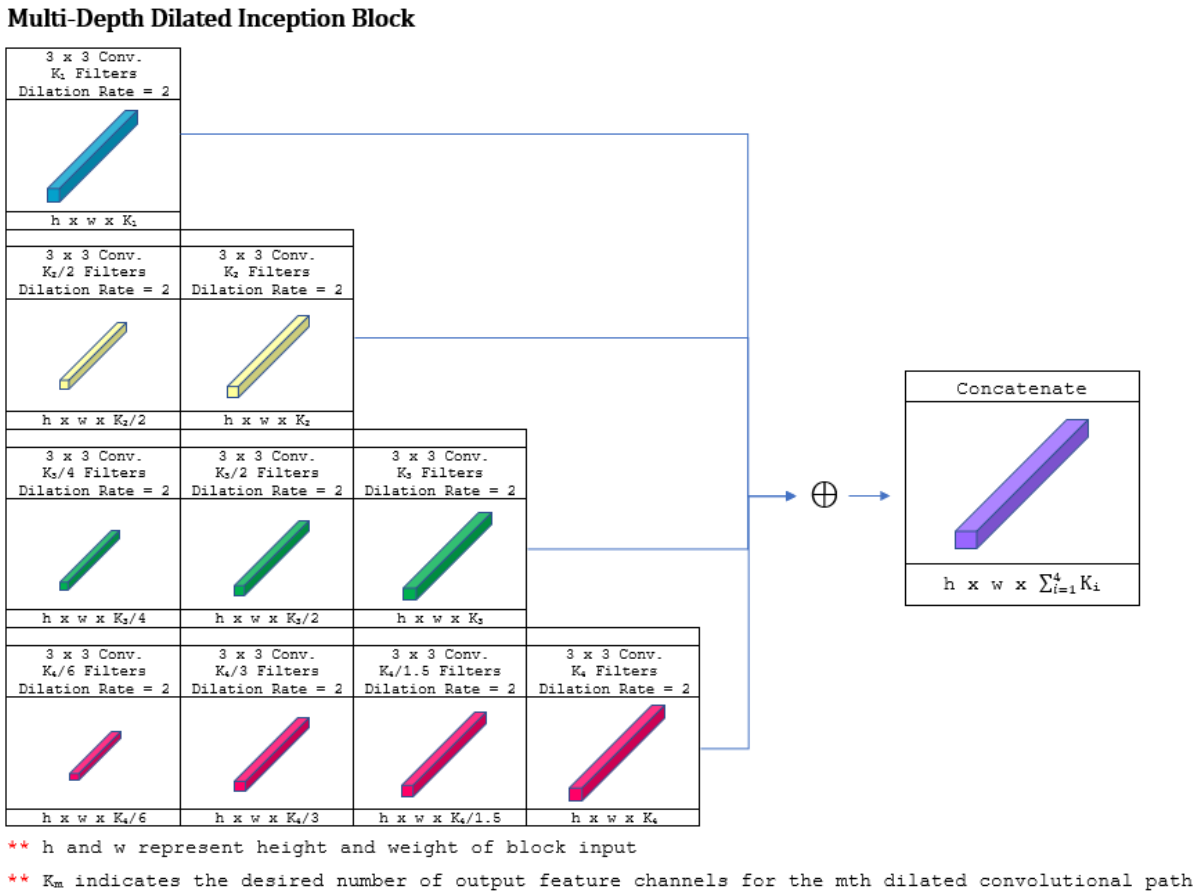


Figure 3.12 Architecture of multi-depth dilated inception block.

Multi-depth dilated inception block plays an important role in encoder part of Double U-Net model as well as skip connections in TransDAU-Net, implying that the number of output channels should vary and increase as the downsampling operation proceeds. Thus, some modifications are made compared to the original MDI blocks to allow the number of output feature channels generated by multi-depth dilated inception block to be more dynamic and dependent on the depth of the encoder's current position. The number of filters per dilated convolution within each path in the block is no longer the same, but increases as the path progresses. All dilated convolutions have twice as many filters as the previous dilated convolution, except that the last dilated convolution in the fourth path has 1.5 times more filters than the previous one. For example, if the number of output feature channels required for the fourth path is 192, the number of filters for the four embodied dilated convolutions will be 32, 64, 128, and 192 respectively. Most importantly, each dilated convolution wrapped in the block has a dilation rate of 2.

3.3.4 Double U-Net

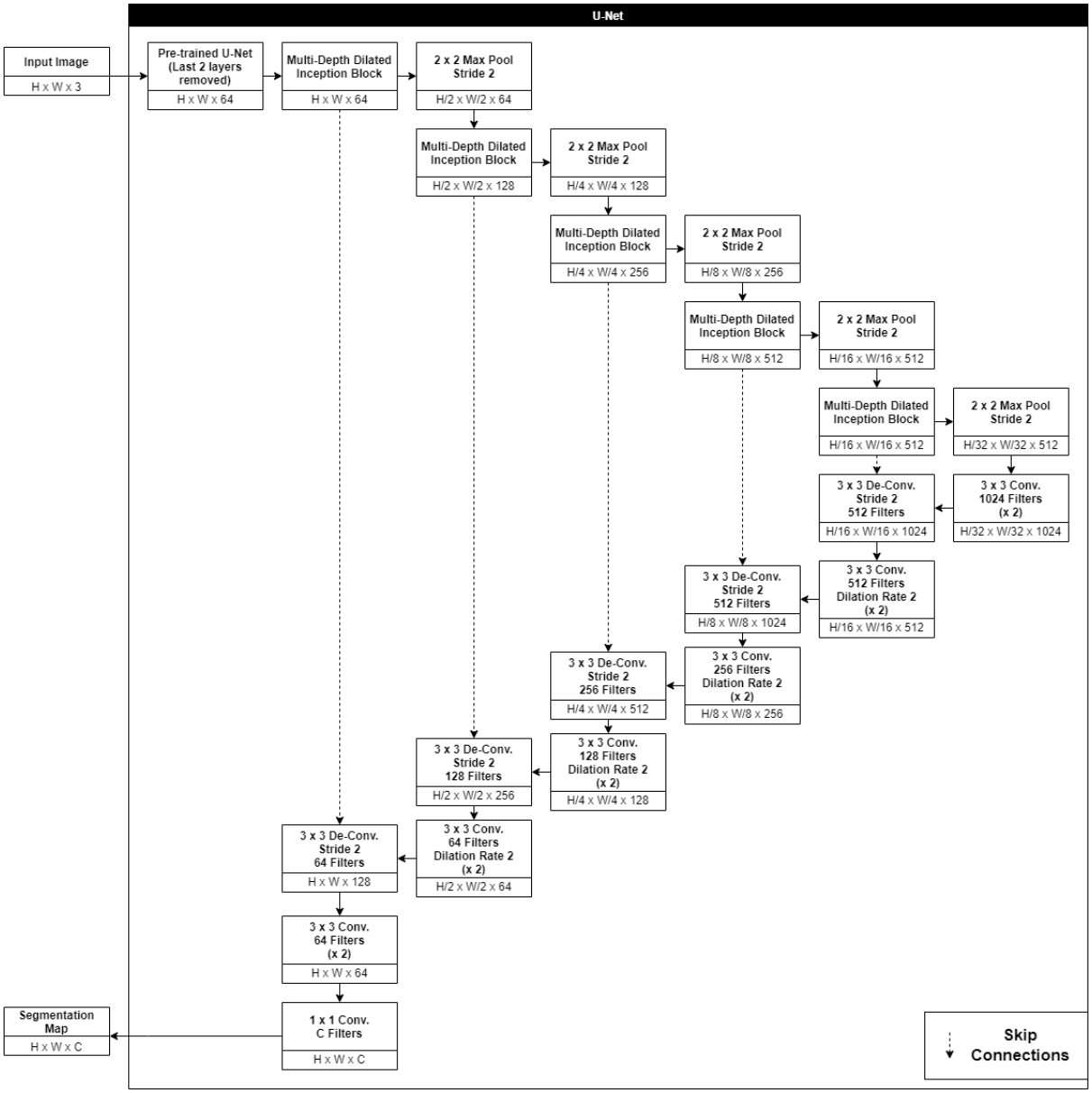


Figure 3.13 Detailed architecture of Double U-Net.

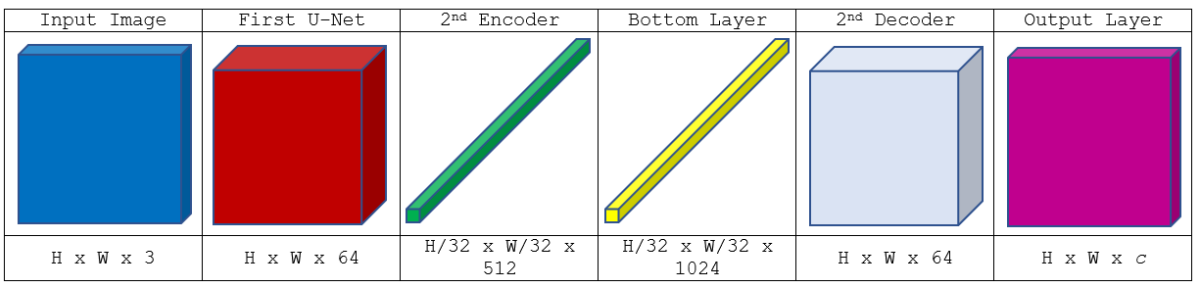


Figure 3.14 Architecture of Double U-Net.

The architecture of Double U-Net consists of two U-Nets, where the first U-Net is derived from the pre-trained U-Net model that has completed its training on the training set, and the second U-Net is built from scratch. The last two layers of the imported pretrained U-Net model, a 3 × 3 convolution and an output layer, are detached such that the first U-Net produces a H × W × 64 activation map for each resized magnetic tile input image. The activation map is then passed to the second U-Net for more explicit computation. Each downsampling block in the second U-Net consists of a multi-depth dilated inception block followed by a 2 × 2 max-pooling layer with stride 2. On the other hand, each upsampling block in the second U-Net consists of one 3 × 3 transposed convolution along with two 3 × 3 dilated convolutions. Similar to the first U-Net, there are five blocks for each contracting and expanding path in the second U-Net, resulting in a total of ten downsampling and upsampling operations for the entire network. The bottom block of second U-Net comprises two dilated 3 × 3 convolutions with 1024 filters. Skip connections are also introduced in the second U-Net to combine the activation maps of the encoder and decoder accordingly. A 1 × 1 convolution with *c* filters is adopted as the output layer at last. All dilated convolutions in the second U-Net have a dilation rate of 2.

3.3.5 SETR

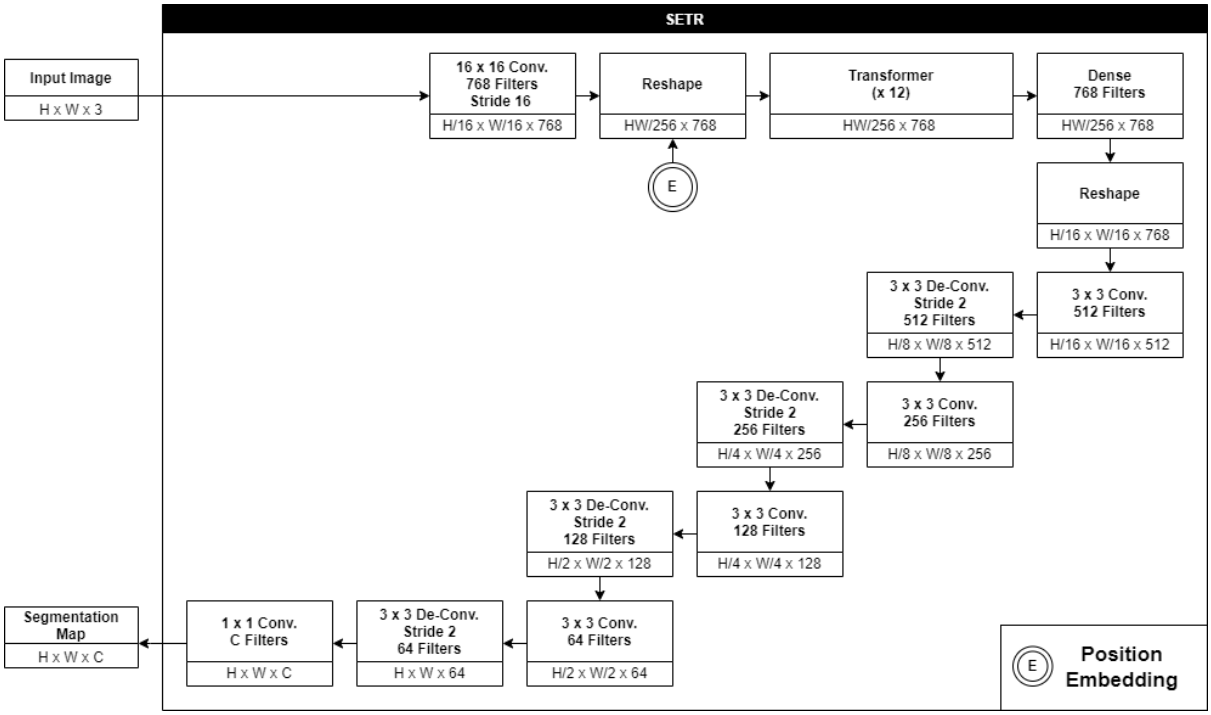


Figure 3.15 Detailed architecture of SETR.

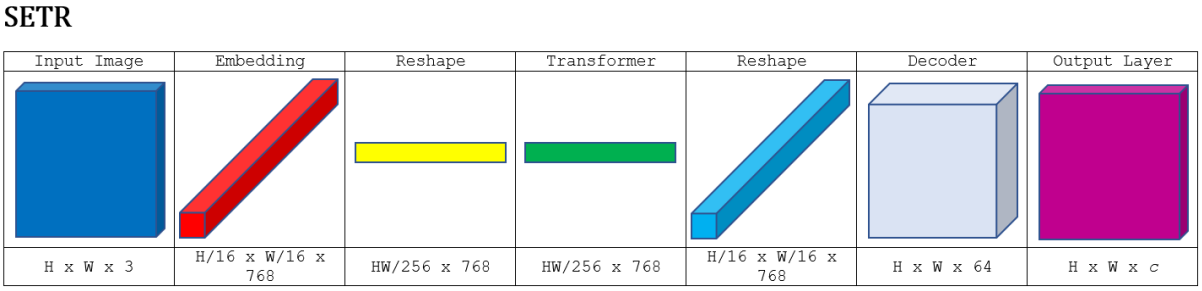


Figure 3.16 Architecture of SETR.

The feature learning process of SETR is a sequence-to-sequence approach. Hereby, the input image will be divided into a grid of image patches, each of which is fixed in size, 16 × 16 in this study. To achieve this, an embedding layer, which is a 16 × 16 convolutional layer with stride 16, is used to extract $\frac{H}{16} \times \frac{W}{16}$ number of image patches. Later, the grid of image patches will be reshaped into a sequence of length $\frac{HW}{256}$. At this stage, positional embeddings are added to the grid to encode the order of each image patch. The embedding sequence is then fed to Transformer-alone encoder containing 12 Transformer layers to learn feature representations while modelling global context at each layer. After that, a final dense layer is employed right after the encoder, and the resulting output maps will be reshaped back to $\frac{H}{16} \times \frac{W}{16}$ and passed to the bottom block containing two 3 × 3 convolutional layers with 512 filters. The following decoding operation is similar to those in typical encoder-decoder architecture, in which feature maps are progressively upsampled to the original input image dimension through 4 upsampling blocks. At last, a 1 × 1 convolution with *c* filters is adopted to perform pixel-level classification. The hidden size or units of the entire Transformer-only encoder, the final dense layer as well as the embedding layer is set to 768.

3.3.6 TransU-Net

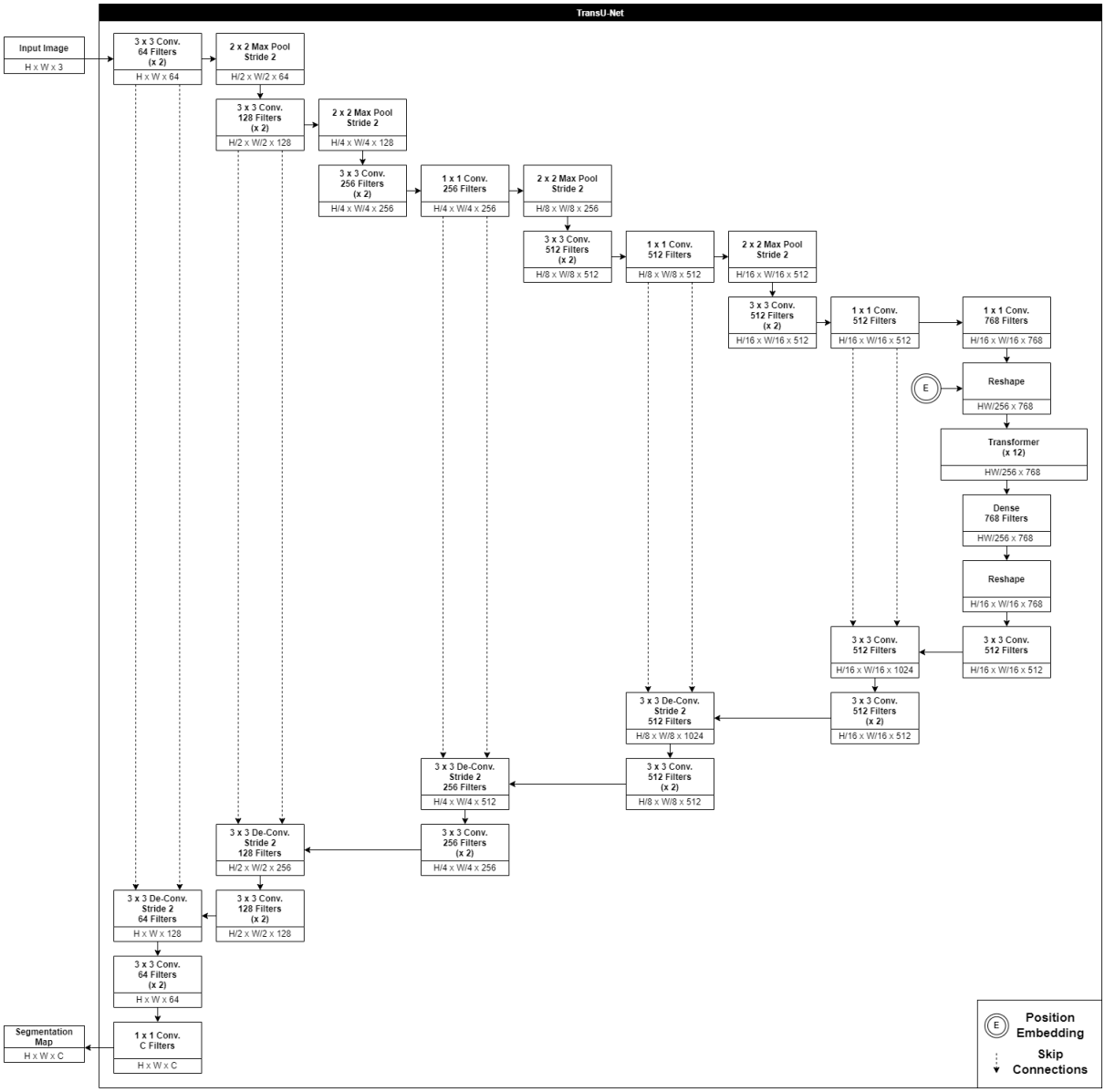


Figure 3.17 Detailed architecture of TransU-Net.

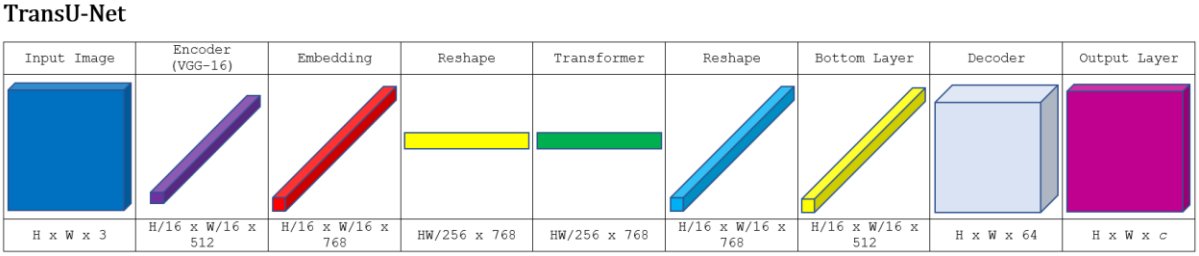


Figure 3.18 Architecture of TransU-Net.

The encoder section of TransU-Net stacks a pretrained VGG-16 model on top of a pure Transformer by removing the last max pooling layer in VGG-16 model to keep the resolution of the feature maps at $\frac{H}{16} \times \frac{W}{16}$. An embedding layer followed by a reshaping operation are then used to divide the features into grid of patches and convert the grid into a sequence respectively. Positional embeddings are added to the grid to encode the order of each image patch. Later, the pure Transformer encoder incorporating 12 Transformer layers takes the sequence as the input and enhances the features representations by enriching each patch with relevant patches in the sequence. Subsequently, a final dense layer followed by a reshaping operation are adopted, resulting in output maps with resolution of $\frac{H}{16} \times \frac{W}{16}$. A bottom layer, 3 × 3 convolution with 512 filters is then introduced. Similar to U-Net, except for the first upsampling block, each upsampling block consists of one deconvolutional layer and two 3 × 3 convolutional layers, and the number of output channels of each upsampling block is 512, 512, 256, 128, and 64 respectively. Skip connections are introduced in TransU-Net, where feature representations generated by the intermediate layers of VGG- 16 model will be brought across and concatenated with the output feature maps of the very first layer in each upsampling block. This hybrid CNN-Transformer architectural design allows the low-level feature maps with rich spatial information generated at the initial encoder blocks (VGG-16 model) to traverse across the network via skip connections, while enforcing the learning of global context at the bottom encoder blocks (Transformer model).

3.3.7 TransDAU-Net

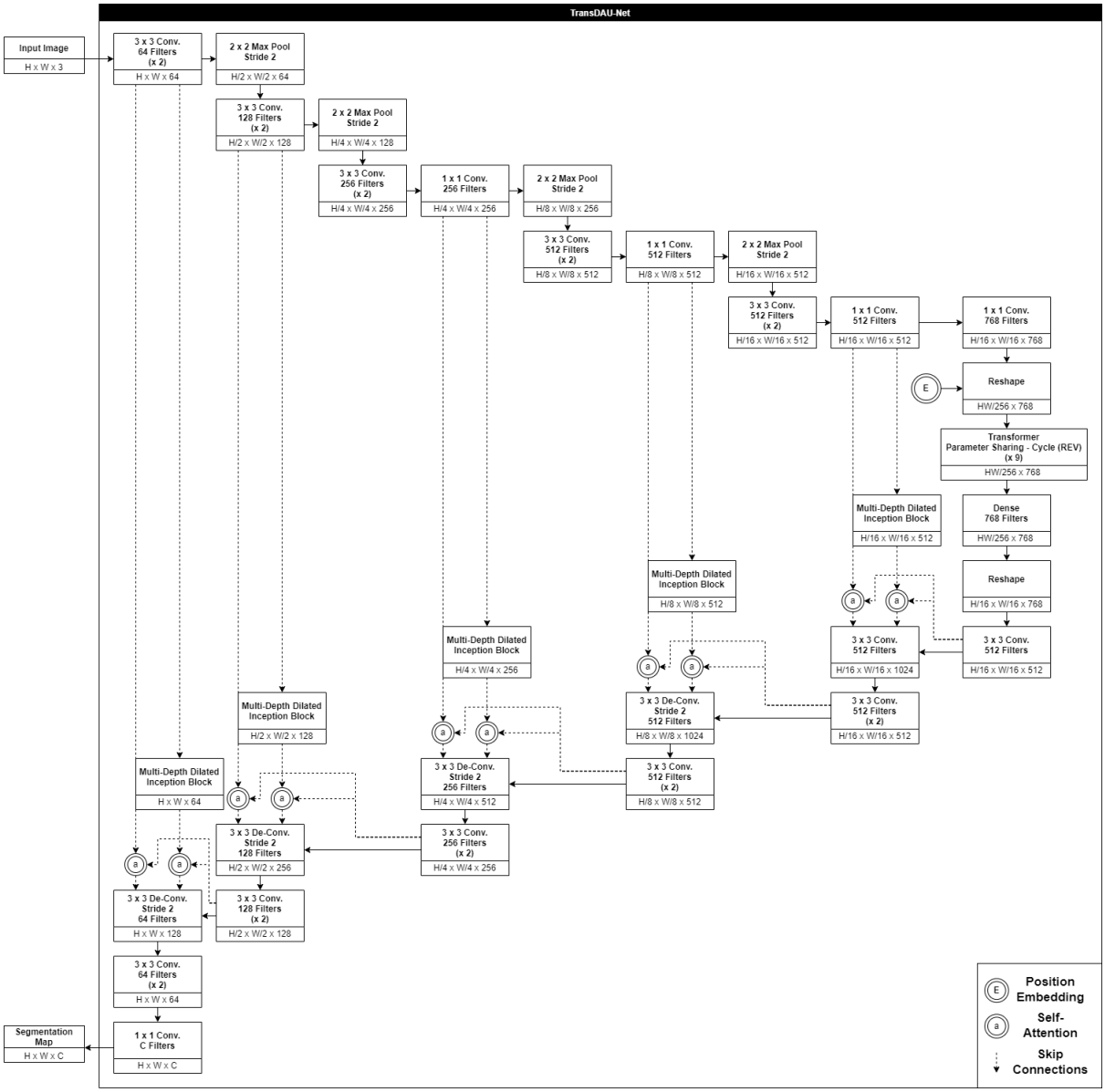


Figure 3.19 Detailed architecture of TransDAU-Net.

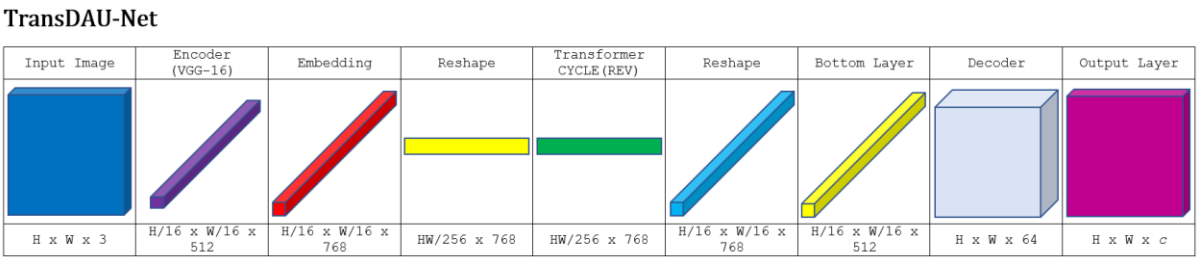


Figure 3.20 Architecture of TransDAU-Net.

TransDAU-Net is an improved model evolved from TransU-Net. In order to reduce the burden imposed by the Transformer encoder in terms of parameter size and computational cost, parameter sharing technique, CYCLE (REV) is implemented to reduce the number of actual Transformer layers while maintaining a similar performance in learning feature representations. In TransDAU-Net, parameters of lower Transformer layers are reused at higher layers in a reverse cyclic order. Simply put, the number of Transformer layers is set to 18, indicating that the first layer shares the same parameters with the 18^{th} layer, the second layer shares the parameters with the 17^{th} layer and so on. This creates a delusion that there are 18 Transformer layers incorporated in the encoder, but in fact only 9 legitimate layers contribute to the learning process. In addition, each skip connection is split into two parallel paths, one of which will be fed to a multi-depth dilated inception block to expand the receptive field. An attention module is inserted at the end of both paths, which takes in the feature maps from the skip connection and next lowest layer of the network as inputs. Number of channels will be halved after passing through the attention module. For example, an attention module that takes in inputs of 512 channels will output an attentive feature representation with 256 channels. Later, the output maps of the two paths as well as the first layer in each corresponding upsampling block are concatenated together, producing representations that embrace multi-scale features and seize precise activations in appropriate regions. Besides, other structural design and configurations of TransDAU-Net are similar to the TransU-Net.

3.3.8 CAM

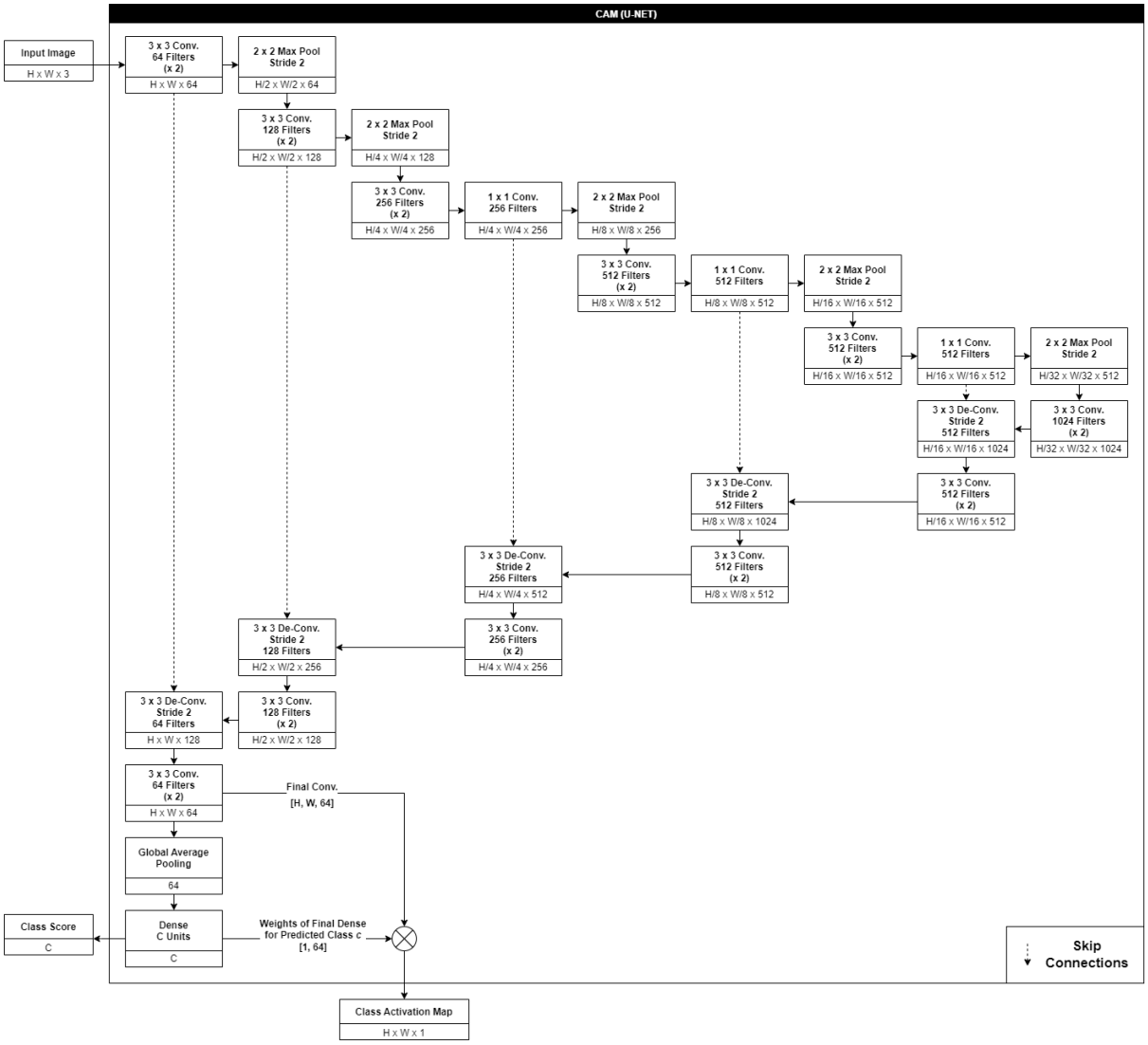


Figure 3.21 Detailed architecture of CAM.

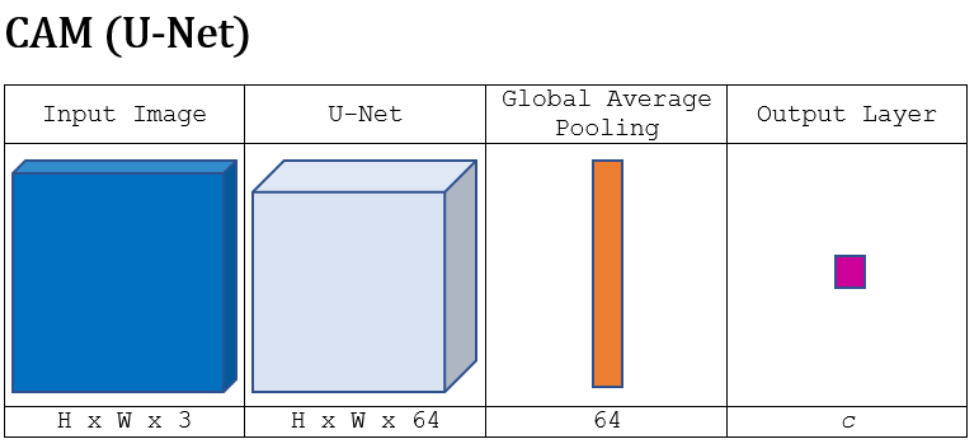


Figure 3.22 Architecture of CAM.

By replacing the last 1 × 1 convolutional layer with *c* filters in U-Net with a global average pooling layer and a fully connected layer with *c* units, CAM model is born. The processing pipeline of the model is modelled as image-level multi-class classification, since in the weakly supervised setting, only image-level ground truth labels are available for model training. The global average pooling layer flattens the output feature maps generated by the U-Net into a vector, and the fully connected layer receives the vector as input and generates an activation vector, with each output unit carrying respective class score and corresponding to a specific class label. After passing through the softmax function, unit with highest class score will be the final predicted class of the model. During inference phase, the specific weights of the connections with highest class score unit as end destination, are multiplied with the output maps of last convolutional layer, followed by a summation operation to aggregate all the feature channels into one. By that, class activation maps (CAMs) are generated, highlighting relevant regions of the predicted class.

3.3.9 SEAM

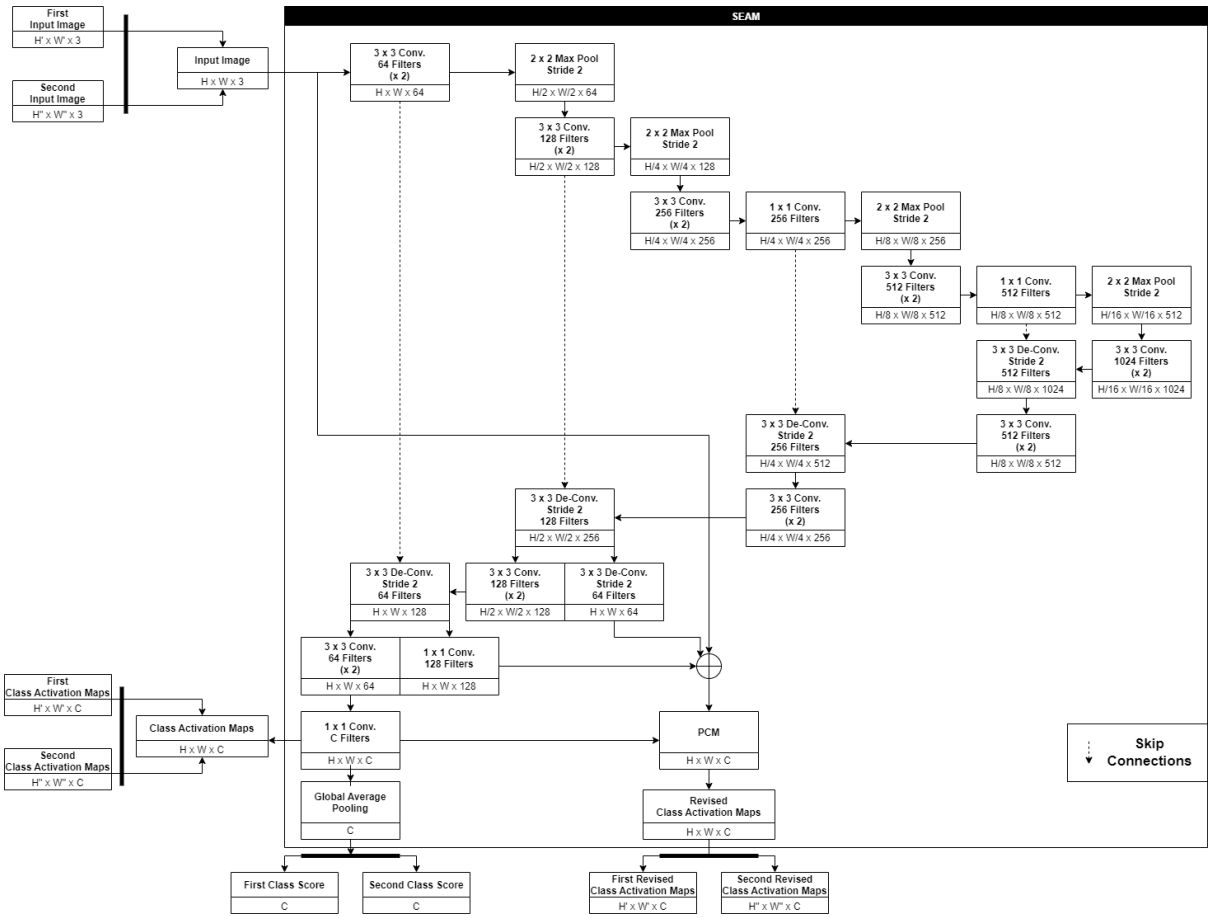


Figure 3.23 Detailed architecture of SEAM.

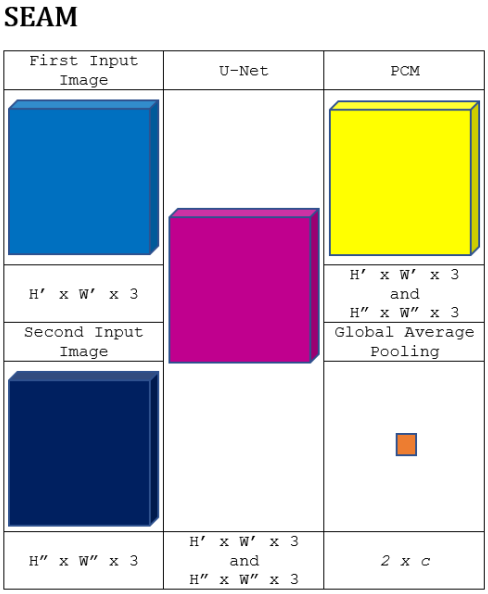


Figure 3.24 Architecture of SEAM.

SEAM model has a Siamese network structure which expecting two inputs of different scales during model training stage. In this study, the dimension of the second input is the half of the first input. A typical U-Net model is adopted as the backbone network, and according to the major process flow, class activation maps (CAMs) are finally generated for each input. By means of that, SEAM model expects the generation of CAMs without performing the weighted sum operation as in CAM model. Since SEAM is built under weakly supervised setting, image-level classification is performed by appending a global average pooling layer to the final 1×1 convolutional layer to generate c-unit class scores. To improve the quality of CAMs, pixel correlation module (PCM) is introduced to model cross-correlation of all pixels in the CAMs. Prior to that, to prepare features for input to the PCM, the feature maps generated from the last and last second concatenation operations with skip connections are fed to a convolutional layer with 128 filters and a 3×3 deconvolutional layer with stride 2 and 64 filters respectively. Later, these two feature maps are concatenated with the input as the feature input to the PCM. PCM will generate an inter-pixel similarity matrix from the feature input and fuse it into another input, the original CAMs. Revised CAMs with stronger inter-pixel correlation are then produced. During the loss computation phase, regularization operations such as equivariant regularization to supervise the consistent prediction of CAMs, and equivariant cross regularization to regularize the revised CAMs with original CAMs from another path, will be carried out. In the inference stage, since both paths share the same parameters, only one path of the Siamese network is used, and the modified CAM will be adopted as the final predicted CAM.

CHAPTER 4 EXPERIMENT/SIMULATION

4.1 General Work Procedure

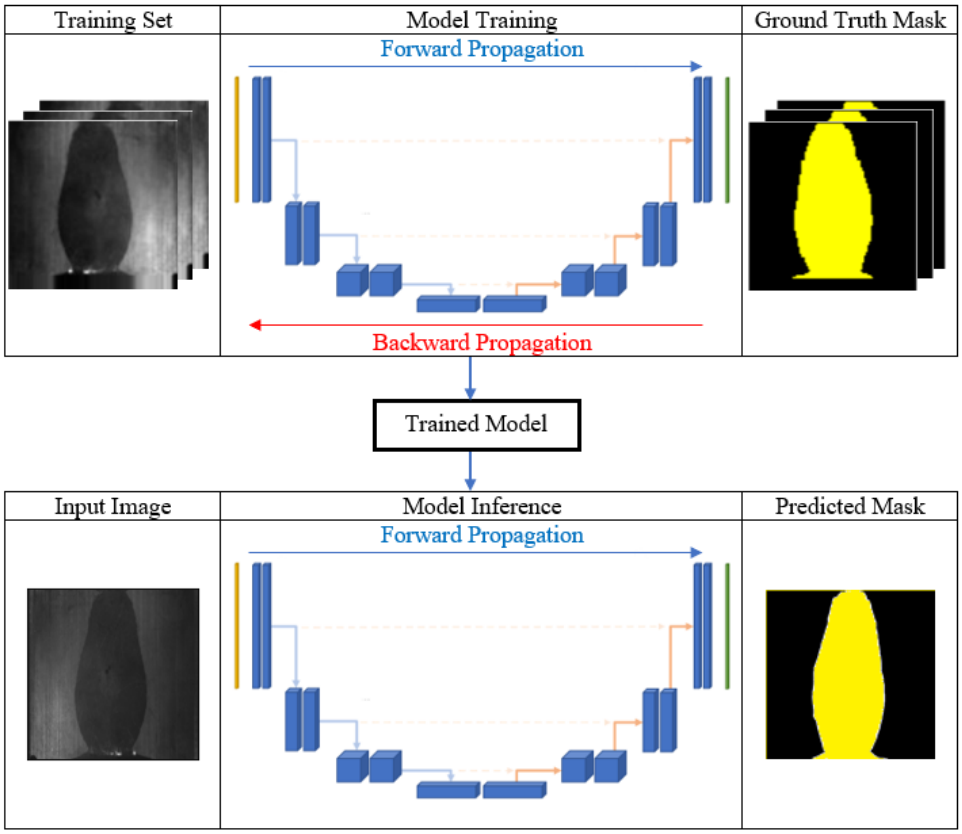


Figure 4.1 General work procedure.

After building the model architecture and defining the model hyperparameters, the model is now ready for training. Prior to that, it must be ensured that the expected input shape of the models is consistent with the shape of input images in the subsequent process. The following are the steps of the model from training to inference:

i. Feed pre-processed training set images as input to the constructed models
ii. Models repeatedly perform forward and backward propagation internally, the former operation predicts the mask outputs and calculates the loss based on the ground truth masks; the latter operation performs gradient descent and updates model parameters, including weight and bias
iii. Save the trained models generated after a certain number of iterations
iv. Load the trained models and perform inference on the input image
v. Visualize the inference result or prediction mask that segment and classify the defects present in the input image

4.2 Visualization and Segmentation

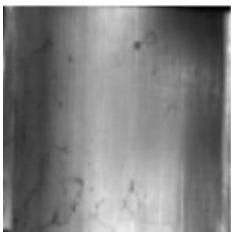
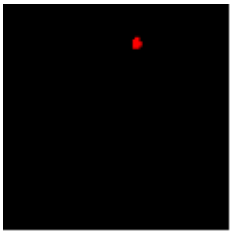
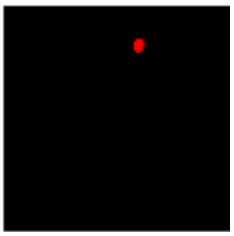
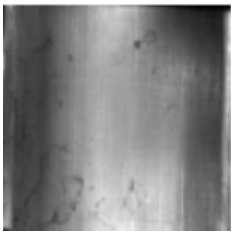
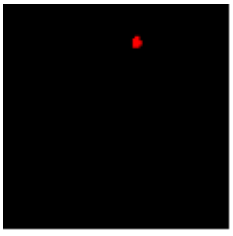
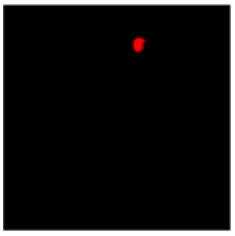
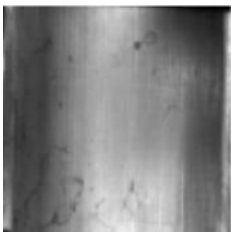
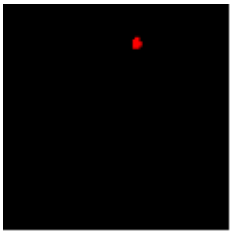
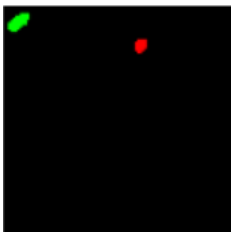
Each defect category in the magnetic tile image is represented by a different color. Blowhole defects are shown in red; break defects are shown in green; crack defects are shown in blue; fray defects are shown in yellow; uneven defects are shown in purple; the irrelevant background (free-of-defect) pixels are shown in black. On the other hand, in production item dataset, defects are shown in yellow and background pixels are shown in deep purple.

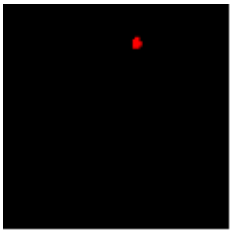
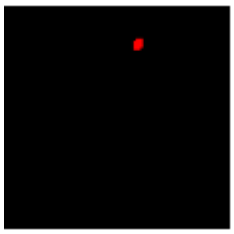
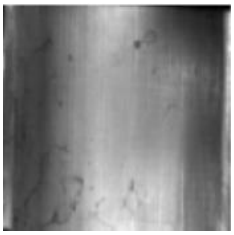
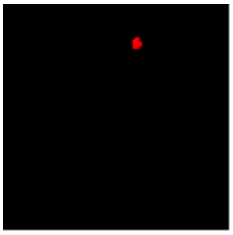
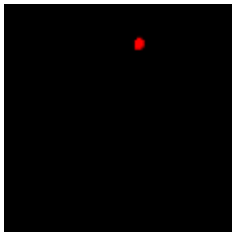
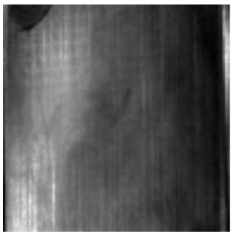
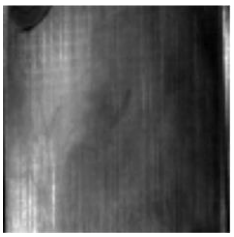
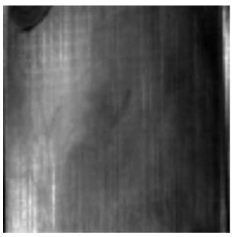
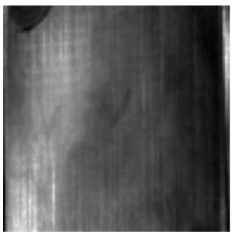
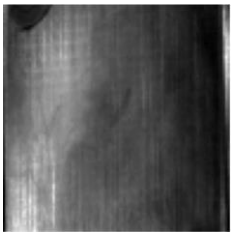
Table 4.1 Color representation of each defect class

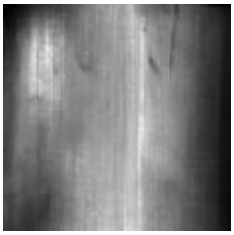
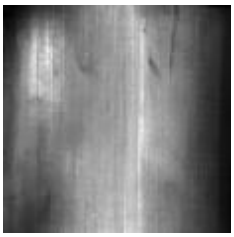
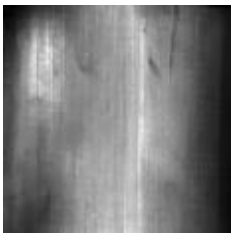
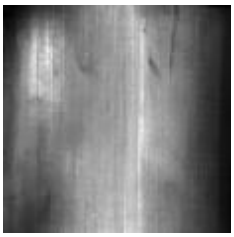
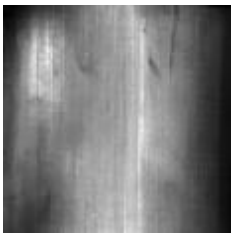
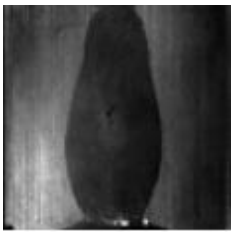
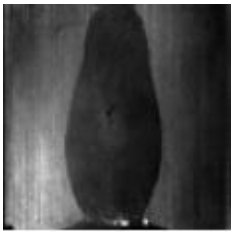
Magnetic Tile Defect Dataset						
Defect Class	Blowhole	Break	Crack	Fray	Uneven	Free
Color						
Production Item Defect Dataset						
Defect Class	Defective	Free				
Color						

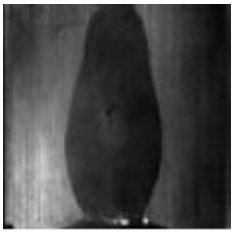
4.2.1 Fully Supervised Segmentation Mask Prediction

Table 4.2 Masks generated by each fully supervised model on magnetic tile images

Defect Class		Input Image	Ground Truth Mask	Predicted Mask
Blowhole	U-Net			
	Double U-Net			
	SETR			

	TransU-Net			
	TransDAU-Net			
Break	U-Net			
	Double U-Net			
	SETR			
	TransU-Net			
	TransDAU-Net			

Crack	U-Net			
	Double U-Net			
	SETR			
	TransU-Net			
	TransDAU-Net			
Fray	U-Net			
	Double U-Net			

	SETR			
	TransU-Net			
	TransDAU-Net			
Uneven	U-Net			
	Double U-Net			
	SETR			
	TransU-Net			

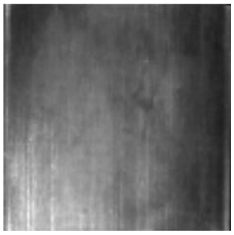
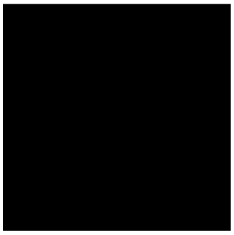
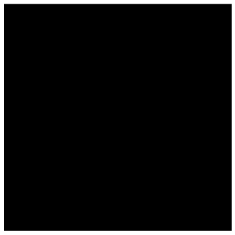
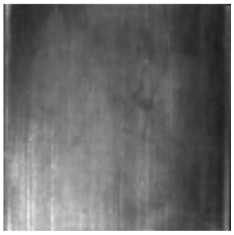
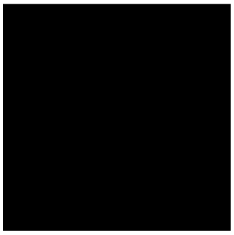
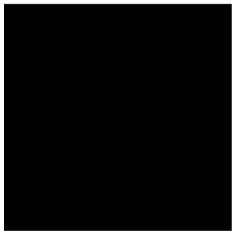
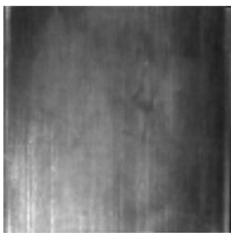
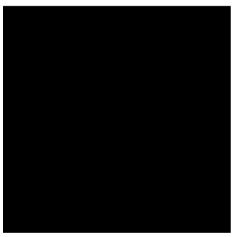
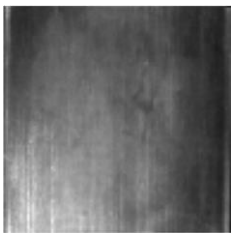
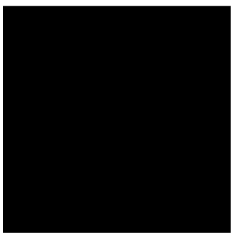
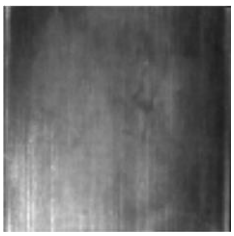
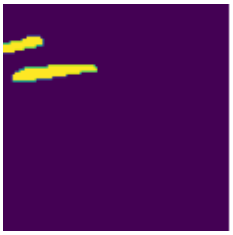
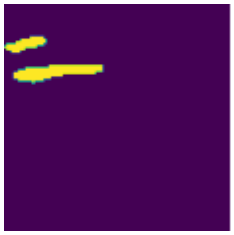
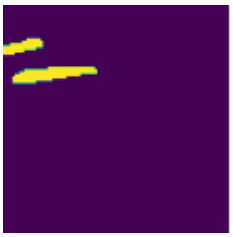
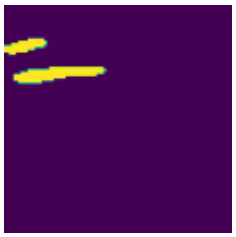
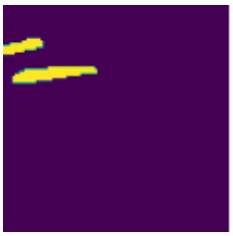
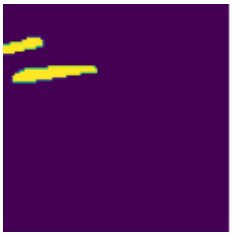
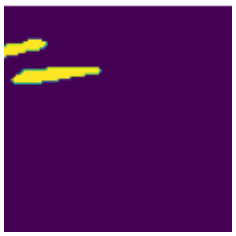
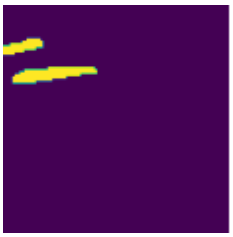
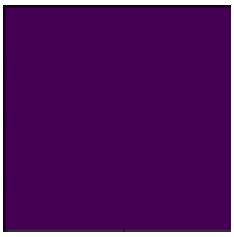
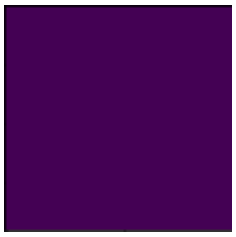
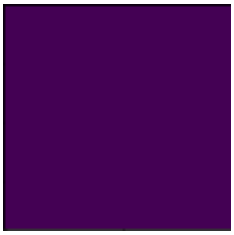
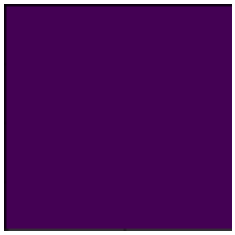
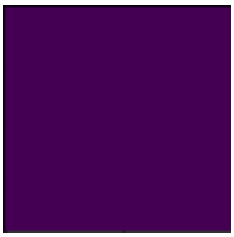
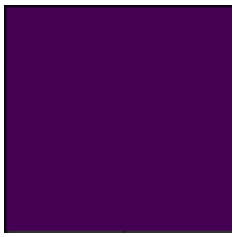
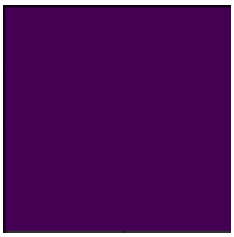
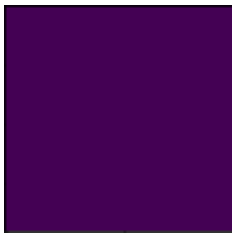
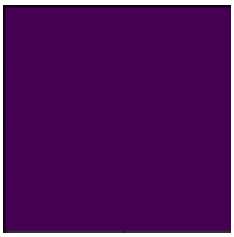
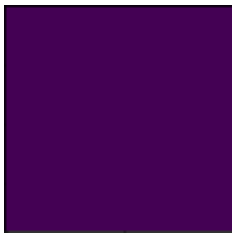
	TransDAU-Net			
Free	U-Net			
	Double U-Net			
	SETR			
	TransU-Net			
	TransDAU-Net			

Table 4.3 Masks generated by each fully supervised model on production item images

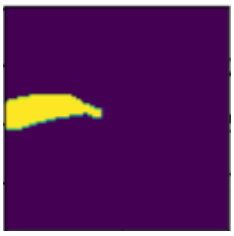
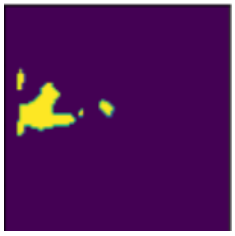
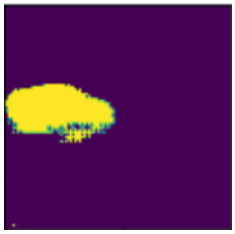
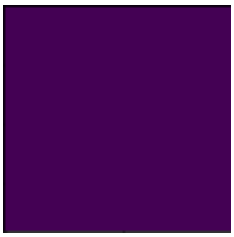
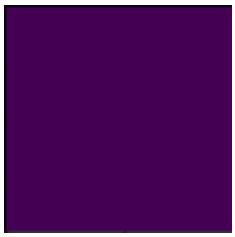
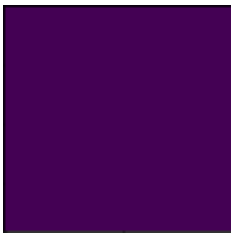
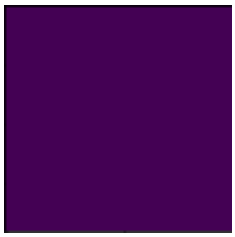
Defect Class		Input Image	Ground Truth Mask	Predicted Mask
Defective	U-Net			
	Double U-Net			
	SETR			
	TransU-Net			
	TransDAU-Net			
Free	U-Net			

	Double U-Net			
	SETR			
	TransU-Net			
	TransDAU-Net			

It is observed that for all the models, slightly larger coverage areas would be predicted for small-scale defect classes including blowhole and break defects in magnetic tile dataset and defects in production item dataset compared to their ground truth annotations. As to large-scale defect classes such as fray and uneven defects in magnetic tile dataset, all models are able to predict adequate segmentation masks. Upon closer inspection, segmentation masks predicted by TransDAU-Net are more similar to the ground truth annotations, with fewer misclassified pixels. Also, there are some pixels that are misclassified as break defects by SETR, as shown in Table 4.2. It is conspicuous that SETR tends to generate lopsided defect segments. Conversely, defect segments generated by U-Net, Double U-Net, TransU-Net and TransDAU-Net are smoother and more balanced.

4.2.2 Weakly Supervised Segmentation Mask Prediction

Table 4.4 Masks generated by each weakly supervised model on production item images

Defect Class		Input Image	Ground Truth Mask	Predicted Mask
Defective	CAM			
	SEAM			
Free	CAM			
	SEAM			

To convert CAMs generated by weakly supervised models into segmentation masks, a min-max normalization operation is first performed to scale the pixel values between 0 and 1. Next, a threshold operation is performed to classify the normalized pixels into the correct class, where the threshold value is set to 0.2 for CAM model and 0.5 for SEAM model. For instance, normalized pixels of CAMs generated by CAM model with value greater than 0.2 are classified as defective. By that, CAMs are converted into segmentation masks. Since weakly supervised models only utilize image-level labels to carry out the training procedure, it is expected that the final predicted segmentation masks are incomparable to those predicted by fully supervised model in terms of accuracy and integrity. It can be clearly seen that the segmentation mask generated by the SEAM model is much superior to the CAM model. Yet, there are still many irrelevant background pixels predicted by SEAM as defective pixels (over-activation problem).

4.3 Experiment Results

4.3.1 Confusion Matrix

Table 4.5 Confusion matrix of U-Net model on magnetic tile validation set

Ground Truth \ Predicted	Blowhole	Break	Crack	Fray	Uneven	Free
Blowhole	3462	48	13	0	0	176
Break	7	40069	3	478	0	2669
Crack	0	13	13360	1	0	990
Fray	0	1553	0	206003	0	4077
Uneven	0	0	0	0	229735	12631
Free	1315	7862	5236	11344	54154	7395073

Table 4.6 Confusion matrix of Double U-Net model on magnetic tile validation set

Ground Truth \ Predicted	Blowhole	Break	Crack	Fray	Uneven	Free
Blowhole	3492	37	10	0	0	160
Break	6	39695	0	298	0	3227
Crack	0	18	13809	0	0	537
Fray	0	1690	0	205518	0	4425
Uneven	0	0	1	0	228382	13983
Free	1331	4993	5634	9454	35758	7417814

Table 4.7 Confusion matrix of SETR model on magnetic tile validation set

Ground Truth \ Predicted	Blowhole	Break	Crack	Fray	Uneven	Free
Blowhole	3429	0	4	0	0	266
Break	3	40110	38	198	3	2874
Crack	2	6	12315	0	0	2041
Fray	23	2056	0	195011	0	14543
Uneven	0	0	1	0	227723	12642
Free	2935	9875	10369	11298	34009	7406498

Table 4.8 Confusion matrix of TransU-Net model on magnetic tile validation set

Ground Truth \ Predicted	Blowhole	Break	Crack	Fray	Uneven	Free
Blowhole	3508	0	11	0	0	180
Break	9	39574	30	150	0	3463
Crack	0	15	13424	2	0	924
Fray	0	81	0	207674	0	3878
Uneven	0	0	2	0	226978	15386
Free	1133	3883	4387	9886	35887	7419808

Table 4.9 Confusion matrix of TransDAU-Net model on magnetic tile validation set

Predicted / Ground Truth	Blowhole	Break	Crack	Fray	Uneven	Free
Blowhole	3307	0	6	2	0	384
Break	11	40481	2	603	68	2061
Crack	12	12	13282	0	0	1057
Fray	0	1104	0	205048	0	5481
Uneven	1	0	0	0	226513	15852
Free	712	5978	4201	8958	33020	7422115

Table 4.10 Confusion matrix of U-Net model on magnetic tile test set

Predicted / Ground Truth	Blowhole	Break	Crack	Fray	Uneven	Free
Blowhole	4124	25	19	0	0	167
Break	0	51051	11	1	36	1513
Crack	1	23	11018	0	10	948
Fray	1	334	51	250264	0	5081
Uneven	0	10	0	2	259453	16771
Free	1581	9901	4901	12714	63455	8190758

Table 4.11 Confusion matrix of Double U-Net model on magnetic tile test set

Predicted / Ground Truth	Blowhole	Break	Crack	Fray	Uneven	Free
Blowhole	4182	15	7	0	0	131
Break	5	50690	14	28	0	1875
Crack	0	24	11407	0	0	569
Fray	19	477	40	249415	0	5780
Uneven	0	5	0	3	260320	15908
Free	1679	6167	5283	10262	35238	8224681

Table 4.12 Confusion matrix of SETR model on magnetic tile test set

Predicted / Ground Truth	Blowhole	Break	Crack	Fray	Uneven	Free
Blowhole	3932	0	3	0	2	398
Break	0	49514	41	0	0	3057
Crack	2	0	10237	0	0	1761
Fray	37	481	0	239387	0	15826
Uneven	0	35	0	0	258178	18023
Free	3293	11143	9256	12613	41188	8205817

Table 4.13 Confusion matrix of TransU-Net model on magnetic tile test set

Ground Truth \ Predicted	Blowhole	Break	Crack	Fray	Uneven	Free
Blowhole	4082	0	27	0	0	226
Break	27	50124	0	49	6	2406
Crack	0	10	11233	0	3	754
Fray	4	39	70	250485	0	5133
Uneven	0	0	0	0	261441	14795
Free	1367	5565	4457	11195	45436	8215290

Table 4.14 Confusion matrix of TransDAU-Net model on magnetic tile test set

Ground Truth \ Predicted	Blowhole	Break	Crack	Fray	Uneven	Free
Blowhole	3947	0	0	5	0	383
Break	8	50830	0	84	1	1689
Crack	16	24	10853	0	0	1107
Fray	12	364	0	249115	0	6240
Uneven	0	13	1	0	257652	18570
Free	896	7461	4122	9543	36574	8224714

Table 4.15 Confusion matrix of U-Net model on production item test set

Ground Truth \ Predicted	Defective	Free
Defective	25526	10412
Free	9135	9207791

Table 4.16 Confusion matrix of Double U-Net model on production item test set

Ground Truth \ Predicted	Defective	Free
Defective	25780	10158
Free	9228	9207698

Table 4.17 Confusion matrix of SETR model on production item test set

Ground Truth \ Predicted	Defective	Free
Defective	21922	14016
Free	11482	9205444

Table 4.18 Confusion matrix of TransU-Net model on production item test set

Ground Truth \ Predicted	Defective	Free
Defective	24282	11656
Free	5903	9211023

Table 4.19 Confusion matrix of TransDAU-Net model on production item test set

Ground Truth \ Predicted	Defective	Free
Defective	26268	9670
Free	8652	9208274

Table 4.20 Confusion matrix of CAM model on production item test set (defective only)

Ground Truth \ Predicted	Defective	Free
Defective	10506	25432
Free	219664	758158

Table 4.21 Confusion matrix of SEAM model on production item test set (defective only)

Ground Truth \ Predicted	Defective	Free
Defective	19217	16721
Free	15650	962172

Since the total number of defect-free pixels in both magnetic tile dataset and production item dataset is significantly higher, it is normal for a relatively great number of pixels to be misclassified as non-defective. TransDAU-Net has lesser true positives for blowhole and uneven defects in magnetic tile dataset, most of which are misclassified as background pixels. However, it achieves highest true positive rate for break defects and background pixels, despite the fact that a significant amount of break defects being falsely predicted as fray defects. On both magnetic tile dataset validation and testing sets, Double U-Net has more true positives for crack classes, while TransU-Net has more true positives for blowhole and fray classes. U-Net has the highest true positive rate for uneven magnetic tile defects, but it tends to falsely activate background pixels as defects compared to other models. It is observed that all models except TransU-Net tend to misclassify break defects as fray defects and vice versa, especially SETR.

On production item test set, TransDAU-Net outperforms all the other models in defect prediction, with the highest true positive rate and lowest false positive rate, while TransU-Net performs best in detecting background pixels with the highest true negative rate and the lowest false negative rate. U-Net and Double U-Net perform very similarly, while SETR is the most inferior model, with worst index at all metrics. Since CAM and SEAM are capable of image-level classification, these two weakly supervised models are only evaluated on images in the test set of production items that contain surface defects. Overall, SEAM tops over CAM in all aspects, especially in not misclassifying background pixels as defects.

4.3.2 Precision Rate, Recall Rate, Pixel Accuracy, Intersection Over Union and Dice Score

Table 4.22 Evaluation results of fully supervised models on magnetic tile validation set

Magnetic Tile Validation Set						
Evaluation Metrics \ Models		U-Net	Double U-Net	SETR	TransU-Net	TransDA U-Net
Precision Rate (%)	Blowhole	72.37	72.31	53.65	75.44	**81.78**
	Break	80.87	85.49	77.06	**90.86**	85.09
	Crack	71.78	70.98	54.19	75.19	**75.94**
	Fray	94.57	95.47	94.43	95.39	**95.54**
	Uneven	80.92	86.46	87.01	86.35	**87.25**
	Free	**99.72**	99.70	99.54	99.68	99.67
Recall Rate (%)	Blowhole	93.59	94.40	92.70	**94.84**	89.40
	Break	92.70	91.83	92.29	91.55	**93.65**
	Crack	93.01	**96.14**	85.74	93.45	92.47
	Fray	97.34	97.11	92.15	**98.13**	96.89
	Uneven	**94.79**	94.23	93.96	93.65	93.46
	Free	98.93	99.24	99.08	99.26	**99.29**
Pixel Accuracy (%)	Blowhole	99.98	99.98	99.96	99.98	**99.99**
	Break	99.84	99.87	99.81	**99.90**	99.88
	Crack	99.92	99.92	99.84	**99.93**	**99.93**
	Fray	99.78	99.80	99.65	**99.82**	99.80
	Uneven	99.16	99.38	**99.39**	99.36	**99.39**
	Free	98.74	99.01	98.71	99.01	**99.03**
Intersection over Union (%)	Blowhole	68.95	69.34	51.47	72.46	**74.55**
	Break	76.03	79.45	72.71	**83.83**	80.45
	Crack	68.11	69.01	49.71	71.42	**71.51**
	Fray	92.19	92.83	87.40	**93.69**	92.70
	Uneven	77.48	82.12	**82.40**	81.57	82.23
	Free	98.66	98.94	98.63	98.95	**98.96**
Dice Score (%)	Blowhole	81.62	81.89	67.96	84.03	**85.42**
	Break	86.38	88.55	84.20	**91.21**	89.16
	Crack	81.03	81.67	66.40	83.33	**83.39**
	Fray	95.94	96.28	93.28	**96.74**	96.21
	Uneven	87.31	90.18	**90.35**	89.95	90.25
	Free	99.33	99.47	99.31	99.47	**99.48**

Table 4.23 Evaluation results of fully supervised models on magnetic tile test set

Magnetic Tile Test Set						
Evaluation Metrics \ Models		U-Net	Double U-Net	SETR	TransU-Net	TransDA U-Net
Precision Rate (%)	Blowhole	72.26	71.06	54.13	74.49	**80.90**
	Break	83.22	88.34	80.94	**89.93**	86.60
	Crack	68.86	68.10	52.40	71.15	**72.47**
	Fray	95.16	96.04	94.99	95.70	**96.28**
	Uneven	80.34	**88.08**	86.24	85.19	87.57
	Free	99.70	99.71	99.53	**99.72**	99.66
Recall Rate (%)	Blowhole	95.13	**96.47**	90.70	94.16	91.05
	Break	**97.03**	96.35	94.11	95.27	96.61
	Crack	91.82	**95.06**	85.31	93.61	90.44
	Fray	97.86	97.53	93.61	**97.95**	97.41
	Uneven	93.92	94.24	93.46	**94.64**	93.27
	Free	98.88	**99.29**	99.06	99.18	**99.29**
Pixel Accuracy (%)	Blowhole	99.98	99.98	99.96	99.98	**99.99**
	Break	99.87	99.90	99.83	**99.91**	99.89
	Crack	99.93	99.93	99.88	**99.94**	**99.94**
	Fray	99.80	99.81	99.67	99.81	**99.82**
	Uneven	99.10	**99.42**	99.33	99.32	99.38
	Free	98.68	**99.07**	98.69	98.97	99.03
Intersection over Union (%)	Blowhole	69.69	69.26	51.28	71.20	**74.94**
	Break	81.16	85.48	77.04	**86.09**	84.05
	Crack	64.88	65.77	48.06	**67.86**	67.31
	Fray	93.23	93.76	89.21	93.82	**93.88**
	Uneven	76.37	**83.58**	81.33	81.27	82.37
	Free	98.59	**99.00**	98.60	98.90	98.96
Dice Score (%)	Blowhole	82.14	81.84	67.80	83.18	**85.67**
	Break	89.60	92.17	87.03	**92.52**	91.34
	Crack	78.70	79.35	64.92	**80.85**	80.46
	Fray	96.49	96.78	94.30	96.81	**96.84**
	Uneven	86.60	**91.05**	89.71	89.67	90.33
	Free	99.29	**99.50**	99.29	99.45	99.48

Table 4.24 Evaluation results of fully supervised models on production item test set

Production Item Test Set						
Evaluation Metrics \ Models		U-Net	Double U-Net	SETR	TransU-Net	TransDA U-Net
Precision Rate (%)	Defective	73.64	73.64	65.63	**80.44**	75.22
	Free	99.89	99.89	99.85	99.87	**99.90**
Recall Rate (%)	Defective	71.03	71.73	61.00	67.57	**73.09**
	Free	99.90	99.90	99.88	**99.94**	99.91
Pixel Accuracy (%)	Defective	99.79	99.79	99.72	**99.81**	99.80
	Free	99.79	99.79	99.72	**99.81**	99.80
Intersection over Union (%)	Defective	56.63	67.08	46.23	58.03	**58.91**
	Free	99.79	99.79	99.72	**99.81**	99.80
Dice Score (%)	Defective	72.31	72.67	63.23	73.44	**74.14**
	Free	99.89	99.89	99.86	**99.90**	**99.90**

Table 4.25 Evaluation results of weakly supervised models on production item test set (defective only)

Production Item Test Set (defective only)			
Evaluation Metrics \ Models		CAM	SEAM
Precision Rate (%)	Defective	4.56	**55.12**
	Free	96.75	**98.29**
Recall Rate (%)	Defective	29.23	**53.47**
	Free	77.54	**98.40**
Pixel Accuracy (%)	Defective	75.82	**96.81**
	Free	75.82	**96.81**
Intersection over Union (%)	Defective	4.11	**37.25**
	Free	75.57	**96.75**
Dice Score (%)	Defective	7.90	**54.28**
	Free	86.09	**98.35**

Table 4.22 depicts that on the magnetic tile validation set evaluation, TransDAU-Net has the highest precision in the prediction of blowhole, crack, fray and uneven defect classes, while achieving the highest recall in prediction of break and defect-free classes. Conversely, TransU-Net outperforms TransDAU-Net as to precision rate for break categories prediction, and recall for blowhole and fray defects prediction. Concerning pixel accuracy, intersection over union and dice score metrics, TransDAU-Net tops over other models in blowhole, crack and defect-free classes prediction, while TransU-Net achieves the best performance in break and fray defects prediction. Although SETR model scores highest on several metrics for predicting uneven defect class, it performs relatively poorly in predicting other defect classes. Table 4.23 illustrates the evaluation results of all models on the magnetic tile test set based on each defect class. This time, TransDAU-Net in blowhole and crack defects prediction; TransU-Net in break and crack defects prediction; Double U-Net in uneven and defect-free classes prediction, outperforms other models with regards to pixel accuracy, intersection over union and dice score metrics. These results suggest that models with a hybrid CNN-Transformer architecture, i.e., TransU-Net and TransDAU-Net, perform better in predicting small-scale defect classes including blowhole, break and crack. However, it can be observed that all models perform relatively poorly in predicting relatively small defects including blowhole and crack, with much higher recall than precision, indicating a lot of false positives occur.

In view of evaluation results of the fully supervised models on production item test set, TransDAU-Net has a lower precision rate but the highest recall rate in surface defect detection compared to TransU-Net. Although TransU-Net scores highest in terms of precision rate in predicting defective pixels, it has much lower recall compared to TransDAU-Net and other CNN-based approaches. In terms of pixel accuracy, intersection over union and dice score metrics, TransDAU-Net tops over other models in defective pixel predictions, while TransU-Net is superior in background pixel predictions, with slightly higher scores than TransDAU-Net. On the other hand, the evaluation results of the weakly supervised model, CAM are not ideal, with dramatically low scores in terms of precision rate, intersection over union and dice score in defective pixel predictions. These results are definitely unacceptable for surface defect detection on real production lines. Conversely, scoring performance of the alternative weakly supervised model SEAM is satisfactory and closer to that of fully supervised models. There is a large gap between the scores of the CAM and SEAM models, indicating that the SEAM framework does improve upon the classic class activation mapping technique adopted in the CAM model.

4.3.3 Mean Intersection Over Union

Table 4.26 Mean IoU of fully supervised models on magnetic tile validation set

Models	U-Net	Double U-Net	SETR	TransU-Net	TransDAU-Net
Mean IoU (%)	80.24	81.95	73.72	**83.65**	83.40

Table 4.27 Mean IoU of fully supervised models on magnetic tile test set

Models	U-Net	Double U-Net	SETR	TransU-Net	TransDAU-Net
Mean IoU (%)	80.65	82.81	74.25	83.19	**83.58**

Table 4.28 Mean IoU of fully supervised models on production item test set

Models	U-Net	Double U-Net	SETR	TransU-Net	TransDAU-Net
Mean IoU (%)	78.21	78.43	72.98	78.92	**79.36**

According to the mean intersection over union (mIoU) metric scores considering true positives, false positives, and false negatives shown above, with a best-of-three-set system, it is undoubtedly that TransDAU-Net has the best overall performance among all fully supervised models. TransU-Net ranks the second, with a slightly lower prediction score than Double U-Net on magnetic tile test set and production item test set. The third and fourth place models in the ranking are Double U-Net and U-Net, which are far behind TransU-Net and TransDAU-Net in terms of score performance. The worst performing model is SETR, probably due to the fact that the mean intersection over union metric takes false positives into account. This strongly implies that improper use of Transformers can lead to poor model performance. Contrary to popular rumours, the Transformer-based model, SETR does not outperform CNN-based approaches, including U-Net and Double U-Net in this study. However, it is worth noting that the TransU-Net and TransDAU-Net with hybrid CNN-Transformer architecture achieve relatively good performance, hinting that a mix of CNN and Transformer methods complement each other's shortcomings. As a result, judicious architectural design of the model is crucial. Models with larger parameter sizes do not pledge optimal performance. Blindly increasing the depth and complexity of the network does not necessarily improve the performance of the model. On the contrary, a simple yet robust network design brings the crowning achievements to the model's predictions.

Table 4.29 Mean IoU of weakly supervised models on production item test set (defective only)

Models	CAM	SEAM
Mean IoU (%)	39.84	**67.00**

As expected, the scoring performance of weakly supervised models is worse than that of ordinary fully supervised models, as it is unreasonable to require weakly supervised models to achieve competitive segmentation performance due to the lack of proper training annotations. Judging by the mean intersection over union metric score on production item test set, performance of SEAM model is commensurate with fully supervised models. On the contrary, the primitive weakly-supervised approach, CAM model is totally incompetent compared to SEAM. This suggests that SEAM framework, as claimed by the authors, resolves all the shortcomings and limitations of basic class activation mapping approach and generates better revised and consistent CAMs with appropriate activations in relevant regions. If only a general orientation is needed to localize surface defects on the production items, SEAM would be a good implementation as the cost and effort of annotating pixel-level ground-truth labels is completely waived.

CHAPTER 5 CONCLUSION

Based on the methodologies and experimental results discussed in previous chapters, it is proved that TransDAU-Net achieves the best overall performance among all the constructed fully supervised models, with mean intersection over union of 83.58% and 79.36% in magnetic tile test set and production item test set evaluations respectively. What makes TransDAU-Net even more commendable is that it has fewer total parameters (113,666,648) compared to TransU-Net (127,443,014) by adopting parameter sharing technique in Transformer encoder, thus making it less burdensome in terms of computational cost. The segmentation performance of TransDAU-Net is not influenced by the reduced parameter size, credited to the multi-depth dilated inception blocks and attention modules introduced into skip connections. Additionally, another custom model built in this study, Double U-Net also achieves decent results, with only slightly poorer performance than TransU-Net and TransDAU-Net. This may be due to Double U-Net combining the advantages of U-Net's sensible architectural design with the flourishing effects of multi-depth dilated inception convolutional blocks. After all, multi-scale features and larger receptive fields are critical to the performance of a model. If one is looking for a robust segmentation model with small parameter size (< 100 million), Double U-Net would be a good choice.

Unexpectedly, Transformer-based model, SETR is far inferior in performance to CNN-based models, including U-Net and Double U-Net. It might be due to the features generated by Transformer-alone encoder only retain global interaction information and lack fine-grained spatial details. Another possibility is that skip connections are not introduced in the network, which combine and embrace fine-grained features with successive convolutional layers in the upsampling path. In any case, the experimental results strongly suggest that the blind use of Transformers may lead to the risk of worse model segmentation performance. By comparing the performance of each model, it can be demonstrated that the hybrid CNN-Transformer architecture leveraging strong global context and fine-grained details is the best one to employ.

Although the segmentation performance of weakly supervised models built in this study, CAM and SEAM are not as good as that of fully supervised models, they do not require ground truth segmentation masks but only relatively easily available image-level labels to supervise the model learning process. Furthermore, the overall performance of the SEAM model is outstanding, reaching the standards of fully supervised models. One might consider trading off segmentation performance to ease the effort of preparing ground truth pixel-level annotations.

REFERENCES

[1] Malaysian Consumer Protection Act 1999 (CPA) s. 68.1 (MY)

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Lecture Notes in Computer Science, pp. 234–241, 2015, doi: 10.48550/arXiv.1505.04597.

[3] S. A. Bala and S. Kant, "Dense Dilated Inception Network for Medical Image Segmentation," International Journal of Advanced Computer Science and Applications, vol. 11, no. 11, 2020, doi: 10.14569/ijacsa.2020.0111195.

[4] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation,", Feb. 2021, doi: 10.48550/arXiv.2102.04306.

[5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, doi: 10.1109/cvpr.2016.319.

[6] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, doi: 10.48550/arXiv.2004.04581.

[7] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, doi: 10.1109/wacv.2018.00162.

[8] Y. Huang, C. Qiu, Y. Guo, X. Wang, and K. Yuan, June. 2018, "Surface Defect Saliency of Magnetic Tile," 2018 IEEE 14th International Conference on Automation Science and Engineering (CASE), doi: 10.1109/coase.2018.8560423.

[9] J. Božič, D. Tabernik, and D. Skočaj, Apr. 2021, "Mixed supervision for surface-defect detection: From weakly to fully supervised learning," Computers in Industry, vol. 129, p. 103459, doi: 10.48550/arXiv.2104.06064.

[10] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, doi: 10.48550/arXiv.1409.4842.

[11] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, “Rethinking semantic segmentation from a sequence-to-sequence perspective with Transformers,” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, doi: 10.1109/CVPR46437.2021.00681.

[12] T. Sho and K. Shun, “Lessons on Parameter Sharing across Layers in Transformers”, Apr. 2021, 10.48550/arXiv.2104.06022.

[13] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, and B. Glocker, “Attention u-net: Learning where to look for the pancreas”, 2018, doi: 10.48550/arXiv.1804.03999.

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: S3Y3	**Study week no.: 3**
Student Name & ID: Loh Xiao 19ACB06100	
Supervisor: Dr Ng Hui Fuang	
Project Title: Automated Visual Defect Detection Using Deep Learning	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Search and download a public defect dataset, KolektorSDD2 with pixel-level annotations
- Explore and preprocess the data in the new dataset by applying techniques such as data augmentation
- Evaluate all the models proposed in FYP 1 on the new dataset

2. WORK TO BE DONE

- Study Transformer-based semantic segmentation models
- Rebuild the studied Transformer-based models in own method

3. PROBLEMS ENCOUNTERED

- All the models proposed in FYP 1 need to be rebuilt and transformed to adapt to the problem domain of new dataset
- Segmentation performance of all models on the new dataset are poor compared to the magnetic tile dataset used in FYP 1

4. SELF EVALUATION OF THE PROGRESS

Preparation of additional dataset is necessary to demonstrate further the performance of all proposed segmentation models.
It's time to work on Transformer-based models, which are yet to the knowledge.

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: S3Y3	**Study week no.: 5**
Student Name & ID: Loh Xiao 19ACB06100	
Supervisor: Dr Ng Hui Fuang	
Project Title: Automated Visual Defect Detection Using Deep Learning	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Learn the concepts and procedure flow of Transformer-based encoders
- Reproduce SEgmentation TRansformer (SETR) model using Tensorflow library
- Evaluate the SETR model on two prepared defect datasets and compare the results with other trained CNN-based models, including U-Net and Double U-Net

2. WORK TO BE DONE

- Search for the reasons why SETR is underperforming
- Explore and study alternative models that utilize Transformers

3. PROBLEMS ENCOUNTERED

- Facing issue in initializing weights of Transformer layers with pretrained ViT
- Prediction results of SETR model are inferior to those CNN-based models, which was unexpected since Transformer-based models were anticipated to outperform CNN-based models in segmentation performance from the outset

4. SELF EVALUATION OF THE PROGRESS

As one of the classic Transformer-based segmentation models, SETR has been successfully extended and implemented in this project.
However, efforts must be made to find the root of the problem of poor SETR model segmentation performance.

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: S3Y3	**Study week no.: 7**
Student Name & ID: Loh Xiao 19ACB06100	
Supervisor: Dr Ng Hui Fuang	
Project Title: Automated Visual Defect Detection Using Deep Learning	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Understand that SETR performance is suboptimal due to naive usage
- Research and construct TransU-Net with hybrid CNN-Transformer encoder
- Evaluate the TransU-Net on two prepared defect datasets and compare the results with other built models

3. WORK TO BE DONE

- Learn the working principles of weakly supervised models using only image-level labels as supervision
- Explore and construct popular weakly supervised models

3. PROBLEMS ENCOUNTERED

- Parameter size of TransU-Net is too large, placing a heavy burden on GPU memory capacity
- Training process of TransU-Net is frequently interrupted due to the tendency to exceed GPU limits

4. SELF EVALUATION OF THE PROGRESS

The objective of comparing CNN-based, Transformer-based, and hybrid CNN-Transformer based models are achieved.
Can start to work on the next objective, which is to explore and implement advanced weakly supervised semantic segmentation approaches.

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: S3Y3	**Study week no.: 9**
Student Name & ID: Loh Xiao 19ACB06100	
Supervisor: Dr Ng Hui Fuang	
Project Title: Automated Visual Defect Detection Using Deep Learning	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Learn how class activation mapping technique is capable of generating masks that highlight relevant regions without the need for pixel-level annotations
- Extend the class activation mapping technique to a classical U-Net to perform defect segmentation under weakly supervised setting

2. WORK TO BE DONE

- Look for another weakly supervised model that complements class activation mapping technique

3. PROBLEMS ENCOUNTERED

- Masks generated by class activation mapping technique are awful and only provide approximate directions about the exact location of the defect
- A lot of pixels are falsely classified by the U-Net model employing the class activation mapping technique

4. SELF EVALUATION OF THE PROGRESS

Performance of class activation mapping technique on KolektorSDD2 dataset is unsatisfactory.
An alternative weakly supervised method should be explored to perform segmentation with similar performance to fully supervised models.

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: S3Y3	**Study week no.: 11**
Student Name & ID: Loh Xiao 19ACB06100	
Supervisor: Dr Ng Hui Fuang	
Project Title: Automated Visual Defect Detection Using Deep Learning	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Research and construct SEAM framework that addresses the limitations of class activation mapping technique
- Evaluate the SEAM framework on KolektorSDD2 dataset and compare the results with the U-Net model that employs the primitive class activation mapping technique

4. WORK TO BE DONE

- Propose and develop an enhanced version of TransU-Net with smaller parameter size but similar segmentation performance
- Investigate and incorporate attention modules into the newly developed model

5. PROBLEMS ENCOUNTERED

- Details of building a complete SEAM model are not explained in the official journal, requiring a deep dive into the complex and lengthy source code
- Building a SEAM framework from scratch is difficult due to its complex design and dopant many additive regularization operations

4. SELF EVALUATION OF THE PROGRESS

The goal of exploring and building weakly supervised defect segmentation models has been achieved. Efforts should now be made to build a custom model.

Supervisor's signature

Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: S3Y3	**Study week no.: 13**
Student Name & ID: Loh Xiao 19ACB06100	
Supervisor: Dr Ng Hui Fuang	
Project Title: Automated Visual Defect Detection Using Deep Learning	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- Incorporates multi-dilated inception block proposed in FYP1 and attention modules into the newly developed model
- Propose and develop a new custom model, TransDAU-Net, that achieves better segmentation performance with smaller parameter size

3. WORK TO BE DONE

- Tidy up the codes, visualizations and prediction results for all models
- Complete the FYP report and submit for Turnitin check

3. PROBLEMS ENCOUNTERED

- Evaluation results of TransDAU-Net on magnetic tile validation set are inferior compared to that of TransU-Net
- Model validation phase and test phase must be done in separate sessions due to GPU limits

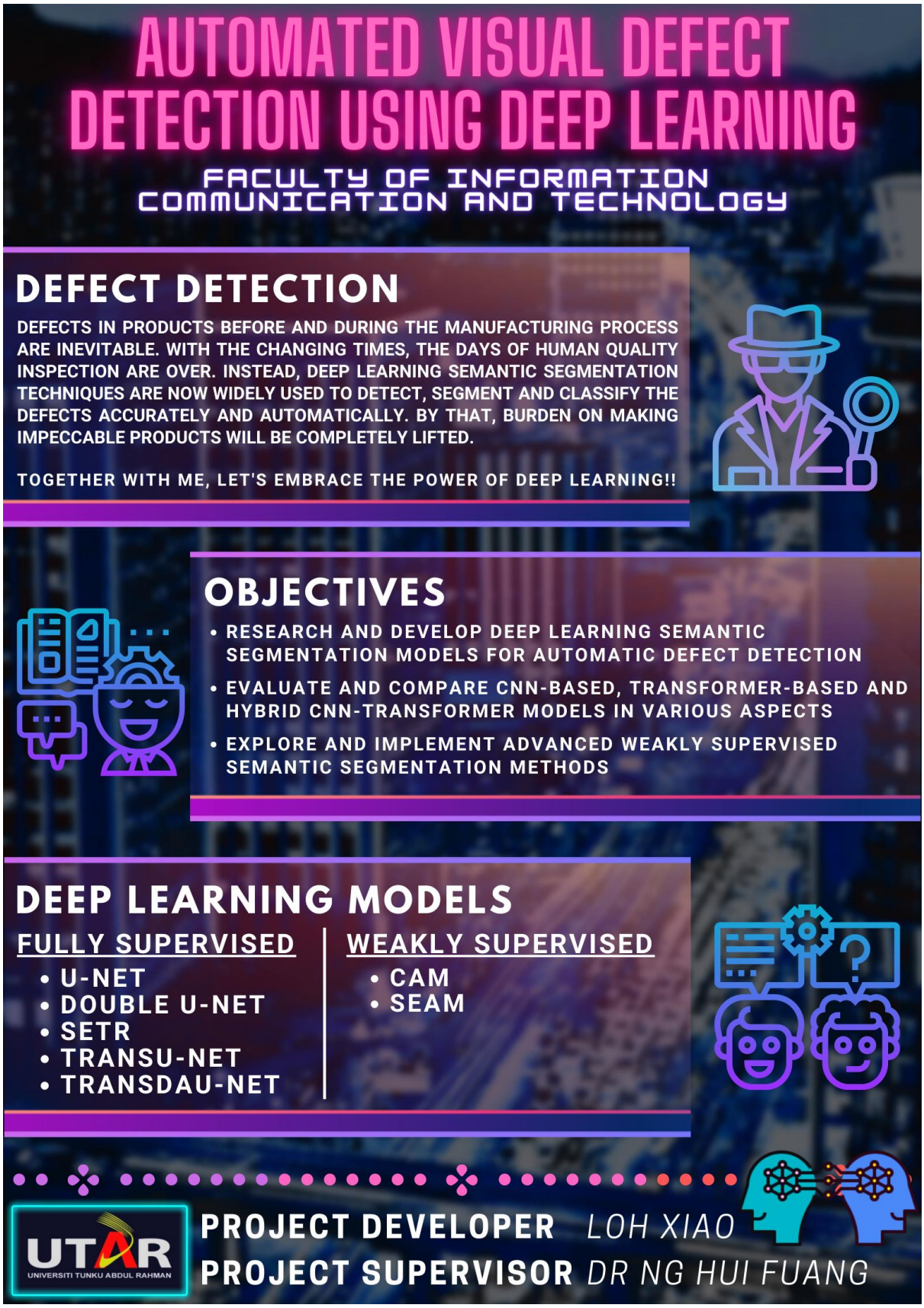
4. SELF EVALUATION OF THE PROGRESS

Everything is on the right track.
Attention should be paid to the upcoming presentation.

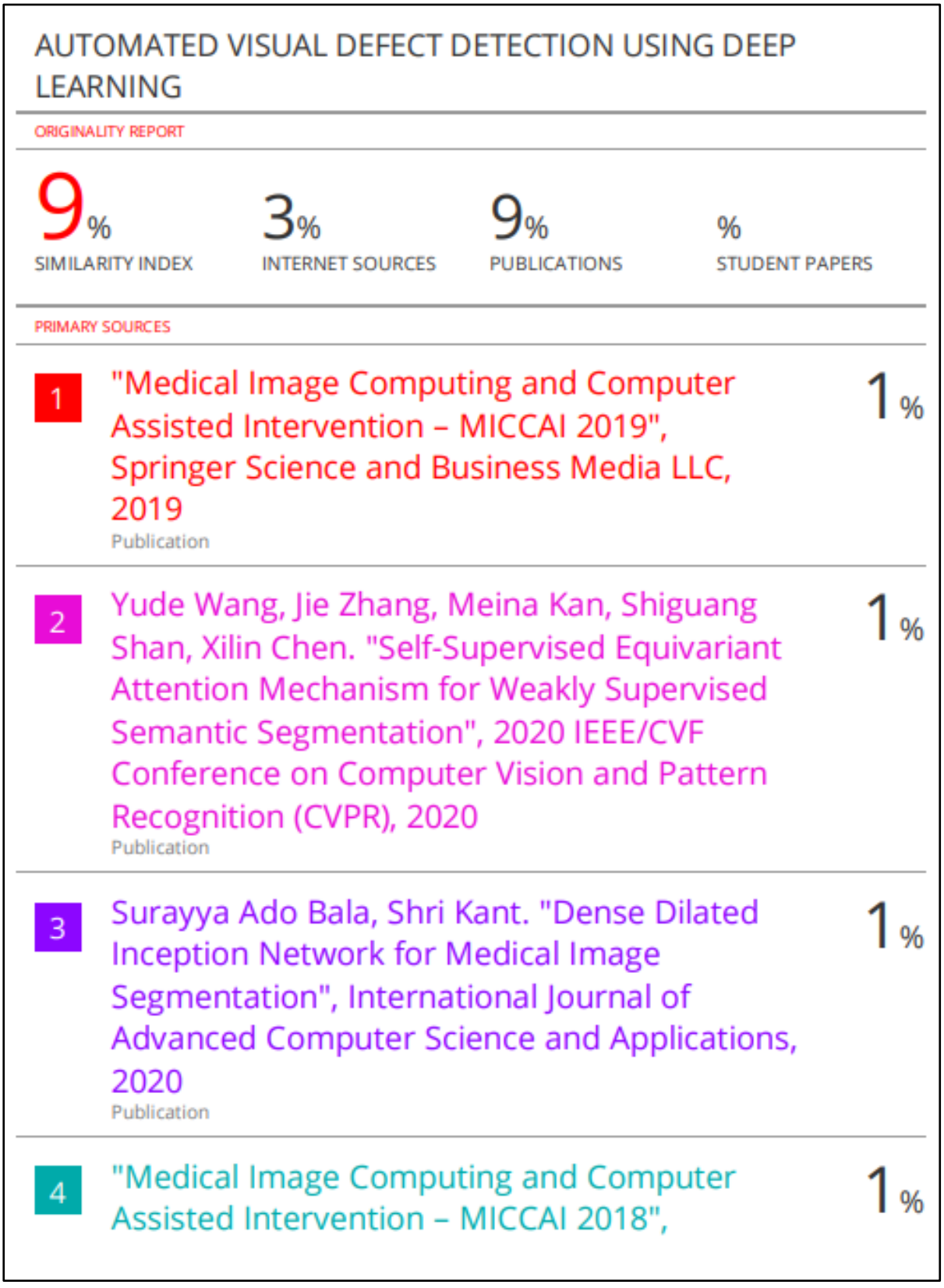
Supervisor's signature

Student's signature

POSTER


AUTOMATED VISUAL DEFECT DETECTION USING DEEP LEARNING
FACULTY OF INFORMATION COMMUNICATION AND TECHNOLOGY
DEFECT DETECTION
DEFECTS IN PRODUCTS BEFORE AND DURING THE MANUFACTURING PROCESS ARE INEVITABLE. WITH THE CHANGING TIMES, THE DAYS OF HUMAN QUALITY INSPECTION ARE OVER. INSTEAD, DEEP LEARNING SEMANTIC SEGMENTATION TECHNIQUES ARE NOW WIDELY USED TO DETECT, SEGMENT AND CLASSIFY THE DEFECTS ACCURATELY AND AUTOMATICALLY. BY THAT, BURDEN ON MAKING IMPECCABLE PRODUCTS WILL BE COMPLETELY LIFTED.
TOGETHER WITH ME, LET'S EMBRACE THE POWER OF DEEP LEARNING!!
OBJECTIVES
• RESEARCH AND DEVELOP DEEP LEARNING SEMANTIC SEGMENTATION MODELS FOR AUTOMATIC DEFECT DETECTION
• EVALUATE AND COMPARE CNN-BASED, TRANSFORMER-BASED AND HYBRID CNN-TRANSFORMER MODELS IN VARIOUS ASPECTS
• EXPLORE AND IMPLEMENT ADVANCED WEAKLY SUPERVISED SEMANTIC SEGMENTATION METHODS
DEEP LEARNING MODELS
FULLY SUPERVISED
• U-NET
• DOUBLE U-NET
• SETR
• TRANSU-NET
• TRANSDAU-NET
WEAKLY SUPERVISED
• CAM
• SEAM
UTAR
UNIVERSITI TUNKU ABDUL RAHMAN
PROJECT DEVELOPER LOH XIAO
PROJECT SUPERVISOR DR NG HUI FUANG

PLAGIARISM CHECK RESULT



AUTOMATED VISUAL DEFECT DETECTION USING DEEP LEARNING

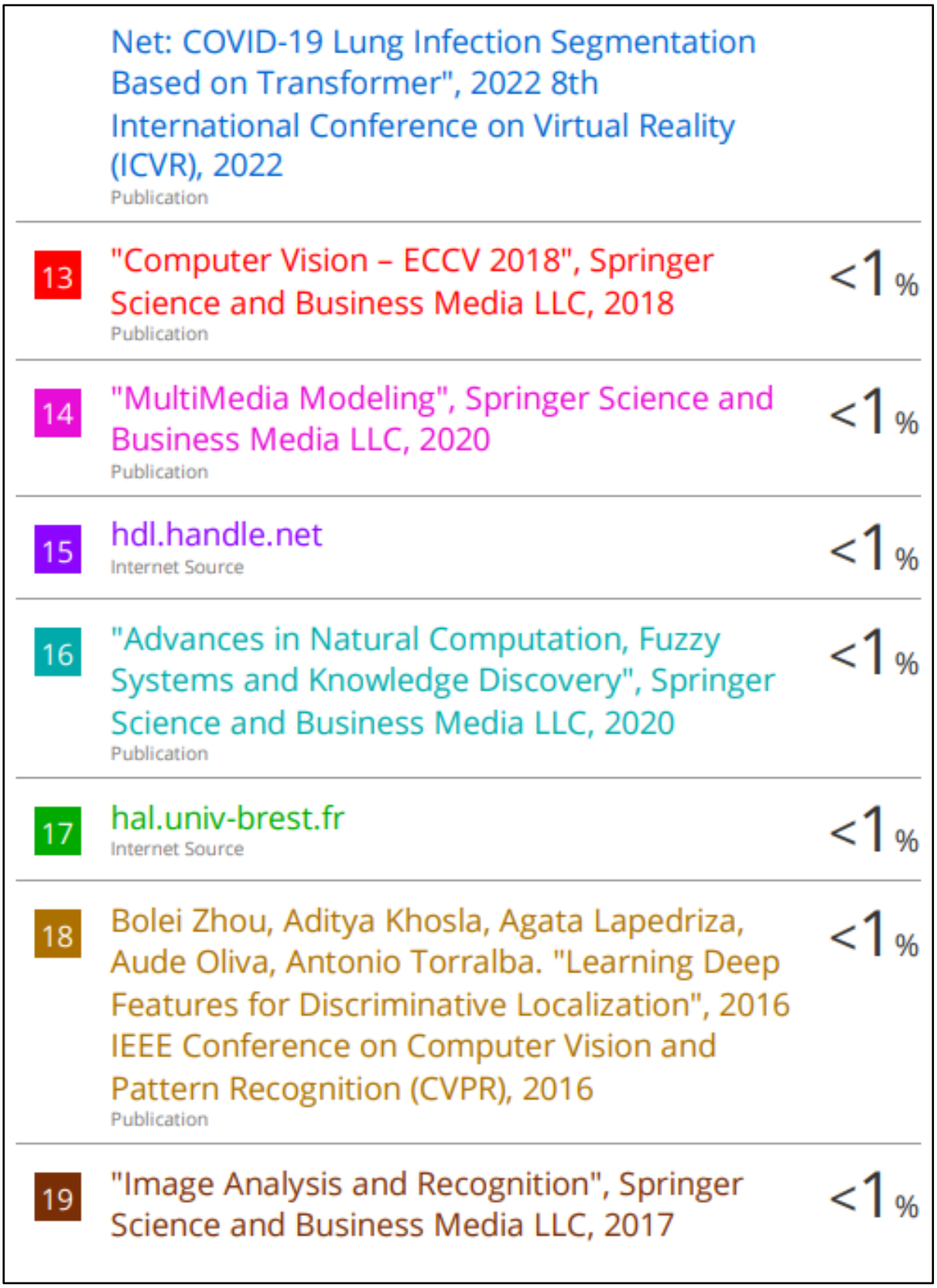
ORIGINALITY REPORT

9%	3%	9%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

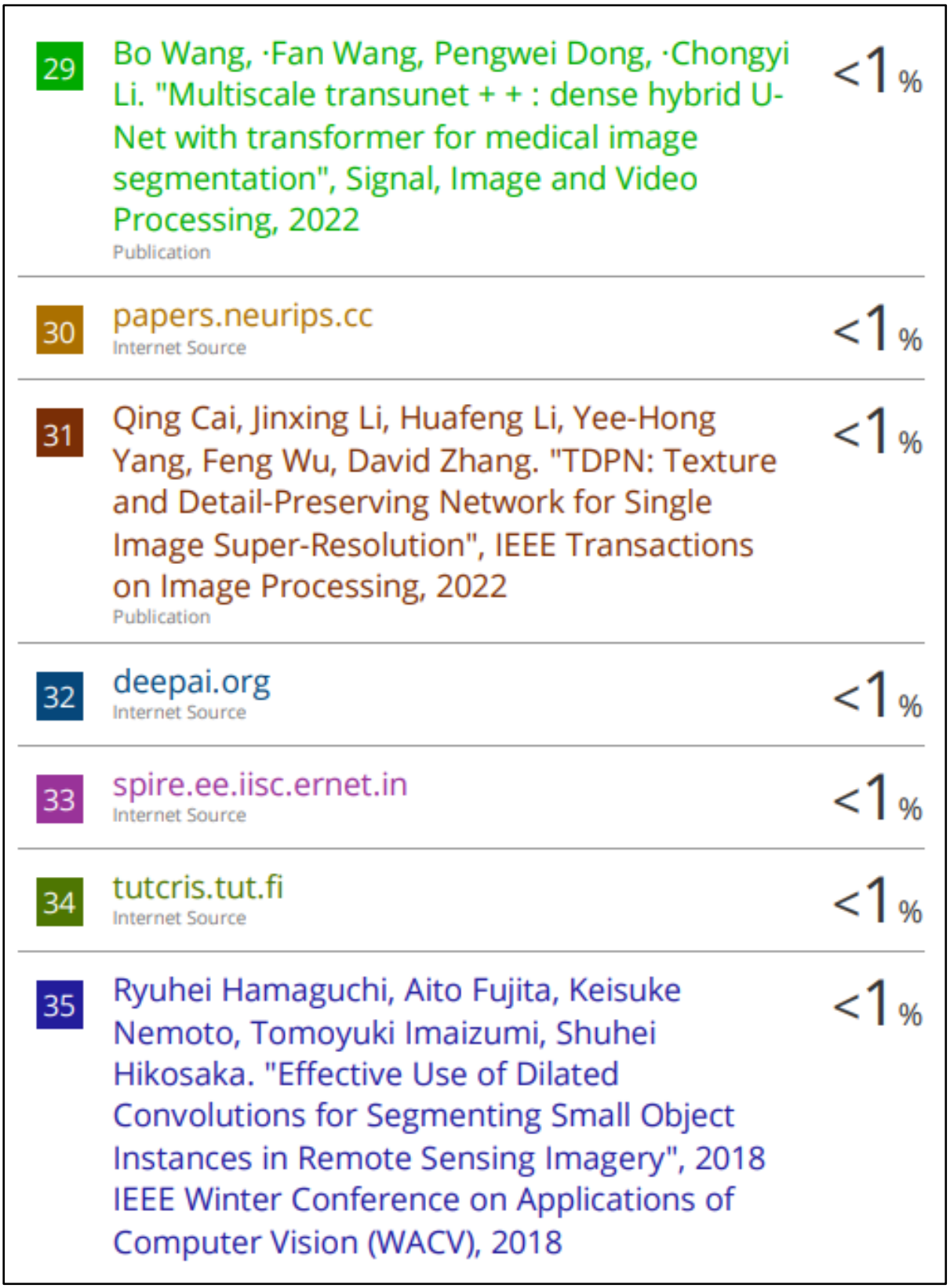
PRIMARY SOURCES

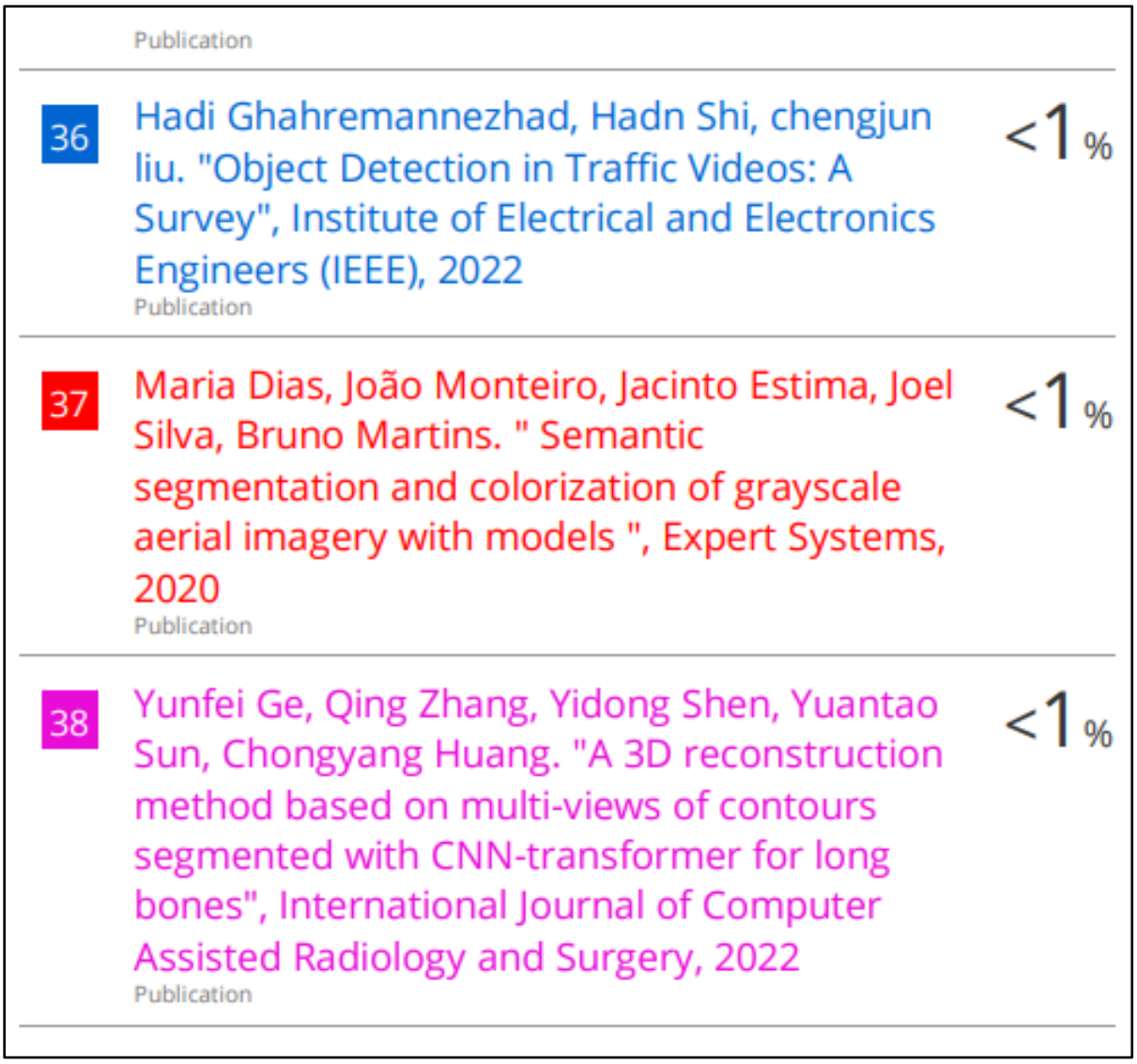
1	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2019", Springer Science and Business Media LLC, 2019 Publication	1%
2	Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, Xilin Chen. "Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation", 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020 Publication	1%
3	Surayya Ado Bala, Shri Kant. "Dense Dilated Inception Network for Medical Image Segmentation", International Journal of Advanced Computer Science and Applications, 2020 Publication	1%
4	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2018",	1%


Springer Nature America, Inc, 2018
Publication
5
"Medical Image Computing and Computer Assisted Intervention – MICCAI 2020", Springer Science and Business Media LLC, 2020
Publication
<1 %
6
d2l.ai
Internet Source
<1 %
7
"Pattern Recognition", Springer Nature, 2019
Publication
<1 %
8
Lecture Notes in Computer Science, 2015.
Publication
<1 %
9
"Pattern Recognition and Machine Intelligence", Springer Science and Business Media LLC, 2019
Publication
<1 %
10
"Computer Vision – ECCV 2016", Springer Nature, 2016
Publication
<1 %
11
Abhinav Valada, Rohit Mohan, Wolfram Burgard. "Self-Supervised Model Adaptation for Multimodal Semantic Segmentation", International Journal of Computer Vision, 2019
Publication
<1 %
12
Jie Zhang, Kunlan Xiang, Jingyi Wang, Jiahao Liu, Mengfei Kang, Zhigeng Pan. "Trans-Inf-
<1 %


Net: COVID-19 Lung Infection Segmentation Based on Transformer", 2022 8th International Conference on Virtual Reality (ICVR), 2022
Publication
13
"Computer Vision – ECCV 2018", Springer Science and Business Media LLC, 2018
Publication
<1 %
14
"MultiMedia Modeling", Springer Science and Business Media LLC, 2020
Publication
<1 %
15
hdl.handle.net
Internet Source
<1 %
16
"Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery", Springer Science and Business Media LLC, 2020
Publication
<1 %
17
hal.univ-brest.fr
Internet Source
<1 %
18
Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba. "Learning Deep Features for Discriminative Localization", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016
Publication
<1 %
19
"Image Analysis and Recognition", Springer Science and Business Media LLC, 2017
<1 %


Publication
20 www.shortscience.org
Internet Source
<1 %
21 "Computer Vision – ECCV 2016", Springer Science and Business Media LLC, 2016
Publication
<1 %
22 Qingyang Li, Ruofei Zhong, Xin Du, Yu Du. "TransUNetCD: A Hybrid Transformer Network for Change Detection in Optical Remote Sensing Images", IEEE Transactions on Geoscience and Remote Sensing, 2022
Publication
<1 %
23 Lecture Notes in Computer Science, 2011.
Publication
<1 %
24 academic.oup.com
Internet Source
<1 %
25 dash.harvard.Edu
Internet Source
<1 %
26 www.mdpi.com
Internet Source
<1 %
27 www.ncbi.nlm.nih.gov
Internet Source
<1 %
28 "Advances in Multimedia Information Processing – PCM 2017", Springer Science and Business Media LLC, 2018
Publication
<1 %


29
Bo Wang, ·Fan Wang, Pengwei Dong, ·Chongyi Li. "Multiscale transunet + + : dense hybrid U-Net with transformer for medical image segmentation", Signal, Image and Video Processing, 2022
Publication
<1 %
30
papers.neurips.cc
Internet Source
<1 %
31
Qing Cai, Jinxing Li, Huafeng Li, Yee-Hong Yang, Feng Wu, David Zhang. "TDPN: Texture and Detail-Preserving Network for Single Image Super-Resolution", IEEE Transactions on Image Processing, 2022
Publication
<1 %
32
deepai.org
Internet Source
<1 %
33
spire.ee.iisc.ernet.in
Internet Source
<1 %
34
tutcris.tut.fi
Internet Source
<1 %
35
Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, Shuhei Hikosaka. "Effective Use of Dilated Convolutions for Segmenting Small Object Instances in Remote Sensing Imagery", 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018
<1 %


Publication
36
Hadi Ghahremannezhad, Hadn Shi, chengjun liu. "Object Detection in Traffic Videos: A Survey", Institute of Electrical and Electronics Engineers (IEEE), 2022
Publication
<1 %
37
Maria Dias, João Monteiro, Jacinto Estima, Joel Silva, Bruno Martins. " Semantic segmentation and colorization of grayscale aerial imagery with models ", Expert Systems, 2020
Publication
<1 %
38
Yunfei Ge, Qing Zhang, Yidong Shen, Yuantao Sun, Chongyang Huang. "A 3D reconstruction method based on multi-views of contours segmented with CNN-transformer for long bones", International Journal of Computer Assisted Radiology and Surgery, 2022
Publication
<1 %

Universiti Tunku Abdul Rahman			
Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.: 1of 1

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	Loh Xiao
ID Number(s)	19ACB06100
Programme / Course	CS
Title of Final Year Project	Automated Visual Defect Detection Using Deep Learning

Similarity	**Supervisor's Comments (Compulsory if parameters of originality exceed the limits approved by UTAR)**
Overall similarity index: 9 % **Similarity by source** Internet Sources: 3 % Publications: 9 % Student Papers: 0 %	
Number of individual sources listed of more than 3% similarity: 0	

Parameters of originality required and limits approved by UTAR are as Follows:

(i) Overall similarity index is 20% and below, and

(ii) Matching of individual sources listed must be less than 3% each, and

(iii) Matching texts in continuous block must not exceed 8 words

Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.

Note: Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

Signature of Supervisor

Name: Dr Ng Hui Fuang

Date: 9 September 2022

Signature of Co-Supervisor

Name: _____

Date: _____

UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY
(KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student ID	19ACB06100
Student Name	Loh Xiao
Supervisor Name	Dr Ng Hui Fuang

TICK (√)	DOCUMENT ITEMS Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
	Front Plastic Cover (for hardcopy)
√	Title Page
√	Signed Report Status Declaration Form
√	Signed FYP Thesis Submission Form
√	Signed form of the Declaration of Originality
√	Acknowledgement
√	Abstract
√	Table of Contents
√	List of Figures (if applicable)
√	List of Tables (if applicable)
√	List of Symbols (if applicable)
√	List of Abbreviations (if applicable)
√	Chapters / Content
√	Bibliography (or References)
√	All references in bibliography are cited in the thesis, especially in the chapter of literature review
	Appendices (if applicable)
√	Weekly Log
√	Poster
√	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)
√	I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report.

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.

(Signature of Student)
Date: 8 September 2022