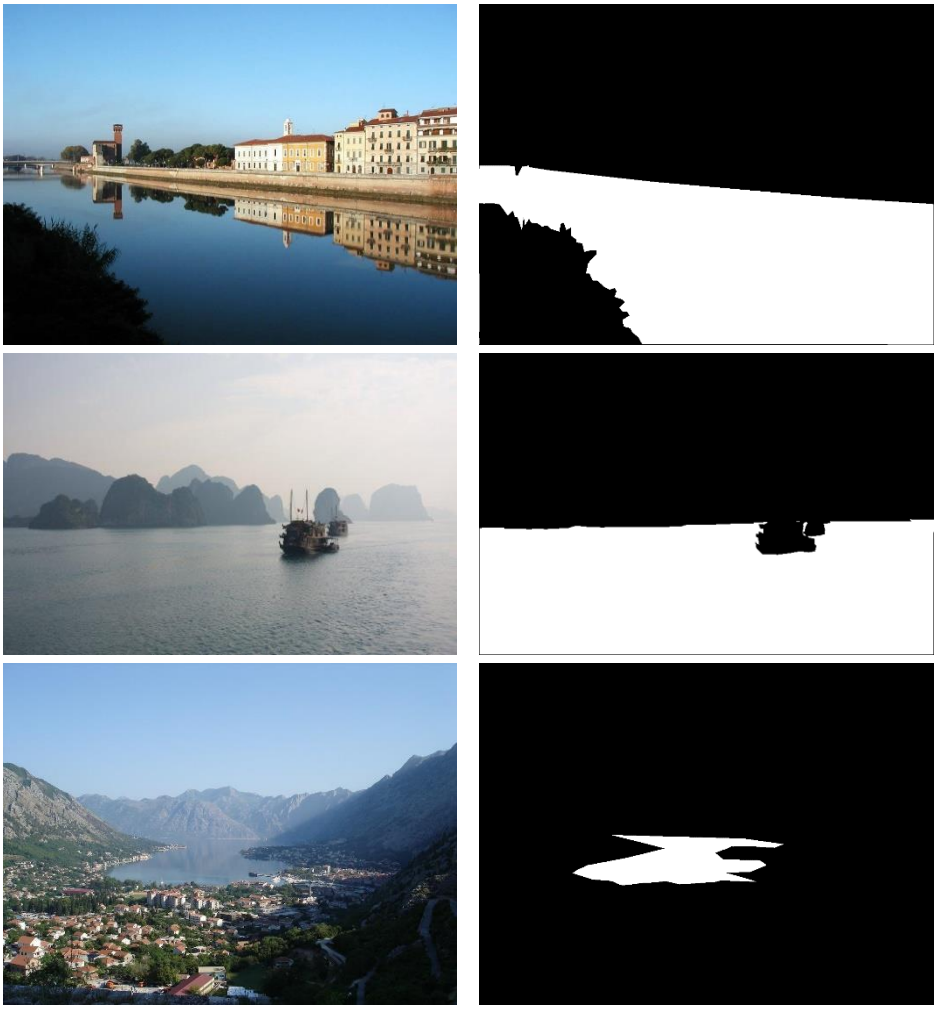
DIP Final Project Report

Group 39 : 109511207 蔡宗儒、109511234 蔡鎧宇

I. Observation on Dataset

我們原先想使用傳統方法來完成此次的 water segmentation，但是經過我們觀察這次給的 dataset 後，我們認為會有幾個問題在傳統方法會變得難以辨識或難以選定使用的 feature，如下圖 1。因此我們最後使用 deep learning 的方式，讓他從 dataset 中學習要使用的 feature，去分辨這些有問題的部分。



a	b
c	d
e	f

Figure 1: (a)(b) 反射讓水面的顏色和 edge 被改變，無法單純使用 edge-based 去分割水的部分。*(c)(d)* 霧氣等容易把物件的邊界模糊掉，edge-based 表現會不好。*(e)(f)* 天空、背景顏色與水面相似，甚至類似水面反射，使用 region-based 或 cluster-based 也難以辨識。

II. Decoder Model

在本次專題中，為了找出合適的模型去訓練，我們使用 4 個常用於 Image Segmentation 的 decoder model，去分析和比較各自的架構差異，並用數據比較和驗證。分別是 Unet[1]、Unet++[2]、Linknet[3]、FPN[4]

A. Unet

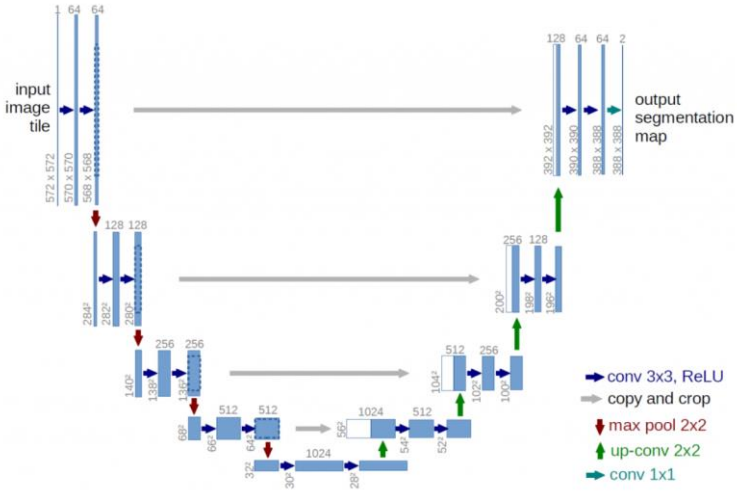


Figure 2: Unet architecture

Unet 為 Deep Learning 在 image segmentation 領域的基礎模型。由於 convolution-based segmentation model 若直接將 downsample 後的 feature map 做 deconvolution upsample，會出現位置資訊丟失的問題，而 Unet 透過在不同大小的 feature map 之間透過 residual & concat 連接起來，保留 downsample 的過程中的一部份位置資訊，傳遞到後面 decoder，大幅度的解決了位置資訊缺失的問題。

B. Unet++

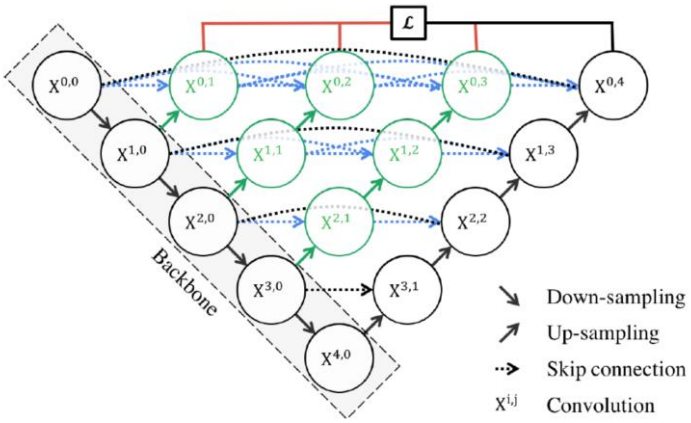


Figure 3: Unet++ architecture

Unet++在概念上是 Unet 的延伸，並在 upsampling 和 residual 的過程中增加了很多的 node，藉此增加模型的複雜度，能夠去學習更多特徵，但是同時卻增加了許多的運算複雜度和參數量，且需要較大的 dataset 和 epoch 數去讓他收斂。

C. Linknet

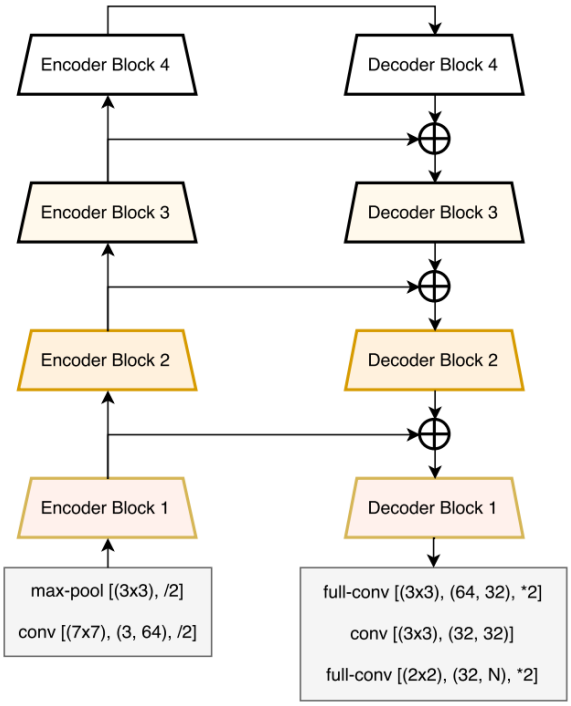


Figure 4: Linknet architecture

Linknet 同樣是針對 Unet 去做優化，在他的論文中提到，原先的 Unet 很多都是拿 VGG19 或 ResNet50 去做 Unet 的 backbone，需要很大的計算量，因此論文改用 ResNet18 當作 backbone 去做訓練以減少計算量。

再來，在原先 Unet 的架構中，在傳遞 Residual 之前，會先做 pooling 在傳到 decoder 做 concat，這個 pooling 層的參數是無法計算梯度去做訓練的，因此會讓梯度不容易傳遞，讓模型不容易收斂，也降低模型的成效。所以 Linknet 是直接將 downsample 和 upsample 中相同大小的 feature map 直接做 concat 合併，並在每個 encoder 對應的 decoder 共用一部份參數，讓 decoder 可以透過和 encoder 相近的參數去還原影像，一來減少參數量的使用也增進了模型的 performance。

D. FPN

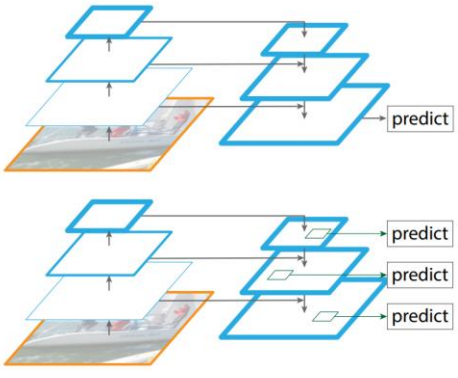


Figure 5: FPN architecture

FPN 原為做 object detection 的模型，但原論文也有做 segmentation 的板。在 Unet 的基礎下，FPN 改變了部分 Unet 的架構，在不同大小的 feature map 分別 predict 一個 mask 出來，並將不同大小的 mask 透過 upsample 生成出相同大小的 mask 後，全部相加再用 conv 和 fc 生成出預測的 mask，相比原先 unet 的架構，FPN 多了在不同大小下預測的 mask，讓模型有多種不同的 receptive field，可以找到在不同大小下的 feature map，不同大小的 object，提升模型的辨識能力。

E. Conclusion

Table 1: Comparison on different decoder model

Decoder	Improvement
Unet	提出在 upsample 和 downsample 的 feature map 中加入 residual 以傳遞位置資訊。
Unet++	增加 upsample 和 residual 中的 conv node 數量，以增加模型的複雜度。
Linknet	使用較輕量化的 backbone，並簡化 Unet 中的 pooling 層以更好的傳遞梯度優化訓練。
FPN	在不同大小的 feature map 都預測一個 mask，讓模型能取得在不同大小上不同強度的資訊。

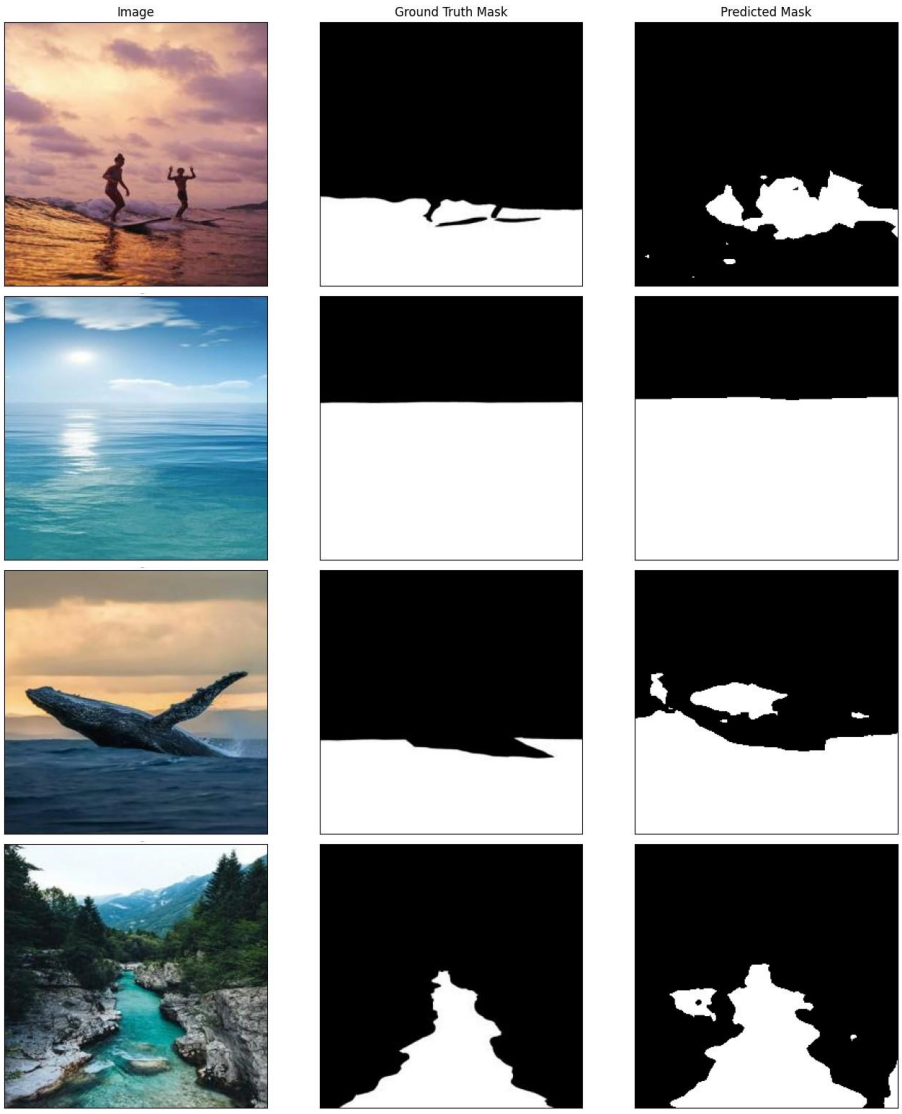
經過以上的分析比較，我們討論出在本次給的圖片水域大小差異很大，若能利用 FPN 在不同大小下的優異的特徵提取能力，最後的表現會最好，最後選擇 FPN 這個 decoder model。

但是由於從頭訓練模型只用 60 張圖太容易出現 overfitting，造成 test 上的預測失準，因此我們上網尋找 FPN 用於 water segmentation 的 pretrained model & weights，最後找到 moblienet_v2 [5]為 backbone 的 FPN pretrained model [8][9]，再針對本次期末專題給的 60 張影像去做 finetune，希望能得到比較好的結果。

III. Result

Table 2: IoU from Demo

Input	IoU
1	0.15
2	0.64
3	0.81
4	0.76
5	0.97
6	0.92
7	0.90
8	0.89
9	0.82
10	0.85
11	0.95
12	0.70
Avg	0.78



a	b	c
d	e	f
g	h	i
j	k	l

Figure 6: (a)(b)(c) image, ground truth and predicted mask from input1. *(d)(e)(f)* from input5. *(g)(h)(i)* from input3 *(j)(k)(l)* from input4

我們從表 2 可觀察出大部分的 predict IoU 都有在 0.7 以上，但是在第一張圖的結果特別的差，因此我們特別去比較第一張圖和其他張圖的差異，試圖去解釋為何分辨的結果那麼差。

由圖 6 可以觀察到，與其他圖比起來，input1 的水面顏色與 dataset 中的較不相同，且有浪花影響，讓水面的邊緣的特徵與我們訓練的 dataset 的特徵分布並不相近，因此讓這張圖的辨識結果較差，若是在其他張水面顏色都較接近藍色，且與周邊物件都有明顯落差的情況下，與我們的訓練 dataset 分布較為接近，得到的辨識結果也會比較好。

IV. Appendix

在使用這些 pretrained model 之外，我們也使用[8]提供的 dataset，自己訓練幾個不同的 decoder + encoder 組合的 pretrain weight，再去對本次的 60 張影像去做 finetune，試著去比較不同組合間的結果。(此為我們額外自己做的比較，並無使用在 demo 上。)

Decoder model 我們選擇我們前面比較過的 4 種 Unet[1]、Unet++[2]、Linknet[3]、FPN[4]，encoder model 我們則選擇了 3 種，分別是前面使用過的 MobileNet_v2[5]，和 EfficientNet-b3[6]、VGG11[7]，與 mobilenet 相比，EfficientNet 為較新的架構，VGG11 為較舊的架構，我們希望以上述模型來做比較，結果如下。

Table 3: IoU from each model using different encoder and decoder

model	efficientnet-b3	mobilenet_v2	vgg11	Avg
FPN	0.83287	0.78079	0.81569	0.80978
Linknet	0.82386	0.79218	0.77852	0.79819
Unet	0.82859	0.75932	0.78720	0.79170
Unet++	0.81038	0.76806	0.81770	0.79871

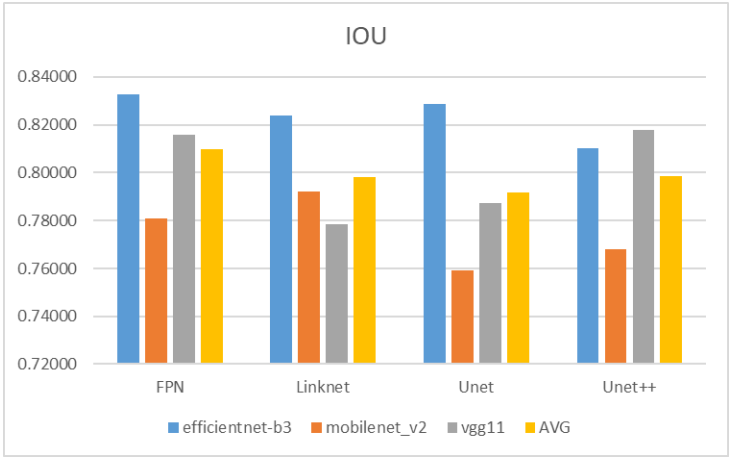


Figure 7: Comparison of each encoder model using the same decoder

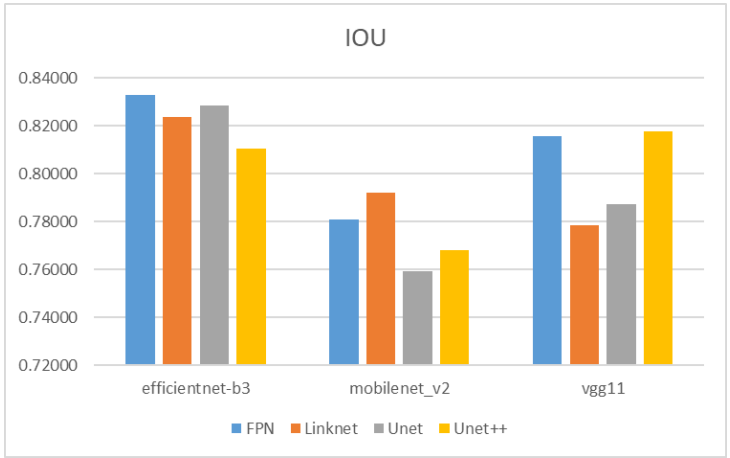


Figure 8: Comparison of each decoder model using the same encoder

從圖 7、8 可看到，EfficientNet 的表現為各個模型中最好，由於它是藉由 compound scaling method，去搜尋模型在什麼樣的寬度、深度、解析度三個維度能得到最高的精確度，在 CNN 影像分類中取得了很好的表現。將其用在 image segmentation 的 backbone 上，也不意外的取得很好的成績。

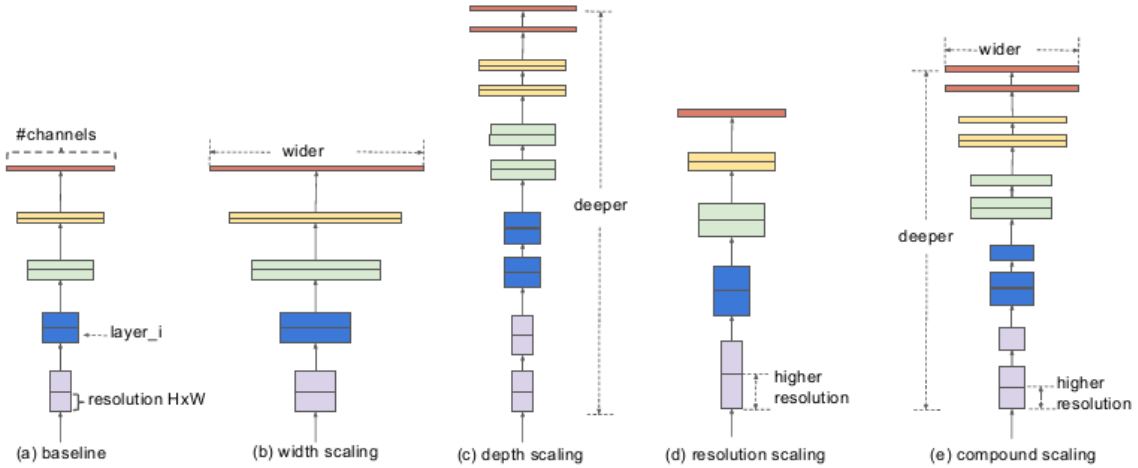


Figure 9: scaling method

而 MobileNet v2 和 VGG11 相比，表現卻幾乎略遜 VGG11，我們推測是由於 MobileNet 主要是在做模型的輕量化，因此犧牲掉了部分的精確度，以換取較低的參數量和 FLOPs，而 VGG11 則有較大的參數量，大約為 9M，相對的 MobileNet v2 只有 2M，因此 VGG11 能在資料中學習到更多的特徵，且複雜度較高較不會受到 overfitting 的影響，最後訓練出來的辨識結果也比較好。

Decoder 以平均來做比較的話，FPN 表現如預期的最好，在各個 encoder 組合下都有較好的精確度，較能從各種不同的大小中去尋找需要的特徵並預測出較精準的 mask。Linknet 跟 Unet++的表現則相當接近，且都大於 Unet 的表現，確實有做到一定程度的優化。最原始的 Unet 則最低。但是在平均上的差異都還在 0.02 以內，主要還是以 encoder 的差異最大。

V. Conclusion

我們最後得到最好的結果是 FPN+EfficientNet b3，IoU 來到最高的 0.83，跟我們預期的相符。第二名為 Unet+EfficientNet b3，Unet 雖然為最基礎的架構，但是在跟較新的模型比較，仍有不差的結果，因此到現在仍被廣泛使用，是有其原因的。Linknet 和 Unet++對 Unet 做了一定程度的改善，但我們在訓練 Unet++時，發現訓練需要的參數量較大，時間也較長，甚至接近其他的 3 倍，可是他的模型雖然複雜度提高，但是 performance 卻沒有高出 Linknet 多少，所以我們認為他並不是一個很好的選擇，在訓練時間和記憶體大小的考量下，還是選擇 FPN 或 Linknet 較為合適。Encoder 則選擇 EfficientNet，可以達到最好的精確度。

VI. Reference

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

[2] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4 (pp. 3-11). Springer International Publishing.

[3] Chaurasia, A., & Culurciello, E. (2017, December). Linknet: Exploiting encoder representations for efficient semantic segmentation. In 2017 IEEE visual communications and image processing (VCIP) (pp. 1-4). IEEE.

[4] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).

[5] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

[6] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

[7] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[8] https://github.com/erdemunal35/WaterSegNets

[9] https://github.com/qubvel/segmentation_models.pytorch