Salient Object Detection: Bas2Net
AML 19/20 - Project Report

Irene Cannistraci 1603090, Giovanni Ficarra 1659089

1 Abstract

Salient Object Detection is an emerging field in Computer Vision, which goal is to identify most relevant objects in natural scenes.
Our proposal is to investigate the combination of two relevant state of the art works in SOD field (*BasNet* [1] and *Res2Net* [2]) into a new one, we called *Bas2Net*. The resulting model requires 40% less training time and leads to better results on different datasets, comparing BasNet and Bas2Net both trained for 100 epochs (instead of the original 100,000 epohcs). Also in terms of memory required to store the models we got a good results by achieving an improvement of 50%.

2 Introduction

Our project is focused on the *Salient Object Detection* task, that aims to emulate human vision system by trying to identify visually distinctive regions in real life scenes. SOD plays an important role in many computer vision tasks, such as visual tracking, robot navigation and content-aware image editing.
It basically consists of two phases, first input images are analyzed to detect the most significant object(s) and then they are accurately segmented based on the previous detection step. In SOD, it is important to understand the global image as a whole, but without neglecting smaller details, which can be very relevant as well, especially in the segmentation step.

Our purpose is to modify the backbone of BasNet, replacing ResNet-34 with Res2Net-50, to see if we could obtain better results.

3 Related works

3.1 BasNet

In 2019 Qin et al. proposed BasNet [1], a neural network with a predict-refine architecture. It is composed by two modules, the first one is a U-Net-like [3] densely supervised Encoder-Decoder network, which deals with saliency prediction, while the second one is a multi-scale residual refinement module, which is in charge of refining the saliency map. Furthermore, they proposed a hybrid loss, which is a combination of Binary Cross Entropy (BCE), Structural SIMilarity (SSIM) and Intersection over Union (IoU) losses.

3.2 Res2Net

In the same year, Gao et al. proposed Res2Net [2], a building block for CNNs, which exploits hierarchical residual-like connections within one single residual block, to achieve multi-scale representation of features at a finer level and enlarge the range of receptive fields for each layer. As the name suggests, this block is derived from ResNet [4]. The Res2Net strategy introduces a new dimension, called *scale*, which is added to depth, width and cardinality, and they proved to be very effective. This new block was applied to many networks, with different goals, always obtaining more or less evident gains in terms of performances.

4 Proposed method

Our idea came out of the reading of the two papers about *BasNet* and *Res2Net*. The authors of the second paper, during the experiment phase, demonstrate that this new block can improve the performances of various models such as *InceptionV3*, *ResNet-50*, *ResNeXt-50* and *DLA-60*. Whereas, as we previously said, the authors of the *BasNet* paper showed how good is their model in SOD by using *ResNet-34*. By looking at these two approaches and at the good results they both obtained, we decided to combine them in order to try to improve the performance of BasNet by using the optimum performances of *Res2Net-50* as backbone.

From a technical point of view, our purpose is to replace ResNet blocks (called `basic blocks` in ResNet-34 and `bottlenecks` in ResNet-50) with the `bottle2necks` which characterizes Res2Net-50. The main difference between these two blocks is that in the second one they introduce intermediate convolutional layers that are combined to produce a multi-scale representation of the input features. These differences can be better appreciated in Figure 1, while figures 2 and 3 compare BasNet and Bas2Net architectures.

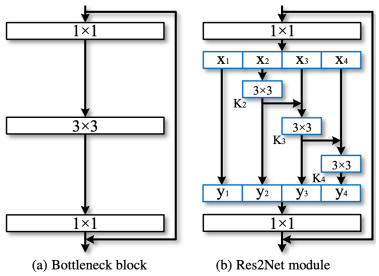

Figure 1: Bottleneck VS Bottle2neck, from Res2Net [2] paper.

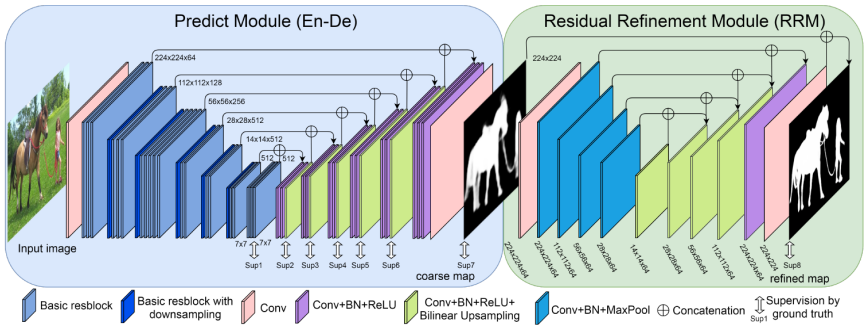

Figure 2: BasNet's architecture, from Res2Net [1] paper.

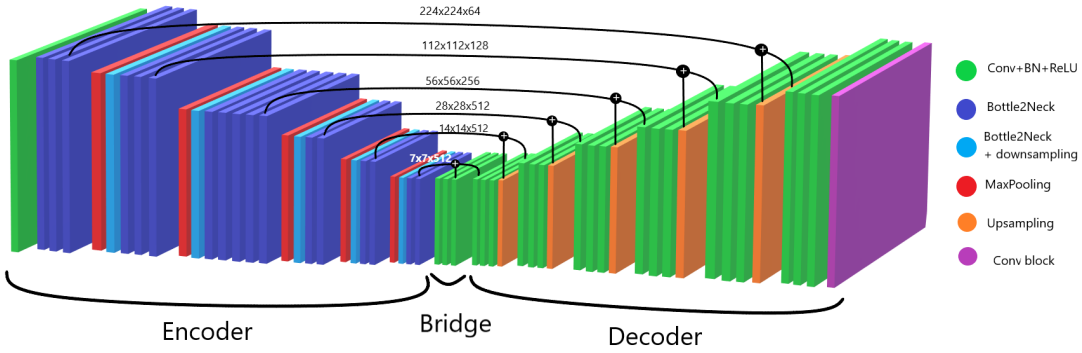

Figure 3: Architecture of Bas2Net's Predict Module.

During this operation, we had to face a problem due to the network structure: each block of the encoder needs to produce an output which fits not only with the next block, but also the specular block in the decoder. When we started our project we had not taken into account this blocking problem since, when we first read the BasNet paper, we thought that the backbone was a "black box". Instead, during this phase, we noticed that the backbone was tailored to the model. So, after replacing the layers, we had to re-fit the new blocks into our model. This made it difficult to use a pretrained backbone.

To be able to compare the results of our model (which we called **Bas2Net**) to those obtained by the original BasNet, we decided to retrain the latter for 100 epochs on the DUTS-TR [5] dataset, the same used in the original paper. Then we did the same with Bas2Net. This choice is due to the fact that BasNet was trained for 100000 epochs, but we can't replicate this scenario with our time and our devices.

5 Dataset and Benchmark

In the original paper the authors trained the model on DUTS and then used six different datasets to test it. In order to better compare our model with theirs, we decided to train Bas2Net on DUTS and then test it on **DUTS** [5], **PASCAL-S** [6] and **ECSSD** [7] (three of the six datasets used by the author of the paper).

- **DUTS** contains 15,572 images and is divided into **DUTS-TR** and **DUTS-TE**. Both training and test images represent natural scenes and have different size; the author of DUTS asserted that it is one of the largest saliency detection benchmark with the explicit training/test evaluation protocol.
- **PASCAL-S** dataset was built starting from segmented images, and applying eye tracking techniques to determine which were the most salient objects. It consists of X images with the relative masks, thus it is one of the biggest datasets for SOD.
- **ECSSD** is a dataset generated starting from the Complex Scene Saliency Dataset (CSSD). The authors, in order to represent the situations that natural images generally fall into, extended CSSD in to a larger dataset (ECSSD) with 1000 images, which includes semantically meaningful but structurally complex images for evaluation.

6 Experimental results

Before reaching a configuration that gave us a working model, we tried different approaches:

1. Replace any block of layers in BasNet (encoder and decoder) with a Bottle2Neck;
2. Reduce the size of the network to better understand how each block was joint to the others;
3. Replace only the actual basic blocks in the encoder (the decoder doesn't contain basic blocks), carefully observing the shapes obtained at each layer by BasNet.

Finally, we got a working configuration, by using the third approach.

6.1 Experiment 1

Our first experiment, mostly motivated by time and device constraints, consisted in training BasNet and Bas2Net with the training-set size reduced from 10,553 to 1000 images and with the number of epochs recuded from 100000 to 100. As shown in Figures 4 and 5, the results obtained by Bas2Net (rs2) were only slightly worse then those given by BasNet (rs1), but we could observe that the training time was non-negligibly shorter with our model: the training of both models was executed on Google Colab with the same GPU (Tesla T4 with 15079MiB of dedicated memory), and BasNet required 5h 12m, while Bas2Net took only 3h 40m. To go deeper into the results, we used the **Binary-Segmentation-Evaluation-Tool** [8], that is the same code used by BasNet's authors to compute the following scores:

	BasNet	Bas2Net
Average MAE	0.084	0.101
Max F-measure	0.753	0.714
Mean F-measure	0.725	0.687

Table 1: Compared scores with 1000 images from DUTS-TE

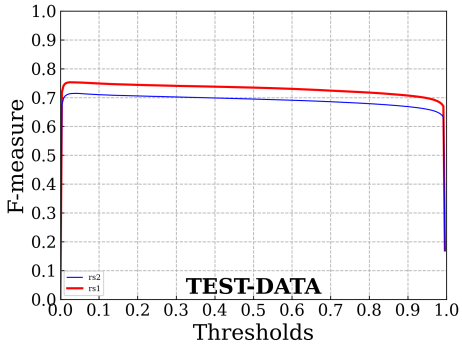

Figure 4: F-measure curves with 1000 images from DUTS-TE.

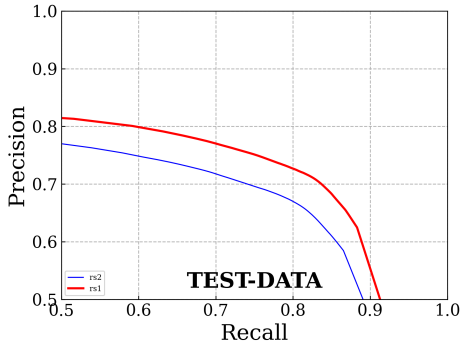

Figure 5: Precision-Recall curves with 1000 images from DUTS-TE.

6.2 Experiment 2

The second experiment consists in training BasNet and Bas2Net with the entire training-set size, that is 10,553, for 50 epochs. Starting from this experiment we used a Microsoft Azure Virtual Machine with a GPU Tesla K80 with 12GB of dedicated memory. Using this configuration BasNet required 66h 30m, while Bas2Net took only 40h 0m. Following are the scores obtained by testing the three different datasets.

6.2.1 DUTS-TE Test

First, we decided to test this model on DUTS-TE [5]. As shown in Figures 6 and 7, the results obtained by Bas2Net were only slightly worse then those given by BasNet, but the training was faster with our model. Following are the scores and the curves we got:

	BasNet	Bas2Net
Average MAE	0.068	0.070
Max F-measure	0.808	0.804
Mean F-measure	0.779	0.772

Table 2: Compared scores with the entire DUTS-TE dataset

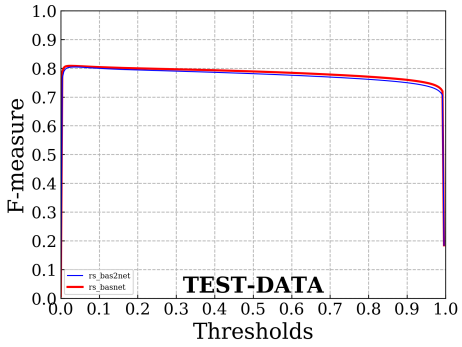

Figure 6: F-measure curves with the entire DUTS-TE dataset.

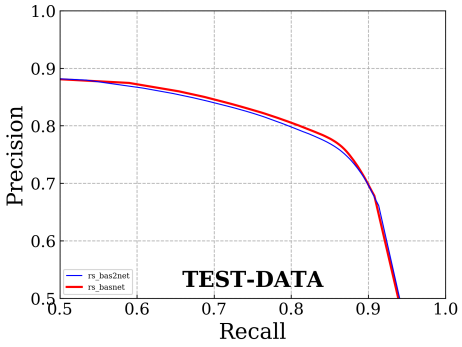

Figure 7: Precision-Recall curves with the entire DUTS-TE dataset.

6.2.2 PASCAL-S Test

Then, we decided to test this model on PASCAL-S [6]. As shown in Figures 8 and 9, the results obtained using this dataset are better than those obtained using DUTS: the MAE is just 0.003 higher by using Res2Net, while the Max F-measure and the Mean F-measure obtained with Bas2Net are respectively just 0.001 and 0.004 smaller than the one obtained with BasNet. Following are the scores and the curves we got:

	BasNet	**Bas2Net**
Average MAE	0.072	0.075
Max F-measure	0.834	0.833
Mean F-measure	0.815	0.811

Table 3: Compared scores with PASCAL-S

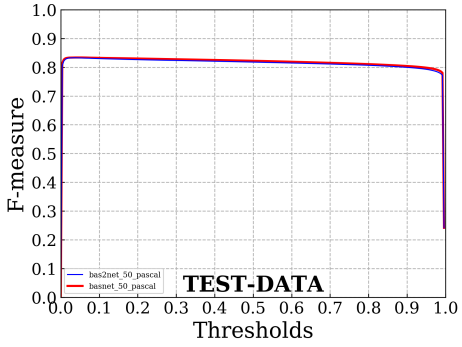

Figure 8: F-measure curves with PASCAL-S.

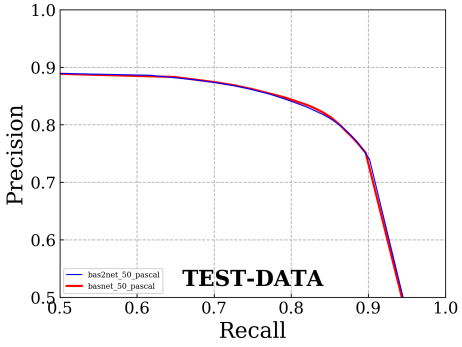

Figure 9: Precision-Recall curves with PASCAL-S.

6.2.3 ECSSD Test

Our last test for this model, was on the Extended Complex Scene Saliency Dataset (ECSSD) [7] which contains 1000 test images. As shown in Figures 10 and 11, the results obtained using this dataset are better than those obtained using both DUTS and PASCAL-S: the MAE is exactly the same for both the models, while the Max F-measure and the Mean F-measure obtained with Bas2Net are respectively just 0.001 and 0.003 smaller than the one obtained with BasNet. Following are the scores and the curves we got:

	BasNet	Bas2Net
Average MAE	0.055	0.055
Max F-measure	0.906	0.905
Mean F-measure	0.890	0.887

Table 4: Compared scores with ECSSD

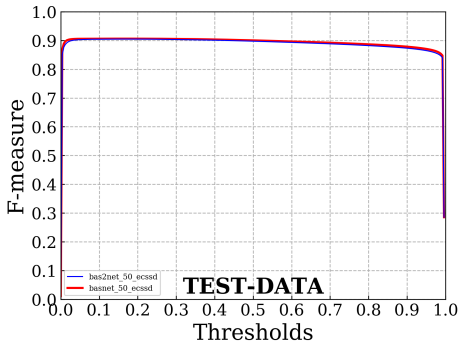

Figure 10: F-measure curves with ECSSD.

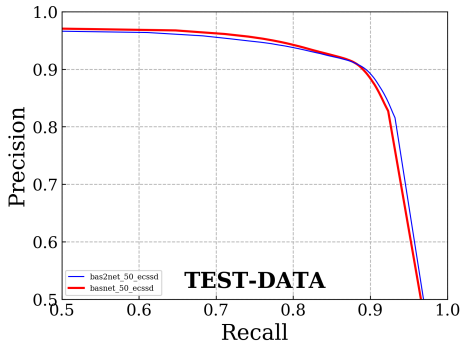

Figure 11: Precision-Recall curves with ECSSD.

6.3 Experiment 3

The last experiment consists in training BasNet and Bas2Net with the entire training-set size, that is 10,553, for 100 epochs. The duration of the training was proportional to the one reported for Experiment 2 6.2.

6.3.1 DUTS-TE test

As for the previous experiment, we first tested this model on the same dataset we used for training, but using the test images (DUTS-TE instead of DUTS-TR). In this case (see Figure 12 and Figure 13), the results obtained by Bas2Net were better than those given by BasNet and the training was faster with our model. Following are the results:

	BasNet	Bas2Net
Average MAE	0.064	0.064
Max F-measure	0.814	0.822
Mean F-measure	0.790	0.791

Table 5: Compared scores with DUTS-TE

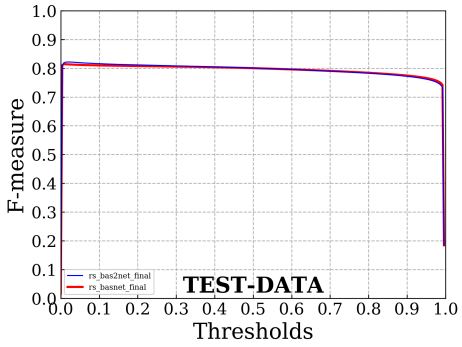

Figure 12: F-measure curves with DUTS-TE.

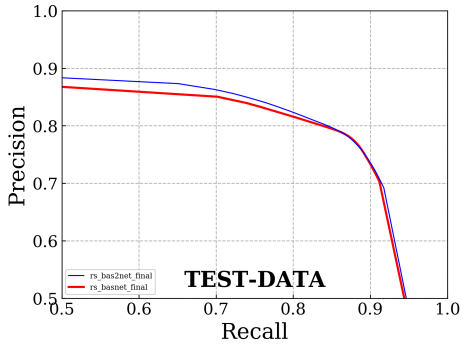

Figure 13: Precision-Recall curves with DUTS-TE.

In figure 14 are show some images from DUTS-TE, with the relative masks predicted by BasNet and Bas2Net.

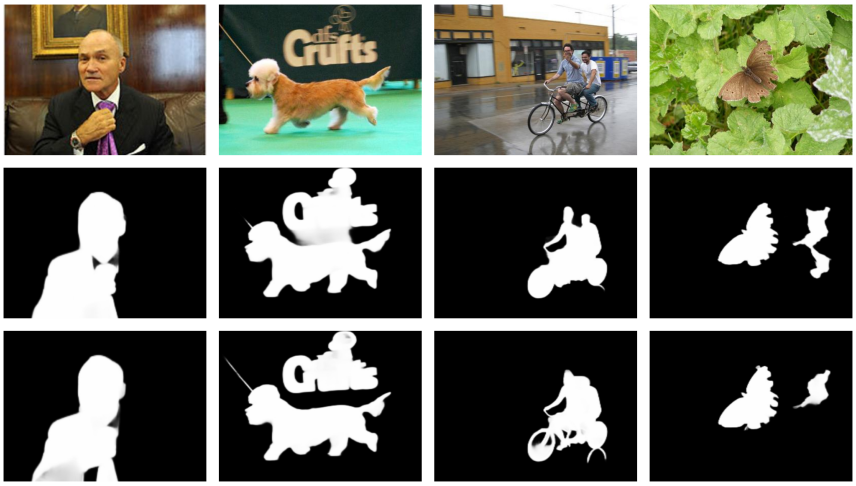

Figure 14: Sample of DUTS-TE images.

6.3.2 PASCAL-S test

Then, we tested this model on the PASCAL-S dataset. Also in this case, see Figure 15 and Figure 16, the results obtained by Bas2Net were a little bit better than those given by BasNet. Following are the results:

	BasNet	**Bas2Net**
Average MAE	0.071	0.070
Max F-measure	0.839	0.839
Mean F-measure	0.819	0.820

Table 6: Compared scores with PASCAL-S

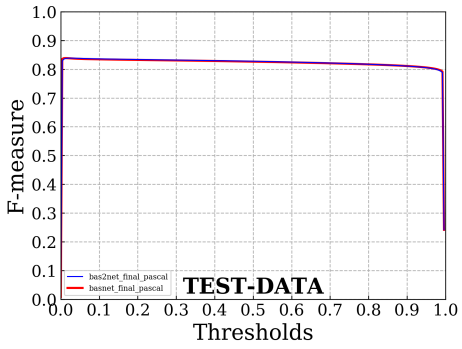

Figure 15: F-measure curves with PASCAL-S.

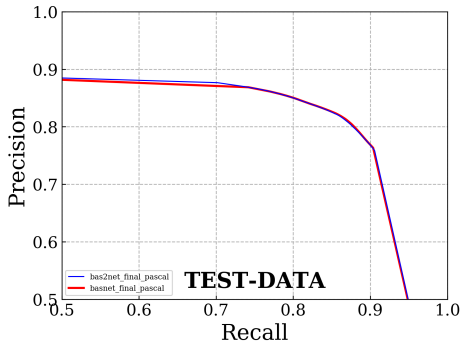

Figure 16: Precision-Recall curves with PASCAL-S.

In figure 17 are show some images from PASCAL-S, with the relative masks predicted by BasNet and Bas2Net.

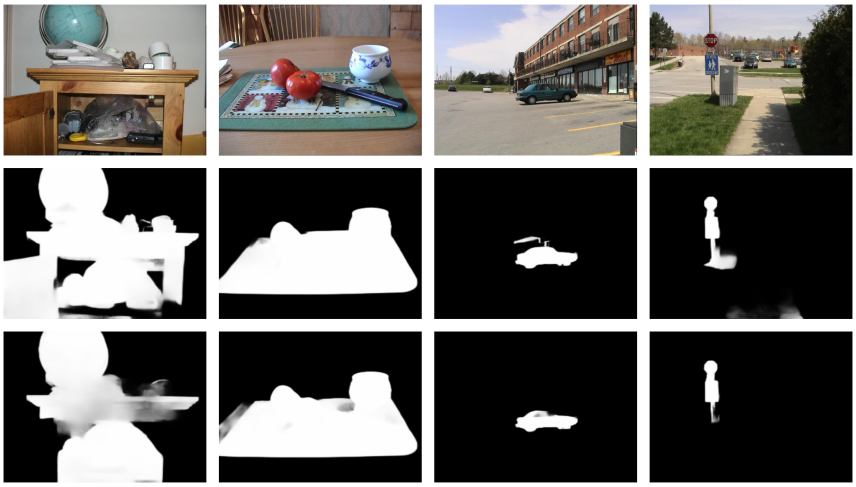

Figure 17: Sample of PASCAL-S images.

6.3.3 ECSSD test

At the end we tested this model on the ECSSD. Also in this case, see Figure 18 and Figure 19, the results obtained by Bas2Net were better than those given by BasNet, except for the Mean Absolute Error that is 0.001 higher with Bas2Net. Following are the results:

	BasNet	**Bas2Net**
Average MAE	0.048	0.049
Max F-measure	0.914	0.920
Mean F-measure	0.898	0.903

Table 7: Compared scores with ECSSD

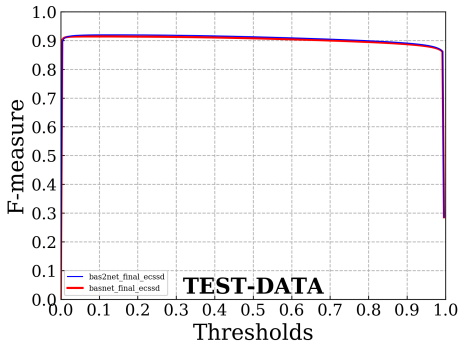

Figure 18: F-measure curves with ECSSD.

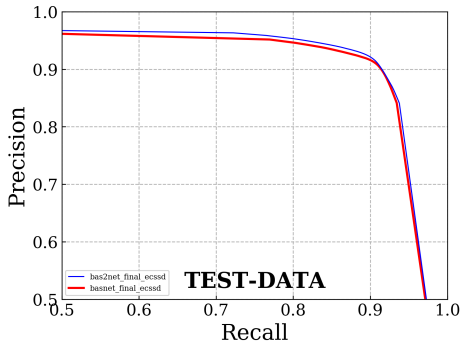

Figure 19: Precision-Recall curves with ECSSD.

In figure 20 are show some images from ECSSD, with the relative masks predicted by BasNet and Bas2Net.

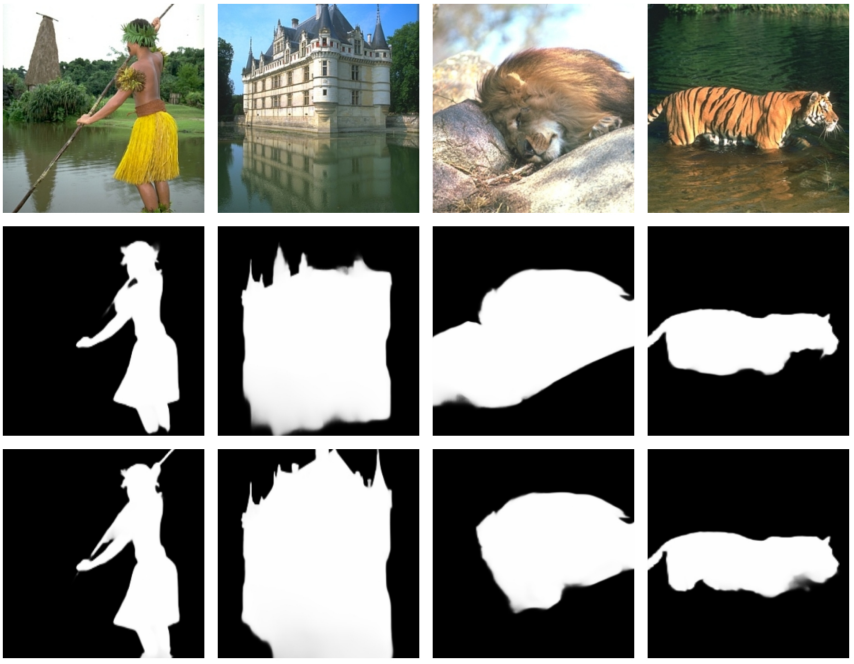

Figure 20: Sample of ECSSD images.

7 Conclusions

This was the first time we faced a similar task and we are now aware of the fact that it is very complex to do, especially if the model you are trying to modify is not written by yourself. What we learned and what we would like to share and keep in mind for the future is, first of all, to proceed step by step, by starting from the replacement of small parts of the model and see what happens. If possible, try to simplify the model parameters (such as number of epochs and training images, input size, batch size) in order to do faster test and see the evolution of the changes. Another important point is to visualize the model information such as the size of each layer, so that you can keep track of the model structure while you are changing it.

Our experiments were based mainly on two approaches: the first one consisted in testing both BasNet and Bas2Net trained for 50 epochs and the second one for 100. The results obtained by comparing the models trained for 50 epochs showed that BasNet was a little bit better in terms of performances than our model, but slower. Instead, the results given by the second approach showed that our model was not only faster, but also better in terms of MAE and F-Measure scores. Thus, this scenario could be given by the fact that even if our model has less parameters, it is deeper than BasNet, so it needs more training to perform better. The following table reports a summary of the scores we obtained for both the models:

Model	**DUTS-TE**			**PASCAL-S**			**ECSSD**		
	MAE	*Max-F*	*Mean-F*	*MAE*	*Max-F*	*Mean-F*	*MAE*	*Max-F*	*Mean-F*
BasNet	0.064	0.814	0.790	0.071	0.839	0.819	0.048	0.914	0.898
Bas2Net	0.064	0.822	0.791	0.070	0.839	0.820	0.049	0.920	0.903

Table 8: Experiments summary.

As we can notice by looking at the Table 8, Bas2Net followed the trend of BasNet, giving higher scores on the ECSSD, then on PASCAL-S, then on DUTS-TE. But, on the first one, our F-measure scores were considerably better.

As we previously said, even if in the Experiment 2 (see section 6.2) Bas2Net performed worse, in all ours experiments it was better in terms of training velocity and the models saved during the training phase were smaller w.r.t the models saved during the training of BasNet.

8 Future works

To improve and better understand our results, we think many different paths may be took:

- Retraining Bas2Net with different scales may exploit better the power of Res2Net, leading to even better results;
- Comparing Bas2Net with other state of the art models in SOD, such as PoolNet [9] and Res2Net-PoolNet [10], would make our findings more robust;
- Finally, finding a way to apply transfer learning to Bas2Net, using pretrained weights from the first layers of Res2Net, may significantly improve our results.

References

[1] Xuebin Qin, Zichen Zhang, Chenyang Huang, et al. *BASNet: Boundary-Aware Salient Object Detection*. 2019. URL: https://paperswithcode.com/paper/basnet-boundary-aware-salient-object.

[2] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, et al. *Res2Net: A New Multi-scale Backbone Architecture*. 2019. URL: https://paperswithcode.com/paper/res2net-a-new-multi-scale-backbone.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. URL: https://paperswithcode.com/paper/u-net-convolutional-networks-for-biomedical.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. *Deep Residual Learning for Image Recognition*. 2016. URL: https://paperswithcode.com/paper/deep-residual-learning-for-image-recognition.

[5] Wang Lijun, Lu Huchuan, Wang Yifan, et al. *Learning to Detect Salient Objects with Image-level Supervision*. 2017. URL: http://saliencydetection.net/duts/.

[6] Y. Li, X. Hou, C. Koch, et al. *The Secrets of Salient Object Segmentation*. 2014, pp. 280–287. URL: http://cbs.ic.gatech.edu/salobj/.

[7] Jianping Shi, Qiong Yan, Li Xu, et al. *Hierarchical Image Saliency Detection on Extended CSSD*. Vol. 38. 2016, pp. 717–729. URL: http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/dataset.html.

[8] Qin Xuebin, Zhang Zichen, Huang Chenyang, et al. *Binary Segmentation Evaluation Tool*. 2019. URL: https://github.com/NathanUA/Binary-Segmentation-Evaluation-Tool.

[9] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, et al. *Res2Net: A New Multi-scale Backbone Architecture*. 2019. URL: https://paperswithcode.com/paper/a-simple-pooling-based-design-for-real-time.

[10] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, et al. *Res2Net: A New Multi-scale Backbone Architecture*. 2019. URL: https://github.com/Res2Net/Res2Net-PoolNet.