When Machines Find Sites for the Archaeologists: A Preliminary Study with Semantic Segmentation applied on Satellite Imagery of the Mesopotamian Floodplain

Luca Casini
University of Bologna, Department of Computer Science and Engineering
Bologna, Italy
luca.casini7@unibo.it

Valentina Orrú
University of Bologna, Department of History and Cultures
Bologna, Italy
valentina.orru@unibo.it

Nicolò Marchetti
University of Bologna, Department of History and Cultures
Bologna, Italy
nicolo.marchetti@unibo.it

Marco Roccetti
University of Bologna, Department of Computer Science and Engineering
Bologna, Italy
marco.roccetti@unibo.it

ABSTRACT

In the perspective of landscape archaeology, remote sensing is a very important tool that allows to recognize and locate potential sites, which will then be "groundtruthed" through a surface survey. Remote sensing is, unfortunately, a very time-consuming process that scales terribly with the size of the area under investigation. In this paper we explore the possibility of using semantic segmentation models to detect and highlight the presence of archaeological sites present in the Mesopotamian floodplain. Whereas archaeologists usually combine information from a variety of basemaps, including aerial and satellite photos taken from the 1950s onwards, we investigated the possibility of using an easily accessible online maps (in our case, Bing Maps). Trying to build an accessible and lightweight system also dictated the choice of trying pretrained segmentation models and use transfer learning. The preliminary results obtained (from different models and parameters choices), as well as the dataset, its idiosyncrasies and how we can deal with them are discussed in this paper.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**.

KEYWORDS

archaeology, semantic segmentation, mesopotamian floodplain, human-in-the-loop

ACM Reference Format:
Luca Casini, Valentina Orrú, Nicolò Marchetti, and Marco Roccetti. 2022. When Machines Find Sites for the Archaeologists: A Preliminary Study with Semantic Segmentation applied on Satellite Imagery of the Mesopotamian Floodplain. In *Conference on Information Technology for Social Good (GoodIT'22), September 7–9, 2022, Limassol, Cyprus.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3524458.3547121

GoodIT'22, September 7–9, 2022, Limassol, Cyprus

ACM ISBN 978-1-4503-9284-6/22/09.
https://doi.org/10.1145/3524458.3547121

1 INTRODUCTION

Remote sensing indicates the array of technologies and techniques that allows one to observe and survey something from a distance, without directly interacting with such objects or their surroundings. In archaeology, remote sensing has allowed researchers to make discoveries in a non-invasive way, using radar technology, for example, or to detect remotely sites of interest from the analysis of aerial or satellite images. The latter case is the type of remote sensing this paper is referring to. In the context of a collaboration between data scientists and archaeologists, it is interesting to study the feasibility of an intelligent system that learns to detect potential sites by looking at satellite imagery, in pretty much the same way humans are doing right now. From an archaeological viewpoint, this task is a fundamental step, although it is also extremely time consuming, given the size of the areas under observation (hundreds of thousands of square kilometers in the case of the eastern Mesopotamian floodplain in Iraq).

In particular, we are dealing here with a type of archaeological site called in Arabic "tell". They are a typical element of near eastern landscapes. The Arabic word "tell" literally means hill and indicates a stratification of buildings mostly made of mudbricks and debris that, over time, resulted in an actual artificial hill. Given the nature of the Mesopotamian floodplain, these elements tend to emerge visibly from the landscape and to be recognizable from satellite imagery. The shape, size and color can vary considerably, but they generally present an elliptic form with red and brown hues (as they are composed of clay).

Our research built upon two basic reference tools: the first one is public, Bing Maps, which is easily and freely available within QGIS (an open-source GIS software), while the second is about the digitizing in a GIS of the spatial data collected by multiple archaeological surveys from the 1950s to the present day in the southern Mesopotamian floodplain [9].

Before moving to an introduction on the automatic methodology we have tried to employ, it is important to underline how our

vision of a useful automation of the work to be carried out in this scenario is not that of a system that completely replaces the human activity, but rather that of an automatic assistant that can pre-screen thousands of images while pointing the human expert to the more appropriate location to be studied. In turn, the collaboration with the archaeologists can inform the engineering choices for a better system [4].

In a previous work, we have already tried to approach a similar problem, by dividing the geographical area under investigation (which, indeed, was far smaller than the one we are considering here) into small overlapping tiles that we proceeded to classify as containing a site or not. The resulting classification probabilities were then averaged to remove the overlap and used as a tiled heatmap to guide the user [12]. All things considered, we obtained satisfactory results, but the system was unwieldy, and still a bit imprecise.

This time, instead, the task was framed as a problem of semantic segmentation. Semantic segmentation is one of the main challenges in computer vision and consists in learning to predict a kind of mask that corresponds to a certain object of interest in the picture. There could be multiple different classes of interests to be recognized in the picture, with the most common example being different road elements in the camera footage for autonomous driving. For most applications, though, the class of interest can be just one, while the rest of the picture is considered as a background. This is exactly the case of our study, as well as the setup for most medical applications [3].

It must be noted that segmentation is a supervised learning technique that relies on an annotation of pictures provided by human experts. This makes this approach not always viable, as the manual annotation can be very time consuming, and even expensive to carry out. Fortunately, the big data and its annotations of the FloodPlains project has solved the above problem for us.

Our interest, here, has been that of both investigating the viability of the segmentation approach for this particular task and that of comparing its results with those coming from our previous classification-based approach (albeit it was conducted with a smaller dataset). We have also to admit that, given the limited number of examples and their nature, we started this project with some misgivings on its viability. We conclude by noting that other applications of semantic segmentation to remote sensing have been developed that deal with slightly different situations, characterized by a large corpus of annotated maps and several categories to classify. Those categories are typically easier to differentiate, as they refer to more manageable urban scenarios, mostly. In archaeology, segmentation has been already used, but with point clouds data obtained by LIDAR technology [1]. However, this type of data was obtained with a special and expensive equipment, making it not easily available for most applications. Moreover, its tridimensional nature and the level of detail make it quite a different case study than the one we are dealing with.

This paper is structured as follows: Section 2 gives an overview of the methods used, both concerning the dataset and the deep learning models used for segmentation; Section 3 illustrates the results obtained with those methods; Section 4 discusses the results and sheds a light on the next steps of this research project.

2 METHODS

2.1 Dataset

The GIS shapefiles come from the FloodPlains project that contains around 5000 sites which were surveyed by different teams throughout the years, in a vast region spanning more than 100,000 square kilometers, as shown in Figure 1. Using the QGIS software we fixed the centroid of each site and measures its size. We found that most of those sites had a length of around 500 meters, or less. We proceeded then to discard bigger sites, as they were a few exceptions and would hardly fit in a single image. At that point, we assembled a collection of square images of the size of 1000 × 1000 meters, centered on each site. From those, a square of 500×500 meters have been consequently cropped randomly, in order to avoid having the target shapes being always in the center of the image. This was necessary, as in the absence of this strategy, the model could learn an inductive bias towards the central position of the shapes, as it represented a sort of safe bet: Figure 2a shows some examples. The final image resolution was set to 256×256 pixels. Initially we tried using 512×512 but the performance given by the higher detail did not justify the increase in computational time. We have also added images, with no sites, to see if this would have been of some aid to the model in learning to avoid certain special areas, like urban structures and flooded regions. The dataset was split into training and validation sets according to the common proportions 80:20

2.2 Data Augmentation

It is well known that data augmentation is a very useful technique to use in situations in which the size of the dataset is limited, and the cost of collecting more data is quite high. This also represents an important strategy to force the model to learn to be invariant against certain transformations (e.g., mirroring). The type of augmentations that can be appropriate to a specific task highly depends on the objective to achieve, however. In particular, we leveraged the Python library albumentations, which provides a framework for data augmentation to be applied with a random probability at load time [2]. This ensures a different augmentation for each epoch, making the model more flexible. Apart from the random crop we described before, which was always performed, we included three types of augmentations. They are as follows: The first is a random rotation of 90, 180 or 270 degrees. This transformation is non-destructive, and it is useful to teach the model that an archaeological site is recognizable, regardless of the orientation. Similarly, we also applied a random mirroring, either horizontal or vertical covering all the possible symmetries an image can have. Lastly, we applied a (slight) brightness and contrast shift, as the images are not uniform in their lightning conditions, and this should help the model recognize sites in those cases. Figure 2b shows some examples.

2.3 Filtering out bad examples

Some of the shapes in the dataset correspond to sites that are only visible from older maps and have been partially or completely destroyed through the years by anthropic processes. Additionally, some surveys reported the position of known sites that are no

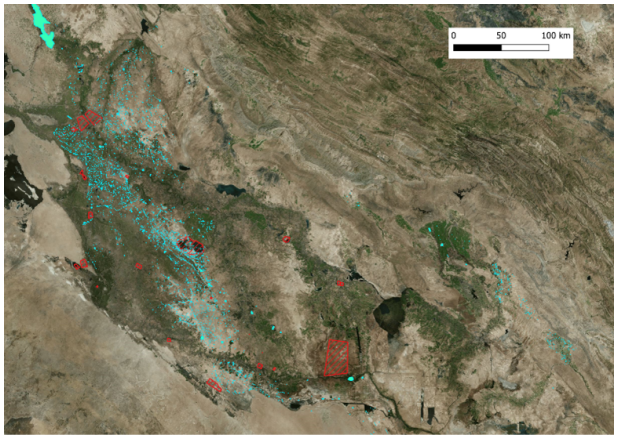


Figure 1: Investigation area. Green shapes represent surveyed sites. Red areas are location where no site can be found, like cities and artificial lakes

longer detectable with a small circle in place of the actual shape to roughly indicate their position.

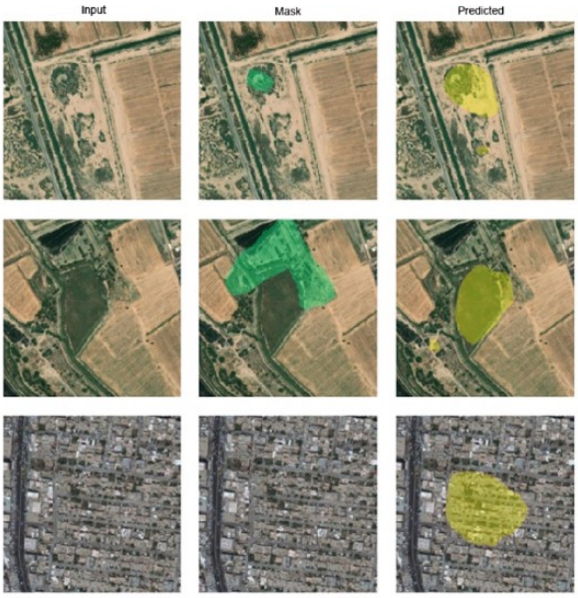
Those examples, shown in Figure 3, are obviously going to skew the model learning process and we tried removing them, as an additional experiment [13]. We took the best model, dropped the 700 elements in the training set that had an IoU score of 0, and then retrained a new model with the same settings. This procedure was a somehow quick and dirty attempt that if promising would be replicated with a human-in-the-loop filtering step.

3 MODELS

All deep learning models for semantic segmentation are based on the same architectural concept of employing both an encoder and a decoder. The encoder is in charge of feature extraction, and at various levels of detail, of reducing the image to smaller and smaller feature maps, essentially learning where to look in the image. The decoder instead plays the role of inflating the feature maps back to the input size, while learning to create the actual mask one wants to predict. All this said, we employed a library of pretrained segmentation models for Pytorch, as the primary goal of this study was to check its feasibility [19]. The library in fact allows the use of different segmentation architectures, that in turn shape the decoder section, and combine them with different encoders for feature extraction.

In our experiments, we used Unet and MAnet as the segmentation architectures, and ResNet and EfficientNet as encoders for feature extraction. Unet is a fully convolutional network architecture introduced in 2015 for semantic segmentation of cellular tissues. The model is characterized by two almost specular encoder and decoder, hence the U shape that gives the name, with connections that go across at the same depth level [16]. MAnet is a model specifically built for image segmentation in remote sensing, featuring attention blocks (made popular by the transformers architecture) and an architectural design aimed at better capturing long range spatial dependencies [7]. ResNet is a very popular and influential deep learning architecture for computer vision introduced in 2015.

(a)



(b)

Figure 2: a) Nine examples of possible augmentations for the same site (green contour); b) prediction from the model trained with no random cropping (ground truth in green, prediction in yellow)

It popularized the idea of skip connections, becoming the state of the art for convolutional neural networks (CNN), and it is often used as a benchmark for new models. We employed a version with 11 million parameters (i.e., resnet18), pre-trained on the ImageNet dataset [18]. EfficientNet is an optimized convolutional network

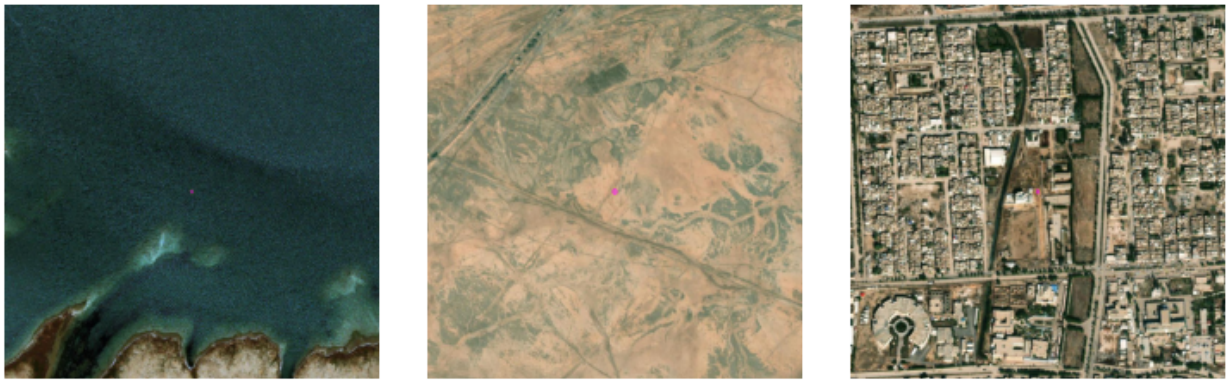


Figure 3: the examples from the filtered images. The sites are either flooded, extremely small or covered by a city

introduced by Google Brain, that features a streamlined architecture thanks to clever design decisions and to the use of a neural architecture search to find the best scaling for depth, width and resolution. We used the B3 model which has a similar amount of parameter to resnet18, while allegedly performing way better [11].

Finally, some words are in order regarding loss functions. Loss functions play a very important role as they are directly responsible for the way the model learns and thus produces its outputs. Among the many alternatives, we used the Intersection-over-Union (IoU) metric. This serves as our performance metric while the formula below shows how it is computed:

$$IoU = \frac{Y \cap \hat{Y}}{Y \cup \hat{Y}} = \frac{TP}{(TP + FP + FN)}$$

where Y is the segmentation mask and $\hat{Y}$ is the predicted mask (TP, FP and FN stand for true positives, false positives, and false negatives).

While IoU can be used a loss function, for numerical differentiation reasons Dice Loss is often preferred.

$$L_{dice} = 1 - \frac{2TP}{2TP + FP + FN}$$

Finally, Focal Loss is a variation of the classical Cross Entropy Loss with the introduction of a mechanism that scales down the contribution of easy to predict elements [8]. For each pixel to classify we have:

$$L_{focal} = -\alpha(1 - p_t)^{\gamma} log(p_t)$$

with

$$p_t = p \cdot \text{target} + (1 - p) \cdot (1 - \text{target})$$

where target corresponds to each pixel in the mask (either 1 or 0) and p to the predicted probability.

4 RESULTS

Models were trained on a GTX 1080ti GPU, with 11Gb of VRAM for 10 or 20 epochs, which corresponds to roughly 30-60 minutes. Longer training is definitely possible but with diminishing returns. Besides, this choice makes the results we achieved easily replicable even with limited resources. Contrary to our expectations, all models performed quite well, with IoU scores, in the validation phase, around the value of 70%, as summarized in Table 1.

Table 1: Validation Performance (per-image IOU)

Architecture	Encoder	Loss	Epochs	IoU
Unet	resnet18	dice	10	0.6823
Unet	efficientnet-b3	dice	10	0.7068
Unet	resnet18	focal	10	0.7221
Unet	efficientnet-b3	focal	10	0.7219
MAnet	efficientnet-b3	dice	10	0.6920
MAnet	efficientnet-b3	focal	10	0.7265
Unet	efficientnet-b3	dice	20	0.7178
MAnet	efficientnet-b3	dice	20	0.7316
MAnet	efficientnet-b3	focal	20	0.7437
MAnet (filtered dataset)	efficientnet-b3	dice	10	0.7246

Not shown in Table 1 are the foreseeable exceptions of the base model (Unet with resnet18 and dice loss) trained either without cropping and negatives, or with only cropping which scores respectively around 50% and 64%.

MAnet seems to provide no significant benefit over Unet, at least in our experiments: their predictions are extremely similar in most cases and so are their scores. MAnet seems to be taking the lead slowly with more training iterations, though. Similarly, Dice Loss and Focal Loss obtain extremely close scores, with the only discernible difference being the output they produce. Dice Loss tends to create masks that are more cohesive and clear-cut, with blocks of high probability that sharply taper off to 0 at the edges, whereas Focal Loss creates hazier prediction maps that change more smoothly. Figure 4 shows some examples of predictions for models with both losses.

Coming to the encoder choice, as expected resnet18 consistently performed worse than efficientnet-b3. However, the gap is not extremely marked in terms of IoU, even though qualitatively it seems to make worse mistakes (e.g., it misses some part of the sites). It is also to notice that we tried compensating the presence of bad

Figure 4: Examples of predictions from the MAnet model with different loss functions used during training, Focal Loss (purple) or Dice Loss (orange). Shown in green is the ground truth. Values close to zero are made transparent for clarity.

examples by removing the worse performing images from the train dataset that obtained a score of 0. Visual inspection showed that they were mostly point-like masks or invisible sites in a flooded region. Retraining a MAnet model with this refined dataset improved the results by around 3%.

5 CONCLUSION AND DISCUSSION

We set out to investigate the feasibility of using semantic segmentation to approach the remote sensing task of detecting archaeological sites from satellite imagery. The results we obtained were pretty satisfactory, even if they leave quite a lot of room for improvement.

The IoU score of around 70% we obtained may not seem like a great result in absolute terms, especially if compared to other semantic segmentation tasks, but it is very good and promising in this archaeological context. To understand this assertive statement, first of all, one should notice how the archaeological task we tackled concerns a type of scenario where visual features are not as evident as in other segmentation situations and even human beings have a hard time tracing the shape of a site.

Second, the dataset we used was quite noisy, with some sites not visible any longer from present-day photos. Additionally, among the sites not traced precisely, common is the situation where the detected shape is an ellipse of roughly the right size of the real target, some other is a point-like circle; all this to emphasize that something was there, but it was not detectable automatically with more precision. Those ambiguous examples should be removed but doing it automatically can be difficult and a human-in-the-loop process would be better suited.

Not only, but since we have recognized from the start that masks are not precise, 100% IoU would mean very little in any case. At the end, 70% is a good result because it is accurate enough to be useful for archaeologists. Also, after some preliminary interaction with the archaeologists, we concluded that a partial mask in the right spot can be sufficient to aid a human detection, and that when the model is wrong it is often wrong in a way that a human expert would also likely be. Finally, the homogeneous performance of the model suggests that key to the success in this work is the dataset and the way it is managed. Leveraging on this aspect, we could expect that pushing in the direction of data quality would yield the most benefits compared to architectural changes.

The next steps will be the application of our method to similar geomorphological areas in different countries, as well as to additional basemaps: especially where archaeological surveys have been limited or non-existent the potential gain seems immense, both for heritage preservation and for planning a sustainable development which includes and does not obliterate heritage when prioritizing economic development. Even in already surveyed areas, we may expect a significant increase in the number of potential archaeological sites. All these improvements require furthering the human-AI collaboration process we started, also employing additional and allied techniques [5, 6, 10, 14, 15, 17].

REFERENCES

[1] Marek Bundzel, Miroslav Jaščur, Milan Kováč, Tibor Lieskovskỳ, Peter Sinčák, and Tomáš Tkáčik. 2020. Semantic segmentation of airborne lidar data in maya archaeology. *Remote Sensing* 12, 22 (2020), 3685.

[2] A Buslaev. 2020. Iglovikov VI Khvedchenya E Parinov A Druzhinin M Kalinin AA. *Albumentations: fast and flexible image augmentations Information* 11, 2 (2020), 125.

[3] Luca Casini and Marco Roccetti. 2020. Medical imaging and artificial intelligence. In *Philosophy of advanced medical imaging*. Springer, 81–95.

[4] Luca Casini, Marco Roccetti, Giovanni Delnevo, Nicolò Marchetti, and Valentina Orrù. 2020. The Barrier of meaning in archaeological data science. In *International Science Fiction Prototyping Conference 2020 (SciFi-It'2020), 23-25 March, 2020, Ghent, Belgium*. 61–65.

[5] Flavio Corradini, Roberto Gorrieri, and Marco Roccetti. 1997. Performance preorder and competitive equivalence. *Acta Informatica* 34, 11 (1997), 805–835.

[6] Stefano Ferretti, Silvia Mirri, Marco Roccetti, and Paola Salomoni. 2007. Notes for a collaboration: On the design of a wiki-type educational video lecture annotation system. In *International Conference on Semantic Computing (ICSC 2007)*. IEEE, 651–656.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[9] Nicolò Marchetti. 2020. FloodPlains Project. https://floodplains.orientlab.net/. The FloodPlains Project has been developed in the framework of the European Union project "EDUU – Education and Cultural Heritage Enhancement for Social Cohesion in Iraq" (EuropeAid CSOLA/2016/382-631), www.eduu.unibo.it, coordinated by Nicolò Marchetti. The ongoing project "KALAM. Analysis, protection and development of archaeological landscapes in Iraq and Uzbekistan through ICTs and community-based approaches," funded by the Volkswagen Foundation and coordinated by N. Marchetti, www.kalam.unibo.it, has allowed a review of our data input and the development of the research presented in this paper. The CRANE 2.0 project of the University of Toronto provided the geospatial servers on which FloodPlains is running.

[10] Gustavo Marfia, Marco Roccetti, Alessandro Amoroso, Mario Gerla, Giovanni Pau, and J-H Lim. 2011. Cognitive cars: constructing a cognitive playground for VANET research testbeds. In *Proceedings of the 4th International Conference on Cognitive Radio and Advanced Spectrum Management*. 1–5.

[11] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*. IEEE, 565–571.

[12] Marco Roccetti, Luca Casini, Giovanni Delnevo, Valentina Orrù, and Nicolò Marchetti. 2020. Potential and Limitations of Designing a Deep Learning Model for Discovering New Archaeological Sites: A Case with the Mesopotamian Floodplain. In *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good* (Antwerp, Belgium) *(GoodTechs '20)*. Association for Computing Machinery, New York, NY, USA, 216–221. https://doi.org/10.1145/3411170.3411254

[13] Marco Roccetti, Giovanni Delnevo, Luca Casini, Nicolò Zagni, and Giuseppe Cappiello. 2019. A paradox in ML design: less data for a smarter water metering cognification experience. In *proceedings of the 5th EAI international conference on smart objects and Technologies for Social Good*. 201–206.

[14] Marco Roccetti, Vittorio Ghini, Giovanni Pau, Paola Salomoni, and Maria Elena Bonfigli. 2001. Design and experimental evaluation of an adaptive playout delay control mechanism for packetized audio for use over the internet. *Multimedia Tools and Applications* 14, 1 (2001), 23–53.

[15] Marco Roccetti, Gustavo Marfia, and Marco Zanichelli. 2010. The art and craft of making the tortellino: playing with a digital gesture recognizer for preparing pasta culinary recipes. *Computers in Entertainment (CIE)* 8, 4 (2010), 1–20.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[17] Paola Salomoni, Catia Prandi, Marco Roccetti, Lorenzo Casanova, Luca Marchetti, and Gustavo Marfia. 2017. Diegetic user interfaces for virtual environments with HMDs: a user experience study with oculus rift. *Journal on Multimodal User Interfaces* 11, 2 (2017), 173–184.

[18] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[19] Pavel Yakubovskiy. 2020. Segmentation Models Pytorch. https://github.com/qubvel/segmentation_models.pytorch.