Deep Semantic Model Fusion for Ancient Agricultural Terrace Detection

Yi Wang[1,2], Chenying Liu[1,2], Arti Tiwari[3], Micha Silver[3], Arnon Karnieli[3], Xiao Xiang Zhu[1], Conrad M Albrecht[2]
[1]*Chair of Data Science in Earth Observation, Technical University of Munich (TUM), Germany*
[2]*Remote Sensing Technology Institute, German Aerospace Center (DLR), Germany*
[3]*The Remote Sensing Laboratory, Institutes for Desert Research, Ben Gurion University (BGU), Israel*

***Abstract*—Discovering ancient agricultural terraces in desert regions is important for the monitoring of long-term climate changes on the Earth's surface. However, traditional ground surveys are both costly and limited in scale. With the increasing accessibility of aerial and satellite data, machine learning techniques bear large potential for the automatic detection and recognition of archaeological landscapes. In this paper, we propose a deep semantic model fusion method for ancient agricultural terrace detection. The input data includes aerial images and LiDAR generated terrain features in the Negev desert. Two deep semantic segmentation models, namely DeepLabv3+ and UNet, with EfficientNet backbone, are trained and fused to provide segmentation maps of ancient terraces and walls. The proposed method won the first prize in the International AI Archaeology Challenge. Codes are available at https://github.com/wangyi111/international-archaeology-ai-challenge.**

***Index Terms*—deep learning, archeology, semantic segmentation**

I. Introduction

Discovering ancient agricultural terraces in desert regions has great importance for both archaeological and anthropological research in the monitoring of long-term climate change. First, the information can be gathered to advance our knowledge of ancient human endeavors. Second, indicating the potential use of surface runoff may also help reveal locations for enhancing future food production. While there is a growing need to discover new agricultural land resources, traditional ground surveys are limited in scale. With the development of modern imaging and machine learning techniques, to discover these regions on a large scale becomes possible [9].

The Negev, Israel, is a subtropical desert with occasional precipitation in the autumn, winter, and spring, while the hot summer is completely dry for almost half a year from about May to October. Thousands of ancient dry stonewalls (named terraces) were built in the central Negev Highlands during ancient times, mainly between the 4th and the 7th centuries, across ephemeral stream channels (wadis, Figure 1). This water harvesting technique is very common in wadi beds with gentle slopes. As a result of the slow water velocity, eroded sediments and nutrients usually settle in the wadi bed and create good agricultural land.

The ancient agricultural terraces in the Negev desert have been abandoned since the 7th century, but many stone terrace walls are still intact. Nowadays, the terraces can be observed from the surface either as exposed stonewalls or as buried stonewalls that can be identified by the above-ground vegetation (Figure 2). In the International AI Archeology Challenge [10], aerial image data including RGB images and terrain feature images generated from LiDAR surveys are provided to detect ancient terraces and walls as a multiclass semantic segmentation task.

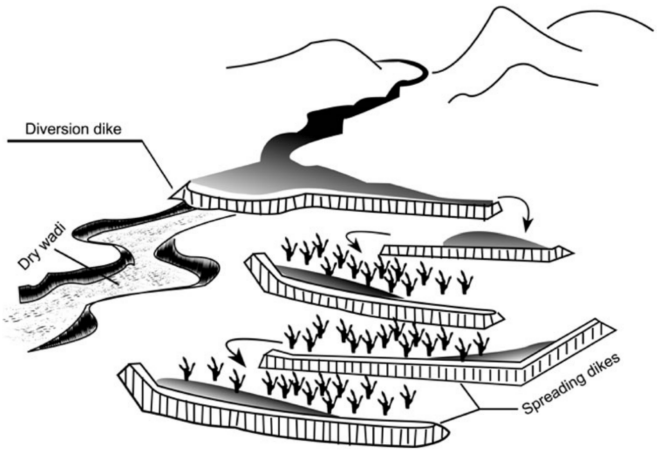


Fig. 1: A typical flood water harvesting system (figure adapted from [1, 2]).

In this paper, we describe the winning solution of the International AI Archeology Challenge. Specifically, we propose a deep semantic model fusion method that combines the soft prediction score of DeepLabv3+ [7] and U-Net [4] to segment the ancient agricultural terraces and walls. Our results verify the promising potential of deep neural networks in the application of archeological landscape recognition.

II. Methodology

In this section, we describe the details of the proposed deep semantic model fusion method. We first separately train a U-Net and a DeepLabv3+ model, each with EfficientNet [8] as the encoder backbone. During inference, the soft prediction scores from each model are fused together to decide the final output class.

A. Semantic segmentation models

(a) Example exposed terrace. (b) Example buried terrace. (c) Example stonewall fence.

Fig. 2: Ground photos of example terraces and walls [10].

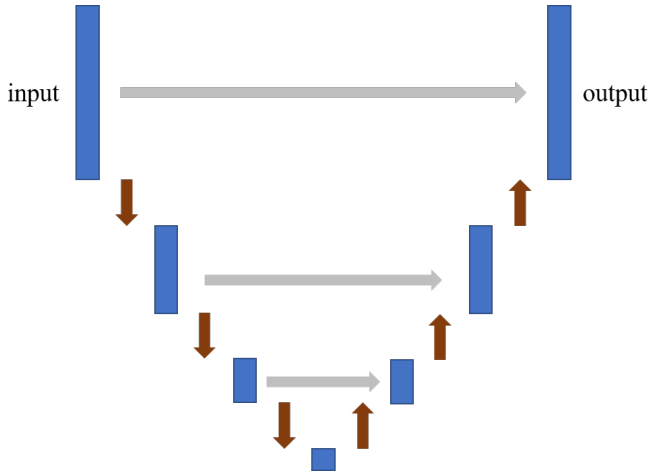


Fig. 3: U-Net [4]. Multi-scale features from the downsampling encoder are concatenated to those from the upsampling decoder. Each blue box corresponds to a multi-channel feature map.

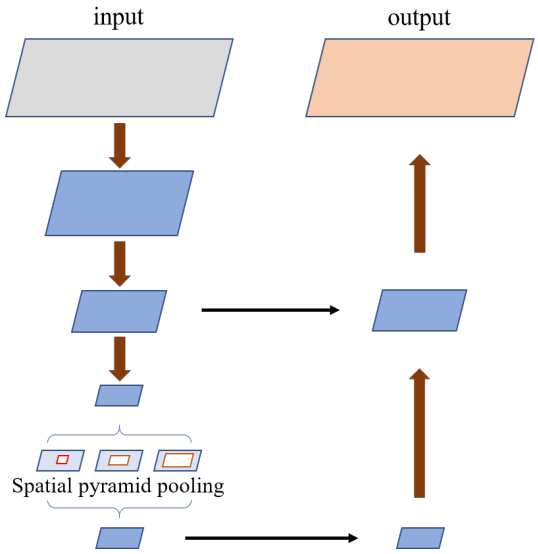


Fig. 4: DeepLabv3+ [7]. The encoder contains a spatial pyramid pooling module [6] that allows extracting features at an arbitrary resolution by applying atrous convolution. The simple decoder helps to recover detailed object boundaries.

1) U-Net: The U-Net [4] architecture was initially designed for biomedical image segmentation. As is shown in Figure 3, the network consists of a contracting path and an expansive path, which gives it the U-shaped architecture. The downsampling encoder path is used to capture the multiscale context in the image, which is originally a stack of convolutional and max pooling layers. The upsampling decoder path is the symmetric expanding path, which is used to enable precise localization using transposed convolutions. Multiscale features from each layer of the encoder are kept and concatenated to the corresponding decoder layer, allowing the network to propagate context information to higher resolution layers.

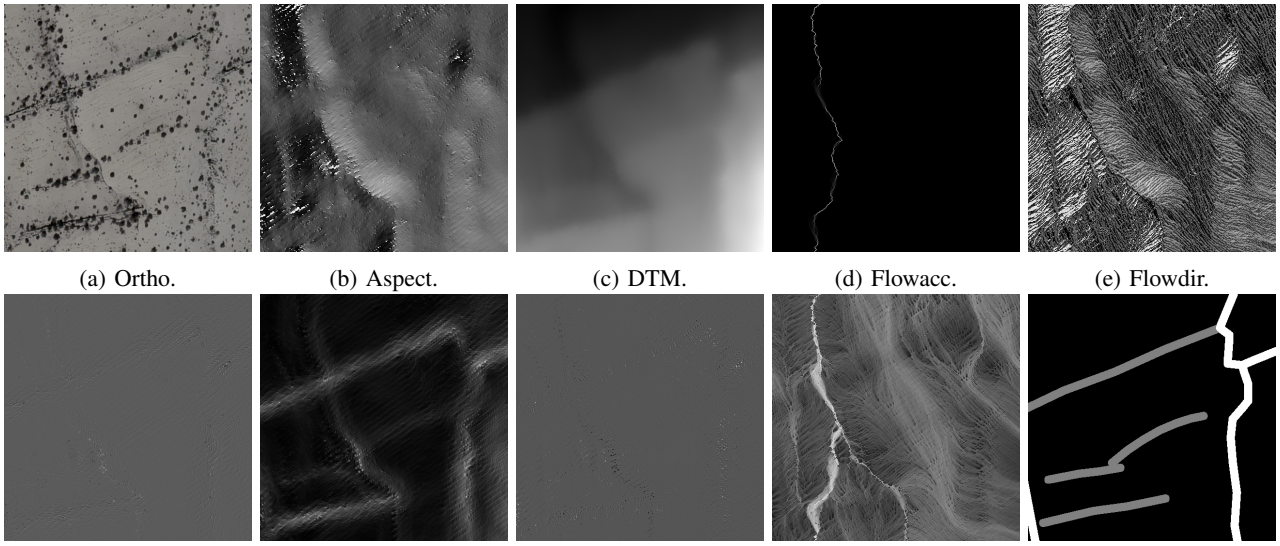
2) DeepLabv3+: The DeepLabv3+ [7] architecture is based on the DeepLab semantic segmentation model series [3, 5, 6]. As is shown in Figure 4, the network has an encoder-decoder structure. Atrous Spatial Pyramid Pooling [5] is used in the encoder to capture multi-scale contextual information. Parallel atrous convolution with different rates is applied in the input feature map, and fused together. Depthwise separable convolution, or atrous separable convolution, helps to reduce the computation complexity while maintaining similar (or better) performance. A simple decoder is used to recover detailed spatial information.

B. Encoder backbones

We choose EfficientNet [8] as the encoder backbones for both U-Net and DeepLabv3+. EfficientNet is a convolutional neural network architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient. Unlike conventional practice that arbitrarily scales these factors, the EfficientNet scaling method uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients.

C. Semantic Model Fusion

After the separate training of both U-Net and DeepLabv3+, we merge the output probability maps from both models to get the final segmentation map:



(a) Ortho. (b) Aspect. (c) DTM. (d) Flowacc. (e) Flowdir.

(f) Pcurv. (g) Slope. (h) Tcurv. (i) TWI. (j) Ground truth.

Fig. 5: Example training data. (a) is aerial RGB image; (b)-(i) are terrain feature images generated from LiDAR survey. (j) is the ground truth mask, where gray represents terraces and white represents walls.

$$\text{Out}_{(i,j)} = \alpha \cdot UNet_{(i,j)} + (1-\alpha) \cdot \text{DeepLabv3plus}_{(i,j)} \quad (1)$$

where i, j indicates each pixel, and α is a weighting parameter.

D. Losses

The model outputs 3 class probabilities: terrace, wall and background. Due to the high unbalancing of foreground (terraces and walls) and backgroud pixels, a naive cross entropy loss may push the model towards predicting only the background. To tackle this issue, we add a DICE loss that optimizes the overlap of predictions and ground truth masks for each class:

$$\text{DiceLoss} = 1 - \frac{2TP}{2TP + FN + FP} \quad (2)$$

where TP, FN, FP represent true positive, false negative and false positive, respectively. The weighted mean of the cross entropy loss and the Dice loss is used as the total loss:

$$\text{TotalLoss} = \beta \cdot \text{ CrossEntropyLoss } + (1-\beta) \cdot \text{ DiceLoss} \quad (3)$$

E. Data augmentations

The data sizes in archeology are relatively small, which may lead to strong overfitting with deep neural networks. To tackle this issue, we introduce various data augmentations to increase the diversity of the training data. Apart from commonly used RandomResizedCrop and RandomHorizontalFlip in natural images, we explicitly add more geometric and color augmentations including RandomVerticalFlip, RandomRotation, RandomAffine and RandomGaussianBlur.

III. Experiments

A. Data

The input data consists of decimeter resolution aerial RGB orthophotos and 8 terrain feature images generated from LiDAR survey. Specifically, the terrain features are Aspect, Digital Terrain Model (DTM), Flow accumulation, Flow direction, P-curve, T-curve, Slope and Topographic wetness index. The ground truth masks contain 3 classes: terrace, wall and background. Figure 5 shows the various features of an example patch. 500 patches with ground truth masks are provided in the training phase, and 200 patches (masks hidden) are used for testing. All images are with the size 512x512.

B. Implementation Details

We split the training data into 400/100 training/validation splits. All features (3 channel RGB + 8 channel terrain features) are used as input to the networks.

We use ImageNet pretrained EfficientNet-B5 as the encoder backbone for both U-Net and DeepLabv3+ models. The model fusion weights α and the loss weighting parameter β are both set to 0.5.

We use the batch size 16, learning rate 0.001, and AdamW optimizer. Training one model on an NVIDIA RTX 3090 takes about 1 hour.

C. Results

1) Qualitative evaluation: Figure 6 shows the predicted segmentation maps for 5 example patches. It can be seen that the segmentation results are very close to ground truth masks from a first look. While it is not easy for none-experts to visually check, most of the terraces and walls are detected successfully by the model. A closer look shows that the exact shape of the objects and the detailed location of the pixels are

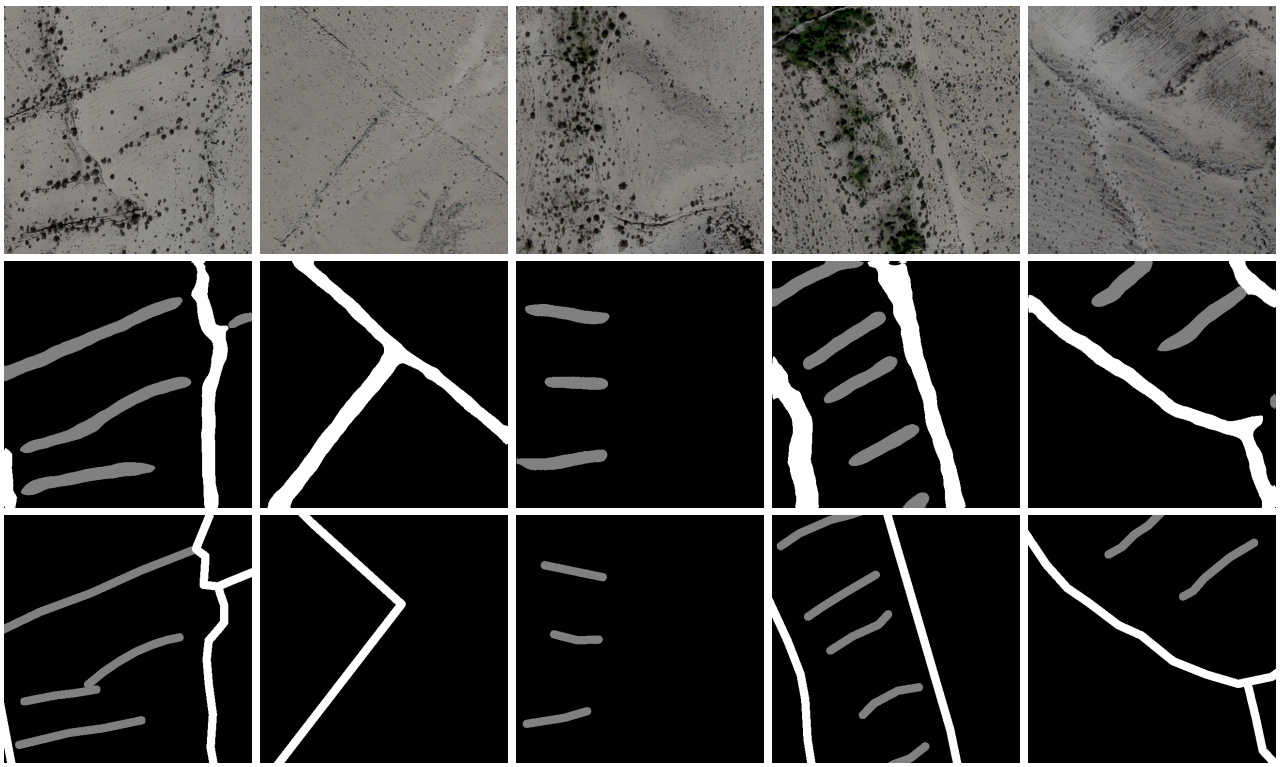


Fig. 6: Qualitative evaluation results of 5 example patches. The columns correspond to the 5 patches; the rows correspond to orthophotos, predicted segmentation maps and ground truth masks, respectively.

TABLE I: Evaluation metrics on the validation set.

	terrace			wall			IoU	mIoU
	precision	recall	F1	precision	recall	F1		
U-Net	**.87**	.12	.21	**.60**	.86	**.71**	.45	.33
DeepLabv3+	.31	**.41**	.35	.53	**.90**	.67	.44	.36
Fusion	.65	.29	**.40**	.57	.89	.70	**.47**	**.39**

still hard to recognize. However, the situation here is different from natural images, where clear boundary information can be defined from the input data. From the aspect of ancient agricultural terraces (especially the buried ones), only the rough structures are possible to be detected even by experts.

2) Quantitative evaluation: The final evaluation score for the challenge is two-calss (foreground only) intersection over union (IoU):

$$IoU = \left| \frac{\{(i,j) : P(i,j) = T(i,j) > 0\}}{\{(i,j) : P(i,j) > 0 \text{ or } T(i,j) > 0\}} \right| \quad (4)$$

where P,T denote the predicted map and ground truth mask. Our proposed method reached a final score of 0.31 on the hidden testing data. Besides that, we consider also precision, recall, F1 score and mean IoU (mIoU) to evaluate the prediction results. Table I reports the evaluation metrics on the validation set, which verify the advantage of the fusion strategy. While U-Net has a higher precision, DeepLabv3+ has a higher recall. The fused model in the end has better scores on the balanced metrics like F1 score and IoU. The general low metric values compared to the impressive qualitative results confirm the difficulty of recognizing detailed terrace boundaries from the input data. Furthermore, it can be seen that terraces are performing worse than walls, which reflects the fact that stonewalls are generally more complete and regular while terraces are usually split into pieces.

IV. Discussion

A. Importance of different features

To efficiently analyze the importance of different features, we perform inference on "modified" input and report the corresponding IoU score. Specifically for each feature channel, we remove it by replacing with zero values, and check the inference score of remaining features. The lower the score compared to all features remained, the more important the removed feature. Compared to separately training a model for each feature, this "test-time-evaluation" is much less costly. The results are shown in Table II, where green channel and the feature "slope" are clearly the two most important features, and the features "flow accumulation" and "flow direction" are the two least important features. This is also consistent with the visual perception, as we can see from Figure 5 that the

TABLE II: Feature importance evaluation. We report the IoU scores after removing one feature during the inference, i.e., replacing the corresponding channel with zero values. The columns represent which feature to remove.

	All features	Red	Green	Blue	Aspect	DTM
IoU	.37	.31	**.27**	.31	.36	.33
	FlowAccum	FlowDir	Pcurv	Slope	Tcurv	TWI
IoU	.36	.34	.34	**.26**	.34	.31

orthophoto and the slope are the two most obvious features that align with the ground truth mask.

B. Ambiguity in archaeological labels

Unlike for natural images commonly used in the computer vision community, the labels for archaeological images are difficult or sometimes impossible to be accurate. There are two main challenges. First, though labels are usually collected from ground surveys by archaeology experts, the ground survey itself can not ensure 100% accuracy because of the extremely long time gap. Second, the exact boundaries of the ancient landscapes are almost impossible to be accurate, especially for those buried terraces that don't exist on the ground any more. For the first challenge, introducing uncertainty quantification in the labeling phase can greatly improve the quality of the training datasets. For the second challenge, new evaluation metrics may be explored. For example, one can give different weights to pixels with different label confidence. Another example could be introducing object-level metrics like comparing the distance between vectorized central lines of the prediction and the target mask.

V. Conclusion

In this work, we present a deep semantic model fusion method for ancient agricultural terrace detection in the Negev desert. We train two types of semantic segmentation models and fuse the predicted probabilities to output the final segmentation map. The experimental results verify the great potential of AI in archeology, but also call for further studies on domain specific characteristics like the ambiguous bundaries of ancient landscapes.

Acknowledgment

This work is supported by the Helmholtz Association through the Framework of Helmholtz AI (grant number: ZT-I-PF-5-01) - Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)".The work of X. Zhu is additionally supported by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001) and by German Federal Ministry for Economic Affairs and Climate Action in the framework of the "national center of excellence ML4Earth" (grant number: 50EE2201C). We thank both IDSI and Helmholtz Information and Data Science Academy (HIDA) for organizing the challenge.

References

[1] N French and J Hussain. "Water Spreading Manual, Range Management Re. 1". In: *Pakistan Range Improvement Scheme, Lahore, Pakistan* (1964).

[2] Albert Rango and Kris Havstad. "Water-harvesting applications for rangelands revisited". In: *Environmental practice* 11.2 (2009), pp. 84–94.

[3] Liang-Chieh Chen et al. "Semantic image segmentation with deep convolutional nets and fully connected crfs". In: *arXiv preprint arXiv:1412.7062* (2014).

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

[5] Liang-Chieh Chen et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

[6] Liang-Chieh Chen et al. "Rethinking atrous convolution for semantic image segmentation". In: *arXiv preprint arXiv:1706.05587* (2017).

[7] Liang-Chieh Chen et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

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

[9] Simon H Bickler. "Machine learning arrives in archaeology". In: *Advances in Archaeological Practice* 9.2 (2021), pp. 186–191.

[10] *International AI Archeology Challenge*. URL: https://www.helmholtz-hida.de/en/events/internationale-ki-challenge-archaeologie/.