Augmentation Invariance and Adaptive Sampling in Semantic Segmentation of Agricultural Aerial Images

Augmentation Invariance and Adaptive Sampling in Semantic Segmentation of Agricultural Aerial Images

Antonio Tavera*[1], Edoardo Arnaudo*[1,2], Carlo Masone[3], Barbara Caputo[1]
[1]Politecnico di Torino, Turin, Italy
[2]LINKS Foundation, Turin, Italy
[3] Consorzio Interuniversitario Nazionale per l'Informatica, Rome, Italy
[1]{first.last}@polito.it
[2]{first.last}@linksfoundation.com

Abstract

In this paper, we investigate the problem of Semantic Segmentation for agricultural aerial imagery. We observe that the existing methods used for this task are designed without considering two characteristics of the aerial data: (i) the top-down perspective implies that the model cannot rely on a fixed semantic structure of the scene, because the same scene may be experienced with different rotations of the sensor; (ii) there can be a strong imbalance in the distribution of semantic classes because the relevant objects of the scene may appear at extremely different scales (e.g., a field of crops and a small vehicle). We propose a solution to these problems based on two ideas: (i) we use together a set of suitable augmentation and a consistency loss to guide the model to learn semantic representations that are invariant to the photometric and geometric shifts typical of the top-down perspective (Augmentation Invariance); (ii) we use a sampling method (Adaptive Sampling) that selects the training images based on a measure of pixel-wise distribution of classes and actual network confidence. With an extensive set of experiments conducted on the Agriculture-Vision dataset, we demonstrate that our proposed strategies improve the performance of the current state-of-the-art method. [1].

1. Introduction

Semantic segmentation, i.e., the task of classifying each pixel of an image into a preset taxonomy of semantic categories, is a fundamental research problem in computer vision and a key technology in many real-world applications. Among these applications, the environmental monitoring from remote aerial images has grown considerably in recent years, with examples of categorization of land cover [14, 32], delineation of wildfire [17], and identification of deforested regions [2]. In this task, as in other computer vision problems, deep learning models have demonstrated promising results, thanks also to an increasing availability of open datasets and large scale collections of aerial images [6, 32]. However, the majority of these deep learning models were originally designed for other use-cases, such as self-driving vehicles [23] and medical imaging [27], and then transferred to the aerial domain without considering its specific characteristics. In particular, we find two peculiarities that set the task of aerial segmentation apart from its autonomous driving counterpart:

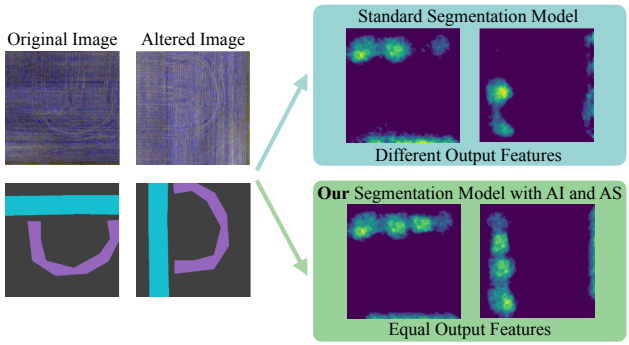

Figure 1. A semantic segmentation model that is not designed to expect changes in points of view, may produce different output features for the same image when seen from different angles. Our technique, on the other hand, makes the model invariant to these viewpoint shifts, encouraging the model to learn more robust representations.

Top-down perspective: In remote sensing the images are collected with a top-down perspective, i.e., from a camera mounted on an aircraft and pointed towards the ground. This remote perspective implies not only a lack of depth

*Equal contribution
[1]Code can be found at: https://github.com/taveraantonio/AIAS.

and reference points in the pictures, but it also allows to capture the same scene with arbitrary rotations around the vertical axis (see Fig. 1). Thus, whereas in autonomous driving datasets [1, 13] the model is bound to experience a well structured organization in the semantic elements of the scene (e.g., the road is expected at the bottom of the image, the sky on the top), this is not true in aerial imagery.

Extreme class imbalance Although the problem of class imbalance in class-wise pixel distributions is typical of semantic segmentation [30], in aerial images this is brought to an extreme because the entities to be recognized range from small vehicles to large natural biomes.

We argue that a semantic segmentation model that is designed to account for these characteristics of the aerial setting can be more effective at the task. Thus, we propose a solution based on two ideas: Augmentation Invariance (AI) and Adaptive Sampling (AS). The first one uses augmentations to guide the model to learn representations that are invariant to shifts in appearance and perspective (e.g., rotations around the vertical axis, as shown in Fig. 1). The second is intended to regularize the training of underrepresented classes by adaptively sampling the training images according to the distribution of pixels and the actual network confidence. These two modules cooperate in an end-to-end training stage.

Summarizing, the contributions of this paper are:

- An Augmentation Invariance technique that is tailored to handle the specific challenges given by the perspective in the aerial data and to help the model to separate semantic information from appearance.
- An Adaptive Sampling approach to address the problem of class imbalance, by dynamically sampling training data based on the current network confidence and the global, pixel-wise class distribution.
- An extensive set of experiments on the Agriculture Vision dataset [11], which is the only agricultural aerial dataset with several semantic classes and complexity. We show what happens when only RGB images are used for training, as well as when NIR data is exploited. Furthermore, an exhaustive ablation study examines the impact of all the solutions introduced. The code will be made available in order to encourage research.

2. Related Work

2.1. Semantic Segmentation

There is a flourishing literature on semantic segmentation, mostly pertaining different network architectures and techniques to capture the global context of a scene. Methods such as FCN [20] include only convolutional layers and use skip connections to incorporate semantic and appearance information from deep and shallow layers, respectively. Most common segmentation models, such as U-Net [27], HRNet [29] and HRNetV2 [31], use an encoder-decoder structure to extract objects and image context at different scales. Multi-scale approaches are also used in solutions such as FPN [19], UperNet [34] and PSPNet [40] to better condition the global context of a scene. DeepLab V2 [7] and V3 [8] use the dilation parameter of convolutional layers and present the ASPP to robustly segment objects throughout many scales. DeepLab V3+ [9] boosts the DeepLab family by adopting an encoder-decoder structure. More recent methods, such as OCR [38] in conjunction with HRNetV2 [31] or SegFormer [35], have demonstrated the effectiveness of Transformers for Semantic Segmentation. All of the above approaches seek to acquire semantics and global context at numerous scales, or in an encoder-decoder fashion, focusing mostly on autonomous driving scenarios. However, agricultural aerial imagery poses challenges that are not addressed by such architectures. With respect to these prior works, to better process the aerial data we guide the network to learn semantic representations that are invariant to the visual distortion and changes of orientation typical in the top-down perspective.

2.2. Aerial Semantic Segmentation

In aerial and remote sensing, the target environment of the semantic segmentation may vary greatly, from urban areas [4, 15, 24] to land cover [6, 14, 32] and agricultural scenarios [11, 22, 36]. These various target environments are generally linked to different applications, each with some specific challenges or requirements. For example, in urban monitoring semantic segmentation is mainly used to identify infrastructures, such as roads [39] and buildings [10]. This requires using high-resolution imagery as input and sometimes also to consider changes in time [21]. In land cover tasks, the main challenges are the extreme difference in the size of different semantic categories and the stark visual differences across different domains. Recent solutions address the first problem using multi-level or multi-scale feature aggregation [37], and the latter resorting to Domain Adaptation approaches [5, 32]. When it comes to agricultural scenarios, classical segmentation solutions are mostly based on vegetation indices such as the NDVI [33], but the current trend is to move away from these handcrafted indices and towards more robust computer vision techniques, such as automated fusion of multi-spectral data [28] and precise crop segmentation [18]. Indeed, agricultural aerial images are rarely limited to the visible spectrum and frequently include other bands, such as Near-Infrared (NIR). In deep learning literature, the most common solutions to jointly exploit RGB and NIR images are the duplication of input weights [11, 25] or multi-modal approaches based

on late or early fusion [36, 37]. Another peculiarity of the aerial data is the fact that the orientation of the camera is arbitrary and uncertain. Although this problem has been addressed in incremental learning [3] and in classification tasks [26], none of the current solutions in Semantic Segmentation consider this issue.

3. Method

Our approach, depicted in Fig. 2, expands the SegFormer architecture [35] with two ideas. Firstly, by minimizing a loss that aligns the pixel embeddings generated by the Transformer network for the original image and for its augmented version, we promote the invariance of the learned semantic representations to photometric distortions and perspective changes that are typical in the aerial setting. Secondly, we introduce an adaptive sampling mechanism to actively select the training samples using prior knowledge on class distribution and actual network confidence. In the rest of this section we first introduce the problem setting and then we detail these two mechanisms.

3.1. Problem Setting

We consider the problem of supervised semantic segmentation of agricultural aerial imagery, where during training we are provided with a collection of tuple $X = \{(x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z})\}$, with $\mathcal{X}$ being the set of RGB images, $\mathcal{Z}$ the set of Near-Infrared (NIR) images and $\mathcal{Y}$ the set of semantic masks that associate to each pixel a class c from a predefined set of semantic classes $\mathcal{C}$. Additionally, we denote as $\mathcal{I}$ the set of pixels in each image and mask, and we define as $\hat{x} \in \hat{\mathcal{X}}$ the four-channel RGB-NIR image obtained by concatenating channel-wise x and z.

Our goal is to find a map $f_\theta : \hat{\mathcal{X}} \rightarrow \mathbb{R}^{|\mathcal{I}|\times|\mathcal{C}|}$, depending on a set of learnable parameters θ, that assigns to each pixel of the RGB-NIR images an individual probability to belong to each semantic category in $\mathcal{C}$. As a baseline, the parameters θ are optimized to minimize the standard cross-entropy loss L_{seg}:

$$L_{\text{seg}}(\hat{x}, y) = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{c\in\mathcal{C}} y_i^c \log(p_i^c(\hat{x})), \qquad (1)$$

where $p_i^c(\hat{x}) = f_\theta(\hat{x})[i, c]$ is the output of the model at a pixel i for the class c, and y_i^c represents the expected result for the same pixel and class.

3.2. Augmentation Invariance

Current state-of-the-art frameworks are developed for the autonomous driving task and they suffer from performance degradation when applied to aerial data. We identify a few factors that have a significant contribution to this degradation:

- aerial images are not constrained to view the environment from a fixed perspective and, in particular, the camera orientation around the vertical axis is free;
- aerial images can display severe distortions due to the angle of the camera;
- there can be significant photometric shifts across different fields.

We propose a mechanism, called Augmentation Invariance (AI), which uses augmentations to guide the model to learn a mapping that is invariant to these shifts in perspective and appearance. This mechanism works as follows. Given an input image $\hat{x}$, we extract the pixel-wise features $f_i(\hat{x})$ originating from the second-to-last layer of the SegFormer architecture at each iteration, explicitly skipping the last layer used for pixel-wise segmentation. Simultaneously, we transform a copy of $\hat{x}$ using both a random selection of geometric augmentations A_g (*horizontal flipping*, *vertical flipping*, *random rotation*), and a random photometric augmentation A_p (*color jitter*). Hereinafter, to simplify the notation, we denote the combination of both augmentations as $A_p \circ A_g = A$. The transformed image $A(\hat{x})$ is also passed through the model to extract the features $f_i(A(\hat{x}))$. Finally, to make the model invariant to shifts in perspective and appearance we impose that the features extracted from the original image $\hat{x}$ are coincident with the features extracted from the transformed image $A(x)$, after reversing the geometric augmentation. We achieve this with a pixel-wise mean squared error loss L_{AI}, which is defined as

$$L_{\text{AI}}(\hat{x}, A(\hat{x})) = \frac{1}{\mathcal{I}}\sum_{i\in\mathcal{I}}(f_i(\hat{x}) - A_g^{-1}\left(f_i(A(\hat{x}))\right))^2 \qquad (2)$$

In (2), A_g^{-1} indicates the inversion of the geometric augmentations performed on $\hat{x}$, which is critical to ensure that we compare the original and augmented features corresponding to the same pixel.

We also maintain the ground truth annotations of the augmented images, so that the same segmentation loss can also be applied to them. The total training loss, denoted by L_{tot}, can be summarised as follows:

$$L_{\text{tot}} = L_{\text{seg}}(\hat{x}, y) + L_{\text{seg}}(A(\hat{x}), A_g(y)) + \lambda L_{\text{AI}}(\hat{x}, A(\hat{x})), \qquad (3)$$

where $A_g(y)$ denotes the same geometric transformation applied to the ground truth annotation y, and λ is a modulating factor.

We remark that the mechanism of augmentation invariance used here is different from a classical data augmentation, because we do not use photometric and geometric transformations just to extend the training dataset with examples not in the original data distribution. Rather, through

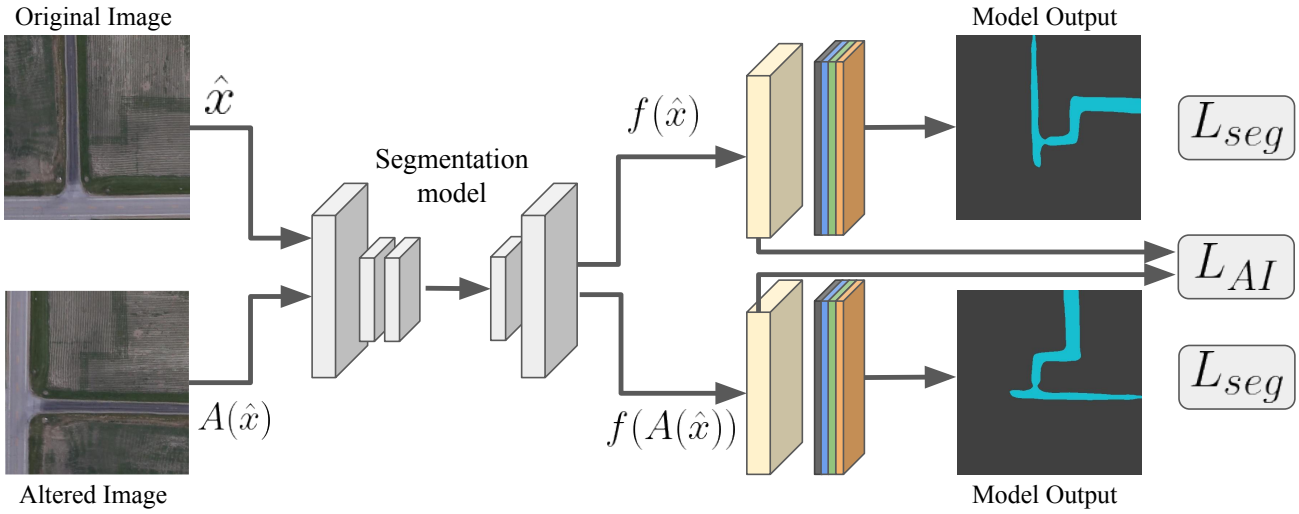

Figure 2. Illustration of the overall framework. The Adaptive Sampling picks a sample, and then an augmented version is generated. Both images are forwarded to the segmentation model, which computes the segmentation loss L_{seg}, whereas the L_{AI} loss forces the model to extract the same features from both the original image and its transformed counterpart.

the loss L_{AI} we also use the original and the transformed images paired together, to give a stronger guidance to the training process.

3.3. Adaptive Sampling

Despite the difficulty of avoiding learning bias from data, one of the primary issues in semantic aerial data is the strong distribution imbalance of semantic classes, with some that appear rarely and others that are extremely frequent. To address this issue, we use an Adaptive Sampling (AS) approach that works together with the Augmentation Invariance. At each iteration, the current sample of images required to train the network is selected according to two pieces of information: the global, pixel-wise class distribution, and the class-wise network confidence. In doing so, the data sampler will dynamically choose images giving priority to those whose categories appear with low frequency and for which the network has the least confidence. Formally, the AS samples a class c with an adaptive probability AS_c defined as:

$$AS_c = \sigma((1 - dist * conf)^{\gamma}), \tag{4}$$

where $dist$ is an array that represent the classes distribution, $conf$ represents the actual class-wise network confidence, σ is a min-max normalization function and γ is a relaxation parameter. Once a semantic category c has been chosen given this dynamically updated probability, an image is picked randomly from a subset of data X_c that contains this class c. To compute (4) we use the following definitions of *dist* and *conf*:

Class Distribution *dist*. Given that we are working in a supervised environment, we can compute a fixed, static distribution estimate as the amount of pixels for each semantic class $c \in \mathcal{C}$ only once, as a preprocessing step. This array, which reflects the distribution of the classes, is normalized in the range $[0, 1]$ and termed *dist*. As a normalization step, we maintain the min-max normalization function. In addition, for each class c, we keep track of the subset of images X_c in which that category is represented.

Network Confidence *conf*. We compute the network confidence for each class during training and store the result in an array with size $|C|$, named $conf$. At each iteration step t, the pixel-wise Softmax probabilities are computed on the current batch of prediction logits. The mean confidence value for each class c is then derived from the available ground truth labels, by averaging pixels belonging to such category. Lastly, the actual network confidence is computed as the exponential moving average of the prior confidence at step $t - 1$:

$$conf_t = \alpha conf_{t-1} + (1 - \alpha) conf_t, \tag{5}$$

where α represents a smoothing factor.

4. Experiments

4.1. Dataset and Metric

We assess the performance of our approach considering the evaluation protocol described in Agriculture-Vision [11]. Due to the unavailability of the test set, we measure performances on the provided validation set. We conduct two sets of experiments: the first uses only RGB images for training and testing, while the second exploits NIR data in conjunction with the RGB images.

Method	Semantic Classes IoU									mIoU
	Background	Double Plant	Drydown	Endrow	Nutrient Deficiency	Planter Skip	Water	Waterways	Weed Cluster	
FCN	69.99	16.91	45.55	0.18	13.66	6.62	42.27	0.52	8.50	22.91
DeepLab V3	66.27	17.01	40.64	9.46	16.40	10.04	17.06	12.29	9.97	22.13
DeepLab V3+	68.55	16.31	46.36	6.46	16.05	4.56	16.61	19.10	13.89	23.10
UperNet	65.84	15.79	38.03	10.12	17.31	11.09	4.47	15.45	16.94	21.67
SFPN	69.65	10.61	49.49	2.70	11.46	4.80	35.68	9.89	11.16	22.83
PSPNet	68.11	16.93	45.77	4.89	18.99	8.54	11.31	17.64	17.20	23.26
HRNetV2	71.21	16.81	55.10	5.22	18.63	13.26	13.03	21.23	14.07	25.39
HRNetV2+OCR	72.42	19.46	56.79	12.31	17.30	21.31	28.36	24.62	18.05	30.07
SegFormer	74.93	33.19	**59.65**	18.28	**31.64**	39.20	77.97	**41.45**	28.31	44.96
Ours	**75.47**	**36.97**	58.49	**22.69**	31.29	**41.39**	**80.23**	40.07	**30.42**	**46.41**

Table 1. Experiments using RGB images for training and testing on the Agriculture Vision dataset.

Method	Semantic Classes IoU									mIoU
	Background	Double Plant	Drydown	Endrow	Nutrient Deficiency	Planter Skip	Water	Waterways	Weed Cluster	
FCN	68.35	9.40	47.57	0.54	15.16	9.97	53.74	0.47	10.17	23.93
DeepLab V3	69.03	19.97	43.94	5.85	23.98	17.86	46.74	29.03	11.36	29.75
DeepLab V3+	68.29	17.18	48.07	7.48	24.17	19.57	19.43	24.58	13.22	26.89
UperNet	67.43	15.63	36.40	10.73	20.37	14.57	34.21	25.28	14.54	26.57
SFPN	68.69	5.99	48.71	0.18	22.74	17.21	44.50	18.30	12.79	26.57
PSPNet	66.92	17.73	29.87	10.24	28.01	18.66	13.90	29.83	11.99	25.24
HRNetV2	71.28	16.99	54.30	4.52	27.90	15.74	21.66	25.47	17.88	28.42
HRNetV2+OCR	72.60	17.98	56.69	11.97	27.91	23.79	48.99	27.73	22.06	34.42
SegFormer	76.17	33.63	58.96	18.92	40.57	38.93	80.56	42.85	27.88	46.50
Ours	**76.19**	**37.32**	**61.75**	**24.57**	**42.75**	**42.01**	**81.32**	**43.71**	**31.75**	**49.04**

Table 2. Experiments using the combination of NIR data and RGB images for training and testing on the Agriculture Vision dataset.

Dataset. Agriculture-Vision is a large-scale aerial farmland imagery collection for semantic segmentation of agricultural patterns. The dataset is made up of images collected from 3432 farmlands around the United States. It consists of 56944 RGB training images, which have been semantically labeled with 9 different categories. The validation set comprises instead 18334 images. In addition, for each training and validation sample, the NIR channel is also provided. Images are already provided in tiled format, 512 × 512 pixels each.

Metric. Following previous works, we adopt the standard average Intersection over Union metric (mIoU) [16] to measure the performance in all the experiments presented in the following sections.

4.2. Implementation Details

Architecture. The segmentation module at the base of our method is the SegFormer architecture [35], using a MiT-B5 encoder pretrained on ImageNet-1k as backbone.

Baselines. We compare our method against multiple baselines, taking into account the majority of state-of-the-art semantic segmentation techniques reported in the literature. The first model we examine is the FCN [20]. We conduct experiments using the DeepLab V3 [8] and DeepLab V3+ [9] models from the DeepLab family. To compare with multi-scale techniques, we consider the FPN [19], UperNet [34], and PSPNet [40]. All of these models are trained using ResNet-50 pretrained on ImageNet as the backbone. We then report results for the HRNetV2 method [31] and its extension with transformer HRNetV2+OCR [38]. HRNetV2-W18 pretrained on ImageNet serves as the backbone for both of them. Lastly, we examine the SegFormer architecture [35], exploiting the standard pretrained encoder MiT-B5 with ImageNet weights, as it represents the baseline for our approach.

Training. To develop our framework and reproduce all the baselines, we leverage the *mmsegmentation* [12] framework, which is based on PyTorch. We train every configuration on two NVIDIA Tesla v100 GPUs with 16GB of RAM each. In terms of dataset augmentation, we employ random resizing with ratio in range (1.0, 2.0), random horizontal and vertical flipping, and random crops resized to 512x512 during training. Considering the evaluation pipeline, we perform inferences on raw data with no further preprocessing. We train all of the baselines and our model for 80k iterations using the AdamW optimizer. The learning rate is set to 6×10^{-5}, the weight decay to 0.01 and the betas are set to (0.9, 0.999). We use a *poly* learning rate decay with a factor of 1.0 and an initial linear warm-up for 1500 iterations. We do not use class-balanced loss or OHEM approaches as in SegFormer [35]. When training using NIR data, we expand the network input to four channels by doubling the input weights of the red channel.

For the Augmentation Invariance (AI) variants, we further alter the available images using horizontal and vertical flipping, random rotation from 0° to 360° with a step of

90°, photometric and perspective distortion with a strength of 0.1. The probability for each transform is set to 0.5. We set the value of λ in Eq. (2) to 0.75 (see Sec. 5.2 for additional details).

Considering a hyperparameters search that compared the following values for $\gamma = \{1, 2, 4, 6\}$ and for $\alpha = \{0.75, 0.85, 0.90, 0.968, 0.99\}$ on both settings, we set $\gamma = 4$ on Eq. (4), and $\alpha = 0.968$ on Eq. (5).

4.3. Results

RGB. The results for this set of experiments are reported in Tab. 1. The results confirm the difficulty of the task, as the averaged mIoU reaches 23.03% when all the baseline approaches minus Transformer-based architectures are considered, while it increases to an average value of 32.68% when OCR and SegFormer are added. With a mIoU of 21.67%, UperNet is the least performing approach. Despite this, it is one of the best in segmenting underrepresented classes, such as *double plant*, *waterways* or *weed cluster*, as it is meant to capture multi-scale information. When using Transformer architectures, far better results can be observed, with a 30.07% mIoU using the HRNetV2+OCR technique and even 44.96% when using SegFormer; the improvement over UperNet is $+8.4\%$ and $+23.29\%$, respectively.

The majority of these strategies is designed for the autonomous driving domain, without considering the specific challenges intrinsic in aerial data. The introduction of our Augmentation Invariance and Adaptive Sampling results in a substantial boost in performance among almost all the semantic classes, especially on the underrepresented ones like *double plant* or *endrow*, yielding a total mIoU of 46.41%, and an improvement of $+24.74\%$ over the least performing UperNet and of $+1.45\%$ over the SegFormer architecture.

NIR-RGB. As expected, when using the additional Near-Infrared data provided by the Agriculture-Vision dataset, we observe performance improvements on all the baselines. The results are summarized in Tab. 2. The average performance obtained from all baselines without considering transformers is 26.76% in terms of mIoU, while it reaches 35.85% on average when the Transformer-based architectures are also considered. Compared to the setting with RGB images only, the measured improvement reaches $+3.73\%$ and $+3.17\%$, respectively. This demonstrates how NIR infrared data enhances the whole training method by adding value and knowledge, in agreement with the literature [36]. With a mIoU of 23.93%, the FCN architecture achieves the lowest score, while the SegFormer architecture represents again the best performing approach among the baselines with a mIoU of 46.50%. The overall improvement in comparison to FCN is $+22.57\%$.

In this set of experiments, our solution appears to be successful, achieving the best performance among all considered approaches, with a mIoU of 49.04%. AI and AS improve the performance in all the semantic classes, with some outliers in underrepresented ones, such as *double plant*, which gains a $+27.92\%$, *endrow*, which gains a $+24.03\%$, *planter skip*, which gains a $+32.04\%$ and *waterways*, which gains a $+43.24$ w.r.t. the least performant FCN. The overall improvement in comparison to FCN that AI and AS allow to reach is of $+25.11\%$. These results and the qualitatives in Fig. 3 confirm the validity and effectiveness of our solution in dealing with the primary challenges raised by this task.

5. Ablation study

5.1. Contribution of each component

In this section, we assess how each proposed component contributes to the overall performance of our method. We investigate four distinct cases: (a) the SegFormer framework, (b) the introduction of our Augmentation Invariance (AI), (c) the introduction of our Adaptive Sampling (AS) technique, and lastly (d) the entire framework, which includes both AI and AS. We report the results in Tab. 3. This table shows how the Augmentation Invariance is critical for providing a boost to the overall framework, thus confirming our conjecture about the specific challenges in agricultural aerial imagery. The achieved boost is $+2.32\%$ in comparison to the baseline architecture. The addition of AS boosts the simple SegFormer design, delivering state-of-the-art results and highlighting the need to address the semantic class imbalance. The combination of AI and AS provides a further improvement, particularly on underrepresented classes, e.g. *double plant* rise of $+3.69\%$ and *endrow* rise of $+5.65\%$ w.r.t. SegFormer, and of $+2.06\%$ and $+3.83\%$ w.r.t. AI, respectively.

5.2. Ablation on λ

The λ hyperparameter is required to determine the intensity of the Augmentation Invariance (AI) loss. The following values of λ are being compared: 0.1, 0.25, 0.5, 0.75, and 1.0. We run the experiments using the NIR-RGB protocol, without applying the Adaptive Sampling, and we report the outcomes in Tab. 4. The best results are obtained when $\lambda = 0.75$. Even though $\lambda = 0.1$ yields the lowest performance, with a difference of 1.13% when compared to $\lambda = 0.75$, the achieved score can still be considered state of the art on its own. In conclusion, even when using sub-optimal hyperparameters, our AI outperforms all the baselines, highlighting the effectiveness of the approach.

Components	Semantic Classes IoU									mIoU
	Background	Double Plant	Drydown	Endrow	Nutrient Deficiency	Planter Skip	Water	Waterways	Weed Cluster	
SegFormer	76.17	33.63	58.96	18.92	40.57	38.93	80.56	42.85	27.88	46.50
SegFormer + AI	76.62	35.26	61.24	20.74	43.45	43.49	80.41	45.10	33.12	48.82
SegFormer + AS	75.89	35.86	59.23	22.5	41.25	40.72	77.98	40.85	30.99	47.25
SegFormer + AI + AS	76.19	37.32	61.75	24.57	42.75	42.01	81.32	43.71	31.75	**49.04**

Table 3. Ablation study showing the effectiveness of the AI and AS components on the NIR-RGB setting.

λ	Semantic Classes IoU									mIoU
	Background	Double Plant	Drydown	Endrow	Nutrient Deficiency	Planter Skip	Water	Waterways	Weed Cluster	
0.1	76.60	33.92	60.24	18.84	41.92	41.28	82.23	42.45	31.70	47.69
0.25	76.54	35.26	60.70	20.55	42.22	43.84	80.60	43.16	33.25	48.46
0.5	76.48	35.79	59.71	20.34	42.65	40.03	81.12	44.52	32.00	48.07
0.75	76.62	35.26	61.24	20.74	43.45	43.49	80.41	45.10	33.12	**48.82**
1	76.57	34.42	60.25	20.32	41.95	40.03	82.14	43.51	32.22	47.93

Table 4. Ablation study on the influence of λ on the NIR-RGB setting.

6. Conclusions

Limitations. When we apply Adaptive Sampling to the Agriculture-Vision dataset, we see a modest drop in performance on some categories, such as *planter skip* or *waterways*. This is because the difference in absolute pixel counts between these categories (excluding the background class) does not appear to be extremely significant, limiting the influence of the AS technique on the final outcome. Moreover, we note that the adopted training configuration might not be optimal for this setting, therefore hyperparameter changes, such as a higher iteration count, may limit or completely solve these issues.

Conclusion. In this paper we address the problem of Semantic Segmentation for agricultural aerial images. Aside from the standard issues in semantic segmentation, delineating patterns in aerial imagery poses additional challenges such as how to leverage the additional multi-modal data that comes with the visible spectrum, the imbalance in class-wise pixel distribution, and the changes in point of view. We offer two approaches to address these challenges in an end-to-end trainable framework: an Augmentation Invariance solution that forces the model to learn semantic representations that are invariant to the point-of-view shifts typical in aerial imagery, and an Adaptive Sampling solution that addresses the problem of class imbalance by actively sampling the training images based on their class-wise pixel distribution and the current network confidence.

We propose a comprehensive series of experiments and ablation studies on the Agriculture-Vision dataset and we prove how our methods considerably increase the performance of the actual state-of-the-art models, especially on underrepresented classes.

Future research will look into the feasibility of developing a plug-and-play technique for bringing Augmentation Invariance and Adaptive Sampling to any segmentation backbone. Furthermore, we will examine the influence of various types of augmentations employed to force the model to be agnostic. We also intend to evaluate our method in a variety of contexts, including Domain Adaptation and Domain Generalization.

References

[1] E. Alberti, A. Tavera, C. Masone, and B. Caputo. Idda: A large-scale multi-domain dataset for autonomous driving. *IEEE Robotics and Automation Letters*, 5(4):5526–5533, 2020. 2

[2] RB Andrade, GAOP Costa, GLA Mota, MX Ortega, RQ Feitosa, PJ Soto, and Christian Heipke. Evaluation of semantic segmentation methods for deforestation detection in the amazon. *ISPRS Archives; 43, B3*, 43(B3):1497–1505, 2020. 1

[3] Edoardo Arnaudo, Fabio Cermelli, Antonio Tavera, Claudio Rossi, and Barbara Caputo. A contrastive distillation approach for incremental semantic segmentation in aerial images. *arXiv preprint arXiv:2112.03814*, 2021. 3

[4] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journ. Phot. Rem. Sens.*, 140:20–32, 2018. 2

[5] Nadir Bengana and Janne Heikkilä. Improving land cover segmentation across satellites using domain adaptation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1399–1410, 2020. 2

[6] Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. Landcover. ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1102–1110, 2021. 1, 2

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions*

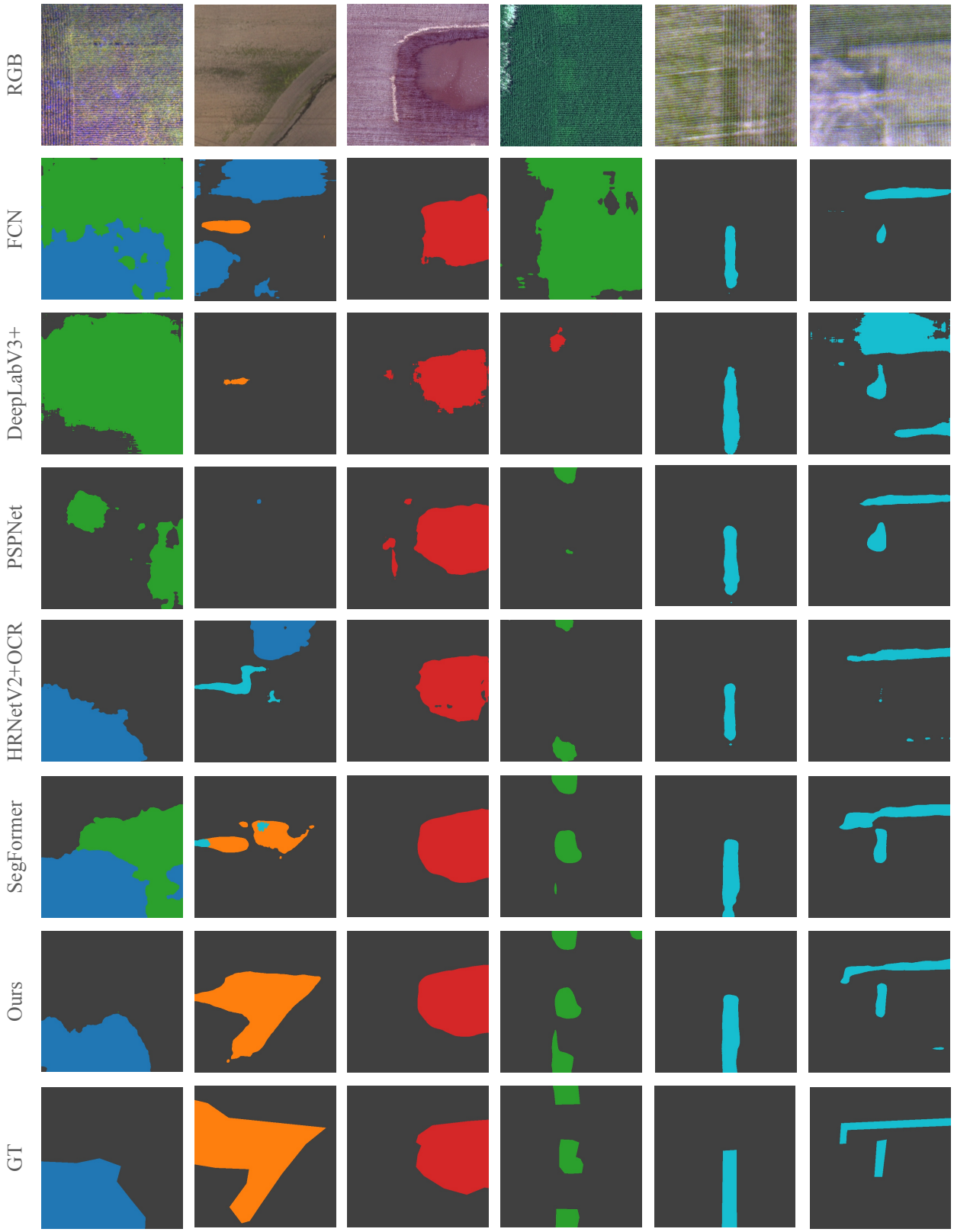

Figure 3. Qualitative results on the validation set of the Agriculture-Vision dataset.

on Pattern Analysis and Machine Intelligence, 40:834–848, 2018. 2

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017. 2, 5

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2, 5

[10] Dominic Cheng, Renjie Liao, Sanja Fidler, and Raquel Urtasun. Darnet: Deep active ray network for building segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7431–7439, 2019. 2

[11] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, et al. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2828–2838, 2020. 2, 4

[12] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 5

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[14] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018. 1, 2

[15] Foivos I. Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journ. Phot. Rem. Sens.*, 162:94–114, 2020. 2

[16] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111:98–136, 2014. 5

[17] Alessandro Farasin, Luca Colomba, and Paolo Garza. Double-step u-net: A deep learning-based approach for the estimation of wildfire damage severity through sentinel-2 satellite data. *Applied Sciences*, 10(12):4332, 2020. 1

[18] Mulham Fawakherji, Ali Youssef, Domenico Bloisi, Alberto Pretto, and Daniele Nardi. Crop and weeds classification for precision agriculture using context-independent pixel-wise segmentation. In *2019 Third IEEE International Conference on Robotic Computing (IRC)*, pages 146–152. IEEE, 2019. 2

[19] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 2, 5

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2015. 2, 5

[21] Ning Lv, Chen Chen, Tie Qiu, and Arun Kumar Sangaiah. Deep learning and superpixel feature extraction based on contractive autoencoder for change detection in sar images. *IEEE transactions on industrial informatics*, 14(12):5530–5538, 2018. 2

[22] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 2229–2235. IEEE, 2018. 2

[23] Shervin Minaee, Yuri Y. Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1

[24] Keiller Nogueira, Mauro Dalla Mura, Jocelyn Chanussot, William Robson Schwartz, and Jefersson A. dos Santos. Learning to semantically segment high-resolution remote sensing images. In *Int. Conf. Pattern Recog.*, pages 3566–3571, 2016. 2

[25] Bin Pan, Zhenwei Shi, Xia Xu, Tianyang Shi, Ning Zhang, and Xinzhong Zhu. Coinnet: Copy initialization network for multispectral imagery semantic segmentation. *IEEE Geos. Rem. Sens. Lett.*, 16(5):816–820, 2019. 2

[26] Kunlun Qi, Chao Yang, Chuli Hu, Yonglin Shen, Shengyu Shen, and Huayi Wu. Rotation invariance regularization for remote sensing image scene classification with convolutional neural networks. *Rem. Sens.*, 13(4), 2021. 3

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 1, 2

[28] Hao Sheng, Xiao Chen, Jingyi Su, Ram Rajagopal, and Andrew Ng. Effective data fusion with generalized vegetation index: Evidence from land cover segmentation in agriculture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 60–61, 2020. 2

[29] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *CoRR*, abs/1904.04514, 2019. 2

[30] Antonio Tavera, Fabio Cermelli, Carlo Masone, and Barbara Caputo. Pixel-by-pixel cross-domain alignment for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1626–1635, January 2022. 2

[31] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep

high-resolution representation learning for visual recognition. *TPAMI*, 2019. 2, 5

[32] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 1, 2

[33] John Weier and David Herring. Measuring vegetation (ndvi & evi). *NASA Earth Observatory*, 20, 2000. 2

[34] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 2, 5

[35] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 3, 5

[36] S. Yang, S. Yu, B. Zhao, and Y. Wang. Reducing the feature divergence of rgb and near-infrared images using switchable normalization. In *IEEE Conf. Comput. Vis. Pattern Recog. Work.*, pages 206–211, jun 2020. 2, 3, 6

[37] Qinglie Yuan, Helmi Zulhaidi Mohd Shafri, Aidi Hizami Alias, and Shaiful Jahari bin Hashim. Multiscale semantic feature optimization and fusion network for building extraction using high-resolution aerial images and lidar data. *Rem. Sens.*, 13(13), 2021. 2, 3

[38] Yuhui Yuan, Xiaokang Chen, Xilin Chen, and Jingdong Wang. Segmentation transformer: Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019. 2, 5

[39] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 2

[40] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017. 2, 5