Automatic Grassland Degradation Estimation Using Deep Learning

Xiyu Yan[1,*], **Yong Jiang**[1,2], **Shuai Chen**[3,*], **Zihao He**[1], **Chunmei Li**[4], **Shu-Tao Xia**[1,2], **Tao Dai**[1,2], **Shuo Dong**[4] and **Feng Zheng**[5,†]
[1]Dept. of Computer Science and Technology, Tsinghua University
[2]PCL Research Center of Networks and Communications, Peng Cheng Laboratory
[3]Baidu, Inc.
[4]Dept. of Computer Technology and Applications, Qinghai University
[5]Dept. of Computer Science and Engineering, Southern University of Science and Technology
yanqy17@mails.tsinghua.edu.com, {jiangy, xiast}@sz.tsinghua.edu.com, zhengf@sustech.edu.cn

Abstract

Grassland degradation estimation is essential to prevent global land desertification and sandstorms. Typically, the key to such estimation is to measure the coverage of indicator plants. However, traditional methods of estimation rely heavily on human eyes and manual labor, thus inevitably leading to subjective results and high labor costs. In contrast, deep learning-based image segmentation algorithms are potentially capable of automatic assessment of the coverage of indicator plants. Nevertheless, a suitable image dataset comprising grassland images is not publicly available. To this end, we build an original Automatic Grassland Degradation Estimation Dataset (AGDE-Dataset), with a large number of grassland images captured from the wild. Based on AGDE-Dataset, we are able to propose a brand new scheme to automatically estimate grassland degradation, which mainly consists of two components. 1) Semantic segmentation: we design a deep neural network with an improved encoder-decoder structure to implement semantic segmentation of grassland images. In addition, we propose a novel Focal-Hinge loss to alleviate the class imbalance of semantics in the training stage. 2) Degradation estimation: we provide the estimation of grassland degradation based on the results of semantic segmentation. Experimental results show that the proposed method achieves satisfactory accuracy in grassland degradation estimation.

1 Introduction

In recent years, massive grassland ecosystem has undergone degradation because of climatic variations and overgrazing, thus resulting in multifarious ecological problems, such as desertification and sandstorms [Zhan *et al.*, 2017]. Therefore, how to estimate the stage of grassland degradation accurately is of top priority for protecting grassland ecosystem from desertification.

The emergence of indicator plants is an important sign of grassland degradation [Zhao *et al.*, 2004]. Many countries have successfully used specific plant species as indicators for estimating grassland degradation [Mansour *et al.*, 2016; Mansour *et al.*, 2012]. Our case study of degrading grassland in Qinghai-Tibet Plateau demonstrates that as the grassland degrades, the coverage of *Stellera chamaejasme* (*SC*) gradually accumulates. Thus, *SC* is regarded as the indicator plants for grassland degradation. Specifically, grassland would go through five degradation stages before desertification, with the coverage of *SC* building up in each stage [Zhao *et al.*, 2004], as shown in Table 1. Thus, it is intuitive to estimate the grassland degradation stage based on the coverage of *SC*. However, existing methods rely heavily on observations of human eyes and manual labor, thus leading to subjective results and high labor costs, which is undesirable in practice. Consequently, there is an urgent need for developing an effective and efficient method to automatically estimate the grassland degradation stage without human any interactions.

To do this, we attempt to leverage deep learning to calculate automatically the coverage of *SC* in real-world grassland images based on a semantic segmentation algorithm, and then estimate the stage of grassland degradation by the coverage of *SC* based on the results of recognition. Many challenges stand in the way of achieving an automatic estimation of degradation. First, existing public datasets [Mottaghi *et al.*, 2014; Cordts *et al.*, 2016; Caesar *et al.*, 2018; Ros *et al.*, 2016] contain substantially insufficient grassland images and thus fail to provide us with enough samples to train the network. In addition, the aerial or satellite images used in the studies of remote sensing and environmental sciences [Wang *et al.*, 2018] are not high-resolution enough to capture such a tiny target as *SC*.Moreover, due to the particularity of the grassland scene, capturing images with semantic class imbalance is inevitable. Finally, existing semantic segmentation networks cannot handle directly the complex task with these challenges.

To this end, we first design a deep neural network to implement semantic segmentation that could accurately segment the foreground (*SC*) from the background (grassland

*Equal Contribution
†Contact Author

Degradation Stage	Coverage of *SC*
I	0%-19%
II	20%-39%
III	40%-59%
IV	60%-79%
V	80%-95%

Table 1: The relationship between stage of grassland degradation and coverage of *Stellera chamaejasme* (*SC*).

and other elements in a grassland scene) of the grassland image at the pixel level. Aiming at the problem of sample insufficiency, we create a labeled dataset for automatic grassland degradation estimation with ground-level grassland images captured from Qinghai-Tibet Plateau. To alleviate the problem of class imbalance, we combine the advantages of reducing the class imbalance in Focal Loss [Lin *et al.*, 2017] and increasing class distance in smoothed Hinge loss [Rennie and Srebro, 2005], and propose an original Focal-Hinge loss function. Next, we calculate the coverage of *SC* in the grassland area through the analysis of the results from semantic segmentation of grassland images and accordingly determine the degradation stage according to the relationship between stage and coverage (Table 1).

Through these two steps, we manage to automatically estimate grassland degradation based on deep learning. To the best of our knowledge, we are the first to leverage deep learning techniques to solve ecological problems regarding grassland ecosystem. To be more specific, we propose a brand new scheme for grassland degradation estimation using semantic segmentation by a deep neural network. Moreover, we design a Focal-Hinge loss function to train the proposed network for addressing the problem of class imbalance. Experiments of our scheme on the Automatic Grassland Degradation Estimation Dataset (AGDE-Dataset) are carried out and reveal satisfying estimation results, which substantiate the feasibility and prospect to solve the problem of automating grassland degradation estimation leveraging deep learning.

2 Related Work

2.1 Plant Identification

Image-based plant identification is one of the most promising solutions towards furthering botanical taxonomy, as illustrated by the wealth of research regarding this topic [Cerutti *et al.*, 2011; Kebapci *et al.*, 2011; Goëau *et al.*, 2016].

Although we need to identify indicator plants for grassland degradation, our task is much more demanding. First, the deep learning-based plant recognition of these researches are more of image classification. However, in our task concerning automatic grassland degradation estimation, we also need to tell the spatial information like the locations and areas of them in an image. In addition, these algorithms only recognize plants in an image with discernible plants and background. However, the indicator plants often lurk in the vast expanse of grassland, making themselves indistinguishable in an grassland image.

With this regard, to precisely figure out the proportion of the indicator plants in an image, we propose a semantic segmentation methodology.

2.2 Plant Density Estimation

In recent years, there has been a lot of research into plant density estimation based on image processing [Liu *et al.*, 2017a; Liu *et al.*, 2017b; Jin *et al.*, 2017]. For example, [Liu *et al.*, 2017b] takes wheat plant images by a high-resolution RGB camera and train Artificial Neural Networks with 10 manually extracted features to estimate the number of plants. [Jin *et al.*, 2017] captures images by a UAV and train a Support Vector Machine with 13 hand-crafted features to identify wheat plant.

In fact, plant density estimation entails the quantification of the plant within a given unit area, which is highly biased by plant distribution. However, plant coverage refers to a relative area covered by the plant species in a plot, the calculation of which is more complex than that of quantification, since the area would not necessarily scale with the quantity of plant. In addition, both [Liu *et al.*, 2017b] and [Jin *et al.*, 2017] treat the density estimation problem as object classification, which might work with manually extracted features. In contrast, our task is based on semantic segmentation, where hand-engineered features are not feasible, so we automatically extract features with the designed deep network.

2.3 Semantic Segmentation

Semantic segmentation necessitates object classification at the pixel level. Recently, there are many fabulous semantic segmentation models such as FCN [Long *et al.*, 2015], SegNet [Badrinarayanan *et al.*, 2017], DeepLab-v3 [Chen *et al.*, 2018a], PSPNet [Zhao *et al.*, 2017] and etc. Leong first proposed Fully Convolutional Networks (FCN) [Long *et al.*, 2015] that is a convolutional network for dense prediction without a fully-connected layer. This model makes it possible to segment images at any size effectively, and it is much faster than traditional methods based on patch classification. However, an obtrusive problem using convolutional neural networks for semantic segmentation is that pooling layers enlarge the receptive field, aggregating contextual information while discarding location information. Therefore, in order to solve this problem, an encoder-decoder architecture is devised. The encoder gradually reduces the spatial dimensions by pooling layers, while the decoder restores the target details and spatial dimensions step by step. Among such architectures, U-Net [Ronneberger *et al.*, 2015] is a very efficient one, whose semantic segmentation model employs the architecture of encoder-decoder based on a fully convolutional neural network. At present, semantic segmentation is widely applied to the geographic information system, unmanned vehicles [Menze and Geiger, 2015], medical image analysis [Zhang *et al.*, 2017], robots and etc. Endowed with the power of the designed encoder-decoder deep network, we manage to figure the coverage of the indicator plants in grassland images by the results of semantic segmentation.

3 The AGDE-Dataset

Since existing open image datasets contain inadequate grassland images, and even fewer datasets would cover *SC* images,

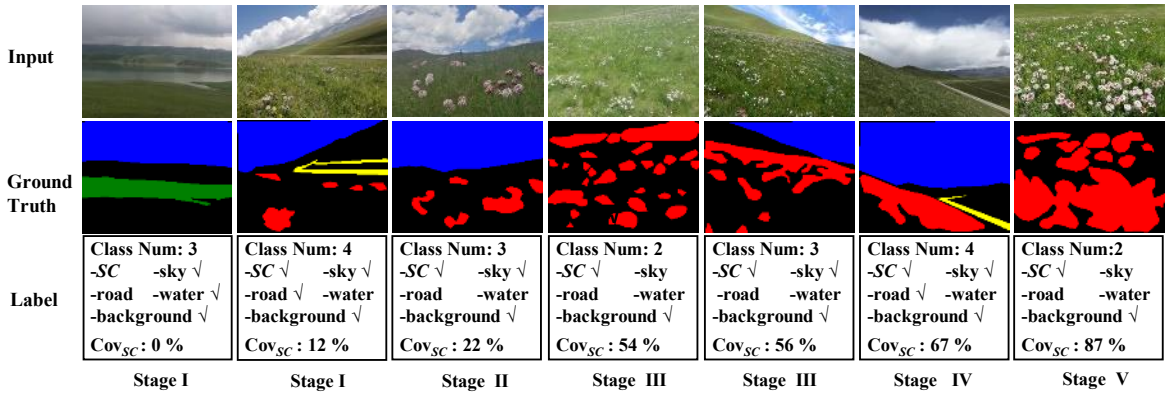

Figure 1: The labeled AGDE-Dataset with ground truth of semantic segmentation and the information of degradation stage. Cov_{SC} in label information represents the coverage of *SC* in the grassland image.

we create a labeled dataset – Automatic Grassland Degradation Estimation Dataset (AGDE-Dataset).

First of all, we capture a large number of images from grasslands on the Qinghai-Tibet Plateau, from which we sift 2,895 images and scale them down. The sizes of the resized images range from 5KB to 38KB.

Next, we manually label pixels belonging to each of the five semantic categories – grassland (background), *SC*, sky, water, and road – for every image with an open annotation tool – LabelMe [Russell *et al.*, 2008]. Due to the limitation on the plateau, there are inevitably quantitative differences in different semantic categories in the dataset. The number of these five semantic categories are 2,895, 2,888, 156, 48 and 32 respectively. In addition, the degradation stage for each image in AGDE-Dataset is labeled according to the coverage of *SC* (Table 1) by calculating the proportion of the area of *SC* in images, which will be articulated in Section 4.2. We show several labeled examples of our dataset in Figure 1.

Finally, we randomly divide the dataset into a training set and a test set in the ratio of 2,095:800, which is detailed in Table 2. All these RGB images are eventually padded to the resolution of 256×341. Although the dataset is not very big, the amount of the training set is enough for our network training.

4 Proposed Method

We address the problem of automatic grassland degradation stage estimation following two steps: 1) semantic segmentation: designing a deep network to implement semantic segmentation for grassland images, and training it using the proposed novel Focal-Hinge loss function; 2) degradation estimation: figuring out the coverage of *SC* according to the semantic segmentation results and further estimating the degradation stage of grassland.

Dataset	Number of Images in Each Stage					
	I	II	III	IV	V	**Total**
Train Set	295	500	500	500	300	**2,095**
Test Set	100	200	200	200	100	**800**

Table 2: The number of images of each stage in AGDE-Dataset.

4.1 Semantic Segmentation for Grassland Scene

Network Architecture

The proposed network is based on the classic encoder-decoder network architecture without the fully-connected layers [Badrinarayanan *et al.*, 2017] and we improve it by adding the refined cross connections as shown in Figure 2.

The encoder network consists of 5 encoder convolution groups. Each *encoder* in the encoder network contains several Convolution, Batch Normalization, and ReLU (Conv + BN + ReLU) layers with stride 1. Following that, max pooling with a 2×2 kernel and stride 2 (non-overlapping window) is performed. Therefore, the input image is totally down-sampled 2^5 (32) times through the encoder network. Symmetrically, the decoder network also contains 5 decoder up-sampling groups. Each *decoder* contains 1 deconvolution layer and 3 Conv + BN + ReLU layers with stride 1. The 5 *decoders* upsample the last *encoder* to 32 times, so the size of the entire network output is equal to that of the input image.

In addition, in order to enrich the representation of the encoder, we connect symmetrically the feature maps of the decoder layers to the encoder layers by refined convolution groups. For example, the feature maps of the 4th *encoder* are connected to the 5th *decoder* by the 4th refined cross connection unit, and the feature maps of the 1st *encoder* are connected to the 2nd *decoder* by the 1st refined cross connection unit. The cross connections utilize low-level features and prevent gradient disappearance in the underlying gradient. Each of refined convolution groups contains 3 Convolution layers, Batch Normalization, ReLU, and Dropout (Conv + BN + ReLU + Dropout) layers with stride 1 [He *et al.*, 2016], which is similar to that in DenseNet [Huang *et al.*, 2017].

Training

We train the network on AGDE-Dataset (Section 3). However, we face the challenge of the segmentation of the 5 imbalanced classes – *SC*, sky, road, water, and background, with *SC* predominating. The problem of class imbalance is very common in semantic segmentation. In order to alleviate this problem, we devise a novel loss function – Focal-Hinge loss as the objective function for training the network.

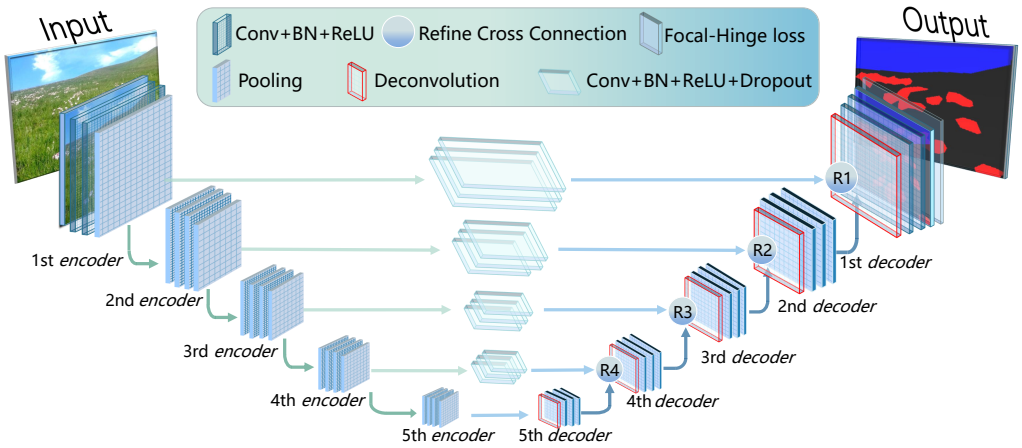

Figure 2: An illustration of the network architecture in this paper. Each of the 5 *encoder* layers implements down-sampling with several convolution layers and max pooling, and then each of the 5 *encoder* layers performs up-sampling with several deconvolution layers. The feature maps of the decoder layers are connected to the corresponding that of encoder layers by refined convolution groups, each of which contains Conv + BN + ReLU + Dropout layers with stride 1.

Focal-Hinge Loss

The following of this subsection details how the loss function is derived.

First, in order to classify each pixel in an image, we consider the Cross Entropy loss with a *sigmoid* activation function

$$p_t(y_{pix}) = \frac{1}{1+\exp(-y_{pix})}, \tag{1}$$

where y_{pix} is a heatmap pixel value of the network output, and $p_t(y_{pix})$ is the probability of ground truth class. The Cross Entropy loss is represented as

$$\mathrm{CE}(p_t) = -\log(p_t). \tag{2}$$

However, $\mathrm{CE}(p_t)$ with *sigmoid* is not capable of correctly classifying some pixels into the minority semantic classes in the training set, due to the class imbalance problem faced by many classic semantic segmentation models [Long *et al.*, 2015; Huang *et al.*, 2015; He *et al.*, 2017]. Considering class imbalance and easy sample overwhelming, we substitute the classic Cross Entropy loss with a new Focal Loss [Lin *et al.*, 2017] to reduce the weight of easy samples. In this way, during the training process, the model focuses more on hard samples. Focal Loss is represented as

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t), \tag{3}$$

where γ is a focusing parameter. If $\gamma > 0$, the relative loss for correctly-classified examples ($p_t > 0.5$) would decrease.

Nevertheless, we find that although $\mathrm{FL}(p_t)$ with *sigmoid* in Eq. (3) alleviates the problem of the easy sample overwhelming, the scoring on hard samples is still far from satisfying because the scores of two classes are too close to each other. In other words, the boundaries of segmentation results are not clear-cut, with one class mixed with another. In this regard, naturally, we would consider the smoothed Hinge loss [Rennie and Srebro, 2005], as shown in Eq. (4), which is always utilized in maximizing the classification interval in SVM, to make the final score more discriminative.

$$\mathrm{HL}(y_{pix}) = \begin{cases} \frac{1}{2} - t\cdot y_{pix} & \text{if} \quad t\cdot y_{pix} \leqslant 0, \\ \frac{1}{2}(1-t\cdot y_{pix})^2 & \text{if} \quad 0 < t\cdot y_{pix} < 1, \\ 0 & \text{if} \quad t\cdot y_{pix} \geqslant 1, \end{cases} \tag{4}$$

where t stands for the ground truth class to which y_{pix} corresponds.

Nonetheless, the smoothed Hinge loss itself does not tackle the problem of class imbalance. Therefore, considering these two problems – insufficient class distance and class imbalance, we propose a Focal-Hinge loss (FH) as

$$\mathrm{FH}(y_{pix}) = \begin{cases} N_t(1-p_t(y_{pix}))^{\gamma}(\frac{1}{2} - t\cdot y_{pix}) & \text{if} \quad t\cdot y_{pix} \leqslant 0, \\ \frac{1}{2}N_t(1-p_t(y_{pix}))^{\gamma}(1-t\cdot y_{pix})^2 & \text{if} \quad 0 < t\cdot y_{pix} < 1, \\ 0 & \text{if} \quad t\cdot y_{pix} \geqslant 1, \end{cases} \tag{5}$$

where $N_t(1-p_t)^{\gamma}$ represents the FL with a *sigmoid* activation function, and N_t denotes the reciprocal of the number of classes t in one image, and γ denotes the hyper-parameter. FH is the combination of FL with *sigmoid* and smoothed HL. On the one hand, for the part of FL with *sigmoid*, the network output can be normalized to [0,1], so that the classes with more samples can receive a severer penalty based on the probability. On the other hand, the use of smoothed HL ensures that a larger score distance is obtained. We experimentally substantiate the effectiveness of proposed FH compared with CE and FL with *sigmoid* for alleviating class imbalance in Section 5.1.

The back propagation gradient of Focal-Hinge loss (FH) is

$$\frac{\partial \mathrm{FH}}{\partial y_{pix}} = \begin{cases} N_t(1-p_t)^{\gamma}(t - \frac{1}{2}\gamma\cdot p_t - \gamma\cdot t\cdot p_t\cdot y_{pix}), & \\ \qquad\qquad\qquad\qquad\qquad \text{if} \quad t\cdot y_{pix} \leqslant 0, & \\ N_t(1-p_t)^{\gamma}(t\cdot y_{pix} - 1)[t + \frac{1}{2}\gamma\cdot p_t(1-t\cdot y_{pix})], & \\ \qquad\qquad\qquad\qquad\qquad \text{if} \quad 0 < t\cdot y_{pix} < 1, & \\ 0, \qquad\qquad\qquad\qquad\qquad \text{if} \quad t\cdot y_{pix} \geqslant 1, & \end{cases} \tag{6}$$

where p_t represents the function $p_t(y_{pix})$. With this brand new loss function, we manage to alleviate the effect of

class imbalance while keeping the distance between different classes large enough.

4.2 Estimating Grassland Degradation Stage

Inputting a grassland image $\boldsymbol{x}$ into the fully trained deep network, the output image $\boldsymbol{y}$ with the results of semantic segmentation is obtained. However, to finish the grassland degradation stage estimation, we further process the image semantic segmentation results. First, we calculate the coverage of *SC* in grassland images. Second, we estimate the grassland degradation stage according to the corresponding relationship between the coverage and the stage (Table 1).

Define of the Coverage of *SC*

There are totally 5 semantic categories: background, *SC*, sky, water, and road in AGDE-Dataset. To obtain the coverage of *SC* (Cvg_{SC}), we calculate the proportion of the areas of *SC* and grassland background in one image, which is represented as

$$Cvg_{SC} = \frac{A_{SC}}{A_{im} - A_s - A_r - A_w}, \tag{7}$$

where A_{SC} denotes the area of *SC* in the image and A_{im} denotes the area of the entire image. The areas of sky, road, and water are represented by A_s, A_r, and A_w respectively. The denominator $(A_{im} - A_s - A_r - A_w)$ stands for the area of grassland background. Given the semantic segmentation results, A_{SC}, A_s, A_r, and A_w are obtained easily by quantifying pixels in an image. Thus, we could acquire the Cvg_{SC} on the grassland by Eq. (7).

Estimation of Degradation Stage

We construct a mapping relationship between *SC* coverage and the degradation stage (Table 1) [Zhao *et al.*, 2004] to obtain the degradation estimation. In this way, we accomplish the automatic estimation of grassland degradation stage leveraging deep learning.

5 Experiments

In order to demonstrate the effectiveness of our proposed scheme, we evaluate the results from two aspects. First, we evaluate the performance of semantic segmentation. To be more specific, we compare the Focal-Hinge loss function with several other loss functions. In addition, we also showcase the competence of our network with comparison to other classic semantic segmentation networks – FCN [Long *et al.*, 2015], SegNet [Badrinarayanan *et al.*, 2017], and DeepLab-V3 [Chen *et al.*, 2018b]. Second, to test the performance of our scheme on grassland degradation stage estimation, we show its success rate on stage estimation.

5.1 Evaluation of Semantic Segmentation

Implementation Details and Evaluation Criteria

We set the parameters of networks – FCN-8s, FCN-16s, FCN-32s, SegNet, and DeepLab-v3 – according to that specified in their original papers. Besides, γ in Eq. (5) and Eq. (6) are set to 2. The size of the input images is padded to 256×341, which is the largest size of images in the dataset all experiments are conducted on a GTX1080Ti. More detailed experimental parameters are specified in *Supplementary Materials*.

The evaluation metrics are the Pixel Accuracy (PA), Mean Pixel Accuracy (MPA), Intersection over Union (IoU) and Mean Intersection over Union (MIoU).

Evaluation of Focal-Hinge Loss

Under the same conditions, using the proposed network, we juxtapose our Focal-Hinge loss with the baseline ones: *Sigmoid* + Cross Entropy loss (CE) and *Sigmoid* + Focal Loss (FL). The evaluation results are shown in Table 3. We can see that the Focal-Hinge loss (FH) achieves remarkably better performance, in terms of MPA and MIoU. It is notable that FH achieves satisfactory results on minority semantics.

Therefore, the results demonstrate that Focal-Hinge loss function outperforms other losses in semantic segmentation of grassland images and effectively alleviates the problem of class imbalance.

Evaluation of the proposed network

Employing the Focal-Hinge loss, we further compare the semantic segmentation performance of our network with that of FCN-32s, FCN-16s, FCN-8s [Long *et al.*, 2015], SegNet [Badrinarayanan *et al.*, 2017], and DeepLab-v3 [Chen *et al.*, 2018b]. The evaluation results are shown in Table 4. Results show that our network, designed with a lighter structure, outperforms other networks. In addition, the proposed method outperforms FCN and SegNet on minority semantics (road and water).

The visual results are displayed in Figure 3. From the results, we can see that our method outperforms other deep networks, especially on minority semantics.

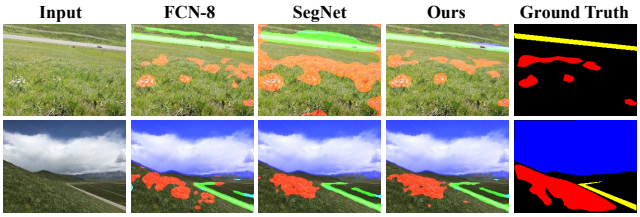

Figure 3: Visual results of semantic segmentation comparing to other deep networks.

5.2 Evaluation of Degradation Stage Estimation

We evaluate the performance of grassland degradation estimation on the test set of AGDE-Dataset, which covers 800 images with 5 degradation stages (Table 2). The success rate δ_c^i in each stage i is calculated by the ratio of the number of correctly estimated images N_c^i to the total number of images N_t^i in the stage:

$$\delta_c^i = \frac{N_c^i}{N_t^i}, \qquad i \in \{\text{I, II, III, IV, V}\}. \tag{8}$$

Accordingly, the error rate δ_e^{ij} of our scheme in each stage i mistaken as stage j is calculated by:

$$\delta_e^{ij} = \frac{N_e^{ij}}{N_t^i - N_c^i}, \quad j \in \{\text{I, II, III, IV, V}\} \setminus \{i\}, \tag{9}$$

Method	PA				MPA	IoU				MIoU
	SC	sky	road	water		*SC*	sky	road	water	
Sigmoid + CE	0.472	**0.792**	0.296	0.231	0.358	0.449	**0.785**	0.242	0.207	0.337
Sigmoid + FL	0.508	0.788	0.399	0.359	0.411	0.477	0.781	0.292	0.296	0.369
FH	**0.543**	0.757	**0.439**	**0.388**	**0.426**	**0.499**	0.753	**0.303**	**0.303**	**0.372**

Table 3: Results of different loss functions using proposed network measured by PA and IoU.

Method	PA				MPA	IoU				MIoU
	SC	sky	road	water		*SC*	sky	road	water	
FCN -32s	0.454	0.722	0.284	0.229	0.338	0.432	0.720	0.238	0.202	0.318
FCN -16s	0.455	0.767	0.311	0.297	0.366	0.433	0.763	0.253	0.258	0.341
FCN -8s	0.495	0.792	0.326	0.336	0.389	0.466	0.766	0.258	0.283	0.369
SegNet	0.463	**0.801**	0.345	0.208	0.391	0.453	**0.780**	0.253	0.265	0.370
DeepLab-v3	**0.551**	0.742	0.422	0.363	0.421	0.459	0.743	0.299	0.298	0.370
Ours	0.543	0.757	**0.439**	**0.388**	**0.426**	**0.499**	0.753	**0.303**	**0.303**	**0.372**

Table 4: Results of different neural networks using Focal-Hinge loss function measured by PA and IoU.

Stage	Success Rate	Error Rate					
		I	II	III	IV	V	other
I	90%	—	100%	0	0	0	0
II	88%	17%	—	83%	0	0	0
III	96%	0	25%	—	75%	0	0
IV	85%	0	0	37%	—	63%	0
V	84%	0	0	0	38%	—	62%
Avg	89%	3%	25%	24%	23%	13%	12%

Table 5: Success rates and error rates of automatic grassland degradation stage estimation.

where N_e^{ij} represents the number of images in stage i mistaken as stage j, and $N_t^i - N_c^i$ represents the total number of mistaken images of stage i. The success rates and error rates of the test set of AGDE-Dataset are shown in Table 5, from which we can see that the success rates on each degradation stage are remarkably satisfactory and the main error estimations tend to occur in two adjacent stages.

The visual results of automatic grassland degradation estimation are shown in Figure 4. We can see that the segmentation results and predicted stage label approximate the ground truth.

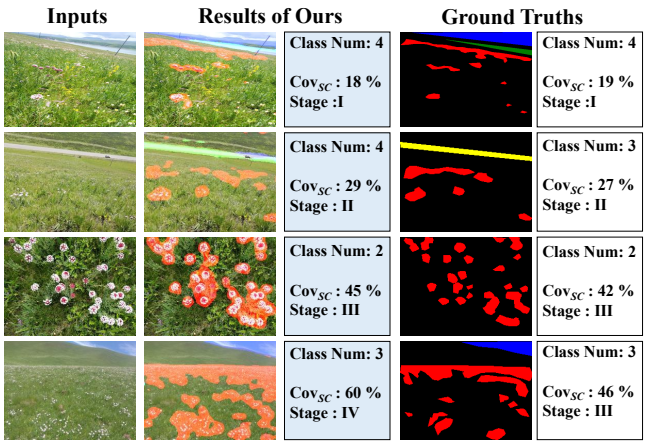

Figure 4: Visual results of semantic segmentation and grassland degradation estimation.

6 Discussion and Future Work

The main contribution of this paper is to provide a scheme to achieve automatic grassland degradation estimation leveraging deep learning. Specifically, we design a deep network especially for semantic segmentation of grassland images. In addition, due to the insufficiency of grassland image samples in public datasets, we capture a large number of grassland images and build a labeled grassland dataset named AGDE-Dataset. Moreover, as for the problem of class imbalance in the dataset, we devise a new Focal-Hinge loss function. Then we calculate the coverage of indicator plants for degradation using the results of semantic segmentation of grassland images and accordingly determine the degradation stage by the mapping of between coverage and stage. Experimental results on AGDE-Dataset indicate that the proposed deep learning based method achieves an remarkably satisfactory result regarding automatic grassland degradation estimation. We hope that our model will be embraced at early date by grassland preservers on automatic estimation of grassland degradation stage.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grant 61771273, the R&D Program of Shenzhen under Grant JCYJ20180508152204044, the research fund of PCL Future Regional Network Facilities for Large-scale Experiments and Applications (PCL2018KP001), and the Program for University Key Laboratory of Guangdong Province (Grant No. 2017KSYS008).

Special thanks for the support of the Research Program of Science and Technology Department of Qinghai Province (Grant No. 2016-ZJ-774), and also for Prof. Li Chunmei and the students from Qinghai University: Dong Shuo, Pi Wei, and Li Zhao, they have made a great contribution to the collection and annotation of the dataset.

References

[Badrinarayanan *et al.*, 2017] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for scene segmentation. *TPAMI*, PP(99):2481–2495, 2017.

[Caesar *et al.*, 2018] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018.

[Cerutti *et al.*, 2011] Guillaume Cerutti, Laure Tougne, Antoine Vacavant, and Didier Coquin. A parametric active polygon for leaf segmentation and shape estimation. In *ISVC*, pages 202–213. Springer, 2011.

[Chen *et al.*, 2018a] L. C. Chen, G Papandreou, I Kokkinos, K Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2018.

[Chen *et al.*, 2018b] Liang Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *CVPR*, 2018.

[Cordts *et al.*, 2016] 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.

[Goëau *et al.*, 2016] Hervé Goëau, Pierre Bonnet, and Alexis Joly. Plant identification in an open-world (lifeclef 2016). In *Conference and Labs of the Evaluation forum*, pages 428–439, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *TPAMI*, PP(99):1–1, 2017.

[Huang *et al.*, 2015] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. Densebox: Unifying landmark localization with end to end object detection. *Computer Science*, 2015.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[Jin *et al.*, 2017] Xiuliang Jin, Baret Liu, Shouyangand Frédéric, Hemerlé Matthieu, and Comarb Alexis. Estimates of plant density of wheat crops at emergence from very low altitude uav imagery. *Remote Sensing of Environment*, 198:105–114, 2017.

[Kebapci *et al.*, 2011] Hanife Kebapci, Berrin Yanikoglu, and Gozde Unal. Plant image retrieval using color, shape and texture features. *The Computer Journal*, 54(9):1475–1490, 2011.

[Lin *et al.*, 2017] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *TPAMI*, PP(99):2999–3007, 2017.

[Liu *et al.*, 2017a] S. Liu, F. Baret, D. Allard, X. Jin, B. Andrieu, P. Burger, M. Hemmerlé, and A. Comar. A method to estimate plant density and plant spacing heterogeneity: application to wheat crops. *Plant Methods*, 13(1):38, 2017.

[Liu *et al.*, 2017b] Shouyang Liu, Fred Baret, Bruno Andrieu, Philippe Burger, and Matthieu Hemmerlé. Estimation of wheat plant density at early stages using high resolution imagery. *Frontiers in Plant Science*, 8:739, 2017.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[Mansour *et al.*, 2012] Khalid Mansour, Onisimo Mutanga, Terry Everson, and Elhadi Adam. Discriminating indicator grass species for rangeland degradation assessment using hyperspectral data resampled to aisa eagle resolution. *Isprs Journal of Photogrammetry & Remote Sensing*, 70(2):56–65, 2012.

[Mansour *et al.*, 2016] Khalid Mansour, Onisimo Mutanga, Elhadi Adam, and Elfatih M. Abdel-Rahman. Multispectral remote sensing for mapping grassland degradation using the key indicators of grass species and edaphic factors. *Geocarto International*, 31(5):477–491, 2016.

[Menze and Geiger, 2015] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, pages 3061–3070, 2015.

[Mottaghi *et al.*, 2014] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam Gyu Cho, Seong Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pages 891–898, 2014.

[Rennie and Srebro, 2005] Jason DM Rennie and Nathan Srebro. Loss functions for preference levels: Regression with discrete ordered labels. In *IJCAI workshop*, pages 180–186, 2005.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[Ros *et al.*, 2016] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, pages 3234–3243, 2016.

[Russell *et al.*, 2008] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[Wang *et al.*, 2018] Yuwei Wang, Zhenyu Wang, Ruren Li, Xiaoliang Meng, Xingjun Ju, Yuguo Zhao, and Zongyao Sha. Comparison of modeling grassland degradation with and without considering localized spatial associations in vegetation changing patterns. *Sustainability*, 10(2):316–, 2018.

[Zhan *et al.*, 2017] Wang Zhan, Xiangzheng Deng, Song Wei, Zhihui Li, and Jiancheng Chen. What is the main cause of grassland degradation? a case study of grassland ecosystem service in the middle-south inner mongolia. *Catena*, 150:100–107, 2017.

[Zhang *et al.*, 2017] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason Mcgough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *CVPR*, pages 3549–3557, 2017.

[Zhao *et al.*, 2004] Chengzhang Zhao, Shengyue Pan, Cuiqin Yin, and Xuebin He. Study on vegetation community's structure of degraded grassland of noxious and miscellaneous grass type. *Journal of Desert Research*, 24(4):507–511, 2004.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 6230–6239, 2017.