BOUNDARY-AWARE BIAS LOSS FOR TRANSFORMER-BASED AERIAL IMAGE SEGMENTATION MODEL

Yan Zhang[1], *Xue Jiang*[1], *Siqi Liu*[2], *Bo Hu*[1], *Xinbo Gao*[1*]

Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications[1],
School of Software & Microelectronics, Peking University[2]

ABSTRACT

Inspired by the tremendous success of the transformer-based model in natural language processing (NLP), many efforts introduce the transformer-based model into the image processing tasks. However, naive transformer models have to downsample the image resolution to satisfy computational restrictions, thus discarding the local information, which is catastrophic for high-performance remote sensing image segmentation. Hence, this paper proposes a novel trainable boundary-aware bias loss function to enhance transformer-based models of extracting local information. On the Challenging ISPRS Potsdam dataset, two representative transformer-based models achieve remarkable performance improvements, proving the effectiveness of the proposed method.

Index Terms— Aerial Image Segmentation, Loss Function, Transformer, Deep Learning, Remote Sensing

1. INTRODUCTION

Aerial image segmentation aims to accurately label each pixel with its corresponding class, which is critical for aerial image understanding. As a downstream task of semantic segmentation, the traditional Fully Convolutional Network (FCN)[1] and its mutations are widely migrated into the aerial image segmentation task. Extending with FCN, UNet [2] and Deeplab[3]achieves much better performance on estimating the boundary area.

However, the intrinsic defect of convolution layer prevents the CNN-based model from modeling global information, seriously affecting the segmentation accuracy. Recently, the transformer model[4] has drawn significant attention from the image processing community due to its effectiveness in modeling global information. Moreover, simply replacing the CNN backbone with a vanilla ViT[5] achieves state-of-the-art performance on several computer vision tasks. Inspired by the advantages of transformers, Liu et.al.[6] proposes a windowed self-attention to enhance performance on semantic segmentation tasks. However, the aerial image segmentation task does not directly benefit from the transformer based model due to the pool performance on the aerial image's boundary area. This is because, firstly, the patching operation in a transformer module drastically decreases the resolution of the input feature, inevitably discarding the local information and vastly deteriorating the segmentation performance at the boundary area. Secondly, the self-attention mechanism within the transformer is position-agnostic, and therefore the additional manual position information imposes the adaptation issues during the feature learning process. Islam et.al.[7] indicate that the CNN model can implicitly learn the position information, partially explaining why some transformer-based models[8] still achieve acceptable results on computer vision tasks without explicitly considering position embedding.

At the same time, many researchers make great efforts to design advanced transformer architectures, the loss function for the transformer-based model has been given less attention. cross-entropy[9] is still the most common loss function for aerial image segmentation. To improve the model's boundary area performance, Michael et.al.[10] proposed a weighted mechanism employing various loss functions and forces the model to learn more local information. Furthermore, chen et.al.[11] suggest a Contour loss to leverage the boundary and object area imbalance. Finally, inspired by the perceptive loss[12], chen et.al.developed the Semantic Edge-Aware strategy (SEMEDA)[13] loss to measure the structural feature difference. Nevertheless, current loss functions are optimized for CNN-based models and thus do not pose an appealing option for transformer-based models due to the transformer's characteristics. Therefore, we propose a boundary-aware bias loss (BB loss) designed explicitly for transformer-based models, which comprises a trainable boundary-aware loss (TB loss) and a class-aware bias loss (CB loss). The TB loss explicitly teaches the model with the position information, while the CB loss power the transformer-based model with class-aware bias inductivity. The experiment results conducted on the challenging ISPRS Potsdam Dataset demonstrate the effectiveness of the proposed BB loss.

THIS WORK WAS SUPPORTED IN PART BY THE NATIONAL NATURAL SCIENCE FOUNDATION OF CHINA UNDER GRANTS 62101084, 62171340, AND 62036007, IN PART BY THE CHONGQING NATURAL SCIENCE FOUNDATION OF CHINA UNDER GRANTS CSTC2021JCYJ-MSXMX0847

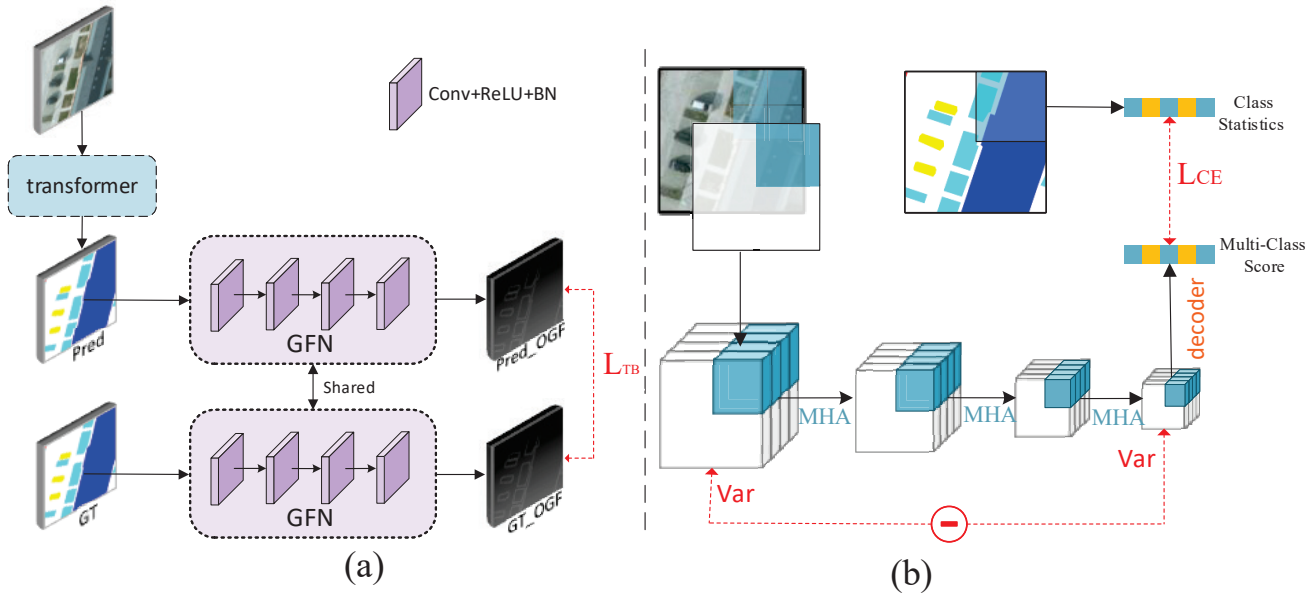

Fig. 1. Overview of the BB loss. (a) and (b) describe the workflows of TB loss and CB loss, respectively.

2. BOUNDARY-AWARE BIAS LOSS

2.1. Problem formulation

This section analyzes the current problems of transformer-based segmentation models under the scope of aerial image segmentation tasks. Compared with the CNN-based model, transformer uses the self-attention mechanism to extract features from aerial images, the computation of self-attention is computed as Eq.1.

$$MHA(X) = Q(x) \otimes K(x) \otimes V(x) \tag{1}$$

One critical problem of the self-attention mechanism is position-agnostic, this character makes the transformer-based model cannot model the position dependence, which is a key factor for extracting high-quality semantic information. Also, compared with a convolutional layer, the computational cost of a multi-head attention layer within a transformer model is much more expensive. Hence, another problem is that the transformer-based model has to down-sample the input feature map into several lower-resolution patches, which weakens the local information of input aerial image and causes boundary mislabeled and blurring predictions. To overcome both problems, we propose the boundary-aware bias loss (BB loss) for transformer-based aerial image segmentation models. As presented in Eq.2, the BB loss composes a trainable boundary-aware loss (TB loss) and a class-aware bias loss (CB loss). The corresponding details of these two loss functions are elaborated in later sections.

$$L_{BB} = L_{TB} + L_{CB} \tag{2}$$

2.2. Trainable Boundary-aware Loss Function

For enhancing the transformer-based model with position information and local information, the latest approaches[14] propose mixed architectures to compensate the transformer with a CNN-based model for extracting high-quality features. Opposing these trends, the proposed trainable boundary-aware loss function is designed to train the transformer models utilizing CNN models. Extending our earlier work [15], the TB loss contains a Gradient Field Net (GFN) and metric functions. The entire framework of applying the TB loss is similar to generative adversarial learning, with the segmentation model (segmentor) taking the generator's role and GFN the discriminator's. As illustrated in Fig. 1 (a), the GFN involves a simple 4-layer CNN architecture with an activation function and is pre-trained on the task that transforming the segmentation mask to the Oientational Gradient Field (OGF). The orientational gradient field is computed as Fig.2, the position embedding is calculated through Eq. 3, where v, h indicates the position coordination. The OGF is compute as Eq.4, where C indicates the Canny boundary detection method, and PE is the position embedding map. α is the balanced parameter and is set as 0.5 in this paper.

$$PE(v,h) = Sin(\frac{v}{2000}) + Cos(\frac{h}{2000}) \tag{3}$$

$$OGF = C(I) + \alpha PE \tag{4}$$

When training the transformer-based with TB loss, the prediction and ground truth are independently fed into the GFN. As shown in Eq.5, The TB loss is used to measure the differences of the hierarchical feature maps by Jaccard Coefficient. Where i indicates the index of layer in the GFN, θ and

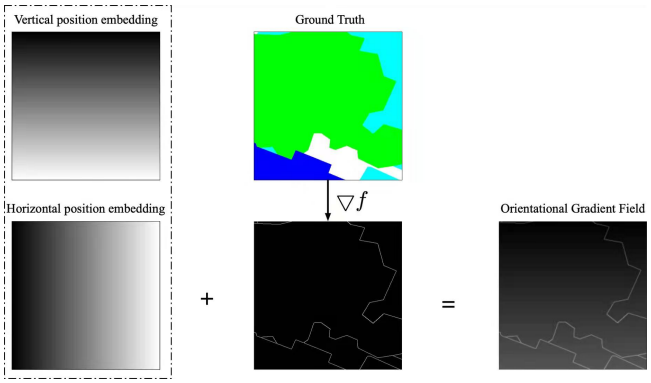

Fig. 2. The diagram of generate orientational gradient filed.

$\hat{\theta}$ are the mediate feature maps with the input of prediction and ground truth.

$$TB_{loss} = \sum_{i} Jaccard(\theta^{(i)}, \hat{\theta}^{(i)}) \tag{5}$$

It should be mentioned that the parameter of GFN is fixed during training the transformer-based model. By transforming the segmentation map into the orientational gradient field map, the GFN learns the structural local and position information. This information can be embedded into the transformer-based segmentation model with the standard optimization algorithm and improve the performance of the transformer-based segmentation model on the boundary areas.

2.3. Class-aware Bias Loss Function

TB loss can significantly improve the performance of the transformer-based model at the boundary area but is weak in extracting tiny objects, since the boundary area of tiny objects is negligible. Also, the self-uses cross product to model the long-range dependencies from query Q, key K, and value V. This operation effectively enlarges the receptive field at the expense of diminishing the ability to conduct bias. To solve that, we design a CB loss to enhance the transformer-based model with bias conduction ability. The proposed CB loss is design to optimize the transformer-based backbone. Assuming that the mediate feature map of the transformer backbone is θ(n), the CB loss is calculated as Eq.6.

$$L_{CB} = \sum_{i=1}^{N} L_{CE}(r_i, \hat{Y}_i) + |Var(\theta^{(4)}) - Var(\theta^{(1)})| \tag{6}$$

The first part of CB loss is an auxiliary multi-label cross-entropy loss. *N* is the number of the feature map pixel on the last level. The bottom feature map is feed into an extra 1×1 convolutional layer and gets a multi-class prediction *r*. The channel of *r* is the same as the default class number, and *i* indicates the index of the pixels on the bottom feature map. As shown in Fig. 1(b), y is the multi-class label transformed from the segmentation ground truth with the corresponding stride. For a general 4-layer transformer-based model, the stride is equal to 16. In this case, y is the multi-class label with patching size 16. In other word, the first item of CB loss is a patched multi-class classification loss function, which helps the transformer-based backbone learn the inductive bias of the input images. Also, local information is dropped rapidly as the backbone depth increases due to the nature of the self-attention. Therefore, we design the second part of the proposed CB loss, which applies the hierarchical feature map's variance to relieve local information loss. Just like literature[16] points out that the variance is an efficient but straightforward metric to measure the local context of feature maps. Hence this work utilizes the absolute variance difference between the first and the last feature map as a loss function to encourage the backbone model to learn local information and thereby has a better performance on small object segmentation.

Combining with the TB loss and CB loss, the proposed BB loss is extremely easy-to-use and flexible. When training a segmentor from scratch, the BB loss utilizes standard cross-entropy with an experimentally defined ratio of 0.4. Also, fine-tuning models with BB loss require only a few iterations to achieve a remarkable performance boost. The efficiency and effectiveness of BB loss are discussed in the next section.

Table 1. **Ablation and adaptation study Potsdam Dataset.**

Cross-Entropy	√	√	√	√	√	√
counter loss		√				
SEMEDA			√			
jaccard high loss				√	√	√
TB loss					√	√
CB loss						√
mIoU(%)	74.04	73.78	74.20	74.38	75.02	75.33

3. EXPERIMENTS

3.1. Dataset and Metrics

Potsdam is a typical historical city with large buildings, narrow streets, and other typical remote sensing objects. The ISPRS Potsdam Semantic Labeling dataset has been manually classified aerial images from Potsdam into five land cover categories, impervious surface, building, low vegetation, trees, cars. Every high-resolution aerial image (6000 × 6000 pixels) is sliced into 144 sub-images of resolution 500 × 500 pixels, and we use random flip and rotation to augment the dataset.The first 24 aerial images are exploited for training the model and the remaining 14 for testing it.

Following the commonly accepted evaluation strategy, we employ the mean intersection over union (mIoU) and the overall accuracy(OA)to evaluate the performance of the segmentation model. mIoU computes the average ratio between the intersection and the union of two sets (ground truth and

Table 2. Performances of SOTA Models with and without BB loss on the Potsdam Dataset.

Method	Loss	Imp.surf	Building	Low veg.	Tree	Car	mIoU	OA
SERT	CE	83.09	90.96	73.03	74.57	81.96	80.72	89.64
SERT	CE+BB	84.6	92.2	74.83	76.55	**83.5**	82.336	90.43
Segformer	CE	84.5	92.34	74.91	76.3	81.45	81.9	90.1
Segformer	CE+BB	85.9	93.1	75.92	**77.1**	82.14	82.83	91.17
Swin-Unet	CE	84.65	92.62	74.76	75.86	82.33	82.04	90.2
Swin-Unet	CE+BB	**86.54**	**94.01**	**76.2**	77.08	83.2	**83.41**	**91.51**
conformer-Unet	CE	84.5	92.34	74.91	76.3	81.45	81.9	90.1
conformer-Unet	CE+BB	85.9	93.1	75.92	77.1	82.14	82.83	91.17

*The best entire in each column is marked in bold.

prediction). OA is the ratio of correctly predicted pixels and total number of them. For a fair comparison, all models are trained by a stochastic gradient descent optimizer (SGD) with a fixed learning rate of 1e-3.

3.2. Ablation and Adaptation Study

This section, studies the effectiveness of the proposed BB Loss and its adaptation with other commonly used pixel-wise loss functions. We use the FCN architecture as the baseline with the backbone of ViT. For a fair comparison, the ablation experiments are conducted on the ISPRS Potsdam dataset with the corresponding experimental results listed in Table 1. Due to the excellent feature parsing afforded by the vision transformer backbone, the naive FCN model can easily achieve 76.52% IoU on the challenging ISPRS Potsdam dataset. Training together with Dice loss and Jaccard hinge loss, a further 0.78% performance boost can be observed on the FCN model. Considering the proposed TB Loss, rarely 1% mIoU performance improvement is obtained on theFCN model. Finally, combining the FCN model with the TB loss and the CB loss achieves the best results surpassing the naive FCN model, with a 1.67% mIoU score. Also, Fig. 3 highlights that model trained with TB loss manages less blur and presents sharper boundary predictions. Moreover, the proposed CB loss significantly relieves the mislabeled issue in small areas.

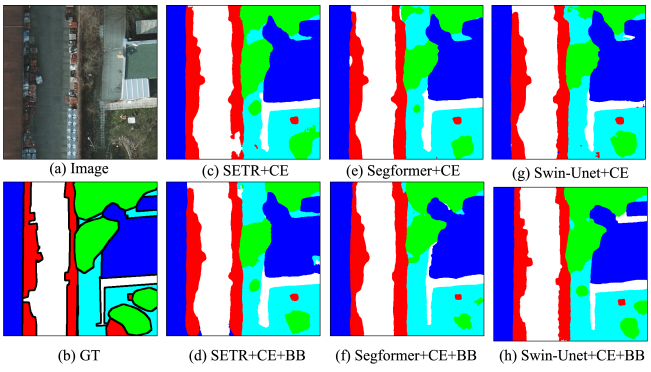

Fig. 3. Visualization Comparisons on ISPRS Potsdam Dataset.

3.3. SOTA Comparisons

We also challenge the proposed transformer-based backbone model against several typical aerial image segmentation models on the ISPRS Potsdam dataset, with the corresponding results presented in Table 2. For fair comparison, all backbones networks are pre-trained on ImageNet[17]. From table 2, various levels of improvements are observed from the models trained with the proposed BB loss. With the backbone of ViT, SETR model rarely attains an additional 1.6% mIoU boosts trained by the proposed method. This is mainly because the decoder of SETR is a simple 2-layer CNN, and therefore, the proposed BB loss forces the transformer-based encoder to extract features containing more local information. With more complex architecture, the Segformer trained by CE loss achieves 81.9% on the mIoU score. Training the Segformer with the proposed BB loss can get further 0.9% mIoU increasement. Benefit the great ability of swin transformer and encoder-decoder architecture, the very recent Swin-Unet achieve 82.04 mIoU without any tricks. With the proposed BB loss, Swin-Unet reports a new SOTA, 83.41% mIoU on Potsdam dataset. It should be noted that due to our method being restricted by the available computer resources, the potential of the proposed BB loss has not been fully revealed, as performance improvement may be obtained through more iterations, advanced optimizer schemes, and ingenious learning rate schedule.

4. CONCLUSION

This paper presents a boundary-aware bias loss function that is appropriate for transformer-based aerial image segmentation models. The proposed BB loss integrates a trainable boundary-aware loss function and a class-aware bias loss function, assisting the model to efficiently characterize the local contextual and discriminative bias features in aerial images. Experimental results indicate that the proposed BB loss can effectively improve the segmentation performance of transformer-based model.

5. REFERENCES

[1] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] 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.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[7] Md Amirul Islam, Sen Jia, and Neil DB Bruce, "How much position information do convolutional neural networks encode?," *arXiv preprint arXiv:2001.08248*, 2020.

[8] Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei, "Contextual transformer networks for visual recognition," *arXiv preprint arXiv:2107.12292*, 2021.

[9] Rasha Alshehhi, Prashanth Reddy Marpu, Wei Lee Woon, and Mauro Dalla Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 139–149, 2017.

[10] Michael Kampffmeyer, Arnt-Borre Salberg, and Robert Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 1–9.

[11] Zixuan Chen, Huajun Zhou, Xiaohua Xie, and Jianhuang Lai, "Contour loss: Boundary-aware learning for salient object segmentation," *arXiv preprint arXiv:1908.01975*, 2019.

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[13] Yifu Chen, Arnaud Dapogny, and Matthieu Cord, "Semeda: Enhancing segmentation precision with semantic edge aware loss," *Pattern Recognition*, vol. 108, pp. 107557, 2020.

[14] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye, "Conformer: Local features coupling global representations for visual recognition," *arXiv preprint arXiv:2105.03889*, 2021.

[15] Yan Zhang, Weiguo Gong, Jingxi Sun, and Weihong Li, "Web-net: A novel nest networks with ultra-hierarchical sampling for building extraction from aerial imageries," *Remote Sensing*, vol. 11, no. 16, pp. 1897, 2019.

[16] Zhang Dong, Zhang Hanwang, Tang Jinhui, Hua Xiansheng, and Sun Qianru, "Self-regulation for semantic segmentation," *arXiv preprint arXiv:2108.09702*, 2021.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.