ICCEPM 2022
The 9th International Conference on Construction Engineering and Project Management
Jun. 20-23, 2022, Las Vegas, NV, USA

Compound Loss Function of semantic segmentation models for imbalanced construction data

Wei-Chih Chern[1]*, Hongjo Kim[2], Vijayan Asari[3], and Tam Nguyen[4]

[1] *Department of Electrical and Computer Engineering, University of Dayton, 300 College Park Ave, Dayton, OH 45469, USA,* E-mail address: chernw1@udayton.edu
[2] *Department of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea,* E-mail address: hongjo@yonsei.ac.kr
[3] *Department of Electrical and Computer Engineering, University of Dayton, 300 College Park Ave, Dayton, OH 45469, USA,* E-mail address: vasari1@udayton.edu
[4] *Department of Computer Science, University of Dayton, 300 College Park Ave, Dayton, OH 45469, USA,* E-mail address: tnguyen1@udayton.edu

Abstract: This study presents the problems of data imbalance, varying difficulties across target objects, and small objects in construction object segmentation for far-field monitoring and utilize compound loss functions to address it. Construction site scenes of assembling scaffolds were analyzed to test the effectiveness of compound loss functions for five construction object classes---workers, hardhats, harnesses, straps, hooks. The challenging problem was mitigated by employing a focal and Jaccard loss terms in the original loss function of LinkNet segmentation model. The findings indicates the importance of the loss function design for model performance on construction site scenes for far-field monitoring.

Key words: Data Imbalance, Semantic Segmentation, Compound Loss Function, Safety Monitoring

1. INTRODUCTION

Training data is essential to train deep learning models in supervised learning for desirable performance. However, it is difficult to collect the same quantity of training samples for target classes. An imbalanced dataset can hurt the model’s performance to the classes with less instances/pixels. Moreover, an object class with more complicated and irregular shapes is hard for a model to learn its invariant representations in comparison to objects with more consistent shapes. Lastly, smaller objects are also challenging to recognize due to the visual feature disappearance for convolutional neural networks for far-field monitoring.

Recognizing workers and their personal protective equipment (PPE) is the most important step to build up a robust safety monitoring system. However, varying object sizes of PPE increase the difficulty for convolutional neural networks (CNNs) to have consistent performance across different object classes of PPE. This problem can be exacerbated if the dataset is imbalanced among target object classes. Previous studies have utilized detection [1, 2] and instance

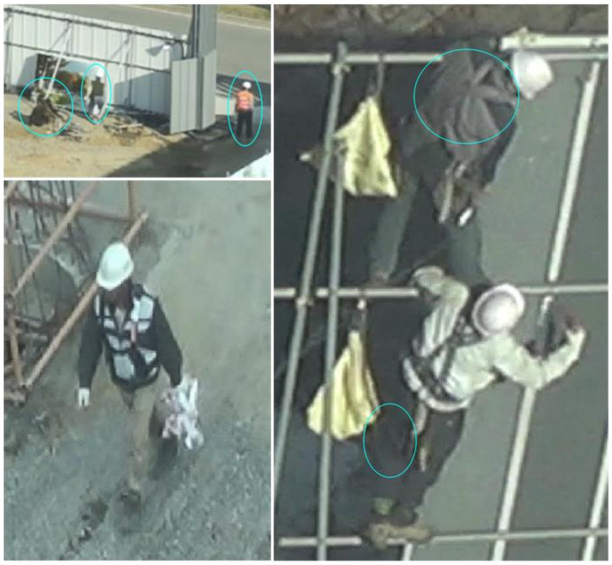

Figure 1. Example of collected YUD-COSA images. The varying object sizes, color similarity, and object thickness create difficulties in workers and PPE recognition for far-field safety monitoring. Limited optimization in loss functions can lead to unreliable monitoring.

segmentation models [3, 4] to recognize PPE and workers in images. However, there has been a lack of detailed investigation on addressing semantic segmentation performance with imbalanced datasets, varying size objects, and small objects for far-field monitoring setups as shown in Figure 1. Although many studies [5, 6] address challenges to safety monitoring and small objects detection, the effectiveness of the methods for semantic segmentation has remained unknown. The goal of this study is to address the challenges from the optimization perspective with a new loss function design to semantic segmentation models for construction site safety monitoring.

For experiments, a new dataset was collected from construction site scenes where workers are installing scaffolds, fences, and more. The dataset is named YUD-COSA and has five object classes including workers, hardhats, safety harnesses, safety straps, and safety hooks. As shown in Figure 1, YUD-COSA contains various distances of worker images that are used for training, and the testing sets contains one scaffold scene which is used to test the performance on recognizing the personal protective equipment. YUD-COSA contains of a total of 695 training images and 107 testing images for which the images were recorded by camcorders placed at height with tilted top down angles from 10 different scenes at Yonsei University. It was found that a compound loss function composed of Jaccard and focal losses for LinkNet can improve the performance of smaller and more challenging objects. Additionally, the CamVid (Cambridge-driving Labeled Video Database) dataset was also experimented to illustrate the generality of compound loss functions for other applications. The experiment structure is shown in Figure 2.

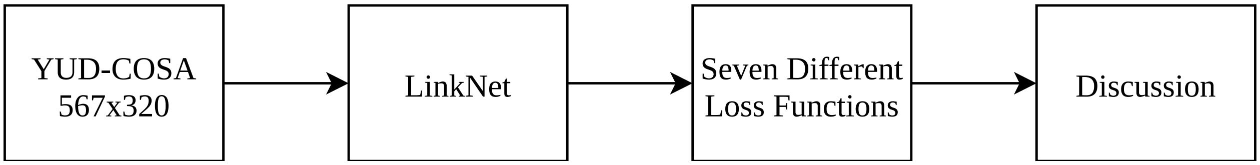

Figure 2. Overview of the method

2. IMPORTANCE OF LOSS FUNCTIONS

Loss functions play a significant role in optimization for modern deep learning. Using an inappropriate loss function would result in poor overall performance. In this study, different type of loss functions were experimented, which are cross entropy loss, focal loss [7], dice loss [8, 9], and Jaccard loss [10]. Cross entropy loss and focal loss are distribution-based loss (DBL) functions in which the overall loss value is accumulated from each pixel. A disadvantage of DBL is being sensitive to the distribution of the dataset. DBL can be easily affected by data imbalance that classes with more instances or pixels can dominate and deteriorate the training performance. On the other hand, dice loss and Jaccard loss are region-based loss (RBL) functions in which the loss value is calculated based on the degree of overlap between the ground truth mask and the predicted mask. Thus, RBL has better ability in working with data imbalance. The equations for the experimented loss functions are as shown below:

$$\text{CrossEntropyLoss}(p_t) = -\log(p_t), \tag{1}$$

$$\text{FocalLoss}(p_t) = -(1 - p_t)^{\gamma} \log(p_t), \tag{2}$$

$$\text{DiceLoss}(g_t, p_t) = 1 - \frac{2(g_t \cap p_t)}{g_t + p_t}, \tag{3}$$

$$\text{JaccardLoss}(g_t, p_t) = 1 - \frac{g_t \cup p_t}{g_t \cap p_t}, \tag{4}$$

Data imbalance between target object classes creates a situation that certain classes were trained and learned more frequently than the others. As the result, the model would learn and perform better on the classes with more instances/pixels and the opposite for the other classes. As shown in Table 1, YUD-COSA dataset also presents the data imbalance, where the worker class dominates in both the number of pixels and instances. The most extreme difference in YUD-COSA is between the worker and the safety strap, where the worker class is x36 more in pixels and x3.75 more in instances than the safety strap class.

Table 1. The statistics of YUD-COSA dataset.

Object Class	Total Pixels	Total Instances
Background	1,258,247,279	-
Worker	18,413,625	2,648
Safety Strap	510,487	706
Safety Hardhat	1,669,464	2,426
Safety Harness	1,268,090	1,021
Safety Hook	266,095	1,190

3. EXPERIMENT AND RESULTS

The proposed method was evaluated using mean intersection over union (mIoU) as shown in Equation 5. The input resolution for LinkNet was 567x320. The LinkNet model pre-trained with ImageNet was used for transfer learning. Among seven different loss combinations experimented, Jaccard loss + focal loss scored the highest mIoU score of 0.748, as shown in Table 2. Each pixel is assigned to an object index with highest confident scores using arguments of maxima (argmax) from the softmax function to obtain the segmentation result. This loss combination scored the highest on two out of three harder objects, which are safety strap and safety harness. The samples of the segmentation results are shown in Figure 3.

Table 2. Per class IoU scores on YUD-COSA. The best performance of each class is highlighted in **boldface**.

Loss Function	Background	Worker	Strap	Hardhat	Harness	Hook
Focal	99.46%	77.95%	17.20%	82.19%	42.91%	44.57%
Cross Entropy	99.42%	77.08%	11.79%	82.90%	42.92%	40.51%
Cross Entropy+Focal	99.51%	79.17%	23.87%	80.97%	47.80%	41.42%
Dice	**99.60%**	80.47%	45.00%	86.73%	61.36%	**64.18%**
Dice+Focal	99.50%	79.65%	49.24%	87.97%	60.26%	63.99%
Jaccard	99.50%	79.90%	46.30%	**88.02%**	62.76%	61.82%
Jaccard+Focal	99.53%	**80.87%**	**51.21%**	87.90%	**63.25%**	63.98%

Table 3. Mean IoU scores of each loss function setup. The performance of combination loss is marked in **boldface**.

Data	Focal	Cross Entropy	Cross.+ Focal	Dice	Dice+Focal	Jaccard	J.+Focal
YUD-COSA	60.71%	59.10%	**61.7%**	72.9%	**73.46%**	73.06%	**74.80%**
CamVid	56.5%	59.85%	**61.45%**	61.19%	**63.31%**	61.21%	**63.89%**

$$mIoU = \sum_{n=1}^{C} \frac{IoU_n}{C}, \qquad (5)$$

To further validate the performance improvement from the compound loss function, Cambridge-Driving Labeled Video Database [11] (CamVid) was also used for the experiment. As shown in Table 3, both YUD-COSA and CamVid scored better when combining focal loss and region-based loss. For YUD-COSA dataset, when adding focal loss with Jaccard loss, the hardest object safety strap scored 10.6% higher in comparison to using Jaccard loss only.

4. CONCLUSION

This study presents a compound loss function by combining two different types of loss functions together, which improved mIoU score without complicated pre- or post- processing. It was found that region-based Jaccard loss function and Dice loss function are good at handling imbalanced construction data, and focal loss is good at shifting learning focus to not-well-learned classes during training. As shown in Table 3, the compound loss function of Jaccard loss and focal loss delivers a 10 percent improvement in IoU score to safety straps, which is considered as the most challenging object class in the YUD-COSA due to the irregular shapes and thinness. On the other hand, experimented distribution-based loss functions can be easily affected by object classes with more instances such as workers and hardhat, resulting in poor performance to objects which are harder and have less instances as shown in Table 1.

Improvements to the overall IoU scores were discovered on the CamVid dataset as well. Adding the focal loss function increased the mIoU score by 2.68% and 2.12% respectively to Jaccard and Dice loss. With CamVid dataset, the compound loss functions are proven to shift the focus toward not well-learned classes during training. In addition, the discovery is not limited to the construction-

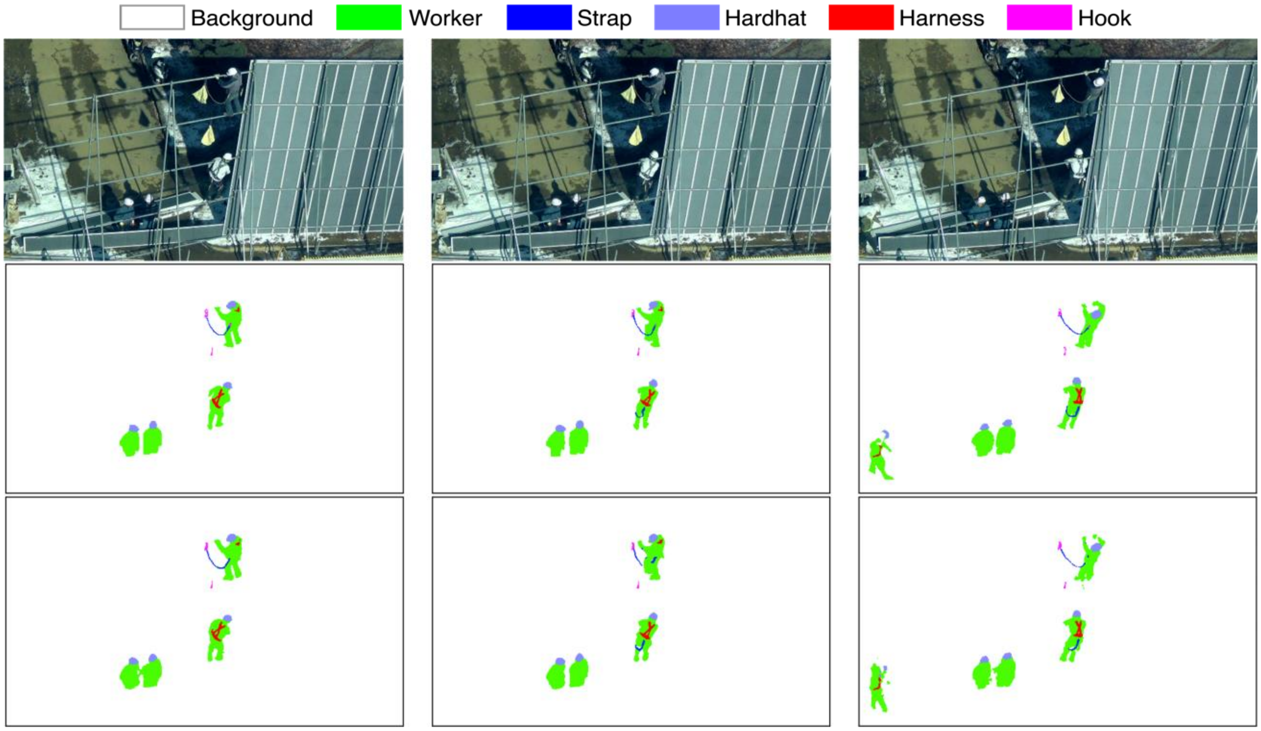

Figure 3. The segmentation results with Jaccard loss and focal loss in construction site scenes. The inference results demonstrate the robustness of the compound loss function. From top to bottom, the first row: three different testing images, the second row: the ground truth of the three testing images, the third row: the prediction results from the combination of Jaccard loss and focal loss.

related datasets. The concept of compound loss functions encourages the identification of issues and challenges in varied datasets and applications for better optimization results for deep learning models. The compound loss design could be further investigated in future study to minimize the performance gaps between easier and harder objects.

ACKNOWLEGEMENTS

This research was conducted with the support of the “2021 Yonsei University Future-Leading Research Initiative (No.2021-22-0037)” and the "National RD Project for Smart Construction Technology (No.22SMIP-A158708-03)" funded by the Korea Agency for Infrastructure Technology Advancement under the Ministry of Land, Infrastructure and Transport, and managed by the Korea Expressway Corporation.

REFERENCES

[1] Nipun D. Nath, Amir H. Behzadan, and Stephanie G. Paal. Deep learning for site safety: Real-time detection of personal protective equipment. Automation in Construction, 112:103085, 2020.
[2] Jixiu Wu, Nian Cai, Wenjie Chen, Huiheng Wang, and Guotian Wang. Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset. Automation in Construction, 106, 2019. ISSN 0926-5805.
[3] T. Truong, A. Bhatt, L. Queiroz, K. Lai, and S. Yanushkevich. Instance segmentation of personal protective equipment using a multi-stage transfer learning process. In 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 1181–1186, 2020.

[4] An Xuehui, Zhou Li, Liu Zuguang, Wang Chengzhi, Li Pengfei, and Li Zhiwei. Dataset and benchmark for detecting moving objects in construction sites. Automation in Construction, 122, 2021. ISSN 09265805.

[5] Lim, J.-S., Astrid, M., Yoon, H.-J., & Lee, S.-I. (2021). Small object detection using context and attention. In 2021 international conference on artificial intelligence in information and communication (icaiic) (p. 181-186).

[6] Luo, H., Liu, J., Fang, W., Love, P. E. D., Yu, Q., & Lu, Z. (2020). Real-time smart video surveillance to manage 19 safety: A case study of a transport mega-project. Advanced Engineering Informatics, 45, 101100.

[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection, 2018.

[8] Thorvald Julius Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biologiske skrifter / Kongelige Danske videnskabernes selskab: bd. 5, nr. 4. I kommission hos E. Munksgaard, 1948. ISBN 0366-3312.

[9] FaustoMilletari,NassirNavab,andSeyed-AhmadAhmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016.

[10] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin del la Société Vaudoise des Sciences Naturelles, 37: 547–579, 1901.

[11] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A highdefinition ground truth database. Pattern Recognition Letters, 2008.