Weakly supervised pseudo label generation for construction vehicle segmentation

W.-C. Chern[1], V. Asari[1], T. Nguyen[2], and H. Kim[3]

[1]Department of Electrical and Computer Engineering, University of Dayton, U.S.A
[2]Department of Computer Science, University of Dayton, U.S.A
[3]Department of Civil and Environmental Engineering, Yonsei University, Republic of Korea

chernw1@udayton.edu, vasari1@udayton.edu, tnguyen1@udayton.edu, hongjo@yonsei.ac.kr,

Abstract -

Segmentation tasks in computer vision have been adopted in various studies in the civil engineering domain to provide accurate object locations in images. However, preparing annotation to train segmentation models is a time consuming and costly process, which hinders the use of segmentation models in vision-based applications. To address the problem, this study proposes a fusion model integrating self-supervised equivariant attention mechanism (SEAM) and sub-category exploration (SC-CAM) to generate pseudo labels in the form of polygon annotation from bounding box annotation that is relatively easy to obtain. To test the performance of the fusion model, a public data set—Advanced Infrastructure Management Group (AIM) dataset—for construction object detection was selected to generate pseudo labels; the effectiveness of pseudo labels was measured by the segmentation performance of a feature pyramid network (FPN) trained with the pseudo labels. FPN showed the mean intersection over union (mIoU) score of 86.03%, demonstrating the potential of the proposed fusion model to reduce the manual annotation efforts in preparing training data for segmentation models.

Keywords -

Weakly supervised learning; Semantic segmentation; Pseudo labels; Training data preparation

1 Introduction

Semantic segmentation is an important task in vision-based analysis, providing pixel-level annotation to represent exact object boundaries. There are various applications of segmentation tasks in civil engineering domain such as construction site monitoring and infrastructure damage assessment.

The quantity and quality of training data greatly affects the performance of semantic segmentation models. The time spent for annotation shows a positive correlation to semantic segmentation's performance [1]. This is a general phenomenon in deep learning applications which require a lot of training data to acquire better performance and generalization capability. However, polygon annotation for semantic segmentation can be overly time consuming in comparison to preparing annotation for object detection and classification tasks. To minimize the time and cost of annotation, weakly supervised learning can be used. In weakly supervised learning, imperfect data are used to train segmentation models. For example, segmentation models can be trained with bounding box annotation by treating the entire box as a target class region. Likewise, segmentation masks of an object of interest can be generated from images with class activation mapping (CAM) to generate pseudo labels for training segmentation models.

CAM often represents a discriminative part of the object only, instead of localizing the entire object, resulting in producing incomplete pseudo labels. Previous studies have utilized consistency regularization on CAMs [2], and a self-supervised task [3] to generate accurate pseudo labels for the Pascal Visual Object Classes Challenges 2012 (VOC2012) [4]. Nevertheless, there has been a lack of detailed investigation on difficult datasets in which the target classes have visual similarity. VOC2012 contains 20 distinctive object classes from animals, transportation vehicles, furniture, and etc. Thus, it presents clear visual differences between target classes for pseudo label generation. On the other hand, a dataset contains 20 different vehicle brands can be more challenging to generate pseudo labels covering entire vehicles. Because models tend to focus on the most distinctive part of vehicles such as the logo of vehicles rather than captures the entire body. This issue can also be found in the AIM dataset having construction machines only. To address the challenge, this study propose a novel architecture integrating SEAM and SC-CAM features to generate accurate pseudo labels.

For experiment, the AIM dataset is used, which contains five object classes—dump trucks, excavators, loaders, mixer trucks, and rollers—with a total of 2,721 images and 2,873 vehicle instances as shown in Table 1. Examples of the AIM dataset is shown in Figure 1. AIM dataset's construction vehicle classes share a considerable degree of visual similarity. For example, wheels and front windows between mixer trucks and dump trucks look similar;

Figure 1. Examples of AIM dataset. The construction vehicles from top to bottom and left to right are: loader, mixertruck, dumptuck, excavator, and roller.

Table 1. AIM Dataset Statistics.

Vehicle Types	Number of Instances
Dumptruck	762
Excavator	413
Loader	714
Mixertruck	632
Roller	352
Sum	2,873

wheels between loaders and excavators show similar appearances. Construction vehicles from the AIM dataset also possess highly distinct features, such as the boom of excavators, the drum of rollers, the mixing drum of mixer trucks, the dump box of dump trucks, and the bucket of loaders, which can lead to over-attention to the distinctive part in CAM. This can result in low-quality pseudo labels for semantic segmentation as the trained classification models may not learn the entire appearance of objects for classification. Instead, the models only pay attention to the most discriminative features from the target objects, resulting in poor pseudo labels which do not cover entire target objects. It was found that the fusion model of self-supervised equivariant attention mechanism (SEAM) [2] and weakly-supervised semantic segmentation via subcategory exploration (SC-CAM) [3] with random erase data augmentation can mitigate this issue to improve the quality of the pseudo labels. The experimental results showed that the feature pyramid network (FPN) [5] for semantic segmentation model recorded the mIoU score of 86.03%.

This study propose a novel model combining two different architectures (SEAM & SC-CAM) along with the data augmentation of random erase [6] to improve the quality of the pseudo labels as shown in Figure 2.

2 Methodology

The proposed fusion model incorporates methods that encourage models to pay attention to more foreground regions for better pseudo labels from SEAM and SC-CAM as shown in Figure 2. SEAM architecture modifies the self-attention module for weakly supervised pseudo labels generation. SC-CAM approaches the same goal by asking models to distinguish differences within the same class, asking models to look at the entire objects for sub-category classification. A common way to generate pseudo labels for semantic segmentation is to use classification models and their CAM [7]. CAM visualizes the attentions of classification models, and it highlights the visual features contributed to classification. However, CAM itself is hardly used as training data for semantic segmentation as the pseudo labels usually cover the most discriminative features only as shown in Figure 3. mIoU is used as the performance metric to evaluate the performance of segmentation models as it measures the degree of overlap between ground truth and predicted masks across all target classes. It is formulated as follows:

$$IoU = \frac{g_t \cap p_t}{g_t \cup p_t}, \tag{1}$$

$$mIoU = \sum_{n=1}^{C} \frac{IoU_n}{C}, \tag{2}$$

where g_t represents ground truth masks, p_t represents predicted masks, and mIoU is an average of IoU scores from all target classes.

2.1 Self-supervised Equivariant Attention Mechanism

SEAM architecture proposed a pixel correlation module (PCM) and a smaller-scaled branch siamese network to teach classification models to pay more attention to the entire region of target objects. PCM adopts the concept of self-attention [8] to extract contextual information. It takes features maps from two convolutional blocks and the RGB image to form the self-attention map that is then used to correct original CAM. The formula of PCM can be formulated as:

$$y'_i = ReLU\left(\frac{\sum_{\forall j} e^{\theta(x_i)^\top \theta(x_j)} \cdot \phi(y_j)}{\sum_{\forall j} e^{\theta(x_i)^\top \theta(x_j)}}\right), \tag{3}$$

where x represents the feature maps input, y' represents the refined CAM, y represents the original CAM, θ, and ϕ are two 1x1 convolutions as embedding functions. The refined CAM further normalized by the sum of the attention map and a ReLU function.

Wang et al. [2] also discovered a unique phenomenon of CAM that classification models' attention on target objects

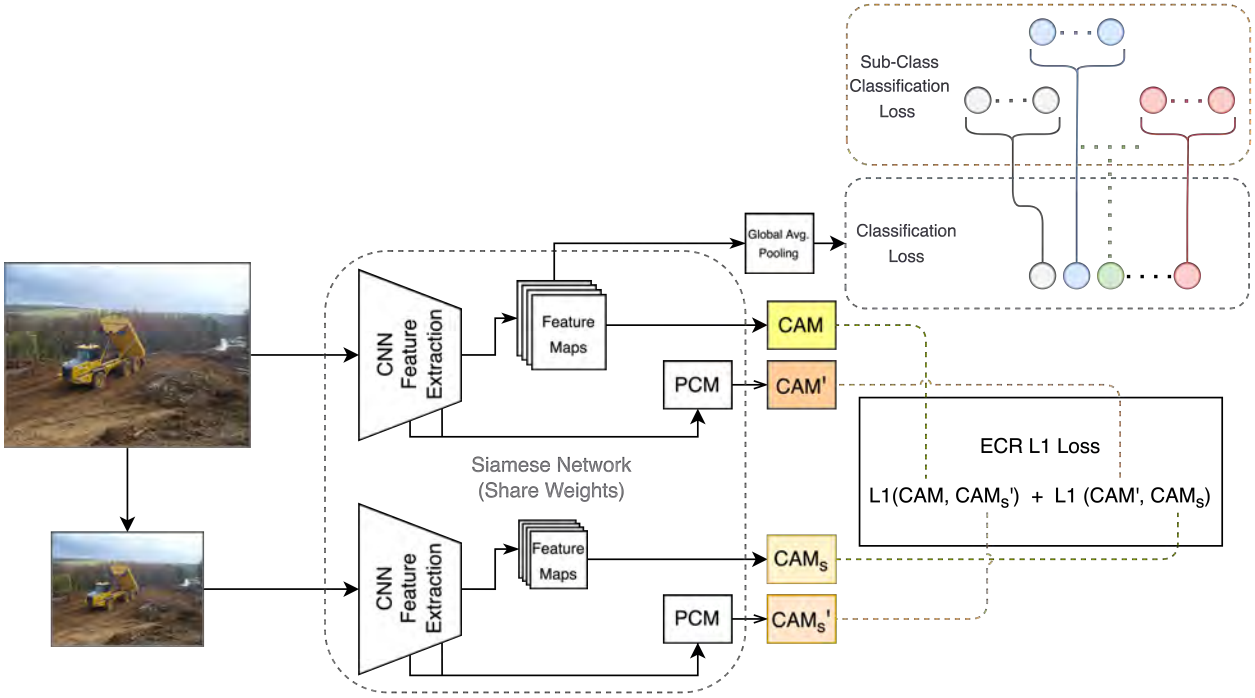

Figure 2. An overview of SEAM + SC-CAM architecture for weakly supervised pseudo labels generation for semantic segmentation.

Figure 3. Examples of models paying attention to the most discriminative features. The construction vehicles from top to bottom and left to right are: roller, mixer truck, excavator, and dump truck

is affected by an input resolution. That is, as input resolution is reduced, the output of its corresponding CAM tends to cover the entire foreground regions more. As the result, SEAM creates a siamese network that takes two input images—an original-scale image, and a down-scaled image. The two images generate two sets of CAM. The two sets including the original CAM and refined CAM are then regulated by an equivariant cross regularization (ECR) function which is used as a loss function. This function is formed by two L1 norms as follow:

$$ECR = ||CAM - CAM'_s||_1 + ||CAM' - CAM_s||_1, \quad (4)$$

where CAM represents outputs from inputs of the original scale, CAM' represents outputs refined by PCM, CAM_s represents output from inputs of the smaller scale, and CAM'_s represents output refined by PCM from inputs of the smaller scale. ECR uses CAM and CAM_s as ground truth to guide CAM' and CAM'_s (output by PCM), because CAM and CAM_s are guided by the classification labels. In another words, CAM and CAM_s are used to optimize PCM using L1 functions to extract contextual features.

2.2 Sub-Category Exploration

In addition to use CAM from downscaled inputs to guide the original inputs' CAM, this study adopts an idea of SC-CAM called sub-category exploration which also encourage classification models to pay more attention to the entire foreground regions. Instead of training a regular classification model with CAM to visualize and output the features as pseudo labels, SC-CAM also cluster each target class into K sub-clusters for training as follow:

$$\text{Class} = \{C_1, ..., C_N\}, \quad (5)$$

$$\text{Class}_K = \{C_{11}, C_{12}..., C_{1K}, ..., C_{N1}, C_{N2}..., C_{NK}\}, \quad (6)$$

where N represent number of classes, K represents the number of sub-cluster to each class. The total number of classes for sub-cluster will be $N \times K$ classes. The sub-class clustering is conducted by K-Mean clustering after images were encoded into feature vectors by a pre-trained ResNet model.

As the result, models are trained with both one-hot labels by Equation 5 and 6. With sub-category during training, models not only learn to recognize differences between original target classes, but also forced to pay more attention to the entire images to distinguish differences between each sub-cluster as shown in Figure 2 at the top-right corner.

2.3 Post-Processing Step

Although the fusion model of SEAM and SC-CAM can improve the quality of CAM, it may still fail on covering some foreground regions. This study follows SEAM and SC-CAM's post-processing procedure of applying dense condition random field (CRF) [9] to the output CAMs.

CRF is able to improve the CAM results from the fusion model as shown in the Figure 4. The output of CRF will then be used as pseudo labels to train FPN models.

3 Experimental Results

In this study VOC2012 was combined with the AIM dataset to increase the visual diversity of the dataset. However, the segmentation performance was only evaluated on the AIM dataset. The complete experiment steps are as follow:

1. Train the fusion model as shown in Figure 2 with classification labels from VOC12 & AIM datasets.
2. Generate pseudo labels as shown in the third column in Figure 4 for the AIM dataset only.
3. Train the FPN semantic segmentation model with the generated pseudo labels.
4. Evaluate FPN's mIoU score with human-annotated ground truth segmentation labels of the AIM dataset.

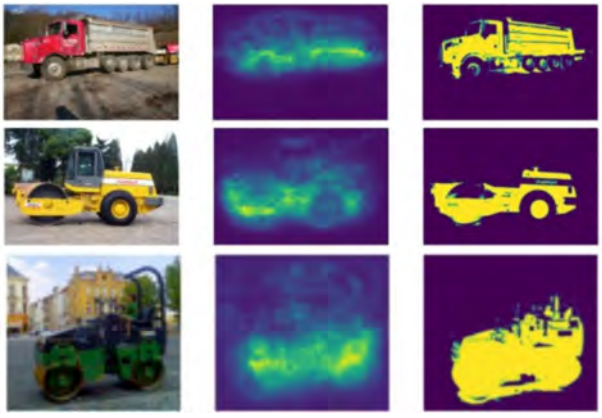

Figure 4. Examples of applying CRF to the fusion model's CAM results. From left to right column: RGB images, CAMs, and CRF results.

The prediction results of the FPN trained with pseudo labels generated from the fusion model is shown in Figure 5. The results demonstrate that the trained FPN model can capture boundaries of loaders, rollers, and dumptrucks. The mIoU score for the FPN model trained with human-labeled ground truth is also shown as the baseline for performance comparison as shown in Table 2. The proposed method achieves mIoU of 86.03%. This result indicates that segmentation models can be trained without human-annotated segmentation masks.

Table 2. Performance of the trained FPN models. (Second column) The FPN trained with pseudo labels from the fusion model. (Third column) The FPN trained with ground truth labels.

Object Class	P.L. Scores	G.T. Scores
Background	80.78	97.55
Dumptruck	87.19	97.55
Excavator	83.44	97.57
Loader	83.72	97.4
Mixertruck	92.71	99.5
Roller	88.39	99.1
mIoU	86.03	98.11

4 Conclusion

The study aims to increase the quality of semantic segmentation pseudo labels in a weakly-supervised manner with classification labels only. To achieve this, this study proposed a fusion model integrating SEAM and SC-CAM to generate pseudo labels and conducts experiments with the AIM dataset. The proposed method demonstrates the ability to generate effective pseudo labels for semantic segmentation models. The FPN model trained with the pseudo labels achieved the mIoU score of 86.03%.

A potential improvement for better pseudo label quality can be made through the change of loss function to the ECR function as shown in Equation 4. ECR contains two L1 norm functions to regulate PCM for extracting contextual features which can generate better pseudo label quality. However, L1 norm function is a distribution-based function where the error is accumulated from all CAM pixels. Thus, the L1 loss function can be dominated by the classes which contain more instances (data imbalance) such as the dump truck class as shown in Table 1. Region-based loss functions such as Jaccard loss [10], or Dice loss [11] can be used to replace the L1 norm functions in ECR

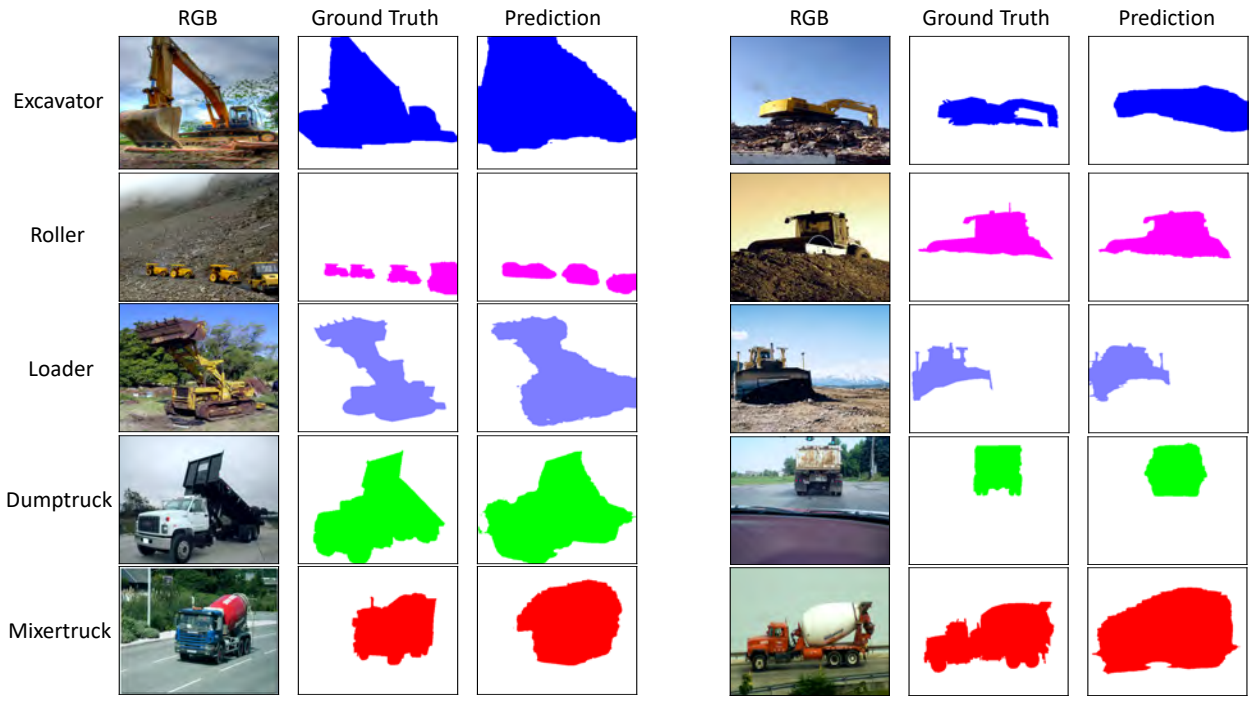

Figure 5. Prediction results of the FPN model trained by pseudo labels.

function as region-based loss functions are commonly used in segmentation tasks.

In addition, Zlateski et al. [1] points out that mixing a small amount of human-labeled fine annotations with a majority of coarse annotations can reach similar performances as using fine annotation. Therefore, it is expected the performance will be better if a small amount of fine annotation is included.

Acknowledgment

This research was conducted with the support of the "2021 Yonsei University Future-Leading Research Initiative (No. 2021-22-0037)" and the "National R&D 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] Aleksandar Zlateski, Ronnachai Jaroensri, Prafull Sharma, and Fredo Durand. On the importance of label quality for semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1479–1487, 2018. doi:10.1109/CVPR.2018.00160.

[2] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12272–12281. IEEE Computer Society, 2020. doi:10.1109/CVPR42600.2020.01229.

[3] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8988–8997, 2020. doi:10.1109/CVPR42600.2020.00901.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012.

[5] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.

[6] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *CoRR*, abs/1708.04896, 2017.

[7] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. doi:10.1109/CVPR.2016.319.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[9] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[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] 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.