Generalization of Construction Object Segmentation Models using Self-Supervised Learning

Yeji Hong [a], Wei Chih Chern [b], Tam Nguyen [c], and Hongjo Kim [d]

[a]Lyles School of Civil Engineering, Purdue University, United States
[b]Department of Electrical and Computer Engineering, University of Dayton, United States
[c]Department of Computer Science, University of Dayton, United States
[d]Department of Civil and Environmental Engineering, Yonsei University, South Korea

E-mail: hong385@purdue.edu, chernw1@udayton.edu, tnguyen1@udayton.edu, and hongjo@yonsei.ac.kr

Abstract

To evaluate the safety of construction site workers, deep learning models recognizing workers and safety equipment in construction site images are widely used. However, it is frequently observed that deep learning models based on supervised learning methods do not work well for unseen data in other domains having different visual characteristics. To address this issue, a novel method for generalizing semantic segmentation models was proposed. This method adopts two strategies: a domain adaptation method based on self-supervised learning and a copy-paste data augmentation. Source domain data with annotations (workers and hardhats) and target domain data without annotations are used for model training in a self-supervised learning scheme. The proposed model showed an improved generalization capability in semantic segmentation without annotation data of the target domain.

Keywords –

Semantic segmentation; Domain adaptation; Self-supervised learning; Copy-paste data augmentation

1 Introduction

Computer vision technology has been actively investigated nowadays to analyze jobsite contexts from construction site images for safety management [1], [2]. For example, by identifying workers [3], [4] and their personal protective equipment [5]–[8], it is possible to determine whether the workers comply with safety rules for specific tasks. Convolutional neural network-based architectures, originated from LeNet-5 [9], show superior performance in visual tasks than other traditional recognition models due to the capability of representation learning from training data in the supervised learning manner. However, supervised learning-based models generally perform well on data similar to a source domain where training images come from and do not work well for unseen data in a different domain. If the training data do not include various visual characteristics of workers and safety equipment, the model is likely to fail in recognizing the same target objects in construction site scenes having different visual characteristics than the training data. Although this problem can be solved by collecting more training data for a new scene, generating a large amount of construction site annotation data consumes a lot of time and money since fine segmentation masks for workers and safety equipment should be annotated carefully. To address this problem, a new domain adaptation method was proposed for semantic segmentation models. The proposed domain adaptation strategy includes two main components: (1) self-supervised learning and (2) copy-paste data augmentation for the model generalization to new data.

2 Related works

For jobsite safety management, previous studies have presented vision-based monitoring methods based on machine learning algorithms. In particular, supervised learning-based models have been utilized to recognize workers and personal protective equipment for ensuring the safety of workers exposed to numerous hazards. Fang et al. [6] trained a Faster R-CNN based Non-hardhat-use worker detection model using 81,000 images collected from various weather, illumination, and visual range situations. Fang et al. [7] trained a worker detection network and harness classification network using 693 positive images having workers with a harness, and about 5,000 negative images having workers without a harness. Similarly, a significant number of training images are required to train a convolutional neural network-based model to have a robust recognition performance with respect to construction site scenes having diverse visual variations.

The preparation of such dataset is time-consuming and labor-intensive especially for semantic segmentation tasks since it entails a manual annotation task requiring high precision on segmentation masks. Training data preparation is often performed again to apply a model to a new target domain. To explore the potential of minimizing data preparation efforts on a new domain, this study proposes a novel domain adaptation method which generalize a pre-trained model on a source domain to a target domain.

3 Methodology

3.1 Domain Adaptive Semantic Segmentation

The proposed method includes two steps: the first step is to train a semantic segmentation model using only a source domain dataset $X_s = \{x_s\}$ with annotations, and the second step is to adapt the model using a target domain dataset $X_t = \{x_t\}$ without annotations. The proposed adaptation method adopted prototypical pseudo label denoising Zhang et al. [10] which is a self-supervised learning technique. In the second step, pseudo labels are generated from the target domain using the model trained with a source domain in the first step, and representative features called prototypes are calculated. There are some noises in the pseudo labels caused by the difference between the source and the target domains. The noises of the pseudo labels are reduced using the distance between the prototype and each feature.

The loss function of the semantic segmentation model here starts from the categorical cross-entropy (CE) loss:

$$l_{ce}^{t} = -\sum_{i=1}^{H\times W}\sum_{k=1}^{K}\hat{y}_t^{(i,k)}\log\left(p_t^{(i,k)}\right), \qquad (1)$$

where $\hat{y}_t^{(i,k)}$ is a pseudo label and $p_t^{(i,k)}$ represents the softmax probability of the i^{th} pixel of the target data x_t classified to the k^{th}class. The pseudo labels are adjusted using representative features called prototypes and the conversion function which outputs the hard labels from the probability, as $\hat{y}_t = \xi(p_t)$. This process is formulated as follows:

$$\hat{y}_t^{(i,k)} = \xi\left(\omega_t^{(i,k)} p_t^{(i,k)}\right), \qquad (2)$$

The weight $\omega_t^{(i,k)}$ is calculated as the softmax of the feature distance between the prototype and the feature point:

$$\omega_t^{(i,k)} = \frac{\exp\left(-\left\|\tilde{f}(x_t)^{(i)} - \eta^{(k)}\right\|\right)}{\sum_{k'}\exp\left(-\left\|\tilde{f}(x_t)^{(i)} - \eta^{(k')}\right\|\right)}, \qquad (3)$$

where $\tilde{f}$ and $\eta^{(k)}$ represents the momentum encoder of the feature extractor f and the prototype of the k^{th}class, respectively. The prototype is calculated by the following equation:

$$\eta^{(k)} = \frac{\sum_{x_t\in X_t}\sum_i f(x_t)^{(i)} * \mathrm{I}\left(\hat{y}_t^{(i,k)} == 1\right)}{\sum_{x_t\in X_t}\sum_i \mathrm{I}\left(\hat{y}_t^{(i,k)} == 1\right)}, \qquad (4)$$

where I is the indicator function. The prototype changes in each iteration as the weights of the feature extractor f are updated.

The total loss function in the domain adaptation network for the second step is as follows:

$$l_{total} = l_{ce}^{s} + \alpha l_{ce}(p_t, \hat{y}_t) + \beta l_{ce}(\hat{y}_t, p_t) + \gamma_1 l_{kl}^{t} + \gamma_2 l_{reg}^{t}, \qquad (5)$$

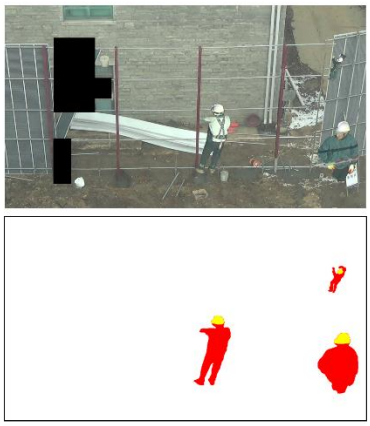
where l_{kl}^{t} and l_{reg}^{t} denotes a Kull-back—Leibler divergence loss and a regularization loss, respectively, for making features compact, which enables denoising easier. α, β, γ_1, and γ_2 are hyper-parameters for each loss term.

3.2 Copy-paste Data Augmentation

Training data with diverse visual features help deep learning models to avoid overfitting and generalize well on unseen data. To increase the visual diversity in the training data which do not have annotations of the source domain, a copy-paste data augmentation method was employed. Figure 1 shows the copy-paste data augmentation framework. First, one of the target domain images is randomly selected. Second, the area containing the foreground is removed. Lastly, instances from the source domain data are pasted into the background of the target domain. In this way, deep learning models can learn the features of the target domain background and the diversified boundaries between the foreground and the background.

(a) Random target domain image selection

(b) Foreground deletion



(c) Copy-paste foreground in source domain

Figure 1. Copy-paste data augmentation process. The upper part of (c) is an augmented image, and the lower part shows its annotations.

3.3 Domain Adaptation Index (DAI)

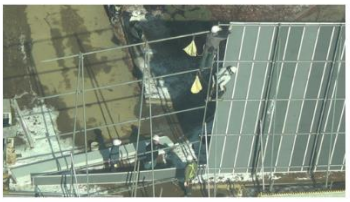
Domain Adaptation Index (DAI) is developed for evaluating the domain adaptation performance of segmentation models, as shown in Equation (6). This index evaluates how well a model performs on the target domain without target domain annotations. That is, the model will be trained with images from the source and target domains, but annotations are only provided by the source domain. When DAI is 1, the domain adaptive model achieves the same performance as the model trained with the target domain data and its annotations.

$$\mathrm{DAI} = \frac{Performance\ of\ a\ model\ trained\ without\ target\ domain\ annotations}{Performance\ of\ a\ model\ trained\ with\ target\ domain\ annotations} \quad (6)$$

4 Experiments

In this study, one source domain dataset and two target domain datasets were used in experiments. All datasets were collected from three videos of scaffold installation operations at the Sinchon Campus of Yonsei University. Image samples are shown in Figure 2. Target domain 1 differs only in scale from the source domain, and target domain 2 differs in the workers' appearance and the background. DeepLabV2 [11] was used as the semantic segmentation model.

The semantic segmentation performance of target domains 1 and 2 is shown in Table 1 and Table 2, respectively. In Table 1 and Table 2, Source-only is a DeepLabV2 trained with the source domain data only, and Upper bound is a DeepLabv2 trained using the target domain data with their annotations. The performance is shown in four measures: the first and the second measures are the Intersection of Union (IoU) of workers and hardhats, respectively, the third measure, mIoU, is a mean of IoU for all target classes, and the last measure is DAI obtained by dividing the mIoU of the model by the mIoU of the Upper bound.



(a) Source domain

(b) Target domain 1

(c) Target domain 2

Figure 2. Sample images of each dataset

It was observed that the generality of the segmentation model was improved by the proposed method. When the domain adaptation and copy-paste data augmentation were applied, mIoU increased compared to Source-only in both target domains 1 and 2.

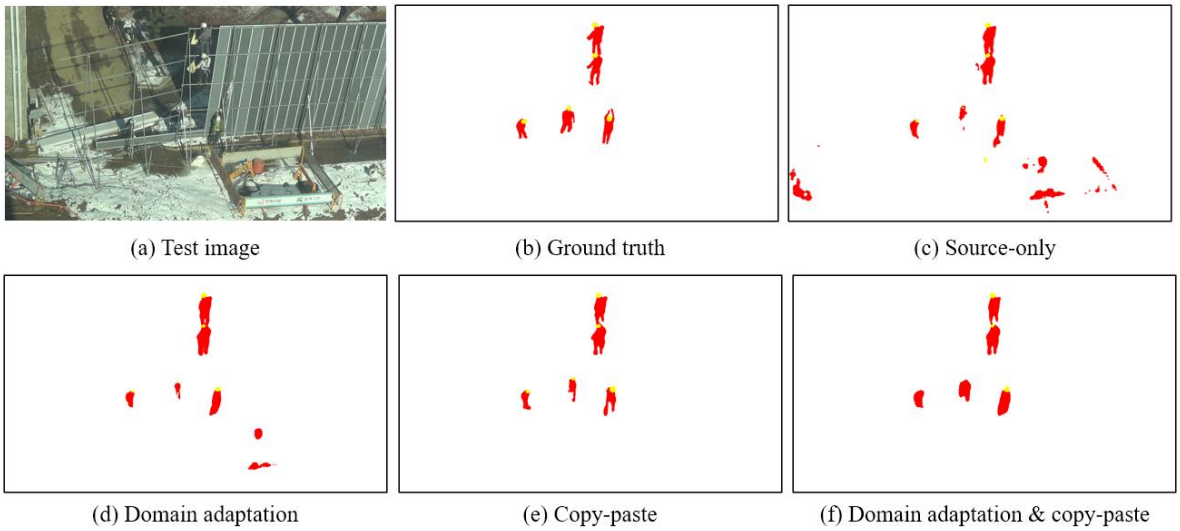


Figure 3. Example test image and the segmentation results of target domain 1

The highest DAI for target domains 1 and 2 were 81.10% and 84.61%, respectively. Especially for target domain 2, which has different visual characteristics from the source domain, mIoU were improved 8.78 and 27.49 percentage points, respectively.

Table 1. Semantic segmentation performance of target domain 1 (D represents the domain adaptation strategy, and C represents the copy-paste strategy).

	D	C	IoU_{worker}	IoU_{hardha}	mIoU (%)	DAI (%)
Source-only	-	-	42.01	59.93	50.97	68.16
Ours	√	-	54.28	59.63	56.96	76.17
	-	√	62.10	59.20	60.65	81.10
	√	√	63.95	55.21	59.58	79.67
Upper bound	-	-	76.25	73.31	74.78	-

Table 2. Semantic segmentation performance of target domain 2 (D represents the domain adaptation strategy, and C represents the copy-paste strategy).

	D	C	IoU_{worker}	IoU_{hardha}	mIoU (%)	DAI (%)
Source-only	-	-	45.82	37.71	41.77	50.47
Ours	√	-	49.68	51.42	50.55	61.08
	-	√	65.66	72.86	69.26	83.69
	√	√	70.19	69.85	70.02	84.61
Upper bound	-	-	85.11	80.41	82.76	-

Experimental results for target domain 1 are shown in Figure 3. The Source-only model incorrectly predicted a large number of pixels in the background as the worker category. Although the noise was partially removed by the domain adaptation model, it missed some parts of hardhats. This result accounts for the increment of the IoU_{worker} and the decrement of the $IoU_{hardhat}$ between the source-only model and the domain adaptation model. The copy-paste model well identified the workers' arms and legs, and all the hardhats. When domain adaptation and copy-paste methods were applied together, a few numbers of hardhats were lost and the shape of the workers, which resulted in 3.99 percent point decrement of $IoU_{hardhat}$.

Figure 4 shows a test image, its ground truth, and the segmentation results of four different segmentation models. Unlike target domain 1, where the background scene was the same but only different in scale to the background of the source domain, scenes of target domain 2 differs from the source domain scene. The source-only model misclassified many background pixels. The domain adaptation model removed the misclassified noise, and as a result, $IoU_{hardhat}$ was improved by 13.71 percent points. The copy-paste strategy worked effectively for the segmentation of target domain 2. It is conjectured that the copy-paste augmentation contributes to generating boundary information between target objects and the target domain background. The domain adaptation strategy decreased the $IoU_{hardhat}$ while improving the IoU_{worker}, resulting in the increase in mIoU by 0.76 percent points.

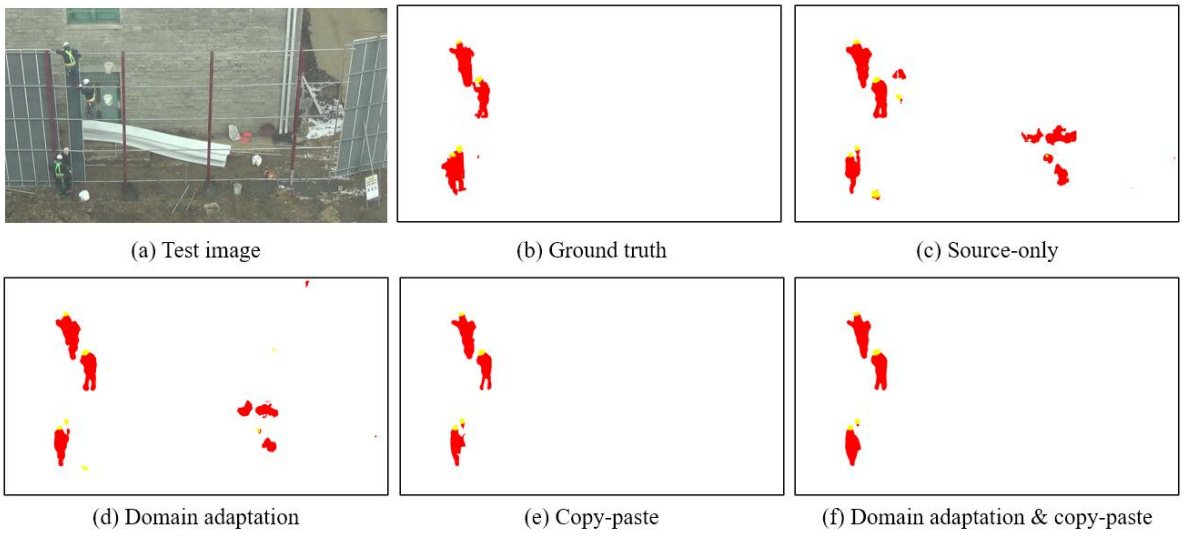


Figure 4. Example test image and the segmentation results of target domain 2

5 Conclusion

This study aims to increase the generalization capability of a semantic segmentation model for workers and hardhats. To address the generalization problem, a self-supervised learning-based domain adaptation and a data augmentation method were experimented for scaffolding installation site images.

The proposed methodology showed the potential of generalizing a semantic segmentation model without additional annotation efforts. The model with the copy-paste strategies achieved a 9.68 percent point increment in mIoU compared to the model trained only with source data for target domain 1. The model incorporating the domain adaptation and the copy-paste strategies achieved a 28.25 percent points increment in mIoU compared to the model trained only with source data for target domain 2. This improvement is remarkable in that the model did not use new annotations from the target domain.

Future study will be conducted to reduce the remaining gap between the highest mIoU of the proposed model and the upper bound model. Additional target domains will also be tested in future study to validate the effectiveness of the proposed method.

Acknowledgement

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.21SMIP-A158708-02)" 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] M. Zhang, R. Shi, and Z. Yang, "A critical review of vision-based occupational health and safety monitoring of construction site workers," *Saf. Sci.*, vol. 126, no. February, p. 104658, Jun. 2020, doi: 10.1016/j.ssci.2020.104658.

[2] B. Sherafat *et al.*, "Automated Methods for Activity Recognition of Construction Workers and Equipment: State-of-the-Art Review," *J. Constr. Eng. Manag.*, vol. 146, no. 6, p. 03120002, 2020, doi: 10.1061/(asce)co.1943-7862.0001843.

[3] H. Kim, K. Kim, and H. Kim, "Vision-Based Object-Centric Safety Assessment Using Fuzzy Inference: Monitoring Struck-By Accidents with Moving Objects," *J. Comput. Civ. Eng.*, vol. 30, no. 4, p. 04015075, 2016, doi: 10.1061/(asce)cp.1943-5487.0000562.

[4] S. Bang, Y. Hong, and H. Kim, "Proactive proximity monitoring with instance segmentation and unmanned aerial vehicle-acquired video-frame prediction," *Comput. Civ. Infrastruct. Eng.*, vol. 36, no. 6, pp. 800–816, Jun. 2021, doi: 10.1111/mice.12672.

[5] M. W. Park and I. Brilakis, "Construction worker detection in video frames for initializing vision trackers," *Autom. Constr.*, vol. 28, pp. 15–25, 2012, doi: 10.1016/j.autcon.2012.06.001.

[6] Q. Fang *et al.*, "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos," *Autom. Constr.*, vol. 85, no. May 2017, pp. 1–9, 2018, doi: 10.1016/j.autcon.2017.09.018.

[7] W. Fang, L. Ding, H. Luo, and P. E. D. Love, "Falls from heights: A computer vision-based approach for safety harness detection," *Autom. Constr.*, vol. 91, no. February, pp. 53–61, 2018, doi: 10.1016/j.autcon.2018.02.018.

[8] B. E. Mneymneh, M. Abbas, and H. Khoury, "Vision-Based Framework for Intelligent Monitoring of Hardhat Wearing on Construction Sites," *J. Comput. Civ. Eng.*, vol. 33, no. 2, p. 04018066, Mar. 2019, doi: 10.1061/(ASCE)CP.1943-5487.0000813.

[9] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.

[10] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical Pseudo Label Denoising and Target Structure Learning for Domain Adaptive Semantic Segmentation," no. 2, pp. 12414–12424, 2021, doi: http://arxiv.org/abs/2101.10979.

[11] 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," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.