17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Weakly Supervised Domain Adaptation using Super-pixel labeling for Semantic Segmentation

Masaki Yamazaki[1], Xingchao Peng[2], Kuniaki Saito[2], Ping Hu[2], Kate Saenko[2], Yasuhiro Taniguchi[1]
[1]Honda R&D Co.,Ltd, Japan
[2]Boston University, USA

Abstract

Deep learning for semantic segmentation requires a large amount of labeled data, but manually annotating images are very expensive and time consuming. To overcome the limitation, unsupervised domain adaptation methods adapt a segmentation model trained on a labeled source domain (synthetic data) to an unlabeled target domain (real-world scenes). However, the unsupervised methods have a poor performance than the supervised methods with target domain labels. In this paper, we propose a novel weakly supervised domain adaptation using super-pixel labeling for semantic segmentation. The proposed method reduces annotation cost by estimating a suitable labeling area calculated from the Entropy-based cost of a previously learned segmentation model. In addition, we generate the new pseudo-labels by applying fully connected Conditional Random Field model over the pseudo-labels obtained using an unsupervised domain adaptation. We show that our proposed method is a powerful approach for reducing annotation cost.

1. Introduction

Current deep convolutional neural networks for segmentation require a large amount of labeled training data to achieve good results. Furthermore, their performance seems to scale linearly with an exponential increase of training data [1]. However, densely annotating image for segmentation is very expensive and time-consuming. For example, each Cityscapes [2] image on average takes about 90 minutes to annotate. To overcome the limitation, generating densely annotated images from rendered scenes, such as the Grand Theft Auto V (GTA5) [3] is useful. However, the large appearance gap (e.g., illumination, pose, and image quality) across simulated / real domains significantly degrades the performance of synthetically trained models.

In light of the above issues, unsupervised domain adaptation for segmentation (CBST) [4] has been recently proposed to solve a domain gap between simulated (GTA5) and real-world (Cityscapes) domains, where pseudo-annotated unlabeled samples are added to the training set with no human cost at all. However, the unsupervised methods have a poor performance than the supervised methods with target domain labels, because the unsupervised methods assume that the pseudo-annotated unlabeled samples are labeled correctly.

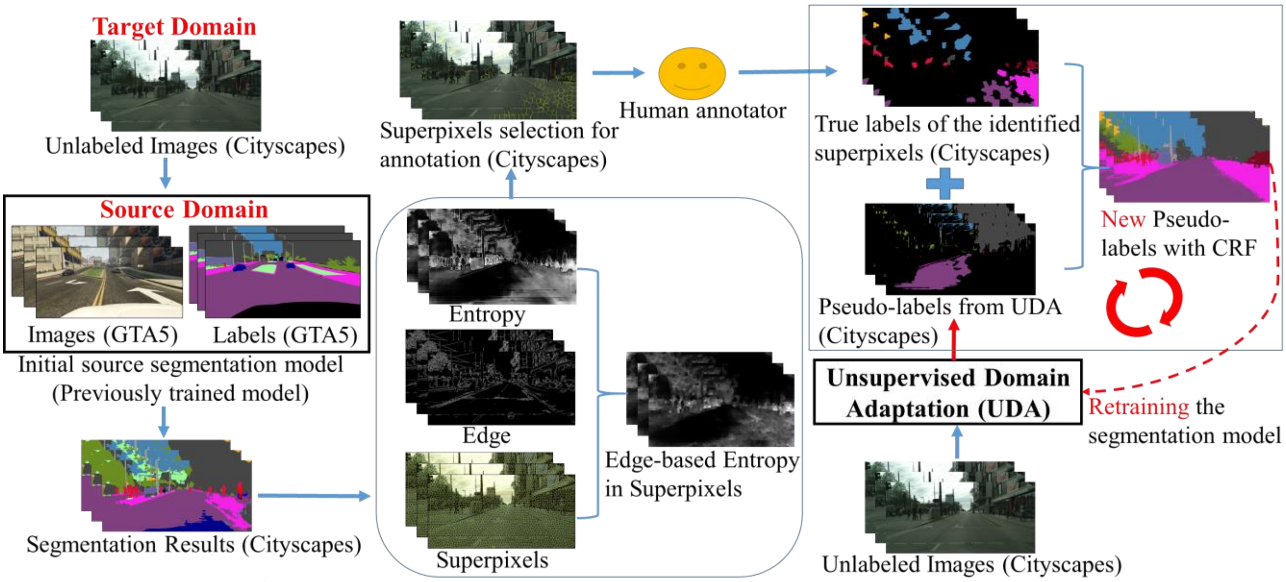
Weakly-supervised supervision for segmentation have been known to reduce annotation cost on unlabeled data. Previous works for annotation have used point-clicks [5, 6, 17], scribbles [6, 7], Pixel-level Rectangle [22], Pixel-level Block [8] to train semantic segmentation networks. Weakly-supervised domain adaptation for segmentation from synthetic data with pixel-level labels, and real-world scenes with only bounding-box labels has been recently proposed [9]. However, these methods don't maximize the performance for segmentation but minimize human annotation effort.

In this paper, we propose a novel weakly supervised domain adaptation using super-pixel labeling for semantic segmentation, in which a human only has to hand-label a few, automatically selected, areas within an unlabeled image. The proposed framework reduces annotation cost by estimating a suitable labeling area calculated from the Entropy-based cost of a learned CNN and by generating the new pseudo-labels using fully connected Conditional Random Field (CRF) [10]. Comprehensive experiments show that the proposed method could reduce the annotation effort to 10%, while keeping 95% of the mean Intersection over Union (mIoU) of a model that was trained with the fully annotated training set of Cityscapes.

2. Related Works

In this section, we briefly review the important works about the two most related tasks: unsupervised domain adaptation and weakly-supervised supervision for reducing annotation cost on unlabeled data.

Unsupervised Domain Adaptation. Unsupervised domain adaptation is to transfer discriminative knowledge from one fully labeled source domain to unlabeled target domain [11]. Adversarial learning based methods reduce the gap between source and target domain [12, 13, 14]. Another important strategy for unsupervised domain adaptation [4] is based on self-training [15, 16] by learning labeled source samples and target data with pseudo-labels generated from a learned segmentation model. However, given a sufficient amount of data, models trained in a supervised way outperform any unsupervised method. Because it is not possible to completely guarantee the correctness of the generated pseudo-labels.



Target Domain
Unlabeled Images (Cityscapes)
Source Domain
Images (GTA5)
Labels (GTA5)
Initial source segmentation model
(Previously trained model)
Segmentation Results (Cityscapes)
Superpixels selection for annotation (Cityscapes)
Human annotator
Entropy
Edge
Superpixels
Edge-based Entropy in Superpixels
True labels of the identified superpixels (Cityscapes)
Pseudo-labels from UDA (Cityscapes)
New Pseudo-labels with CRF
Unsupervised Domain Adaptation (UDA)
Retraining the segmentation model
Unlabeled Images (Cityscapes)

Figure 1. Illustration of the proposed weakly supervised domain adaptation in semantic segmentation training.

Weakly-supervised supervision for segmentation. Weakly-supervised supervision for segmentation such as point-clicks [5, 6, 17], scribbles [6, 7], Pixel-level Rectangle [22], Pixel-level Block [8] have been known to reduce annotation cost on unlabeled data. Annotations such as point-clicks or scribbles or Pixel-level Rectangle/Block are faster to acquire than polygon annotation, which leads to a larger and more varied dataset at the same cost. Recently, some weakly supervised methods [9, 17, 25, 26, 27] are presented to save the costs of annotating ground truth. Wang et all [9] proposed weakly-supervised domain adaptation for segmentation from synthetic data with pixel-level labels, and real-world scenes with only bounding-box labels. However, these methods are not effective labeling for maximizing the performance of a semantic segmentation method.

3. Proposed method

3.1 Weakly-supervised domain adaptation for segmentation

Given a labeled source dataset and unlabeled target dataset, our objective is to reduce the annotation cost on the unlabeled target data for adapting a segmentation model trained on a labeled source domain to a target domain. In the case of unsupervised domain adaptation, the target ground truth labels are not available. Unsupervised domain adaptation (CBST) [4] is carried out by alternately generating a set of pseudo-labels in target domain, and then fine tuning network based on these pseudo-labels and labeled source data. Jointly learning the segmentation model and optimizing pseudo-labels on the unlabeled target data are naturally difficult as it is not possible to completely guarantee the correctness of the generated pseudo-labels. The pseudo-labels are generated from the confident predictions of the segmentation model. Therefore, the pseudo-label generation has the missing value from the less confident predictions of the segmentation model.

To overcome these problems, we propose a novel weakly supervised domain adaptation for effective pixel-level labeling in semantic segmentation. This minimizes human annotation effort while maximizing the performance of semantic segmentation. The proposed method reduces the annotation cost by estimating a suitable labeling area on the unlabeled target dataset calculated from the Entropy-based cost of a segmentation model learned on the labeled source dataset. A human annotator will hand label a few of the suitable labeling area and much more labeling areas will be obtained automatically using the pseudo-labels generated from an unsupervised domain adaptation method (CBST) [4]. We propose three different strategies (Entropy, Edge, and Super pixel) for estimating a suitable labeling area in an image.

Our proposed method starts after training the initial source network over a labeled source dataset. We present effective labeling strategies to reduce annotations at pixel-level using the initial source network. At a pixel-level, for each candidate image in the unlabeled target dataset, we identify the most uncertain super-pixels for annotations. The uncertain super-pixels are identified using uncertainty measures computed at a pixel-level (described later in this section). The human annotator will provide the true labels of the identified super-pixels.

We combine the true labels annotated by a human annotator with the pseudo-labels generated from an unsupervised domain adaptation method (CBST) [4]. To further improve the pseudo-label, we apply fully connected Conditional Random Field (CRF) model [10] over the pseudo-labels with the true labels of the identified pixels. The final pseudo-labels image is used for retraining the segmentation model in the unsupervised domain adaptation framework. The overview of our proposed method is given in Fig. 1.

3.2 Effective labeling strategies

We describe three different strategies (Entropy, Edge, and Super pixel) for computing uncertain super-pixels for annotations in an unlabeled target image. We are computing both information measures for each pixel location individually given the a-posteriori probability distributions from an initial segmentation model trained on a labeled source domain. The uncertain super-pixels in the unlabeled target dataset are then annotated by a human annotator. Fig.2 shows the uncertain super-pixels using the three different strategies.

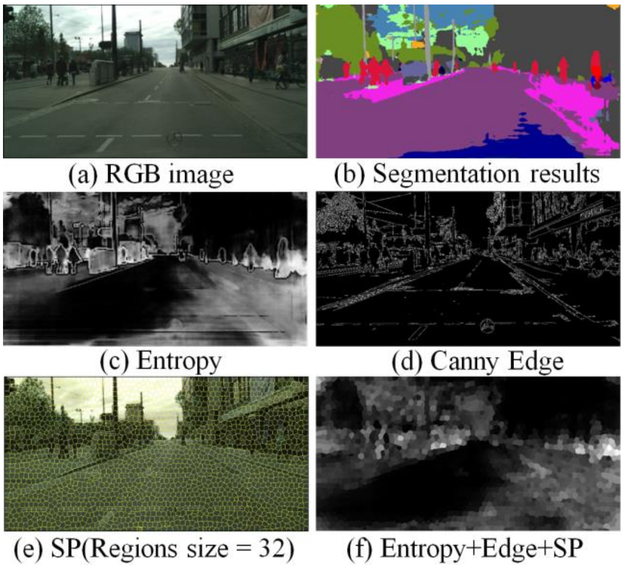

(a) RGB image
(b) Segmentation results
(c) Entropy
(d) Canny Edge
(e) SP(Regions size = 32)
(f) Entropy+Edge+SP

Figure 2. Visualization of the uncertain pixels using the three different strategies (Entropy, Edge, Super pixel).

Entropy. Entropy is the most widely used information measure for computation of uncertainty [18]. Here, the data with the highest positive impact on the model’s performance is estimated to be the one where the posterior probability distribution produces the highest entropy. Entropy is used as a measure of uncertainty, since its value is maximized when the model assigns each considered class the same probability and very small if the model is sure about its decision. The entropy $H^{(u,v)}$ at each pixel is computed as follows (see Fig.2 (c)).

$$H^{(u,v)} := -\sum_c p_c^{(u,v)}\big(f(x)\big)\log\left(p_c^{(u,v)}\big(f(x)\big)\right) \qquad (1)$$

where $p_c^{(u,v)}\big(f(x)\big)$ denotes the probability score map belonging to the class c at a specific pixel position (u,v) in an unlabeled image x. This probability distribution is obtained from the initial segmentation model $f()$ trained on a labeled source domain.

Edge based Entropy. The edge pixels inherently have high uncertainty, because the misclassification rate for pixels at object boundaries/edges is more when compared to the other pixels in the image. To consider edge pixels for annotation, we give a higher weight to the entropy of edge pixels. We use a Canny edge detector to identify edge pixels (see Fig.2 (d)), and the weighted entropy computed for edge pixels in a given image x is obtained as follows.

$$H_e := -\sum_{(u,v)}\sum_c w p_c^{(u,v)}\big(f(x)\big)\log\left(p_c^{(u,v)}\big(f(x)\big)\right) \qquad (2)$$

where $w > 1$ is the weight given to the edge pixels. For other pixels it is set to 1.

Superpixels. In semantic segmentation, the neighboring pixels are highly likely to have a close relationship and share similar information. Therefore, they are likely to belong to the same semantic class. However, the entropy of each pixel is calculated independently without considering this relationship. In order to take advantage of the spatial correlation in images, we use superpixels (SP) in an image. We use SLIC (Simple Linear Iterative Clustering) [19] to computing the superpixels in a given image (see Fig.2 (e)), and define the entropy at the superpixel level as the sum of its pixel entropies (see Fig.2 (f)).

3.3 Superpixel annotation

Superpixel annotations enable workers to mark a group of visually related pixels at once. This can reduce the annotation time for background regions and objects with complex boundaries. We select the superpixels in order of high edge based entropy scores at the superpixel (which we described in section 3.2), and finally request their respective labels from a human annotator. Our superpixel annotation interface is given in figure 3. The human annotator just choose a class label and click the superpixel on the image to annotate.

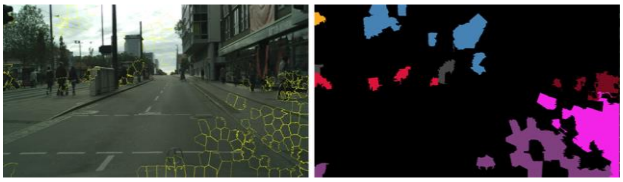


Figure 3. Superpixel annotation UI. Annotators are given several yellow areas to annotate by click.

3.4 Generating the New Pseudo-labels with CRF

We combine the true labels annotated by a human annotator with the pseudo-labels generated from an unsupervised domain adaptation method (CBST) [4]. The unsupervised domain adaptation method is carried out by alternately generating a set of pseudo-labels corresponding to large selection scores (i.e., softmax probability) in target domain, and then fine tuning network based on these pseudo-labels and labeled source data. The generated pseudo-labels has the missing value from the less confident predictions of the network model. Especially, the initial segmentation model trained on a labeled source domain lead to miss-predictions for generating pseudo-labels. This is because the large appearance gap (e.g. appearance, scale, geological position, illumination, camera, etc.) across simulated/real domains significantly degrades the performance of the trained models.

To further improve the pseudo-labels, we apply CRF model [10] over the pseudo-labels with the true labels annotated by a human annotator. The CRF establishes pairwise potential on all pairs of pixels in a given image. The CRF potentials incorporate smoothness terms that maximize label agreement between similar pixels, and can integrate more elaborate terms that model contextual relationships between object classes. In semantic segmentation, the neighboring pixels potentially have a close relationship and belong to the same semantic class. The CRF propagate the true label information to the neighboring pixels. Fig. 4 shows the generated new pseudo-labels with CRF and the generated pseudo-labels by CBST in each round (The model training are repeated for multiple rounds). The black pixels in the pseudo-labels are assigned appropriate labels by CRF.

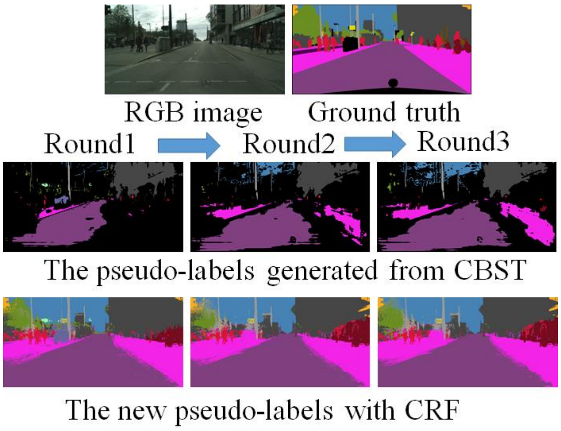

RGB image
Ground truth
Round1
Round2
Round3
The pseudo-labels generated from CBST
The new pseudo-labels with CRF

Figure 4. The examples of the new pseudo-labels with CRF and the pseudo-labels generated from CBST in each round.

4. Experimental Results

In this section, we evaluate our proposed method for semantic segmentation. We focus on adaptation cases from GTA5 [3] / SYNTHIA [20] to Cityscapes [2]. We evaluate the average mean Intersection over Union (mIoU) calculated on the validation dataset of Cityscapes. We use SYNTHIA-RAND-CITYSCAPES subset including labeled 9,400 (760×1280) images. GTA5 dataset includes annotated 24,966 (1052×1914) images captured from the GTA5. Cityscapes dataset contains 2,975 (1024×2048) finely annotated training images along with 500 validation images.

We compare other weakly supervised supervisions for segmentation: point-clicks [6], scribbles [6], Pixel-level Rectangle [22], Pixel-level Block [8], and Coarse annotations [2]. We only annotate 10% and 50% pixel budget in each Cityscapes image using original ground truth for Pixel-level Rectangle, Pixel-level Block, and Our method (in order of high entropy scores). All weakly supervised supervisions show performance by using CBST [4] with ResNet-38 [23] as unsupervised domain adaptation in GTA5 / SYNTHIA to Cityscapes. The networks were pre-trained on ImageNet [21]. SGD has been used to train all the models by MXNET [24]. As parameters for CBST, we use the default values from the authors' public implementation. The performance is evaluated over a hold-out validation dataset.

Table1 gives experimental results of mIoU calculated on the validation dataset of Cityscapes (from GTA5). The baseline result (Fine annotation) is obtained using CBST with full ground truth (annotating all the pixels). Table2 gives experimental results of mIoU calculated on the validation dataset of Cityscapes (from SYNTHIA). Table3 gives the results of annotation average time in each Cityscapes image for all weakly supervised supervisions. Our method-10% pixel budget of full dataset can reduce the annotation effort to 10%, while keeping 95% of mIoU of a model that was trained with the fully annotated training set of Cityscapes.

Annotation methods	mIoU
Point-clicks [6]	46.4
Scribbles [6]	49.4
Pixel-level Rectangle -50% [22]	56.5
Pixel-level Rectangle -10% [22]	52.4
Pixel-level Block -50% [8]	56.4
Pixel-level Block -10% [8]	52.3
Coarse annotation [2]	50.7
Our method -50%	57.1
Our method -10%	54.9
Fine annotation (Full Supervision) [2]	57.3
CBST [4] (Unsupervised Learning)	45.2

Table 1. Weakly-supervised segmentation performance from GTA5 to Cityscapes.

Annotation methods	mIoU
Point-clicks [6]	42.7
Scribbles [6]	43.1
Pixel-level Rectangle -50% [22]	48.5
Pixel-level Rectangle -10% [22]	45.1
Pixel-level Block -50% [8]	48.4
Pixel-level Block -10% [8]	45.0
Coarse annotation [2]	43.7
Our method -50%	49.0
Our method -10%	46.9
Fine annotation (Full Supervision) [2]	49.3
CBST [4] (Unsupervised Learning)	42.5

Table 2. Weakly-supervised segmentation performance from SYNTHIA to Cityscapes.

Annotation methods	Time
Point-clicks [6]	1 min
Scribbles [6]	2 min
Pixel-level Rectangle [22]	8 min
Pixel-level Block [8]	7 min
Coarse annotation [2]	7 min [2]
Our method	7 min
Fine annotation (Full Supervision) [2]	90 min [2]

Table 3. Annotation average time in each Cityscapes image.

5. Conclusion

We have proposed a novel weakly supervised domain adaptation using super-pixel labeling for semantic segmentation, in which a human only has to hand-label a few, automatically selected, areas within an unlabeled image. We have demonstrated our method’s performance on Cityscapes. We show that combining the Entropy-based cost of a learned CNN and the Pseudo-labels with CRF from an unsupervised domain adaptation method is a powerful approach for reducing annotation cost. This will help further research in other segmentation task such as instance segmentation.

References

[1] C. Sun, A. Shrivastava, S. Singh, A. Gupta: “Revisiting unreasonable effectiveness of data in deep learning era,” *IEEE International Conference on Computer Vision*, pp.843-852, 2017.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele: “The cityscapes dataset for semantic urban scene understanding,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp.3213-3223, 2016.

[3] S. R. Richter, V. Vineet, S. Roth, V. Koltun: “Playing for data: Ground truth from computer games,” *The European Conference on Computer Vision*, pp.102-118, 2016.

[4] Y. Zou, Z. Yu, B. V. K. V. Kumar, J. Wang: “Unsupervised domain adaptation for semantic segmentation via class-balanced self-training,” *The European Conference on Computer Vision*, pp.289-305, 2018.

[5] A. Bearman, O. Russakovsky, V. Ferrari, L. Fei-Fei, “What's the point: Semantic segmentation with point supervision,” *The European Conference on Computer Vision*, pp.549–565, 2016.

[6] M. Tang, F. Perazzi, A. Djelouah, I. B. Ayed, C. Schroers, Y. Boykov: “On Regularized Losses for Weakly-supervised CNN Segmentation,” *The European Conference on Computer Vision*, 2018.

[7] D. Lin, J. Dai, J. Jia, K. He, J. Sun: “Scribblesup: Scribble-supervised convolutional networks for semantic segmentation,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp.3159–3167, 2016.

[8] H. Lin, P. Upchurch, K. Bala: “Block Annotation: Better Image Annotation With Sub-Image Decomposition,” *IEEE International Conference on Computer Vision*, 2019.

[9] Q. Wang, J. Gao, X. Li: “Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes,” *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 2019.

[10] P. Krahenbuhl, V. Koltun: “Efficient inference in fully connected crfs with gaussian edge potentials,” *Advances in Neural Information Processing Systems*, 2011.

[11] G. Wilson, D. J. Cook: “A survey of unsupervised deep domain adaptation”, *arXiv preprint*, 2019.

[12] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, B. Scholkopf: “Covariate shift and local learning by distribution matching,” *MIT Press*, pp.131-160, 2009.

[13] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell: “Deep domain confusion: Maximizing for domain invariance,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp.2962-2971, 2017.

[14] M. Long, Y. Cao, J. Wang, M. Jordan: “Learning transferable features with deep adaptation networks,” *International Conference on Machine Learning*, pp.97-105, 2015.

[15] O. Chapelle, B. Scholkopf, A. Zien: “Semi-supervised learning,” *IEEE Transactions on Neural Networks*, pp.542-542, 2009.

[16] X. Zhu: “Semi-supervised learning literature survey,” *Technical Report*, pp.1530, 2005.

[17] S. Obikane, Y. Aoki: “Weakly supervised domain adaptation with point supervision in histopathological image segmentation,” *Asian Conference on Pattern Recognition*, 2019.

[18] B. Settles, M. Craven: “An analysis of active learning strategies for sequence labeling tasks,” *The conference on empirical methods in natural language processing*, pp.1070-1079, 2008.

[19] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Ssstrunk: “Slic superpixels compared to state-of-the-art superpixel methods,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.2274-2282, 2012.

[20] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, A. M. Lopez: “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp.3234-3243, 2016.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et.al.: “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, pp.211-252, 2015.

[22] R. Mackowiak, P. Lenz, O. Ghori, F. Diego, O. Lange, C. Rother: “CEREALS Cost-effective region-based active learning for semantic segmentation,” *The British Machine Vision Conference*, 2018.

[23] Z. Wu, C. Shen, A.v.d. Hengel: “Wider or deeper: Revisiting the resnet model for visual recognition,” *arXiv preprint*, 2016.

[24] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, Z. Zhang: “Mxnet A flexible and efficient machine learning library for heterogeneous distributed systems,” *arXiv preprint*, 2015.

[25] G. Praveen, E. Granger, P. Cardinal: “Deep weakly-supervised domain adaptation for pain localization in videos,” *IEEE International Conference on Automatic Face Gesture Recognition*, 2020.

[26] S. Tan, J. Jiao, W.S. Zheng: “Weakly supervised open-set domain adaptation by dual-domain collaboration,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[27] S. Yang, C. Zhangjie, L. Mingsheng, W. Jianmin: “Transferable curriculum for weakly-supervised domain adaptation,” *AAAI Conference on Artificial Intelligence*, 2019.