Multimodal Unsupervised Car Segmentation via Adaptive Aerial Image-to-Image Translation

Haohong Lin [*1] **Hanjiang Hu** [*1] **Peide Huang** [*1] **Zhepeng Cen** [*1]

Abstract

Traffic flow monitoring has become a crucial yet challenging tasks in intelligent cities and autonomous driving. As a crucial step to HD-Map understanding, image segmentation under diverse bird's-eye-view scenarios has become increasingly attractive to multidisciplinary researchers. How to perform semantic segmentation while generalizing over multimodal background noise is challenging yet fruitful. In this paper, we propose a semi-supervised image segmentation framework via image-to-image translation by using certain domain knowledge in the birds-eye-view traffic perception. Our framework can be summarized as a hierarchical segmentation. First, we design an image translation model to transform different styles of the road to a unified style. Second, we perform an adaptive vehicle detection based on a non-learnable segmentation algorithm. Empirical results show our methods outperform some fine-tuned methods, and we rank top 10 of the small vehicle segmentation (in IoU metrics) among all the models on the leaderboard.

1. Introduction

Image segmentation is of crucial importance in massive visual understanding systems. The major task is to partition input images into several segments or objects of interest. Formulated as a pixel-wise image classification problem, traditional methods apply brutal-force thresholding(Otsu, 1979), K-means clustering(Dhanachandra et al., 2015), activate contours(Kass et al., 2004), or Markov Random Field(Plath et al., 2009) to segment the image data into different objects according to some handcrafted features. These traditional computer vision models, which use low-level features like contours to conduct the segmentation, are inclined to neglect semantic styles in the images.

However, the accuracies of the above works are not satisfying. With the prosperity of deep learning, a lot of image segmentation frameworks have been proposed (Minaee et al., 2020), including the encoder-decoder models(Badrinarayanan et al., 2017), the generative adversarial networks (GANs)(Goodfellow et al., 2014), and regional convolutional networks (R-CNN)(Ren et al., 2015)(He et al., 2017), as well as some dilated convolutional networks(Yu & Koltun, 2016). These methods, built on convolutional neural networks, are using different scales of receptive fields to capture hierarchical information in the entire images. Still, there exist two major parts of challenges in the image segmentation with current deep learning models.

The first challenge is the requirement of a full set of labels to train the data-driven segmentation framework, which is not always an efficient or even accessible way in real-world applications, since labeling each frame of the traffic flow can be extremely time-consuming or even impossible.

The second challenge is the multimodal nature of different entities. For example, in the image classification task of the traffic flow, the background of the road can have different styles, and the illumination can differ due to the weather. In order to achieve good segmentation results, one potential way is to build a model that could encode such multimodal styles, then transfer them to some unified style. Some existing works, like the image translation work, (Huang et al., 2018) are trying to deal with such multimodality, but are not fully applicable in our setting, since it transforms the style from one multimodal distribution to another, which does not lower the difficulty of image segmentation since it does not give out *invariant* representation of the image.

In this project, we propose a weakly-supervised label-efficient image segmentation framework via image-to-image translation techniques between different domains, aiming at disentangling foreground contents (car objects of our interests) with background style (different roads) to achieve satisfying detection effects within a given patch of the image. Our method takes diverse vehicles/backgrounds as input, then extract the content and separates it from the styles, and

[*]Equal contribution [1]Carnegie Mellon University. Correspondence to: Haohong Lin <haohongl@andrew.cmu.edu>, Hanjiang Hu <hanjianghu@cmu.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

makes a style transfer to extract it from the background. The extracted results will be evaluated by intersection over union (IoU) metrics. We plan to train and compare our model performance with region proposal-based image segmentation baselines, including Faster-RCNN (Ren et al., 2015), and Mask-RCNN(He et al., 2017).

2. Background and Related Work

We are relating our work with two common settings in computer vision: image segmentation and image-to-image translation. We summarize the related works below and develop our own methods in the following sections.

Image Segmentation with Deep Learning. It's a common way to formulate image segmentation as a classification problem. For instance, semantic segmentation performs a pixel-level classification of different objects of interest, which is a harder task compared to general image classification since we need to assign each pixel a prediction at the object level. A lot of deep learning-based methods have been proposed in the past few years to tackle this problem, including Convolutional models with graphical models, R-CNN-based models (Ren et al., 2015)(He et al., 2017), Generative adversarial networks, (Goodfellow et al., 2014) and so on. These methods, however, usually require fully-supervised training to guarantee a satisfying performance in image segmentation, while a full set of labels are sometimes inaccessible in some real-world applications.

Unsupervised image-to-image Translation. Thanks to the rapid development of deep generative models (DGMs), Image-to-image translation starts with conditional generative adversarial networks (GANs), which is later used in generating high-resolution images in the selected target domain. Based on the different demands in the target domain(Wang et al., 2018), different attend to preserve different properties to guarantee the quality of image translation, including pixel-level features(Shrivastava et al., 2017)(Bousmalis et al., 2017), semantic-level features(Taigman et al., 2016), class label features(Bousmalis et al., 2017), and pairwise sample contrastive features(Benaim & Wolf, 2017). Among all the related literature, Unsupervised Image-to-Image Translation (UNIT)(Liu et al., 2017) and Multimodal Unsupervised Image-to-Image Translation (MUNIT)(Huang et al., 2018) propose a framework that separate the content from style, then use the GAN-based loss in both the image and latent space to train the image translation model that could handle the multimodal settings across different domains. In this work, our model shares a similar basic setting and training fashion, while modifying the structure of latent space to adapt to our image segmentation tasks.

3. Problem Setup and Methodology

We are aiming to conduct an image classification for every pixel. Thus this can be formulated as multiple binary-classification tasks, where the input of the model $x \in \mathbb{R}^{N\times N\times 3}$ is an RGB image, and the output $y \in \mathbb{R}^{N\times N}$ is a binary matrix that contains the information of whether a pixel belongs to the vehicle or not. Notice here we are not doing classification between different types of vehicles, thus this is not as complex as object detection works. On the contrary, we will give out an exact segmentation in the shape of vehicles rather than a perpendicular bounding box in the object.

The overall pipeline is represented in 1. Inspired by the image translation works, we propose a disentangled representation in content and style of feature space. We first translate all the images with different backgrounds into a similar target domain with identical background (e.g. green or black background). To enforce the bidirectional encoder-decoder model to capture the 'content' information, which is the foreground car while separating all the irrelevant 'style' information, which is the multimodal road backgrounds, we design two groups of loss in both pixels space and feature space. The process of this encoder-decoder framework can be written as:

$$\text{Encoding: } c = \phi(c|x) \triangleq \phi_c(x), \quad s = \phi(s|x) \triangleq \phi_s(x) \tag{1}$$

$$\text{Decoding: } \hat{x} = f_c(c) = f_c(\phi_c(x)) \tag{2}$$

After we get a style-free birds-eye-view image with identical background color, we apply the threshold masking in the HSV color domain to get a raw mask for downstream image segmentation. Since in practice, this mask is not very robust due to the flaws in reconstruction, we use some domain knowledge about cars (near rectangular, among a certain range in size from the birds-eye-view map) to design another automatic adaptive erosion-dilation algorithm (AEDA) to improve the quality of the mask. After this procedure, the model can directly get contours from the mask and conduct segmentation. The process can be formulated as:

$$\text{Thresholding: } m = f_{threshold}(\hat{x}) \tag{3}$$

$$\text{Adjusting: } \hat{m} = f_{AEDA}(m) \tag{4}$$

$$\text{Segmentation: } y = f_{contours}(\hat{m}) \tag{5}$$

3.1. Image Translation for Background Removal

Generally, the image translation modules can have multimodal inputs with different styles of road backgrounds and contents of vehicles (defined as source domain), and we aim to transform them into the same contents with the desired

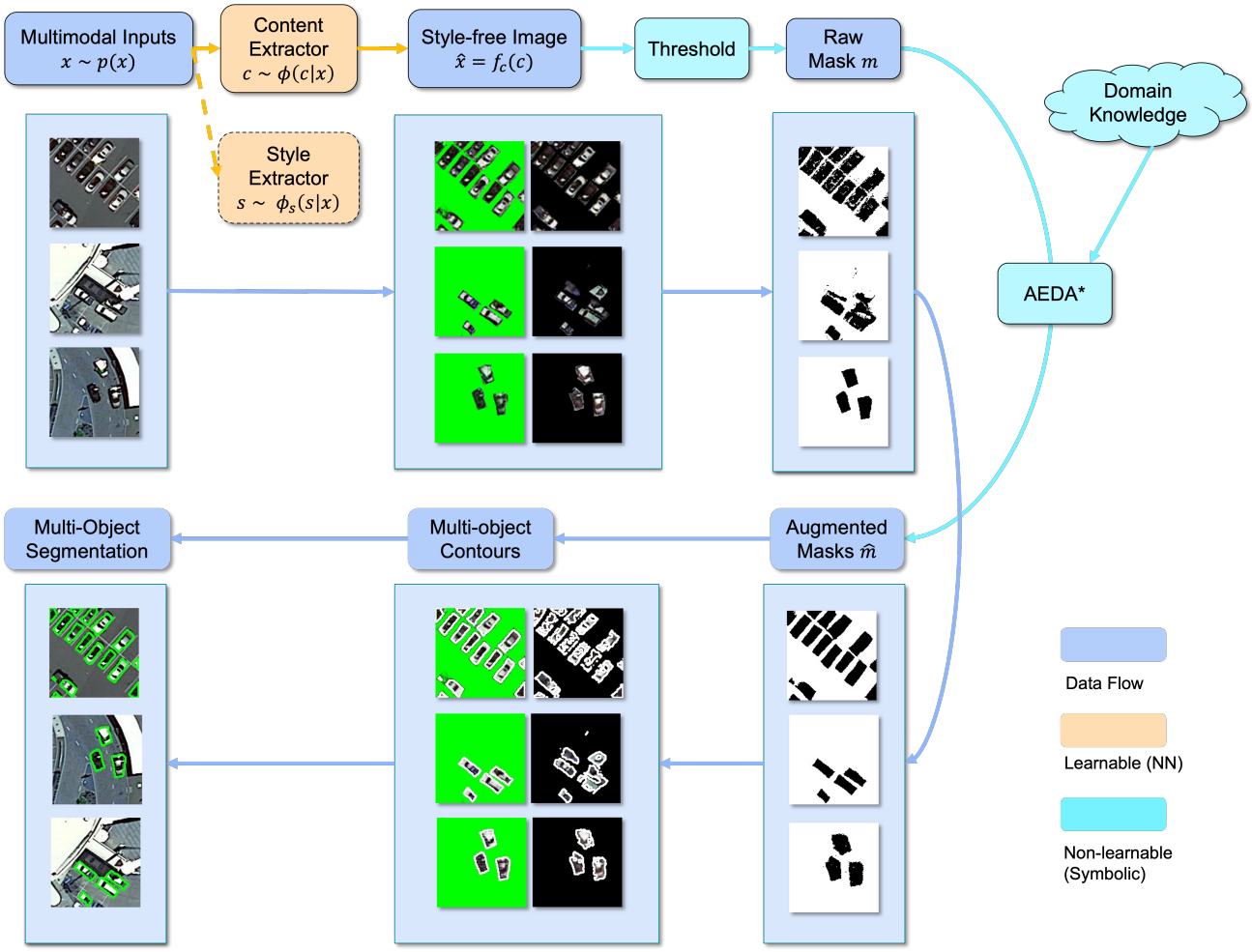

Figure 1. **Pipeline of the Overall Image Segmentation**: We first design content-style disentangled representation modules inspired by the style transfer in image translation works. After this, we are able to drop the style and generate style-free samples with the content-only decoder. After this, we apply domain knowledge in the birds-eye-view vehicles in normal traffic scenes and design proper adaptive masking techniques (non-learnable algorithms) to finish the segmentation tasks. The components of this two-stage framework are further elaborated in the latter part of the methodology section.

style (defined as target domain) (Huang et al., 2018; Liu et al., 2017) that is friendly to downstream segmentation tasks. Specifically, our proposed translation framework is constructed between raw bird-view images with various background roads or entities and foreground vehicles, and the target domain with various foreground vehicles, but with preprocessed zero-valued masks[1] on the background roads as the fixed style, where we want to generate images with only foreground vehicles.

As we can see in Fig. 2 we denote the source domain of bird-view images as domain 1, and the target domain as domain 2, our framework of image translation can be formulated as below: the source domain can be encoded into both style and content latent embedding, while the target domain (with identical background) will only have content-based latent embedding. All the embedding will engage in the reconstruction of raw images, while the cross-domain image translation will be realized by mixing the content of the source domain and the style of the target domain.

The first group of loss is within individual domain 1 and 2, in order to enforce a good quality of image reconstruction in the original pixel space, we have the empirical risk for

[1]Preprocessed by Object Detection Baselines, such as Mask-RCNN.

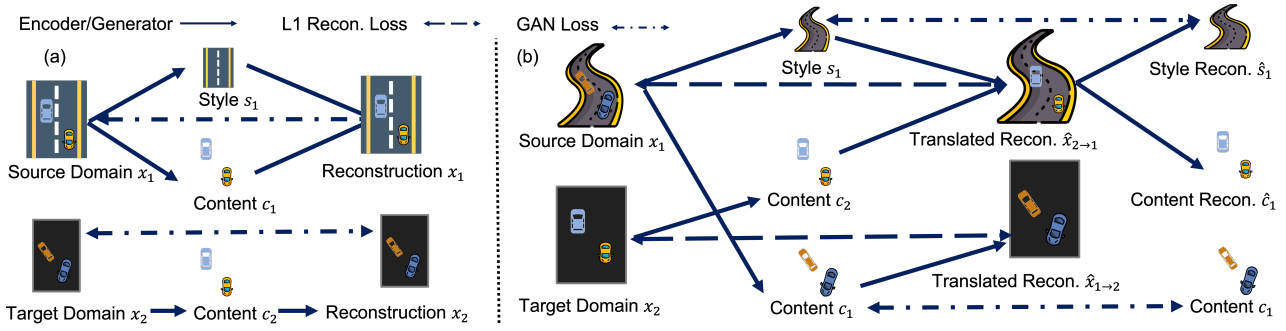

Figure 2. **Zoom-out View of Image Translation**: For the **solid lines**, we use two CNN-based encoders to encode the style and content information of the raw image $x_1 = f(s_1, c_1)$, and a single content decoder for the masked image $x_2 = f(c_2)$ (which is preprocessed by using a certain portion of the label in the training data). Notice that images in this target domain only contain content related to car objects of our interest without different road backgrounds. The training procedure includes a joint optimization in **Part (a)** reconstruction with encoder-decoder framework, and **Part (b)** a content extractor and style-free image translation with feature decomposition and decoding. Two types of **dashed lines** on both left and right side stands for the reconstruction and GAN loss in pixel space and feature space. And we summarize them into two main categories: consistency in image representation and prediction, as well as realism of the extracted style-free images.

Algorithm 1 Adaptive Erosion-Dilation Algorithm (AEDA)

Data: Style-free $\hat{x}$, maximum objects N_{obj}, threshold of background color HSV^-, HSV^+, vehicle size S^-, S^+
Result: Multi-object Segmentation $y \in \mathbb{R}^{N\times N}$
$m \leftarrow x$
$i \leftarrow where(x \in [HSV^-, HSV^+])$ */*Thresholding Raw Masks*/*
$m[i] \leftarrow \mathbf{0}$
$C = \text{Contours}(m)$
$iter = 0$
for $c \in C$ and $iter < N_{obj}$ **do**
  $iter \leftarrow iter + 1$ */*Adaptive Erosion-Dilation Iterations*/*
  if $\text{Size}(c) \notin [S^-, S^+]$ **then**
    while $\text{Size}(c) < S^-$ **do**
      $c \leftarrow \text{Erode}(c)$
      $\hat{m} \leftarrow m.update(c)$
      $C = \text{Contours}(\hat{m})$
    end while
    while $\text{Size}(c) > S^+$ **do**
      $c \leftarrow \text{Dilate}(c)$
      $\hat{m} \leftarrow m.update(c)$
      $C = \text{Contours}(\hat{m})$
    end while
  end if
end for
$y = \text{PolytopeFilling}(C)$

the image reconstruction with L_1 loss terms,

$$\mathcal{L}_{\text{recon}}^{x_1} = \mathbb{E}_{x_1\sim p(x_1)}[\|f^1(\phi_c^1(x_2), \phi_c^1(x_1)) - x_1\|_1] \quad (6)$$

$$\mathcal{L}_{\text{recon}}^{x_2} = \mathbb{E}_{x_2\sim p(x_2)}[\|f^2(\phi_c^2(x_1))) - x_2\|_1] \quad (7)$$

The second group of loss is associated with the image transfer, which is across domain 1 and 2. We have the empirical risk for the content and style reconstruction with L_1 loss terms to guarantee the consistency of image representation between two domains in latent space. Then we design another adversarial loss term using GAN loss to guarantee the translated image in both domains is as realistic as possible

(to fool the discriminator).

$$\mathcal{L}_{\text{recon}}^{s_1} = \mathbb{E}_{s_1\sim\mathcal{N}(0,1),x_2\sim p(x_2)}[\|\phi_s^1(f^1(\phi_c^2(x_2), s_1)) - s_1\|_1] \tag{8}$$

$$\mathcal{L}_{\text{recon}}^{c_2} = \mathbb{E}_{s_1\sim\mathcal{N}(0,1),x_2\sim p(x_2)}[\|\phi_c^1(f^1(\phi_c^2(x_2), s_2)) - \phi_c^2(x_2)\|_1] \tag{9}$$

$$\mathcal{L}_{\text{recon}}^{c_1} = \mathbb{E}_{x_1\sim p(x_1)}[\|\phi_c^2(f^2(\phi_c^1(x_1))) - \phi_c^1(x_1)\|_1] \tag{10}$$

$$\begin{aligned}\mathcal{L}_{\text{GAN}}^{x_2} &= \mathbb{E}_{x_1\sim p(x_1)}[\log(1 - D^2(f^2(\phi_c^1(x_1))))] \\ &+ \mathbb{E}_{x_2\sim p(x_2)}[\log D^2(s_2)]\end{aligned} \tag{11}$$

$$\begin{aligned}\mathcal{L}_{\text{GAN}}^{x_1} &= \mathbb{E}_{s_1\sim\mathcal{N}(0,1),x_2\sim p(x_2)}[\log(1 - D^1(f^1(\phi_c^2(x_2), s_1)))] \\ &+ \mathbb{E}_{x_1\sim p(x_1)}[\log D^1(x_1)]\end{aligned} \tag{12}$$

where domain 1 (raw images) is associated with deconvolution network based decoder $f^1(\cdot,\cdot)$, CNN-based style (background) encoder $\phi_s^1(\cdot)$, content (vehicle) encoder $\phi_c^1(\cdot)$ and discriminator $D^1(\cdot)$. Domain 2 (style-free images) is associated with reconstruction decoder $f^2(\cdot)$ which takes only content as input, as well as CNN-based content (vehicle) encoder $\phi_c^2(\cdot)$ and discriminator D^2.

Finally, the total empirical risk minimaxition optimization problem can be related to all the seven modules, formulated as:

$$\begin{aligned}\min_{f^1,\phi_s^1,\phi_c^1,f^2,\phi_c^2}\max_{D^1,D^2} &\mathcal{L}_{\text{GAN}}^{x_1} + \mathcal{L}_{\text{GAN}}^{x_2} + \lambda_s\mathcal{L}_{\text{recon}}^{s_1} \\ &+ \lambda_c(\mathcal{L}_{\text{recon}}^{c_1} + \mathcal{L}_{\text{recon}}^{c_2}) + \lambda_x(\mathcal{L}_{\text{recon}}^{x_1} + \mathcal{L}_{\text{recon}}^{x_2})\end{aligned} \tag{13}$$

where $\lambda_s, \lambda_c, \lambda_x$ represents the tunable hyperparameters.

3.2. Adaptive Erosion-Dilation Algorithms (AEDA) for Segmentation

After we get a style-free image from our special image translation methods, the images are already friendly to the segmentation task. However, due to the potential flaw in the image reconstruction, we can use some adaptive algorithms by applying domain knowledge in the vehicle size, then properly erode or dilate the masks that we get from thresholding, then get our finalized segmentation results as in 1. The fundamental idea of the AEDA is to guarantee that each detected object is within a range of size.

4. Experiments

4.1. Evaluation metric

We will use the common measure, Intersection over Union (IoU), as our evaluation metric. We denote TP as true positive (pixels with object label that are also correctly predicted as object), TN as true negative (pixels with non-object label that are correctly predicted as non-object), FP as false positive (pixels with non-object label that are falsely predicted as object), FN as false positive (pixels with object label that are falsely predicted as non-object).

IoU is defined as the ratio of the area of overlap between predicted and ground-truth segmentation by the area of their union, i.e.,

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{14}$$

Besides IoU defined above, we also use other common evaluation metric such as accuracy, average precision (AP), recall and F1 Score, which are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{F1 Score} = 2 * \frac{\text{recall} * \text{precision}}{\text{recall} + \text{precision}}$$

4.2. Dataset and Preprocessing

We conduct a quantitative evaluation of the proposed method on DOTA (Xia et al., 2018), an aerial image object detection dataset. The original dataset contains 2806 aerial images from different sensors and aerial platforms. Each image is of the size about 4000×4000 pixels and contains objects exhibiting a wide variety of scales, orientations, and shapes. In this project, we focus on the segmentation of small vehicles only. We first crop the original images to smaller images of size 144×144. We preprocess the cropped images by only retaining the small vehicles in the image and masking the rest of the image with green or black color. We show some image samples in Fig 4. The training set contains two subsets: *trainA* and *trainB*. *trainA* contains 10k masked images and *trainB* contain the corresponding raw image. The test set contains 1k images of the same format.

4.3. Training

Hyperparameter We use a batch-size of 1, Adam learning rate $1e^{-4}, \beta_1 = 0.5, \beta_2 = 0.999$, weight decay $1e^{-4}$, $\lambda_s = 1, \lambda_c = 1, \lambda_x = 10$. For the generator and discriminator, we use 64 filters in the bottom-most layer and MLP of hidden size 256. We train the model for 1M iterations. More hyperparameters are included in the config file in our code. For the AEDA module, in green background, we choose the lower bound of HSV as [50, 100, 100] and the upper bound of HSV as [70, 255, 255]. In black background, we choose the lower bound of HSV as [0, 0, 0] and the upper bound of HSV as [20, 100, 30]. The maximum number of vehicles in one image is 16, and the bound of the pixels of a car given the 144×144 image is $200 \sim 550$ pixels. For better understanding of the training process, we show a smoothed version of the training curve in Fig. 5 using a moving average window of 100 iterations. We can see that with 1 million iterations, both the generator and discriminator gradually achieve convergence.

4.4. Evaluation Results

We compare our methods with three different baselines, k-means clustering (Dhanachandra et al., 2015), Faster-RCNN (Ren et al., 2015)[2], and Mask-RCNN(He et al., 2017) [3]. Faster-RCNN combines Region Proposal Network (RPN) and Fast R-CNN (Girshick, 2015) using the attention mechanism to enable convolutional feature sharing. Mask-RCNN extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition.

Notice that the results we report in K-Means are the implementation with our AEDA methods. Even if we add extra tricks to it, it is not comparable with our deep-learning-based methods illustrated in the table. Our methods show a slight advantage over the

[2]https://github.com/rbgirshick/py-faster-rcnn

[3]https://github.com/facebookresearch/Detectron

Figure 3. Sample images in the DOTA dataset (Xia et al., 2018).

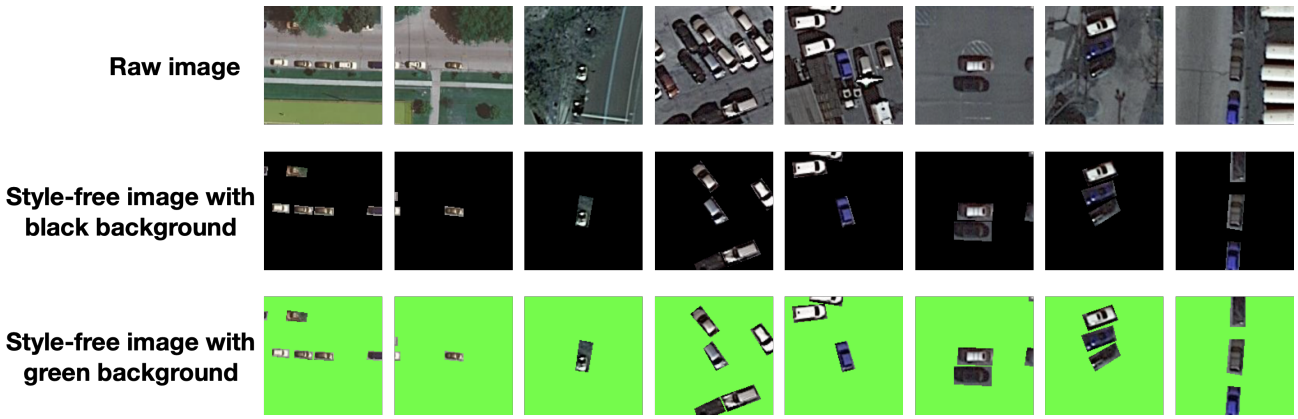

Figure 4. Samples of Training Datasets.

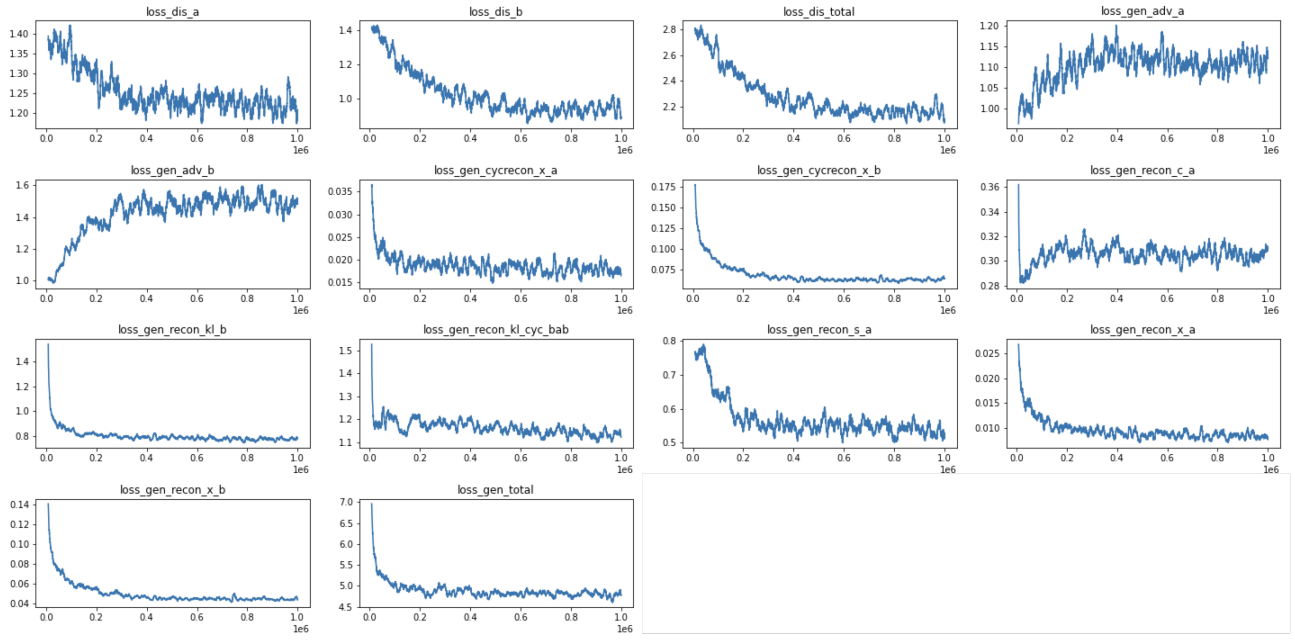

Figure 5. Training loss of reconstruction in image space, feature space, as well as GAN loss. We can see a convergence between generator and discriminator after 1 million iterations.

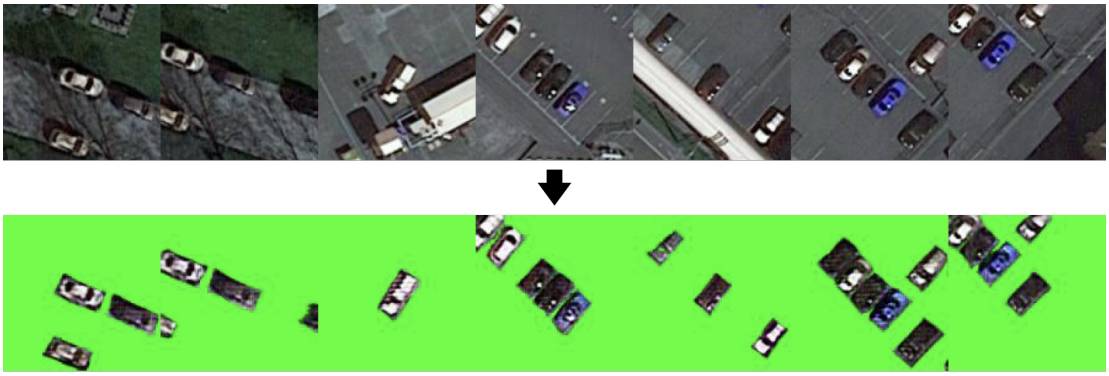

Figure 6. Segmentation results from multimodal source domain to unimodal target domain.

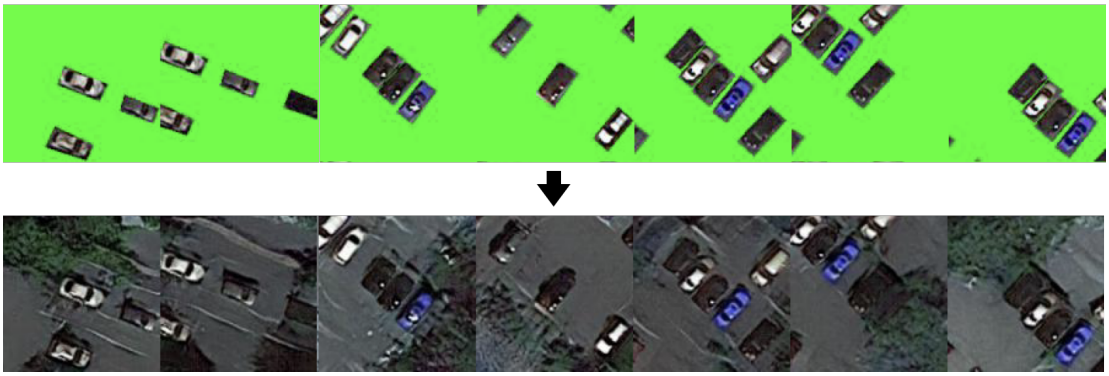

Figure 7. By-product of Translation framework: multimodal road generation for data augmentation

Figure 8. Generated images from the testing dataset.

Table 1. Main results

Methods	IoU	Accuracy	Recall	Precision	F1 Score
K-Means	0.3110	0.9658	0.4059	0.7792	0.4412
Faster RCNN	0.5155	0.9739	0.6377	0.7167	0.6509
Mask RCNN	0.5668	0.9804	0.7660	0.6944	0.7086
Ours	**0.5825**	0.9813	0.7832	0.7073	0.7261

Table 2. Ablation study between different variants of our proposed methods.

Methods	IoU	Accuracy	Recall	Precision	F1 Score
Ours-BM	0.5134	0.9749	0.7709	0.6288	0.6603
Ours-NA	0.5548	0.9683	0.8237	0.6476	0.6917
Ours-HL	0.5155	0.9739	0.6377	0.7167	0.6509
Ours	**0.5825**	0.9813	0.7832	0.7073	0.7261

fine-tuned R-CNN-based methods in the IoU metric, which shows the effectiveness of our image translation and adaptive masking frameworks.

4.5. Ablation Study

In this section, we compare our proposed models with three different variants.

Black Masks (BM) is the variant that we transfer the background style to all black masks. Although zero-value masks are easy for the NN to learn from, there are lots of black-colored vehicles in the normal traffic, therefore, black masks is not as an ideal target domain as the green color. Empirical results show that BM variants have 7% lower in the IoU metric.

No AEDA (NA) is the variant without adaptive erosion and dilation. We can see a performance drop in IoU by around 3%. Therefore, we argue that by applying certain domain knowledge, we can achieve better vehicle segmentation in the birds-eye-view map.

Half label (HL) variant is to demonstrate the data efficiency and the robustness of our proposed model against the missing labels. This setting basically simulates the real-world applications that we can hardly have all the labels at the scene. At the preprocessing stage in Fig. 4, we will randomly mask out 50% of the cars at the scene. The results show that the drop of IoU ia n such setting is still acceptable (comparable with some of our baselines under full labels), indicating that our proposed methods can robustly disentangle the background road and foreground vehicles even without full supervision signals.

5. Conclusion and Future Works

In conclusion, we propose a weakly-supervised image segmentation framework via image-to-image translation between two domains, which successfully disentangle foreground contents (vehicles) with multi-modal backgrounds (parking lots, highway, city roads, etc.). We also optimize our results with an adaptive masking technique. Empirical results demonstrate that our proposed method can successfully extract and separate vehicles from diverse backgrounds.

Thanks to the generative nature of our model, besides the traditional image segmentation task, we can also generate images with roads given some images that contain only vehicles with black or green background. As we see in Fig. 8(b), one by-product of our generative model is that it can be used to generate realistic multimodal images given only vehicle images. Those generated images can serve as good samples for data augmentation in other downstream tasks. Exploring how good are the generated images and how they can facilitate other learning tasks would be an interesting future direction for our work.

Still, we are aware there are some limitation in this work. First, the proposed model is not in a fully end-to-end fashion. In this work, the reason our model works is that we have strong domain kowledge about vehicle shape (a rectangle, in a range of 200 to 550 pixels with the given resolution), while in other domains, such knowledge may be hard to acquire. Therefore, it will be a good idea to replace the AEDA module with a learnable neural networks that can embed domain knowledge as constraints during the segmentation, and the pipeline will then become an end-to-end model that can adapt to multi-class image segmentation or object detection.

References

Badrinarayanan, V., Kendall, A., and Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2481–2495, 2017.

Benaim, S. and Wolf, L. One-sided unsupervised domain mapping. *Advances in neural information processing systems*, 30, 2017.

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3722–3731, 2017.

Dhanachandra, N., Manglem, K., and Chanu, Y. J. Image segmentation using k -means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science*, 54:764–771, 2015.

Girshick, R. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *NIPS*, 2014.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018.

Kass, M., Witkin, A. P., and Terzopoulos, D. Snakes: Active contour models. *International Journal of Computer Vision*, 1: 321–331, 2004.

Liu, M.-Y., Breuel, T., and Kautz, J. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. Image segmentation using deep learning: A survey. *CoRR*, abs/2001.05566, 2020. URL https://arxiv.org/abs/2001.05566.

Otsu, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1): 62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

Plath, N., Toussaint, M., and Nakajima, S. Multi-class image segmentation using conditional random fields and global classification. In *ICML '09*, 2009.

Ren, S., He, K., Girshick, R. B., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39: 1137–1149, 2015.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2107–2116, 2017.

Taigman, Y., Polyak, A., and Wolf, L. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3974–3983, 2018.

Yu, F. and Koltun, V. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122, 2016.