Importance-Aware Semantic Segmentation in Self-Driving with Discrete Wasserstein Training

Xiaofeng Liu[1,4†], **Yuzhuo Han**[1,2†], **Song Bai**[3], **Yi Ge**[4], **Tianxing Wang**[5], **Xu Han**[6], **Site Li**[4], **Jane You**[7], **Jun Lu**[1*]

[1]Beth Israel Deaconess Medical Center, Harvard Medical School, Harvard University;
[2]School of Mathematical Sciences, Dalian University of Technology;
[3]Department of Statistics, University of California, Berkeley;
[4]Carnegie Mellon University; [5]Fudan University; [6]Johns Hopkins University
[7]Department of Computing, The Hong Kong Polytechnic University.
[†]Contribute equally [*]Corresponding Author: jlu@bidmc.harvard.edu

Abstract

Semantic segmentation (SS) is an important perception manner for self-driving cars and robotics, which classifies each pixel into a pre-determined class. The widely-used cross entropy (CE) loss-based deep networks has achieved significant progress w.r.t. the mean Intersection-over Union (mIoU). However, the cross entropy loss can not take the different importance of each class in an self-driving system into account. For example, pedestrians in the image should be much more important than the surrounding buildings when make a decisions in the driving, so their segmentation results are expected to be as accurate as possible. In this paper, we propose to incorporate the importance-aware inter-class correlation in a Wasserstein training framework by configuring its ground distance matrix. The ground distance matrix can be pre-defined following a priori in a specific task, and the previous importance-ignored methods can be the particular cases. From an optimization perspective, we also extend our ground metric to a linear, convex or concave increasing function $w.r.t.$ pre-defined ground distance. We evaluate our method on CamVid and Cityscapes datasets with different backbones (SegNet, ENet, FCN and Deeplab) in a plug and play fashion. In our extenssive experiments, Wasserstein loss demonstrates superior segmentation performance on the predefined critical classes for safe-driving.

Introduction

Semantic segmentation is an importance task in many vision-based applications or systems, such as self-driving, robotics, augmented reality and automatic surgery system (Yang et al. 2018). The goal is to densely assign class label to each pixel in the input image for precisely understanding the scene. Consequently, semantic segmentation can be treated as an image classification task at pixel level. In the past decades, significant amounts of research effort has been spent on this issue (Long, Shelhamer, and Darrell 2015; Paszke et al. 2017; Badrinarayanan, Kendall, and Cipolla 2017).

The recent semantic segmentation method based on deep representation learning with cross-entropy (CE) loss have made considerable success on major open benchmark datasets (Cordts et al. 2016; Brostow, Fauqueur, and Cipolla 2009). For each pixel in the input image, the CE loss compares the prediction with one-hot encoded ground-truth label without considering any connections to other pixels. The final loss is usually calculated as the average of the cumulative CE loss across the entire image, making each pixel contribute equally to the final loss (Liu et al. 2019e; 2019b). This would lead to a problem for different classes with unbalanced representation in the image, due to training probably dominated by the most prevalent class.

Even for the case that the pixels contribute unequally to the final loss, such as assigning large weights to the border of segmented objects (Li et al. 2017), the associated models still encounter challenges in practical applications. Most existing semantic segmentation methods neglect the severity of diverse misclassifications, which may cause unexpected accidents. For example, an accident of Tesla is caused by recognising a white truck as sky, arousing intense discussion of self-vehicle safety [1]. Supposing that the white truck is rec-

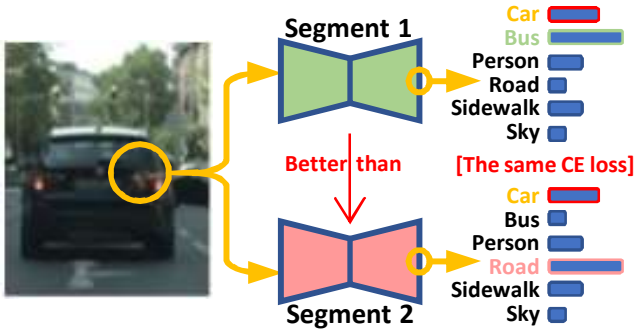

Figure 1: The limitation of CE loss for real-world self-driving system. The ground truth class of the pixel is car i^*. Two possible softmax predictions of the segmenters have the same probability at i^* position. Therefore, both predicted distributions have the same CE loss. However, the top prediction is preferable to the bottom, since the two predictions may result in different severity consequences.

[1]https://www.nytimes.com/2017/01/19/business/tesla-models-autopilot-fatal-crash.html

ognized as a car/bus, the accident could be avoided. Accordingly, it is necessary to investigate the severity of misclassifications in semantic segmentation method.

Figure 1 shows an example to illustrate different severity consequences of misclassifications by using CE loss. For this car image, there are two possible predictions, recognising the car as bus and road by Segment1 and Segment2 respectively. The CE loss cannot discriminate these two softmax probability histograms. With one-hot ground-truth label, CE loss only depends on the prediction probability of the true class. Actually, for self-driving system, the misclassified prediction (Car→Bus) is more expected than the misclassified prediction (Car→Road) in terms of severity. However, when using the CE loss, the classes are assumed to be independent of each other (Liu et al. 2018f). Therefore, the inter-class correlations are not properly exploited. Therefore, the inter-class correlation of (car, bus) should be closer than that of (car, road). This cannot be revealed by CE loss based models.

The importance-aware classification/segmentation (Chen, Gong, and Yang 2018) proposes to define some class groups based on the pre-defined importance of each class. For example, the car, truck, bus are in the most important group, road and sidewalks are in the less important group, and the sky is in the least important group. Then, a larger weight is assigned to the more important group to calculate the loss. Therefore, misclassifying a car into any other classes will receive larger punishment than misclassifying the sky into any other classes. Nevertheless, for a specific class, this method does not incorporate inter-class correlations between this class and any other class in the loss.

Based on the above mentioned analysis, we employ the Wasserstein loss as an alternative to empirical risk minimization. Specifically, we calculate the Wasserstein distance between a softmax prediction histogram and its one-hot encoded ground-truth label. By defining the ground metric based on misclassification severity, classification performance for each pixel can be measured related to inter-class correlations.

The ground metric can be predefined by regarding the severity structure as a priori, $e.g.$, the distance between car and road is larger than car and bus. We further investigate various forms of the ground metric in optimization perspective. In the one-hot label setting, the exact Wasserstein distance can be formulated as a soft-attention scheme of all prediction probabilities and is faster computed than other general Wasserstein distance. For the semantic segmentation with unsupervised domain adaptation using constrained non-one-hot pseudo-label, we can also resort to the fast approximate solution of Wasserstein distance.

The main contributions of this paper are summarized as:

• We propose to render reliable segmentation results for self-driving by considering the different severity of misclassification. The inter-class correlation is explicitly incorporated as a priori to form the ground metric in our Wasserstein training framework. The importance-aware methods can be viewed as a particular case by designing a specific ground metric.

• For either one-hot or constrained target label in self-training-based unsupervised domain adaption setting, we systematically conclude the possible fast solution when a non-negative linear, convex or concave increasing mapping function is applied in ground metric.

• We empirically validate the effectiveness and generality of the proposed Wasserstein training framework which achieves promising performance on multiple challenging benchmarks with different backbone models.

Related Work

Semantic Segmentation

Semantic segmentation provides a comprehensive description of the scene including object category, location and shape details (Badrinarayanan, Kendall, and Cipolla 2017). The deep learning revolution (Liu et al. 2018a; 2019c; Che et al. 2019; Liu et al. 2018b; 2018e) sparked wide interest in deep neural network-based semantic segmentation to replace the conventional methods (Liu et al. 2018c; 2017).

(Long, Shelhamer, and Darrell 2015) introduced a fully convolutional network for pixel or super pixel-wise classification. The conventional approaches usually employ CE loss (Liu et al. 2018d; 2018e; 2019d; 2019a), which equally evaluates the errors incurred by all image pixels/classes without taking into account the different severity-level of different mistakes (Chen, Gong, and Yang 2018).

The importance-aware methods (Chen, Gong, and Yang 2017) argue that the distinction between object/pixel importance need to be taken under consideration. The classes in Cityscapes are grouped as:
Group 4[most important]=$\{\mathrm{Person}, \mathrm{Car}, \mathrm{Truck}, \mathrm{Bus}, \cdots\}$;
Group 3=$\{\mathrm{Road}, \mathrm{Sidewalks}, \mathrm{Train}\}$;
Group 2=$\{\mathrm{Building}, \mathrm{Wall}, \mathrm{Fence}, \mathrm{Vegetation}, \mathrm{Terrain}\}$;
Group 1[least important]=$\{\mathrm{Sky}\}$.

To compute the sum of loss in all pixels, larger weights will be given to the more important group. Consequently, the misclassification of a pixel with ground truth label in group 4 will result in a larger loss than misclassifying the sky to the other classes.

Recently, not only powerful segmentation nets (Chen et al. 2017) have been developed but also the pose-processing strategies are proposed to improve the initial results (Liu, Lin, and Shen 2015). We note that these progress are orthogonal with our method and can be simply added to each other.

Wasserstein Distance

Wasserstein distance is a measure defined between probability distributions on a given metric space (Kolouri, Zou, and Rohde 2016). Recently, it has appealed to a great deal of attention in generative models etc (Arjovsky, Chintala, and Bottou 2017). Due to the significant amount of computation needed to solve the exact distance for general cases, usually, it is difficult to use Wasserstein distance as a loss function. Several methods propose to solve its approximate solution, whose complexity is still in $\mathcal{O}(N^2)$ (Cuturi 2013). (Frogner et al. 2015) applies it for the multi-class multi-label task

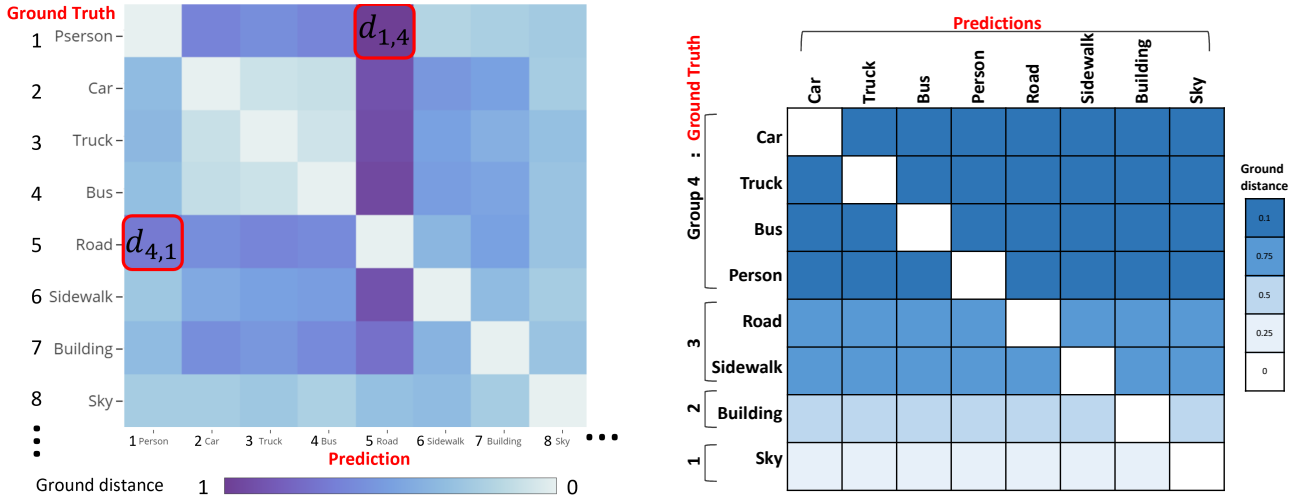

Figure 2: Left: a possible ground matrix for severity-aware segmentation. Right: the ground matrix as an alternative for importance-aware setting.

with a linear model. The fast computing of discrete Wasserstein distance is also closely related to SIFT (Cha and Srihari 2002) descriptor, hue in HSV or LCH space (Cha 2002) and sequence data (Su and Hua 2017).

Recently, several works propose to incorporate the Wasserstein distance as an alternative of cross-entropy loss in the context of deep learning. For example, (Liu et al. 2019e) use it for discrete and modulo classification, e.g., pose estimation. Targeting for the ordinal classification task, (Liu et al. 2019b) propose to incorporate the correlation of health risk-level in a line to alleviate the some what sophisticate neural stick-breaking post-processing in (Liu et al. 2018f).

Inspired by the above works, we further adapted this idea to the severity-aware estimation, and encoded the geometry of label space by means of the ground matrix. We demonstrate that Wasserstein loss can be computed by fast algorithm in our class structure.

Methodology

The target of this work is to learn a segmenter h_θ which is parameterized by θ. It is based on an autoencoder structure. Without loss of generality, suppose it projects a street view image $\mathbf{X} \in \mathbb{R}^{M_x \times M_x \times 3}$ to a prediction of semantic segmentation map $\mathbf{S} \in \mathbb{R}^{M_s \times M_s \times N}$, where N indicates the number of categories that pre-defined by the segmentation dataset. In addition, the spatial size of input $M_x \times M_x$ and output $M_s \times M_s$ are not necessarily the same. We note that the input also not have to be the shape of square in many segmenters. Suppose $\mathbf{s} = \{s_i\}_{i=1}^{N}$ is the pixel-wise prediction of $h_\theta(\mathbf{X})$, *i.e.*, the N classes probability normalized by softmax function. $i \in \{1, \cdots, N\}$ is the index of dimension (categories). Then we can perform learning over a hypothesis space $\mathcal{H}$ of h_θ. Given $\mathbf{X}$ and its target one-hot ground truth label $\mathbf{T} \in \mathbb{R}^{M_s \times M_s \times N}$, typically, learning is a process by empirical risk minimization to solve $\min_{h_\theta \in \mathcal{H}} \mathcal{L}(h_\theta(\mathbf{X}), \mathbf{T})$, with a loss $\mathcal{L}(\cdot, \cdot)$ acting as a surrogate of performance measure. In other words, it is the sum of pixel-wise error in $M_s \times M_s$ positions.

In the context of the self-driving risk minimization, we argue that a good loss function should reflect the properties of the importance of each class. Unfortunately, as the previous statement, cross-entropy (CE)-based loss treat the output dimensions independently (Frogner et al. 2015), ignoring the different severity of misclassification on label space, which is also not adaptive here. Besides, information divergence, Hellinger distance and $\mathcal{X}^2$ distance-based loss are also not the right choices, because it cannot distinguish between predictions.

Let define $\mathbf{t} = \{t_j\}_{j=1}^{N}$ as the target histogram distribution label that can be either one-hot or non-one-hot vector. Assume the class label possesses a ground metric $\mathbf{D}_{i,j}$, which measures the different severity of misclassifying i-th class pixel into j-th class pixel. There are N^2 possible potential outcomes $\mathbf{D}_{i,j}$ in a N class dataset and form a ground distance matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$ (Rüschendorf 1985). When $\mathbf{s}$ and $\mathbf{t}$ are both histograms, the discrete measure of exact Wasserstein loss is defined as

$$\mathcal{L}_{\mathbf{D}_{i,j}}(\mathbf{s}, \mathbf{t}) = \inf_{\mathbf{W}} \sum_{j=0}^{N-1} \sum_{i=0}^{N-1} \mathbf{D}_{i,j} \mathbf{W}_{i,j} \tag{1}$$

where $\mathbf{W}$ is the transportation matrix with $\mathbf{W}_{i,j}$ indicating the mass moved from the i^{th} point in source distribution to the j^{th} target position. A valid transportation matrix $\mathbf{W}$ satisfies:

$$\begin{aligned}
&\mathbf{W}_{i,j} \geq 0; \\
&\textstyle\sum_{j=0}^{N-1} \mathbf{W}_{i,j} \leq s_i; \\
&\textstyle\sum_{i=0}^{N-1} \mathbf{W}_{i,j} \leq t_j; \\
&\textstyle\sum_{j=0}^{N-1} \sum_{i=0}^{N-1} \mathbf{W}_{i,j} = \min(\sum_{i=0}^{N-1} s_i, \sum_{j=0}^{N-1} t_j).
\end{aligned}$$

In mathematics, the Wasserstein or Kantorovich Rubinstein metric or distance is a distance function defined be-

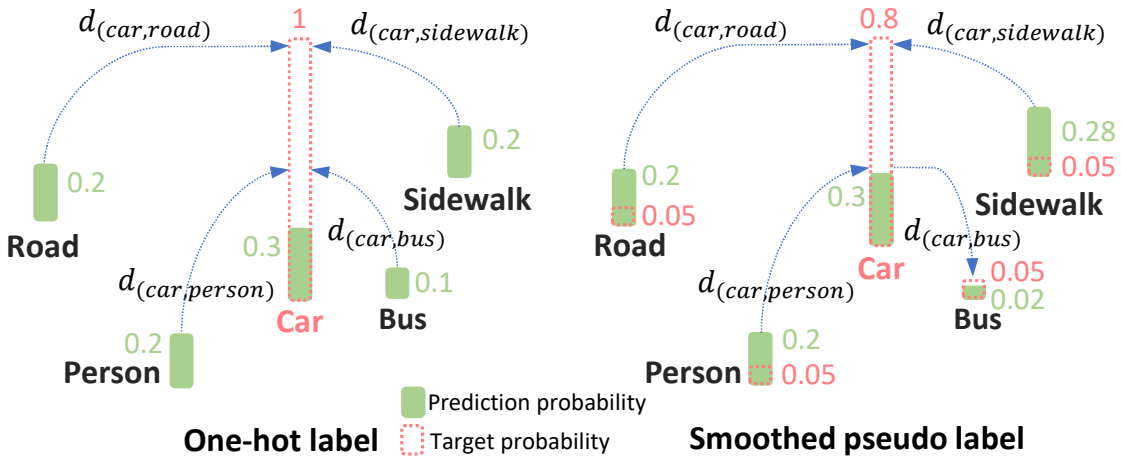

Figure 3: Left: The only possible transport plan in one-hot target case. Right: the transportation in smoothed pseudo label is more complicated, $e.g.$, car→bus.

tween probability distributions on a given metric space. Further, we proposes the Wasserstein distance as a loss function for unsupervised learning depends on a ground metric on the sample space of images, which is an effective distance for image retrieval, since it correlates with human perception. A possible ground distance matrix $\mathbf{D}$ which has considered different levels of importance is shown in Fig. 2.

The Wasserstein distance can be the same as the Earth mover's distance when two discrete histogram distributions with the same masses ($i.e.$, $\sum_{i=0}^{N-1} s_i = \sum_{j=0}^{N-1} t_j$) and choosing the symmetric distance $d_{i,j}$ as $\mathbf{D}_{i,j}$. However, our case is more general and different from this case. The entries in matrix $\mathbf{D}$ are not necessary to be symmetric with respect to the main diagonal. Note that the importance-aware matrix can be achieved by configuring the ground matrix as Fig. 2 right. The groups can be pre-defined by prior knowledge.

This setting is satisfactory for comparing the similarity of SIFT or hue, which do not use a neural network. The previous efficient algorithm usually holds only for $\mathbf{D}_{i,j} = d_{i,j}$. We propose to extend the ground metric in $\mathbf{D}_{i,j}$ as $f(d_{i,j})$, where f is a positive increasing function $w.r.t.$ $d_{i,j}$.

Wasserstein Training with One-hot Target

In the multi-class and one-label classification tasks, the one-hot labeling is a widely-used setting. The distribution of a target label probability is $\mathbf{t} = \delta_{j,j^*}$, where j^* is the ground truth class, δ_{j,j^*} is a Dirac delta, which equals to 1 for $j = j^*$[2], and 0 otherwise.

Theorem 1. *Assuming* $\sum_{j=0}^{N-1} t_j = \sum_{i=0}^{N-1} s_i$, *and* $\mathbf{t}$ *is a one-hot distribution and* $t_{j^*} = 1$(or $\sum_{i=0}^{N-1} s_i$)[3], *there is only one feasible optimal transport plan.*

Following the aforementioned criteria of $\mathbf{W}$, all masses have to be transferred to the cluster of the ground truth label j^*, as illustrated in Fig. 3. Then, the Wasserstein distance between softmax prediction $\mathbf{s}$ and one-hot target $\mathbf{t}$ degenerates to

$$\mathcal{L}_{\mathbf{D}^f_{i,j}}(\mathbf{s}, \mathbf{t}) = \sum_{i=0}^{N-1} s_i f(d_{i,j^*}) \tag{2}$$

We can extend the ground metric in $\mathbf{D}_{i,j}$ as $f(d_{i,j})$, where f can be a linear or increasing function proper, $e.g.$, p^{th} power of $d_{i,j}$ and Huber function. The exact solution of Eq. (2) can be computed with a complexity of $\mathcal{O}(N)$. The ground metric term $f(d_{i,j^*})$ works as the weights $w.r.t.$ s_i, which takes all classes into account following a soft attention scheme (Liu et al. 2018d). It explicitly encourages the probabilities distributing on the neighboring classes of j^*. Since each s_i is a function of the network parameters, differentiating $\mathcal{L}_{\mathbf{D}^f_{i,j}}$ $w.r.t.$ network parameters yields $\sum_{i=0}^{N-1} s'_i f(d_{i,j^*})$.

In contrast, the CE loss in one-hot setting can be formulated as $-1\log s_{j^*}$. Similar to the hard prediction scheme, only a single class prediction is considered resulting in a large information loss (Liu et al. 2018d). Besides, the regression loss with softmax prediction could be $f(d_{i^*,j^*})$, where i^* is the class with maximum prediction probability.

Wasserstein Training with Conservative Target

Deep self-training presents a powerful method for unsupervised domain adaptation in semantic segmentation, which involves an iterative process of predicting on target domain, taking the confident predictions as pseudo-labels for retraining. Obviously, self-training can put overconfident label belief on wrong classes and hence lead to deviated solutions with propagated errors because pseudo-labels can be noisy. (Zou et al. 2019) proposes to construct the soft Pseudo-label, smoothing the one-hot Pseudo-label to a conservative target distribution. With the conservative target label, the fast computation of Wasserstein distance in Eq. (2) does not apply.

[2]We use i, j interlaced for $\mathbf{s}$ and $\mathbf{t}$, since they index the same group of positions in a label set.

[3]We note that softmax cannot strictly guarantee the sum of its outputs to be 1 considering the rounding operation in practice. However, the difference of setting t_{j^*} to 1 or $\sum_{i=0}^{N-1} s_i$) is not significant in our experiments using the typical format of softmax output which has up to 8 decimal places precision.

	Group4							mIoU
	Person	Rider	Car	Truck	Bus	Motor	Bike	
SegNet(Badrinarayanan, Kendall, and Cipolla 2017)	62.8	42.8	89.3	38.1	43.1	35.8	51.9	57.0
+IAL(Chen, Gong, and Yang 2017)	84.1	46.0	91.1	75.9	65.0	22.2	**65.3**	65.7
+$\mathcal{L}_{d_{i,j}}$	86.4	48.7	92.8	78.5	68.2	40.2	62.8	67.4
+$\mathcal{L}_{\mathbf{D}^2_{i,j}}$	87.5	**50.2**	**93.4**	**79.8**	69.5	**42.0**	64.3	**68.0**
+$\mathcal{L}_{\mathbf{D}^{H\tau}_{i,j}}$	**87.6**	49.8	93.2	79.5	**70.3**	41.6	63.6	67.9
ENet(Paszke et al. 2017)	65.5	38.4	90.6	36.9	50.5	38.8	55.4	58.3
+IAL(Chen, Gong, and Yang 2017)	87.7	41.3	92.4	**73.5**	76.2	24.1	69.7	67.5
+$\mathcal{L}_{d_{i,j}}$	90.7	48.7	95.5	70.8	75.3	46.2	73.3	69.1
+$\mathcal{L}_{\mathbf{D}^2_{i,j}}$	90.9	**49.6**	**96.8**	71.4	77.6	**46.3**	**75.1**	69.3
+$\mathcal{L}_{\mathbf{D}^{H\tau}_{i,j}}$	**90.1**	49.5	**96.8**	72.6	**77.8**	46.2	75.0	**69.5**

Table 1: The comparison results of various methods of Cityscapes Group 4 with SegNet and ENet backbone.

Regarding it as a general case of Wasserstein distance and solving its closed-form result with a complexity higher than $\mathcal{O}(N^3)$ cannot satisfy the speed requirement of the loss function. Therefore, a possible solution is to get an approximate result with complexity in $\mathcal{O}(N^2)$. (Cuturi 2013) proposes an efficient approximation of both the transport matrix and the subgradient of the loss, which is essentially a matrix balancing problem that well-studied in numerical linear algebra (Knight and Ruiz 2013). (Cuturi 2013) uses the well-known efficient iterative Sinkhorn-Knopp algorithm.

Monotonic Increasing f w.r.t. $d_{i,j}$ as Ground Metric

Practically, f in $\mathbf{D}^f_{i,j} = f(d_{i,j})$ can be a positive increasing function $w.r.t.\ d_{i,j}$. For simplicity the linear function is satisfactory for comparing the similarity of SIFT or hue (Rubner, Tomasi, and Guibas 2000), which even does not involve neural network optimization.

Convex Function $w.r.t.\ d_{i,j}$ as Ground Metric

Furthermore, we can extend the ground metric as a nonnegative increasing and convex function of $d_{i,j}$. Here, we give some measures[4] using the typical convex ground metric function.

$\mathcal{L}_{\mathbf{D}^\rho_{i,j}}(\mathbf{s}, \bar{\mathbf{t}})$, the Wasserstein measure using d^ρ as the ground metric with $\rho = 2, 3, \cdots$. The case $\rho = 2$ is equivalent to the Cramér distance (Rizzo and Székely 2016). Note that the Cramér distance is not a distance metric proper. However, its square root is.

$$\mathbf{D}^\rho_{i,j} = d^\rho_{i,j} \tag{3}$$

$\mathcal{L}_{\mathbf{D}^{H\tau}_{i,j}}(\mathbf{s}, \bar{\mathbf{t}})$, the Wasserstein measure using a Huber cost function with a parameter τ.

$$\mathbf{D}^{H\tau}_{i,j} = \begin{cases} d^2_{i,j} & \text{if } d_{i,j} \leq \tau \\ \tau(2d_{i,j} - \tau) & \text{otherwise.} \end{cases} \tag{4}$$

Concave Function $w.r.t.\ d_{i,j}$ as Ground Metric.

In real world applications, it may not meaningful to choose the ground metric as the nonnegative, increasing and concave function $w.r.t.\ d_{i,j}$. Noticing that the computing time of obtaining an closed-form solution in the conservative target label case is usually not acceptable. While the step function $f(t) = \mathbf{1}_{t \neq 0}$ (one everywhere except at 0) could be a special case. It can achieve the exact solution with significantly less complexity (Villani 2003). Assuming that the $f(t) = \mathbf{1}_{t \neq 0}$, the Wasserstein metric between two normalized discrete histograms on N bins is simplified to the ℓ_1 distance.

$$\mathcal{L}_{\mathbf{1}_{d_{i,j} \neq 0}}(\mathbf{s}, \bar{\mathbf{t}}) = \frac{1}{2} \sum_{i=0}^{N-1} |\mathrm{s}_i - \bar{\mathrm{t}}_i| = \frac{1}{2} ||\mathbf{s} - \bar{\mathbf{t}}||_1 \tag{5}$$

where $|| \cdot ||_1$ is the discrete ℓ_1 norm. Unfortunately, its efficient computation of closed-form solution is at the cost of losing its ability to differentiate different misclassifications.

Experiments

We show the implementation details and experimental results on two typical self-driving benchmarks ($i.e.$, Cityscapes (Cordts et al. 2016) and CamVid (Brostow, Fauqueur, and Cipolla 2009)). To illustrate the effectiveness of each setting choice and their combinations, we give a series of elaborate ablation studies along with the standard measures. All of the networks are pre-trained with CE loss as their vanilla version. The intersection-over-union (IoU) is defined as:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{6}$$

where TP, FP, and FN denote the numbers of true positive, false positive, and false negative pixels, respectively. Moreover, the mean IoU is the average of IoU among all classes.

[4]We refer to "measure", since a ρ^{th}-root normalization is required to get a distance (Villani 2003), which satisfies three properties: positive definiteness, symmetry and triangle inequality.

	Group3			Group4			mIoU
	Road	Sidewalk	Sign	Car	Pedestrian	Bike	
FCN(Long, Shelhamer, and Darrell 2015)	98.1	89.5	25.1	84.5	64.6	38.6	69.6
+IAL(Chen, Gong, and Yang 2017)	96.3	91.8	21.5	82.2	69.5	57.6	71.2
$+\mathcal{L}_{d_{i,j}}$	98.5	93.2	28.3	87.4	71.3	60.0	72.4
$+\mathcal{L}_{\mathbf{D}^2_{i,j}}$	**98.7**	94.6	**29.7**	89.5	73.4	**60.7**	**72.8**
$+\mathcal{L}_{\mathbf{D}^{H_T}_{i,j}}$	98.5	**95.0**	29.5	**89.7**	**73.5**	60.6	**72.8**

Table 2: The comparison results of various methods on the Group 3/4 of CamVid dataset using FCN as backbone.

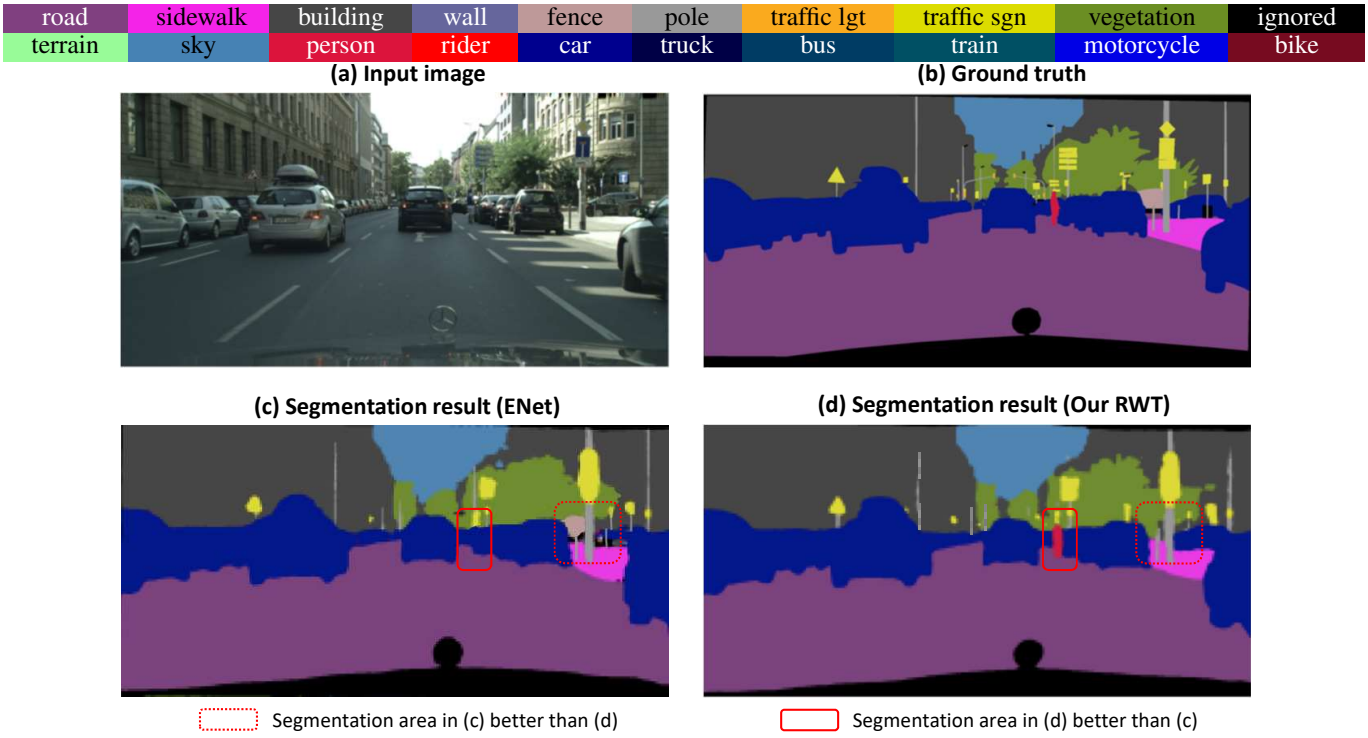

Figure 4: Representative semantic segmentation result of ENet and our Wasserstein training with ENet backbone on Cityscapes dataset. The two image has the same mIoU but the misclassification of the person may lead to more severity result.

Importance-aware SS with One-hot Label

To achieve the importance-aware SS, we first pre-define our ground matrix as Fig. 2. Following the setting in IAL (Chen, Gong, and Yang 2017; 2018), we choose the SegNet (Badrinarayanan, Kendall, and Cipolla 2017) and ENet (Paszke et al. 2017) to be our backbone. We then use IAL and our Wasserstein loss to replace the conventional CE loss in their vanilla version.

For training/validation/testing, the recent Cityscapes dataset contains 2975/500/1525 images respectively. The 19 classes that are most commonly used are selected and grouped as IAL. Table 1 shows that the class in group 4 are segmented with higher IoU when considering the importance of each class. Our Wasserstein loss normally outperforms 2% than IAL, especially apply the convex function $w.r.t.d_{i,j}$. The improvements $w.r.t.$ Motor are more than 15% over IAL.

The CamVid dataset contains 367/26/233 images for training/validation/testing respectively .We use the same setting and measurements as IAL and report the results in the table 2 for a fair comparison. We note that fine-tuning a public available trained FCN segmenter (Long, Shelhamer, and Darrell 2015) with Wasserstein loss is 1.5$\times$ faster than the training of IAL. While the IoU of some relatively unimportant classes may drop, this will have limited impact on driving safety. By introducing a stricter-than-usual objective beyond simple CE loss, we can keep the mean IoU of all classes comparable or even improved. To intuitively present the effectiveness, we provide a representative segmentation example in Fig. 4.

According to above qualitative and quantitative results, we conclude that the proposed importance-aware Wasser-

	Group4							mIoU
	Person	Rider	Car	Truck	Bus	Motor	Bike	
LRENT(Zou et al. 2019)	61.7	27.4	83.5	27.3	37.8	30.9	41.1	46.5
$\mathcal{L}_{d_{i,j}}$	65.4	33.7	88.5	36.2	44.8	39.3	48.4	46.8
$\mathcal{L}_{\mathbf{D}^2_{i,j}}$	65.7	34.0	88.9	36.7	45.3	39.6	49.1	47.0
$\mathcal{L}_{\mathbf{D}^{H\tau}_{i,j}}$	**66.2**	**34.7**	**89.5**	**37.1**	**46.0**	**40.8**	**50.5**	**47.3**

Table 3: The comparison results of various methods on the Group4 of GTA5→Cityscapes unsupervised domain adaptation using DeeplabV2 as backbone.

stein training can improve the segmentation quality of the important objects with a large margin in terms of mIoU metric. Therefore, it is quite suitable for the application of self-driving.

Importance-aware SS with Conservative Label

We further test our method for unsupervised domain adaptation with constrained self-training, i.e., label entropy regularizer (LRENT) (Zou et al. 2019). We compute the approximate Wasserstein distance as the loss. Table 3 shows the performance of GTA5→Cityscapes adaptation and outperforms the CE loss-based LRENT by more than 5% in these important classes consistently. Because the Huber function is more robust to the label noise which is common for the pseudo label in self-learning method. The improvements of $\mathcal{L}_{\mathbf{D}^{H\tau}_{i,j}}$ over $\mathcal{L}_{\mathbf{D}^2_{i,j}}$ are more significant than the one-hot case. This task also indicates that our approach can be considered as a general alternative objective of CE loss. Also it can be employed in a plug and play fashion.

Conclusions

Targeting for the safety driving of self-driving vehicles or robotics, we propose to implement a simple yet effective loss function for semantic segmentation based on the Wasserstein distance. It is an effective alternative of cross-entropy loss for empirical risk minimization. The importance-correlation is given by a ground metric, which can be predefined with expert knowledge. In the Wasserstein training, the importance-ignored task can be regarded simply as a special case of our importance-aware setting. Its effectiveness can be further boosted by using a convex function (e.g., square and Huber) $w.r.t.d_{i,j}$. The fast closed form solution is existed in the one-hot case, and can be used as loss function directly. Besides, the fast approximate solution can also be applied to the case with conservative label which widely exist in self-learning based unsupervised domain adaptation as well. We give extensive experiments to evidence its effectiveness and Wasserstein training achieve the state of the art performance in importance-aware tasks, and also improve the general metric mIoU with a more strictly optimization objective. Although it is originally designed for semantic segmentation tasks, we argue that our framework should has similar applicability to other problems with discrete labels that have different importance-level of misclassfication. In the future, we are planing to apply it on object detection (Ding et al. 2020), and extend it to severity-aware semantic segmentation (Liu et al. 2020).

Acknowledgement

The funding support from National Institute of Health (NIH), National Institute of Neurological Disorders and Stroke (NINDS) (NS061841, NS095986), Fanhan Technology, Youth Innovation Promotion Association, CAS (2017264), Innovative Foundation of CIOMP, CAS (Y586320150) and Hong Kong Government General Research Fund GRF (Ref. No.152202/14E) are greatly appreciated.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI* 39(12):2481–2495.

Brostow, G. J.; Fauqueur, J.; and Cipolla, R. 2009. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters* 30(2):88–97.

Cha, S.-H., and Srihari, S. N. 2002. On measuring the distance between histograms. *Pattern Recognition* 35(6):1355–1370.

Cha, S.-H. 2002. A fast hue-based colour image indexing algorithm. *Machine Graphics & Vision International Journal* 11(2/3):285–295.

Che, T.; Liu, X.; Li, S.; Ge, Y.; Zhang, R.; Xiong, C.; and Bengio, Y. 2019. Deep verifier networks: Verification of deep discriminative models with deep generative models. In *ArXiv*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI* 40(4):834–848.

Chen, B.-k.; Gong, C.; and Yang, J. 2017. Importance-aware semantic segmentation for autonomous driving system. In *IJCAI*, 1504–1510.

Chen, B.; Gong, C.; and Yang, J. 2018. Importance-aware semantic segmentation for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems* 20(1):137–148.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213–3223.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, 2292–2300.

Ding, P.; Kuijper, A.; Huang, S.; Liu, X.; and Jia, P. 2020. Light r-cnn: A simple but efficient r-cnn based on reverse residual model. *arXiv preprint arXiv:1701.07875*.

Frogner, C.; Zhang, C.; Mobahi, H.; Araya, M.; and Poggio, T. A. 2015. Learning with a wasserstein loss. In *NIPS*, 2053–2061.

Knight, P. A., and Ruiz, D. 2013. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* 33(3):1029–1047.

Kolouri, S.; Zou, Y.; and Rohde, G. K. 2016. Sliced wasserstein kernels for probability distributions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5258–5267.

Li, X.; Liu, Z.; Luo, P.; Change Loy, C.; and Tang, X. 2017. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3193–3202.

Liu, X.; Kong, L.; Diao, Z.; and Jia, P. 2017. Line-scan system for continuous hand authentication. *Optical Engineering* 56(3):033106.

Liu, X.; Ge, Y.; Yang, C.; and Jia, P. 2018a. Adaptive metric learning with deep neural networks for video-based facial expression recognition. *Journal of Electronic Imaging* 27(1):013022.

Liu, X.; Kumar, B. V.; Ge, Y.; Yang, C.; You, J.; and Jia, P. 2018b. Normalized face image generation with perceptron generative adversarial networks. In *2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, 1–8.

Liu, X.; Li, Z.; Kong, L.; Diao, Z.; Yan, J.; Zou, Y.; Yang, C.; Jia, P.; and You, J. 2018c. A joint optimization framework of low-dimensional projection and collaborative representation for discriminative classification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, 1493–1498.

Liu, X.; Vijaya Kumar, B.; Yang, C.; Tang, Q.; and You, J. 2018d. Dependency-aware attention control for unconstrained face recognition with image sets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 548–565.

Liu, X.; Zou, Y.; Kong, L.; Diao, Z.; Yan, J.; Wang, J.; Li, S.; Jia, P.; and You, J. 2018e. Data augmentation via latent space interpolation for image classification. In *24th International Conference on Pattern Recognition (ICPR)*, 728–733.

Liu, X.; Zou, Y.; Song, Y.; You, J.; and K Vijaya Kumar, B. 2018f. Ordinal regression with neuron stick-breaking for medical diagnosis. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, 0–0.

Liu, X.; Guo, Z.; Li, S.; You, J.; and B.V.K, K. 2019a. Permutation-invariant feature restructuring for correlation-aware image set-based recognition. In *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Liu, X.; Han, X.; Qiao, Y.; Ge, Y.; Li, S.; and Lu, J. 2019b. Unimodal-uniform constrained wasserstein training for medical diagnosis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 0–0.

Liu, X.; Kumar, B. V.; Jia, P.; and You, J. 2019c. Hard negative generation for identity-disentangled facial expression recognition. *Pattern Recognition* 88:1–12.

Liu, X.; Li, S.; Kong, L.; Xie, W.; Jia, P.; You, J.; and Kumar, B. 2019d. Feature-level frankenstein: Eliminating variations for discriminative recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 637–646.

Liu, X.; Zou, Y.; Che, T.; Jia, P.; You, J.; and B.V.K, K. 2019e. Conservative wasserstein training for pose estimation. In *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Liu, X.; Bai, S.; Li, S.; and You, J. 2020. Reinforced wasserstein training for severity-aware semantic segmentation in autonomous driving. *arXiv preprint*.

Liu, F.; Lin, G.; and Shen, C. 2015. Crf learning with cnn features for image segmentation. *Pattern Recognition* 48(10):2983–2992.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

Paszke, A.; Chaurasia, A.; Kim, S.; and Culurciello, E. 2017. Enet: A deep neural network architecture for real-time semantic segmentation. *ICLR*.

Rizzo, M. L., and Székely, G. J. 2016. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics* 8(1):27–38.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *IJCV* 40(2):99–121.

Rüschendorf, L. 1985. The wasserstein distance and approximation theorems. *Probability Theory and Related Fields* 70(1):117–129.

Su, B., and Hua, G. 2017. Order-preserving wasserstein distance for sequence matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2906–2914.

Villani, C. 2003. *Topics in optimal transportation*. American Mathematical Soc.

Yang, C.; Song, Y.; Liu, X.; Tang, Q.; and Kuo, C.-C. J. 2018. Image inpainting using block-wise procedural training with annealed adversarial counterpart. *arXiv preprint arXiv:1803.08943*.

Zou, Y.; Yu, Z.; Liu, X.; Wang, J.; and B.V.K., K. 2019. Confidence regularized self-training. *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.