Wasserstein Loss based Deep Object Detection

Yuzhuo Han[1†], Xiaofeng Liu[2†], Zhenfei Sheng[4], Yutao Ren[5], Xu Han[2,6],
Jane You[7], Risheng Liu[3], Zhongxuan Luo[1]
[1]School of Mathematical Sciences, Dalian University of Technology
[2]Beth Israel Deaconess Medical Center, Harvard Medical School, Harvard University
[3]School of Software Technology and the International School of Information Science Engineering, Dalian University of Technology
[4]College of Photonic and Electronic Engineering, Fujian Normal University
[5]Wuhan University of Technology [6]John Hopkins Uniersity
[7]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China.
† contribute equally.

Abstract

Object detection locates the objects with bounding boxes and identifies their classes, which is valuable in many computer vision applications (e.g. autonomous driving). Most existing deep learning-based methods output a probability vector for instance classification trained with the one-hot label. However, the limitation of these models lies in attribute perception because they do not take the severity of different misclassifications into consideration. In this paper, we propose a novel method based on the Wasserstein distance called Wasserstein Loss based Model for Object Detection (WLOD). Different from the commonly used distance metric such as cross-entropy (CE), the Wasserstein loss assigns different weights for one sample identified to different classes with different values. Our distance metric is designed by combining the CE or binary cross-entropy (BCE) with Wasserstein distance to learn the detector considering both the discrimination and the seriousness of different misclassifications. The misclassified objects are identified to similar classes with a higher probability to reduce intolerable misclassifications. Finally, the model is tested on the BDD100K and KITTI datasets and reaches state-of-the-art performance.

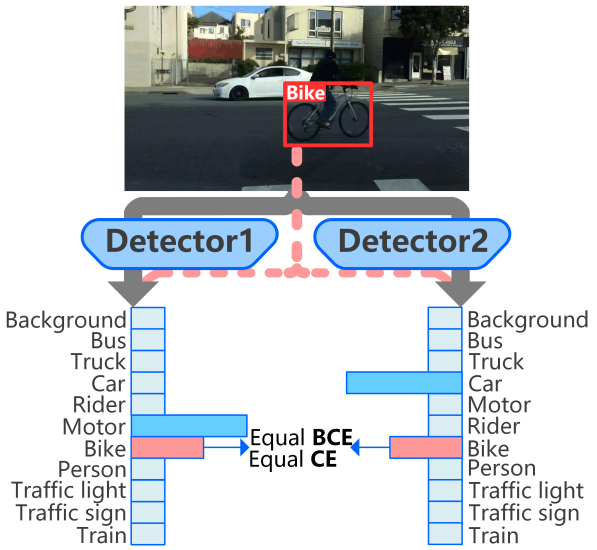

Figure 1. The limitation of BCE/CE loss for object classification. The ground-truth class of the object is 'Bike'. The predicted probability of 'Bike' by Detector 1 and Detector 2 is the same. Therefore, these two detectors have the same BCE/CE loss. However, Detector 1 is preferable to Detector 2, because these two predictions may result in different severity consequences.

1. Introduction

Object detection is a fundamental task in the computer vision field aiming at detecting instances from the surveillance video images. It is meaningful for instance segmentation [40], object tracking, pose estimation, and drone scene analysis etc [21, 25]. A accurate object detection system can be useful in autonomous driving, surveillance, and blind guiding. The framework for object detection consists of bounding boxes proposal, extracting local feature for each bounding box, and classifying objects according to the feature of each bounding box proposal. Existing object detection model focus on detection of certain class instances (e.g. bike, car, bus, person, dog, and cat etc). Attributed by the deep learning [14, 17, 1, 13, 24, 28, 22, 26, 23, 16], object detection task reaches a high-level detection accu-

racy, which is close to the demand of application. Despite many works have been done to improve the detection model, the object detection task faces many challenges such as scale changes, viewpoints, illuminations, and rotations. In addition, the deep learning based method is too computationally intensive and high-demand in hardware. Hence, it has drawn increasing amounts of attention in recent years[36, 10, 12]. Although much work has been performed to improve the detection model, the object detection task still faces many challenges, such as scale changes, illuminations, and rotations. Attributed to the deep learning-based method, the object detection task reaches high detection accuracy, which is closer to the demand of applications. Recently, deep learning-based methods have been used successfully to handle object detection tasks, and many works have been published, including spatial pyramid pooling (SPP) network [8], Fast region-based convolutional network (Fast RCNN) [4], Faster RCNN[32], and YOLO [30]. Most object detection methods neglect the severity of different misclassifications.

As shown in Fig. 1, a 'Bike' in the surveillance image is detected and classified by two detectors. Because these two detectors classify the 'Bike' into the correct category with the same probability value, the same classification loss is obtained if they use the CE/BCE loss function. Nevertheless, classifying the 'Bike' as a 'Car' (Detector 2) would result in the self-driving car making an action not suitable for the current situation. However, classifying the 'Bike' as a 'Motor' (Detector 1) would not lead to serious consequences. Therefore, Detector 1 is safer than Detector 2. Existing methods do not discriminate these two misclassifications. In this paper, we focus on avoiding unacceptable misclassifications caused by CE/BCE loss-based object detection methods.

Based on the problem insights above, we employ the Wasserstein loss as an alternative to empirical risk minimization to improve classification accuracy[27, 29, 18, 15, 19, 20]. 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 the appearance similarity and misclassification severity (e.g., the distance between 'Bike' and 'Car' is larger than 'Bike' and 'Motor'), classification performance for each object can be measured related to inter-class correlations. 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 distances.

The main contributions of this paper are summarized as follows:

- In this paper, we regard classification in object detection as attribute perception problem, which can identify the severity of different misclassifications and guide the deep network to learn more essential attributes of objects for classification.
- We proposed a novel method for the formulation of the Wasserstein loss, which detect the objects from two level. The first level will discriminate objects from the basic attributes like vehicle and person. The second level discriminate the object for the detail.
- Extensive experiments are conducted on challenging benchmarks to validate the effectiveness and generality of the proposed Wasserstein training framework which achieves a promising performance with different backbone models.

2. Related Work

Many works have been published in the past two decades. Deep learning [7, 6, 37, 38] is successfully used in many computer vision task. Object detection is one of the outstanding application of deep network. It has improve the object detection significantly and many methods [4, 10, 30] have been proposed. Girshick et al. [5] proposed the Regions with CNN (RCNN) features structure, which is the first successful deep learning model for object detection. It greatly improved the performance of mean Average Precision(mAP). This method generates region proposals by Selective Search [34].The CNNs is used to extract local region features of a fixed-length for classification by SVM of each class. However, almost of the previous works are based on cross entropy loss for optimization and do not consider the difference of misclassification.

He et al. [5] designed Spatial Pyramid Pooling(SPP) method to deal with the problem of the size of input images and proposed a SPPNet. It broke the constrain of CNN models that the size of input images must be the changeless(e.g. 224x224 in AlexNet [9]). It sufficiently improves the efficiency of feature extraction compared with RCNN. SVM is also selected as classifier in SPPNet. Later, Girshick [4]improved the RCNN method to deal with the time cost problem. They proposed the Fast RCNN model which also use selective search to generate a set of object proposals, but it extract the whole image feature by CNNs instead of extracting the feature for every object proposals. Then it find the corresponding region of interest and divide the region into a H×W grid to do RoI pooling which ensures that the features of each region of equal length. It worth mentioning that Fast RCNN use cross entropy loss to do the classification task. Ren et al. [32] proposed the Faster RCNN based on the RCNN method, which further improved the speed of the deep learning based object detection model. The Faster RCNN is a end-to-end learning framework by combining the process of proposals extraction, classification and bounding box regression benefitting from Region Proposal Network(RPN) and PoI pool-

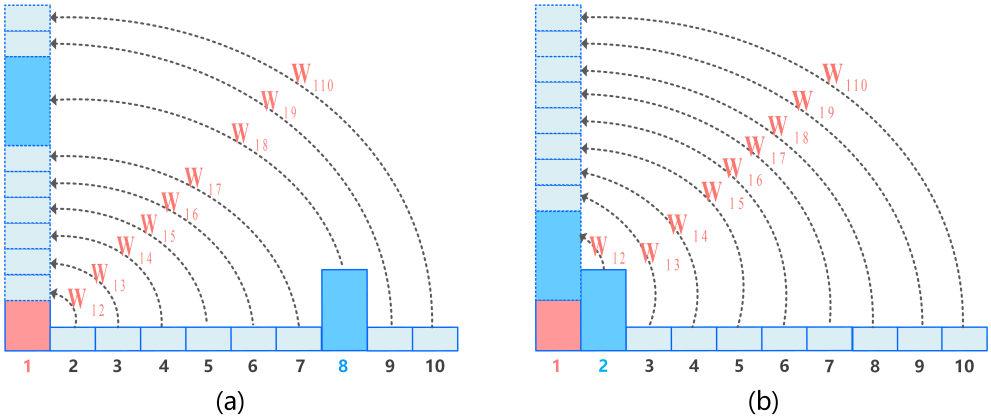

Figure 2. Illustration of the Wasserstein distance. $\mathbf{W}$ implies the distance between categories helps the Wasserstein distance to measure the appearance similarity of different misclassifications [27, 20, 19].

ing. RPN significantly improve the speed of detect region proposals. Faster RCNN also use cross entropy to classify the object of a certain classs. Lin et al. [10] proposed a Feature Pyramid Networks(FPN) based deep network. This framework includes bottom-up pathway, top down pathway and lateral connection. Top-down pathway and lateral connection make it easier to detect multi-scale objects by using deeper features and shallow layer features simultaneously. Faster RCNN with FPN significantly improved the performance of Faster RCNN itself. Joseph Redmon et al. [30] proposed the YOLOv1 deep network which is the first one-stage real-time detector. It divides the image into regions and use one neural network to generate bounding boxes and classify the object for each region at the same time. It use a regression model to classify the object category and predict the bounding box coordinates. Liu et al. proposed a Single Shot MultiBox Detector (SSD) to improve the training and test speed. It predicts the offsets of bounding box and object categories for default boxes of each feature map cell with different ratios and scales. It reached a similar performance with YOLOv3 [31]. Lin et al. [11] have proposed the RetinaNet method which has significantly improved one-stage detection accuracy by introducing a novel loss called "focal loss". Focal loss is committed to solving the problem caused by foreground-background class imbalance and hard examples in training set.

3. Methodology

3.1. Formulation for Object Detection

Given image $\mathbf{I}$ with size $W \times H \times 3$, to solve the object detection problem one should find an effective detector $h(\mathbf{I}, \mathbf{\Theta})$, where $\mathbf{\Theta}$ denotes the parameters. The output of the detector is $\mathbb{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots\ldots, \mathbf{o}_n\}$, and $\mathbf{o}_k = [\mathbf{t}_k; c_k; \mathbf{p}_k]$, where $\mathbf{t}_k = (x_k, y_k, w_k, h_k)$ represents the location of the k-th predicted target, c_k denotes the corresponding confidence score, and $\mathbf{p}_k = [p_{k,0}, p_{k,1}, \cdots, p_{k,N}] \in \mathbb{R}^{1\times N}$ represents a discrete probability distribution. $\mathbf{p}_k$ is activated by an activation function to predict the object category in this bounding box. N is the number of categories in a certain detection dataset.

In this paper, we study the classification problem in the object detection task. Each object in an image is labeled with a one-hot vector for classification and a tuple $\mathbf{t}_k = (x_k, y_k, w_k, h_k)$ for the location. A multi-task loss function can be formulated based on the classification and location labels to train the deep network jointly to locate the bounding boxes and classify the objects:

$$\begin{aligned}\mathbf{L}(\mathbb{O}) =& L_{locat}(\mathbf{T}, \mathbf{T}^*) + L_{conf}(\mathbf{c}^*, \mathbf{c}) + \\ & L_{class}(\mathbf{P}, \mathbf{P}^*),\end{aligned} \tag{1}$$

where $\mathbf{c}^* = \{c_i^*\}$, $c_i^* = 1$ or 0 denotes the ground-truth confidence score, indicating whether there is an object in the predicted bounding box. $\mathbf{T}^* = \{\mathbf{t}_i^*\}$; $\mathbf{t}_i^*$ denotes the ground-truth bounding-box regression offsets. $\mathbf{P}^* = \{\mathbf{p}_i^*\}$ denotes the ground-truth object category. Assuming that the candidate bounding box $\mathbf{t}_i$ is a positive sample belonging to category u, the corresponding one-hot label is represented as $\mathbf{p}_i^* = [0, \cdots, 0, p_{i,u}^*, 0, \cdots, 0]$, $p_{i,u}^* = 1$. The loss function in Eq. (1) consists of three items: the location loss, confidence loss, and classification loss. YOLOv3 takes the sum of the BCE loss as the L_{class} in Eq. (1), and SSD adopts the CE loss. Unfortunately, they treat the output dimensions independently [2], ignoring the different severities of the misclassification and appearance similirity in the label space.

3.2. Wasserstein Distance-based Loss

We formulate the classification problem for object detection based on the assumption that the predicted probability value of the categories, which are more similar to the ground truth, should be larger than the others. The discrete

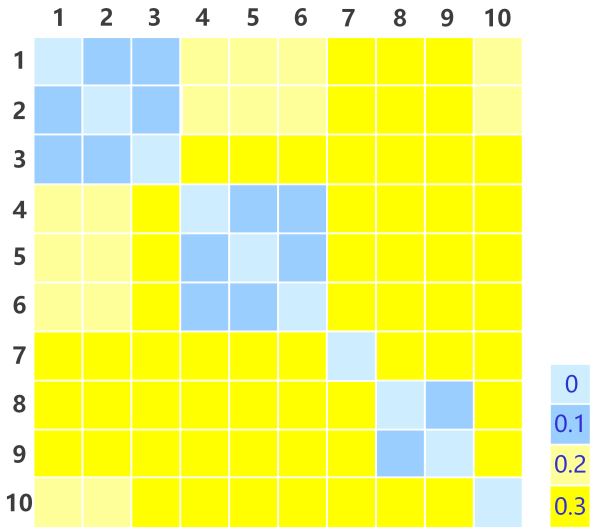

Figure 3. The pre-defined ground matrix for the BDD100K dataset.

Wasserstein distance between two histograms $\mathbf{p}$ and $\mathbf{t}$ is defined as:

$$\mathbf{D}_{\mathbf{W}}(\mathbf{s},\mathbf{t}) = \inf_{\mathbf{\Gamma}} \sum_{m=1}^{N}\sum_{n=1}^{N} \mathbf{W}_{m,n}\mathbf{\Gamma}_{m,n}, \tag{2}$$

where $\mathbf{\Gamma}_{N\times N}$ is the transportation matrix with $\mathbf{\Gamma}_{m,n}$ indicating the mass moved from the i-th position in the source distribution to the j-th position in the target distribution. $\mathbf{W}_{N\times N}$ [33] denotes the ground-distance matrix, and the ground metric $\mathbf{W}_{m,n}$ measures the cost of transporting a unit from the i-th position to the j-th position. $\mathbf{s}$ and $\mathbf{t}$ are often referred as the suppliers set and the consumers set, respectively. We can view $\mathbf{s}$ as the predicted probability distribution for classification and take $\mathbf{t}$ as the ground truth. The valid transportation matrix $\mathbf{\Gamma}$ satisfies:

$$\begin{aligned}
&\mathbf{\Gamma}_{m,n} \geq 0,\\
&\textstyle\sum_{n=1}^{N} \mathbf{\Gamma}_{m,n} \leq s_m;\\
&\textstyle\sum_{m=1}^{N} \mathbf{\Gamma}_{m,n} \leq t_n;\\
&\textstyle\sum_{m=1}^{N}\sum_{n=1}^{N} \mathbf{\Gamma}_{m,n} = \min(\sum_{m=1}^{N} p_m, \sum_{n=1}^{N} t_n).
\end{aligned}$$

The Wasserstein distance between $\mathbf{s}$ and $\mathbf{t}$ is the minimum transportation cost that satisfies the constraints above. In mathematics, the Wasserstein metric is a distance function defined between probability distributions in a given metric space. The Wasserstein distance can be the same as the Earth mover's distance when two discrete histogram distributions have the same masses ($i.e.$, $\sum_{i=1}^{N} s_i = \sum_{i=1}^{N}$) and symmetric matrix $\mathbf{W}$. For object classification, we can define the ground-distance matrix $\mathbf{W}$ by dividing the classes into different groups using prior knowledge and measuring the distance between different groups using a Gaussian filter.

Inspired by the Wasserstein distance, we define the loss function as:

$$\begin{aligned}
\mathbf{L}(\mathbb{O}) = L_{location}(\mathbf{T},\mathbf{T}^*) + L_{conf}(\mathbf{c}^*,\mathbf{c}) + \\
L_{class}(\mathbf{P},\mathbf{P}^*) + \lambda L_{WD}(\mathbf{P},\mathbf{P}^*,\mathbf{W}),
\end{aligned} \tag{3}$$

where $L_{WD}(\mathbf{P},\mathbf{P}^*,\mathbf{W})$ is called the Wasserstein loss.

Theorem 1. *Assume that* $\mathbf{p}_i$ and $\mathbf{p}_i^*$ *are both one-hot histogram distribution*[1], *there is only one feasible optimal transport plan* [35].

According to the criteria for $\mathbf{W}$, all masses must be transferred to the cluster of the ground-truth label position [35]. Then, the Wasserstein distance between softmax prediction $\mathbf{s}$ and one-hot target $\mathbf{t}$ in Eq. (**??**) can be written as an inner product as:

$$L_{WD}(\mathbf{P}_i,\mathbf{P}_i^*,\mathbf{W}) = \sum_i < \mathbf{p}_i - \mathbf{p}_i^*, \mathbf{W}_{u_i} >, \tag{4}$$

where $\mathbf{W}_{u_i}$ denotes the u_i-th row of $\mathbf{W}$.

The ground metric term works as the weights $w.r.t.$ mismatch of two histograms $\mathbf{p}_i - \mathbf{p}_i^*$, which takes all classes into account following a soft attention scheme [14]. It explicitly encourages the probabilities distributed in the neighboring classes of the ground-truth class in the pre-defined ground matrix space. In contrast, the CE loss in the one-hot setting can be a hard attention scheme; only a single class prediction is considered resulting in a large information loss [14].

3.3. Deep Structure of the Proposed Model

The image to be processed is first resized to 416×416 and goes through a convolutional neural network (Darknet-53) for feature extraction. Darknet-53 adopts successive convolutional layers with filters of size 3×3 and 1×1. This network has 52 convolutional layers, 23 residual layers, and a fully connected layer. Three candidate bounding boxes for each anchor represented by three feature maps in the last layer with size $M\times M\times3\times(4+1+N)$, $(M = 13, 26, 52)$ are extracted. Specifically, there are 4 offset parameters for location, 1 parameter for confidence, and a vector of length N for object classification. The cells related to the object category are used to calculate the L_{class} and L_{WD} in Eq. (3).

Table 1. Comparison of mAP for the BDD100K dataset.

	YOLOv3	$W_{0.5}$-y	W_1-y	SSD	$W_{0.5}$-s
mAP	25.8	**28.7**	27.0	33.9	**34.3**

Table 2. Comparison of mAP for the KITTI dataset.

	YOLOv3	$W_{0.5}$-y	W_1-y	SSD	$W_{0.5}$-s	W_1-s
mAP	68.3	69.2	**69.4**	72.8	**74.7**	74.4

4. Experiments

In this section, we evaluate WLOD on the BDD100K [39] and KITTI [3] datasets, and compare it with state-of-the-art methods.

BDD100K: BDD100K is the one of most commonly used datasets for object detection in autonomous driving. It contains 100k images. This dataset is divided into three parts: 70k images for training, 10k images for validation, and 20k images for testing.

KITTI: KITTI is also a dataset for object detection for autonomous driving. It is small compared to the BDD100K dataset. The KITTI dataset contains 7481 training images and 7512 test images.

In the experiments, parameter λ is set to 0.5 ($W_{0.5}$) and 1 (W_1) to analyze the performance of object detection by paying different levels of attention to appearance similarity.

In this dataset, two-dimensional (2D) bounding boxes are annotated for 'Bus', 'Traffic light', 'Traffic sign', 'Person', 'Bike', 'Truck', 'Motor', 'Car', 'Train', and 'Rider'. Extensive experiments are conducted to demonstrate the effectiveness of the proposed model. mAP and AP_{50} (the 0.5-IoU based average precision) for each class of object on the BDD100K are reported for model evaluation. The categories in this dataset are grouped as follows:

Group 1: 1-Bus, 2-Truck, 3-Car
Group 2: 4-Motor, 5-Bike
Group 3: 6-Rider, 7-Person
Group 4: 8-Traffic light, 9-Traffic sign
Group 5: 10-Train.

Group 1 contains all types of vehicles. 'Motor', 'Rider', 'Bike', and 'Person' are usually related to people without protective shells, so we separate them from Group 1. The principles of how to divide the groups is risk-free misclassification in one group. The ground distance metric between two categories in the same group is set equally. Group 1 and Group 2 have factors (wheels) in common; therefore, we assign a smaller distance between them than that between Group 1 and Group 3.

[1]We note that softmax cannot strictly guarantee the sum of its outputs to be 1 considering the rounding operation. However, the difference in setting the true class probability to 1 or the sum of the source distribution probability is not significant in the experiments using the typical format of the softmax output which is accurate to eight decimal places.

Table 1 shows mAP on the BDD100K validation dataset by YOLOv3, SSD, $W_{0.5}$-s (WLOD uses the backbone of SSD and pre-defined $\mathbf{W}$ as in Fig. 3(b)) and $W_{0.5}$-y and W_1-y (WLOD uses Darknet-53 just like YOLOv3 and pre-defined $\mathbf{W}$ as Fig. 3(a)). $W_{0.5}$-y improves the mAP by nearly 3 points. The mAP of $W_{0.5}$-s is also higher than that of SSD.

Objects in the KITTI dataset are labeled with 'Car', 'Van', 'Truck', 'Tram', 'Pedestrian', 'Person (sitting)', Cyclist', and 'Misc'. We do not take the category 'Misc' into account and use three-fold cross-validation on the labeled images in KITTI. These images are randomly divided into three folds (2495, 2493, 2493). We use the same model setting as for the BDD100K dataset for evaluation with the YOLOv3 and SSD methods. The categories in the KITTI dataset are divided into three groups as follows:

Group 1: 1-Car, 2-Van, 3-Truck
Group 2: 4-Tram
Group 3: 5-Cyclist, 6-Pedestrian, 7-Person.

As shown in Table 2, the proposed method improved the mAP obtained by YOLOv3 from 68.3% to 69.4%, and the mAP obtained by SSD from 72.8% to 74.7%.

To intuitively present the effectiveness, we provide eight representative examples in Fig. 4. The images in the first rows are obtained by YOLOv3, and the images in the second rows are obtained by $W_{0.5}$-y. Some 'Truck's are detected and identified as 'Car's by YOLOv3, in an image, while $W_{0.5}$-y classifies them correctly. In another image, there is a 'Rider' sitting on a 'Bike', but it is not detected by YOLOv3. Several images show that YOLOv3 recognizes an object as several classes. For example, 'Bus' is classified as a 'Bus' and 'Truck' at the same time, while $W_{0.5}$-y classifies it as a 'Bus'.

According to qualitative and quantitative results above, we conclude that the proposed appearance similarity aware loss based on the Wasserstein distance can improve the performance of object detection in terms of mAP. In addition, the AP_{50} of common objects obtained by WLOD is much higher than that obtained without the Wasserstein loss. Therefore, the proposed method is suitable for application in self-driving.

5. Conclusion

In this paper, we argue the object detection from a novel angle of view that the CE/BCE loss based on one-hot label will weak the attribute perception of the detector. We explicitly encourage classifying the objects into categories similar to the ground-truth, and suppressing the severity of the misclassification for self-driving with additional Wasserstein loss by a ground matrix. We also increase the predicted probability value of the ground-truth category simultaneously with the stricter overall optimization. The proposed method is demonstrated of effectiveness

Figure 4. The output by YOLO (first line) and $\text{WOLD}_{0.5}$-y (second line).

on the BDD100K and Drone2019 datasets for autonomous driving .

6. Acknowledgement

The funding support from NIH (NS061841, NS095986), 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

[1] Tong Che, Xiaofeng Liu, Site Li, Yubin Ge, Ruixiang Zhang, Caiming Xiong, and Yoshua Bengio. Deep verifier networks: Verification of deep discriminative models with deep generative models. In *ArXiv*, 2019. 1

[2] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015. 3

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 5

[4] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[6] Gewen He, Xiaofeng Liu, Fangfang Fan, and Jane You. Classification-aware semi-supervised domain adaptation. *CVPRW*, 2020. 2

[7] Gewen He, Xiaofeng Liu, Fangfang Fan, and Jane You. Image2audio: Facilitating semi-supervised audio emotion recognition with facial expression image. *CVPRW*, 2020. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(9):1904–16, 2014. 2

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2, 3

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[13] Xiaofeng Liu. Research on the technology of deep learning based face image recognition. In *Thesis*, 2019. 1

[14] Xiaofeng Liu, Kumar B.V.K, Chao Yang, Qingming Tang, and Jane You. Dependency-aware attention control for unconstrained face recognition with image sets. In *European Conference on Computer Vision*, 2018. 1, 4

[15] Xiaofeng Liu, Yi Ge, Lingsheng Kong, Ping Jia, You Jane, and Jun Lu. Unimodal regularized neuron stick-breaking for ordinal regression. In *ArXiv*, 2019. 2

[16] Xiaofeng Liu, Yubin Ge, Chao Yang, and Ping Jia. Adaptive metric learning with deep neural networks for video-based facial expression recognition. *Journal of Electronic Imaging*, 27(1):013022, 2018. 1

[17] Xiaofeng Liu, Zhenhua Guo, Site Li, Jane You, and Kumar B.V.K. Dependency-aware attention control for unconstrained face recognition with image sets. In *ICCV*, 2019. 1

[18] Xiaofeng Liu, Xu Han, Yukai Qiao, Yi Ge, Site Li, and Jun Lu. Unimodal-uniform constrained wasserstein training for medical diagnosis. In *ArXiv*, 2019. 2

[19] Xiaofeng Liu, Yuzhuo Han, Song Bai, Yi Ge, Tianxing Wang, Xu Han, Site Li, Jane You, and Jun Lu. Importance-aware semantic segmentation in self-driving with discrete wasserstein training. *AAAI*, 2020. 2, 3

[20] Xiaofeng Liu, Wenxuan Ji, Jane You, Georges El Fakhri, and Jonghye Woo. Severity-aware semantic segmentation with reinforced wasserstein training. *CVPR*, 2020. 2, 3

[21] Xiaofeng Liu, Lingsheng Kong, Zhihui Diao, and Ping Jia. Line-scan system for continuous hand authentication. *Optical Engineering*, 56(3):033106, 2017. 1

[22] Xiaofeng Liu, BVK Vijaya Kumar, Yubin Ge, Chao Yang, Jane You, and Ping Jia. Normalized face image generation with perceptron generative adversarial networks. In *2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, pages 1–8, 2018. 1

[23] Xiaofeng Liu, BVK Vijaya Kumar, Ping Jia, and Jane You. Hard negative generation for identity-disentangled facial expression recognition. *Pattern Recognition*, 88:1–12, 2019. 1

[24] Xiaofeng Liu, Site Li, Lingsheng Kong, Wanqing Xie, Ping Jia, Jane You, and BVK Kumar. Feature-level frankenstein: Eliminating variations for discriminative recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 637–646, 2019. 1

[25] Xiaofeng Liu, Zhaofeng Li, Lingsheng Kong, Zhihui Diao, Junliang Yan, Yang Zou, Chao Yang, Ping Jia, and Jane You. A joint optimization framework of low-dimensional projection and collaborative representation for discriminative classification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1493–1498. 1

[26] Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *CVPR Workshops*, pages 20–29, 2017. 1

[27] Xiaofeng Liu, Yang Zou, Tong Che, Jane You, and Kumar B.V.K. Conservative wasserstein training for pose estimation. In *ICCV*, 2019. 2, 3

[28] Xiaofeng Liu, Yang Zou, Lingsheng Kong, Zhihui Diao, Junliang Yan, Jun Wang, Site Li, Ping Jia, and Jane You. Data augmentation via latent space interpolation for image classification. In *24th International Conference on Pattern Recognition (ICPR)*, pages 728–733, 2018. 1

[29] Xiaofeng Liu, Yang Zou, Yuhang Song, Chao Yang, Jane You, and BV K Vijaya Kumar. Ordinal regression with neuron stick-breaking for medical diagnosis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2, 3

[31] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(6):1137–1149, 2015. 2

[33] Ludger Rüschendorf. The wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1):117–129, 1985. 4

[34] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 2

[35] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003. 4

[36] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 2

[37] Chao Yang, Xiaofeng Liu, Qingming Tang, and C-C Jay Kuo. Towards disentangled representations for human retargeting by multi-view learning. *arXiv preprint arXiv:1912.06265*, 2019. 2

[38] Chao Yang, Yuhang Song, Xiaofeng Liu, Qingming Tang, and C-C Jay Kuo. Image inpainting using block-wise procedural training with annealed adversarial counterpart. *arXiv preprint arXiv:1803.08943*, 2018. 2

[39] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2018. 5

[40] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. *ICCV*, 2019. 1