


ORIGINAL ARTICLE

Open Access



ML-ANet: A Transfer Learning Approach Using Adaptation Network for Multi-label Image Classification in Autonomous Driving

Guofa Li^{1,2}, Zefeng Ji¹, Yunlong Chang³, Shen Li⁴, Xingda Qu^{1*}  and Dongpu Cao²

Abstract

To reduce the discrepancy between the source and target domains, a new multi-label adaptation network (ML-ANet) based on multiple kernel variants with maximum mean discrepancies is proposed in this paper. The hidden representations of the task-specific layers in ML-ANet are embedded in the reproducing kernel Hilbert space (RKHS) so that the mean-embeddings of specific features in different domains could be precisely matched. Multiple kernel functions are used to improve feature distribution efficiency for explicit mean embedding matching, which can further reduce domain discrepancy. Adverse weather and cross-camera adaptation examinations are conducted to verify the effectiveness of our proposed ML-ANet. The results show that our proposed ML-ANet achieves higher accuracies than the compared state-of-the-art methods for multi-label image classification in both the adverse weather adaptation and cross-camera adaptation experiments. These results indicate that ML-ANet can alleviate the reliance on fully labeled training data and improve the accuracy of multi-label image classification in various domain shift scenarios.

Keywords: Autonomous vehicles, Deep learning, Image classification, Multi-label learning, Transfer learning

1 Introduction

Benefitting from the rapid development of deep learning technologies in recent years, applications based on convolutional neural network (CNN) have been extensively developed for advanced driver assistance systems (ADASs) and autonomous vehicles (AVs) [1–4]. These applications mainly focused on object detection [5, 6], object tracking [7], and semantic segmentation [8]. Among these applications, image classification is the fundamental technology to divide images into different classes for better detection, tracking, and semantic segmentation performances.

1.1 Image Classification

The methods for image classification can generally be categorized into two categories including single-label image classification (SLIC) [9] and multi-label image classification (MLIC) [10]. SLIC assumes that there is only one category of objects in each image. However, naturalistically collected images often contain multiple categories of objects (e.g., vehicles, cyclists, and pedestrians in a single image) in real world. Therefore, MLIC is needed for safety enhancement of ADASs and AVs, and has attracted more attention in real applications.

Early MLIC algorithms mainly include multi-label k-nearest neighbors (ML-KNN) [11], rank support vector machine (rank-SVM) [8], and multi-label decision tree (ML-DT) [13]. ML-KNN [11] uses the maximum of a posteriori estimation (MAP) to determine the set of labels for test samples based on the traditional KNN algorithm. Rank-SVM [12] uses a rank loss function and the corresponding marginal function as constraints for multi-label learning based on SVM. ML-DT [13] uses the

*Correspondence: quxd@szu.edu.cn

¹ Institute of Human Factors and Ergonomics, College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen 518060, China
Full list of author information is available at the end of the article

information gain criterion based on multi-label entropy to construct decision trees recursively. These traditional algorithms were extensively applied in MLIC tasks for object detection in the late 1990s to early 2000s. However, since these methods have high computational complexity and low accuracy in predicting rare categories when the samples numbers are imbalanced, their performances were generally unsatisfactory in terms of classification accuracy.

More recently, deep learning technologies have been proposed in MLIC tasks, and significant classification improvements have been achieved. The powerful non-linear representation capabilities of deep neural networks can learn more effective features from large-scale datasets for better performance. A hypotheses-CNN-pooling (HCP) algorithm was proposed in Ref. [14] based on binarized normed gradients (BING) that used cross-hypothesis and max-pooling to fuse the classification results of all candidate regions to obtain images with complete label information. The results showed that better performance was achieved by fusing all candidate regions. Wang et al. [15] used CNN to extract features of the input images and used recurrent neural networks (RNNs) to reduce the label dependency. The results concluded that the combined CNN-RNN method could effectively identify image classes by modeling the label co-occurrence dependency in a joint image/label embedding space. Besides, Zhang et al. [16] combined CNN to predict small object classes in images, and Song et al. [17] used a deep multi-modal CNN for multi-instance multi-label image classification. Results from both studies supported the effectiveness of their algorithms in the examined MLIC tasks.

An effective deep neural networks model requires a large amount of accurately labeled samples for training, because millions or even billions of parameters need to be learned from the labeled samples. However, reliable labels heavily rely on extensive human labor work [18, 19]. These time-consuming and labor-intensive labeling work hindered the rapid and widespread applications of these technologies in practical applications. To alleviate this problem, transfer learning was developed for solutions [20].

1.2 Transfer Learning

Transfer learning is an effective technique to improve the performances of classifiers in the target domain with the availability of the annotated data in the source domain only. Transfer learning also refers to unsupervised domain adaptation, which can adapt features from labeled source domains to unlabeled target domains, and thereby would greatly reduce the cost of human labeling work [20].

Pan et al. [21] proposed a transfer component analysis (TCA) algorithm. In the subspace of transfer component, the feature distribution discrepancy between the source domain and the target domain was significantly reduced and the separability of the data was retained. Long et al. [22] simultaneously adapted both the marginal and conditional distributions in a principled dimensionality reduction procedure by using joint distribution adaptation (JDA). Their results demonstrated the superiority of JDA in accuracy and efficiency over the compared deep learning and transfer learning methods in classification tasks. Wang et al. [23] developed a balanced distribution adaptation (BDA) and added a balance factor to dynamically measure the importance of edge distribution and conditional distribution to improve the classification accuracy. In Ref. [24], an easy transfer learning (EasyTL) approach was proposed to learn nonparametric transfer features by exploiting intra-domain structures to obtain an image classifier. It was concluded that EasyTL had high computational efficiency and could be directly applied in image classification technologies on resource-constrained devices such as wearables.

To address the time-consuming and labor-intensive limitations of deep learning algorithms in MLIC applications, transfer learning has been introduced to improve the CNN training process and improve the accuracy in MLIC tasks. Yosinski et al. [25] quantitatively analyzed the features from the CNN encoding process, and found that the encoded features were more effective in transfer learning. Based on their experimental findings, Tajbakhsh et al. [26] concluded that when training a deep CNN model in the target domain, it was better to fine-tune a pre-trained source domain CNN model than to retrain the model in the target domain. Zhang et al. [27] proposed a deep transfer network (DTF) framework which used deep neural networks for cross-domain feature distribution matching. The effectiveness of the algorithm was validated in cross-domain multi-class object recognition tasks. Tzeng et al. [28] analyzed the loss of domain confusion and proposed a deep domain confusion (DDC) algorithm to optimize the objective function of maximizing consistency between the source domain and target domain. The experimental results showed that the learned representations were invariant to domain shifts and thus could be used for MLIC tasks.

1.3 Contributions

Although the above-mentioned MLIC algorithms have achieved significant progresses, image classification in complex traffic environments (e.g., hazy or snow weather) based on camera systems is still a challenging task for the development of ADASs and AVs because the generalization capability of the algorithms in real traffic

still needs to be improved and the algorithms are easy to fail in cross-domain adaptation. To solve this problem and meet the requirement of high accuracy for practical applications, we proposed an effective deep adaptive neural network method for MLIC tasks, namely, multi-label adaptation network (ML-ANet). Specifically, ML-ANet leveraged transfer learning to transfer knowledge from a well-labeled domain to a similar but different domain with limited or no labels. To effectively use the labeled data in the source domain, we conducted MLIC supervised learning on the source domain data, and used multiple kernel variants of maximum mean discrepancies to distribute the feature maps of the source and target domains to reduce domain discrepancy. The main contributions of this paper can be summarized as follows.

- (1) We proposed a new deep adaptation network ML-ANet to learn transferable features for adapting models from a source domain (with labelled information) to a different target domain (without labelled information) in MLIC tasks.
- (2) The effectiveness of our proposed ML-ANet in various traffic environments has been demonstrated by extensive experiments on three large-scale driving datasets. This suggests that when being applied in ADASs and AVs, our proposed ML-ANet could make the ADASs and AVs adaptable to all-around-the-clock illuminations in various weather conditions.
- (3) Our proposed ML-Net alleviates the reliance on fully labeled training data, and therefore no extensive labor work will be needed for network development. This would promote the development efficiencies of ADASs and AVs.

2 Proposed Approach

The proposed MLIC approach (i.e., ML-ANet) mainly consists of two sub-networks including the multi-label learning network (ML-Net) and the adaptation network (ANet). ML-Net uses labeled samples from the source domain to train a multi-label classifier for simultaneous multiple labels prediction of an image. ANet embeds the features from the task-specific layer into the reproducing kernel Hilbert space (RKHS) and matches different distributions optimally using the multi-kernels maximum mean discrepancies (MK-MMD) in RKHS. See Figure 1 for the overall framework of our proposed ML-ANet. A detailed description of the proposed method is given in the following subsections.

2.1 ML-Net (Multi-label Learning Network)

Multi-label learning means that each image is associated with multiple class labels simultaneously. Assume that the training set images can be described as $I = \{x_i\}$, where x_i represents image i and its corresponding label vector is $y_i = \{0, 1\}^c$. $y_i^j = 1$ indicates that the j th label exists in image x_i , while $y_i^j = 0$ indicates the missing of the j th label in image x_i . The MLIC task is essentially about learning a mapping function $f : x \rightarrow y$ from the training set $\{(x_i, y_i) | 1 \leq i \leq n\}$. In this paper, we considered the MLIC problem as multiple binary classification problems, which means that the samples with the same label were considered as positive samples (i.e., $y_i = 1$), while the others were considered as negative samples (i.e., $y_i = 0$).

ML-Net trains multi-label classifiers based on labeled data samples from the source domain. Specifically, an image with a size of $224 \times 224 \times 3$ was fed into the ML-Net, and a feature map was extracted through ResNet-50.

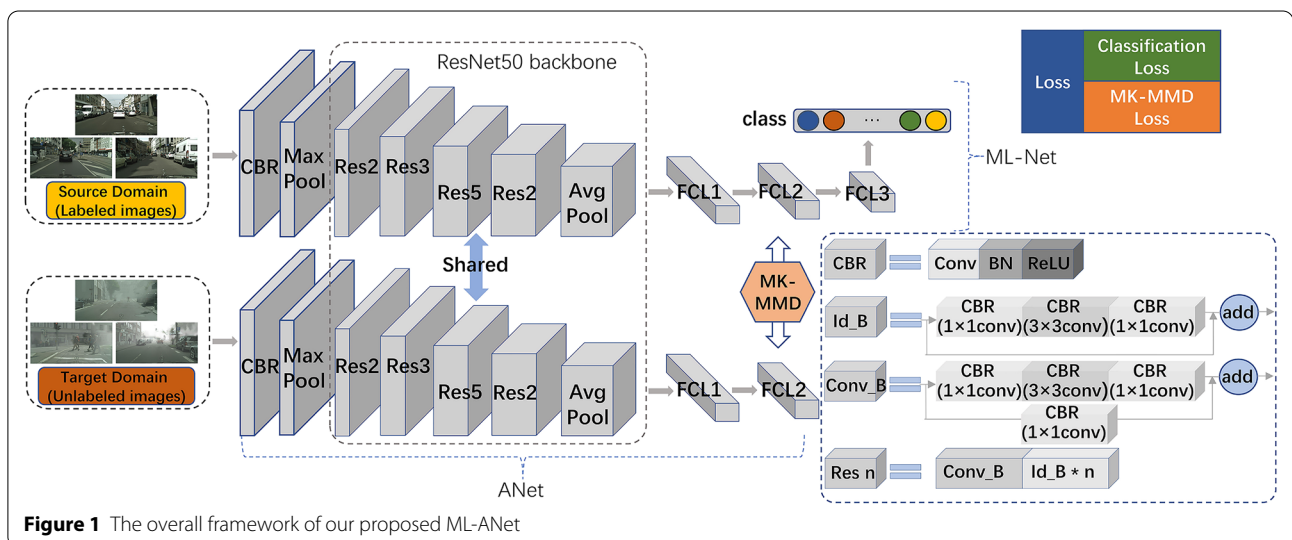


Figure 1 The overall framework of our proposed ML-ANet

As shown in Figure 1, the dimension of feature vector was reduced from 4096 to 2048 by the first fully connected layer (FCL1), from 2048 to 256 by FCL2, and from 256 to the number of examined labels by FCL3. The number of parameters in ML-Net is approximately 24 million, which indicates that it is difficult to learn the large number of parameters directly from the source domain. Therefore, we transferred the pre-trained ResNet-50 model on ImageNet dataset to the ML-Net. We trained the three fully connected layers and fine-tuned the other layers. Finally, we used the sigmoid function to calculate the score for each category and used the binary cross-entropy loss as the multi-label classification loss function. For each minibatch, we calculated the loss using the following formulas:

$$J(\theta) = -\frac{1}{N} \sum_{i=1}^N [y_i \log h_{\theta}(\hat{y}_i) + (1 - y_i) \log(1 - h_{\theta}(\hat{y}_i))], \tag{1}$$

$$h_{\theta}(x_i) = 1/[1 + \exp(-\theta^T x_i)], \tag{2}$$

where N is the number of training samples, $h_{\theta}(x_i)$ denotes the probability of the i th class calculated by the sigmoid function, \hat{y}_i denotes the value of the ML-Net predicted in the i th class, y_i is the ground truth of the i th class, and $y_i \in \{0, 1\}$.

2.2 A-Net (Adaptation Network)

In deep neural networks, the shallow layers learn general features so that the parameters of shallow layers are universal across different tasks, while the parameters of deep layers depend on specific tasks [25]. This inspired us that our proposed network should focus on the deep task-specific layers. Therefore, we proposed an adaptation network (A-Net) to explore the transferability of one domain with labeled information to another domain without labeled information by embedding MK-MMD loss in the last layers. In transfer learning, the domain with labeled information is treated as the source domain, and the domain without labeled information is considered as the target domain. The data from these two domains are usually under different probability distributions.

In this paper, to align different data distributions in the two domains, we introduced a RKHS where the domain discrepancy was measured by using multiple kernel variants of MMD proposed by Gretton et al. [29]. Specifically, A-Net learns transferable features by using MK-MMD to embed the deep features of FCL2 to RKHS, which can optimally match the source and target domain distributions. Figure 2 gives an intuitive example of domain adaptation using multiple Gaussian kernels. For biased datasets (left), a classifier learned from a source domain cannot transfer well to a target domain. By mapping the samples from the source domain and the target domain

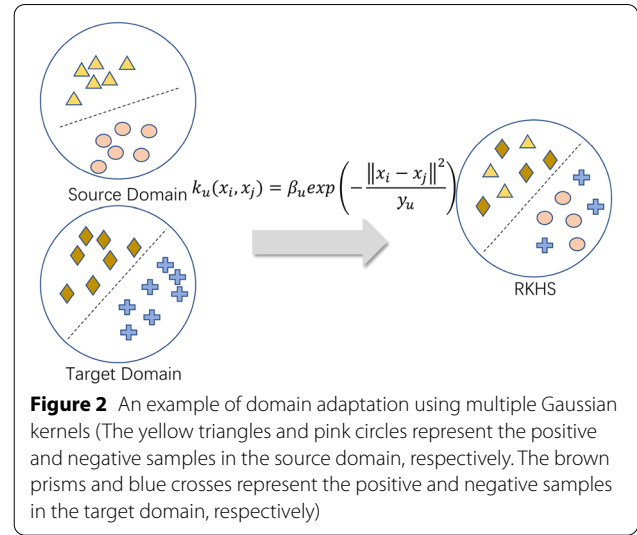


Figure 2 An example of domain adaptation using multiple Gaussian kernels (The yellow triangles and pink circles represent the positive and negative samples in the source domain, respectively. The brown prisms and blue crosses represent the positive and negative samples in the target domain, respectively)

to the RKHS space (right), the distinguished and domain-invariant representations can be learned.

Assuming that $X^S = \{x_1^s, x_2^s, \dots, x_n^s\}$ consists of n samples with the labelled information $Y^S = \{y_1^s, y_2^s, \dots, y_n^s\}$ in the source domain, and $X^t = \{x_1^t, x_2^t, \dots, x_m^t\}$ consists of m samples in the target domain without labels, the source domain and the target domain can be described as $D_s = \{(X^s, Y^s)\}$ and $D_t = \{(X^t)\}$, respectively. The probability distributions of the source and target domain embedded in RKHS are denoted as p and q , respectively. The MK-MMD $d_k(p, q)$ is defined as the distance between the means of probability distributions p and q in RKHS. Hence, the squared formula of MK-MMD can be described as follows:

$$d_k^2(p, q) \triangleq \|E_p[\phi(x^s)] - E_q[\phi(x^t)]\|_{H_k}^2, \tag{3}$$

where H_k is a RKHS with a characteristic kernel k , $E_p[\bullet]$ is the mean of p , $E_q[\bullet]$ is the mean of q , and $\phi(\bullet)$ is a feature mapping function which maps the features from the original feature space to RKHS. In MK-MMD, we denoted K as a particular family of kernels. Hence,

$$K \triangleq \{k = \sum_{u=1}^m \beta_u k_u, \sum_{u=1}^m \beta_u = 1, \beta_u \geq 0, \forall u = \{1, \dots, m\}, \tag{4}$$

where $\{k_u\}$ is the set of u positive definite functions and the constraints on coefficients $\{\beta_u\}$ are imposed to guarantee that the derived kernel k is characteristic.

Given x^s and x^s as independent random variables with distribution p , and x^t and x^t as independent random variables with distribution q , the characteristic kernel $k(\bullet)$ is defined as $k(x^s, x^t) = \langle \varphi(x^s), \varphi(x^t) \rangle$. Hence, the distance

between means of probability distributions p and q can be computed as the expectation of kernel functions:

$$d_k^2(p, q) = \left\| E_{x^s, x^{s'}}[k(x^s, x^{s'})] - 2E_{x^s, x^t}[k(x^s, x^t)] + E_{x^t, x^{t'}}[k(x^t, x^{t'})] \right\|, \quad (5)$$

where $x^{s'}$ is an independent copy of x^s with the same distribution, and $x^{t'}$ is an independent copy of x^t with the same distribution, $E_{x^s, x^t}[\bullet]$ is the mean of $k(x^s, x^t)$. $E_{x^s, x^{s'}}[\bullet]$ and $E_{x^t, x^{t'}}[\bullet]$ are similarly defined.

The purpose of A-Net is to minimize the domain discrepancy between the source and target domains. The domain discrepancy can be measured by the distance between the means of the probability distributions from the source and target domains. Therefore, we have:

$$\min_{D_s, D_t} D_F(D_s, D_t) = \min_{p, q} d_k^2(p, q), \quad (6)$$

where D_s and D_t denote the source and target domains, respectively. $D_F(\bullet)$ denotes the domain discrepancy between the source and target domains in the last fully connected layer of ANet.

2.3 ML-ANet

The loss of ML-ANet consists of the MK-MMD loss and the multi-label classification loss. The objective of minimizing the multi-label classification loss is to improve the distinguishability of features in the source domain, while the goal of minimizing MK-MMD loss is to reduce the discrepancy between the means of the probability distributions in the source and target domains. Thus, we derived the loss function of ML-ANet as:

$$L = J(\theta) + \lambda D_F, \quad (7)$$

where λ is the loss weight parameter and a hyper-parameter, D_F denotes the MK-MMD loss, $J(\theta)$ represents the multi-label classification loss of the source domain, and L is the total loss of the whole ML-ANet, which will be trained by the mini-batch stochastic gradient descent (SGD) algorithm to minimize the loss on the training samples.

Mini-batch SGD is important for the training effect of deep networks, but the calculation of pairwise similarities in mini-batch SGD leads to a computational complexity of $O(n^2)$. To solve this problem, we used an unbiased empirical estimate of MK-MMD proposed by Gretton et al. [29], which can be computed with a complexity of $O(n)$. We used the unbiased empirical estimate to calculate the square form of MK-MMD as follows:

$$d_k^2(p, q) = \frac{2}{n_s} \sum_{i=1}^{n_s/2} g_k(z_i), \quad (8)$$

$$g_k(z_i) = k(x_{2i-1}^s, x_{2i}^s) + k(x_{2i-1}^t, x_{2i}^t) - k(x_{2i-1}^s, x_{2i}^t) - k(x_{2i-1}^t, x_{2i}^s), \quad (9)$$

where n_s denotes the number of variables x^s , z_i is the quad-tuple which is denoted as $z_i \triangleq (x_{2i-1}^s, x_{2i}^s, x_{2i-1}^t, x_{2i}^t)$.

When we train a deep CNN by mini-batch SGD, we only need to consider the gradient of Eq. (8) with respect to each data point x_i . To perform a mini-batch update, we computed the gradient of Eq. (7) with respect to the l th layer parameters θ^l as:

$$\nabla_{\theta^l} = \frac{\partial J(z_i^l)}{\partial \theta^l} + \lambda \frac{\partial g_k(z_i^l)}{\partial \theta^l}. \quad (10)$$

Given that a kernel k is a linear combination of multiple Gaussian kernels $\{k(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / \gamma_u)\}$, the gradient $\partial g_k(z_i^l) / \partial \theta^l$ can be easily calculated by using the chain rule. For instance, the gradient of $k(x_{2i-1}^s, x_{2i}^t)$ in $g_k(z_i^l)$ can be calculated as:

$$\frac{\partial k(x_{2i-1}^s, x_{2i}^t)}{\partial w^l} = - \sum_{u=1}^m \frac{2\beta_u}{\gamma_u} k_u(x_{2i-1}^s, x_{2i}^t) \times (x_{2i-1}^s - x_{2i}^t) \times (x_{2i-1}^{s(l-1)} - x_{2i}^{t(l-1)})^T, \quad (11)$$

where $x_i^l = W^l x_i^{l-1} + b^l$. W^l and b^l represent the coefficient matrix and the bias term from the $(l-1)$ th layer to the l th layer, respectively. In summary, the training process of the entire ML-ANet approach can be described in Algorithm 1.

Algorithm 1 The training process of ML-ANet.

Input: The training datasets $D_s = \{(X^s, Y^s)\}$, $D_t = \{(X^t)\}$, batch size N

Output: The network parameters W_m and W_a for ML-Net and A-Net, respectively.

- 1: **Initialization:** Initialize W_m, W_a with ResNet-50;
 - 2: **while** not converge **do**
 - 3: **for** each training sample $x_i^s \in D_s$ and $x_i^t \in D_t$ **do**
 - 4: Forward pass to obtain the feature representations of x_i ;
 - 5: Back propagate to update the network parameters W_m and W_a via Eq. (10);
 - 6: **end for**
 - 7: **end while**
-

3 Datasets and Experiment

Two adaptation examinations were conducted to verify the effectiveness of our proposed ML-ANet, i.e., adverse weather adaptation and cross-camera adaptation. Whether a detection system can operate faithfully in

different weather conditions is essential for a safe autonomous driving system [30]. This paper mainly addressed the domain shift caused by the conversion between clear weather and hazy weather in the adverse weather adaptation experiment. In the cross-camera adaptation experiment, we examined the effects on alleviating the data bias caused by different resolution and contrast of color cameras under similar weather conditions. The datasets used in our experiment are described in Section 3.1, and the experimental setup is described in Section 3.2.

3.1 Datasets

Naturalistic driving dataset is very important for the development of autonomous driving technologies [31–33]. Three naturalistic driving datasets were used to train and evaluate our proposed ML-ANet, including Cityscapes [34], Foggy Cityscapes [30], and KITTI [35]. Cityscapes dataset is an urban scene dataset for driving scenarios. Foggy Cityscapes dataset is a synthetic foggy dataset from Cityscapes for semantic foggy scene understanding analysis. KITTI dataset is constructed by images collected from real driving in mid-size cities. Though these three datasets cover various urban scenes, the images vary in style, resolution, and illumination between datasets. The main domain divergence between Foggy Cityscapes and Cityscapes is the synthetic fog effect in Foggy Cityscapes, and KITTI has obvious changes in image resolution, illumination, and urban scenes that are not the case in Cityscapes. Figure 3 illustrates the visual differences between Cityscapes, Foggy Cityscapes, and KITTI. In this paper, we used C to represent Cityscapes, F to represent Foggy Cityscapes, and K to represent KITTI. Therefore, the transfer from Cityscapes to Foggy Cityscapes can be denoted as $C \rightarrow F$, and similar expressions can be obtained for the other transfers patterns. In our experiments, we conducted four transfer tasks including $C \rightarrow F$, $F \rightarrow C$, $C \rightarrow K$, and $K \rightarrow C$.



Figure 3 Illustrated images for each dataset (top: KITTI, bottom-left: Cityscapes, bottom-right: Foggy Cityscapes)

3.2 Experiment

The experimental training data consist of source training data with images and their category annotations, and target training data with only images. We extracted three classes (i.e., pedestrians, vehicles, and two-wheelers) from the three datasets for experiments. Table 1 lists the number of sample images for each class in the three employed datasets.

Due to the insufficient number of images in the datasets for reliable training, we randomly used five image augmentation skills to expand the dataset including rotation, shift, contrast, scaling and horizontal flipping [36, 37]. The size of all images in the experiment was resized to $224 \times 224 \times 3$ and we initialized the model with pre-trained weights on ImageNet. Each batch included 32 images from the source and target domains, respectively. We used an optimizer with a momentum of 0.9 and a weight decay of 0.001 in the experiment [38]. Table 2 lists the hyper-parameters used for the training of our ML-ANet. All the experiments were processed on an Intel i5 9600KF (3.70 GHz) with NVIDIA GeForce RTX 2070 GPU.

To validate the effectiveness of our proposed ML-ANet, ML-KNN [11] and five state-of-the-art transfer learning methods (i.e., TCA [21], JDA [22], BDA [23], DDC [24] and DAN [39]) were selected for comparison. All these comparison methods were selected because they have achieved promising MLIC performances with detailed recommended parameters in Refs. [11, 21–24, 28, 39]. In the experiment, we utilized the classification accuracy by following these papers [22, 23, 28, 39] to evaluate the

Table 1 Number of images for each class in the three datasets

Classes	Dataset		
	Cityscapes	Foggy Cityscapes	KITTI
Pedestrian	2343	2343	1779
Vehicle	2832	2832	1689
Two-wheeler	1646	1646	1141
Total	2966	2966	2486

The last line indicates the number of images in each dataset

Table 2 Hyper-parameters for the training of our ML-ANet

Parameter	Value
Total epochs	200
Batch size	64
Initial learning rate	1×10^{-3}
Activation function	ReLU
Optimizer	SGD

effectiveness of our method to reduce the divergency between source and target domains. The results when using different methods are shown in the following section.

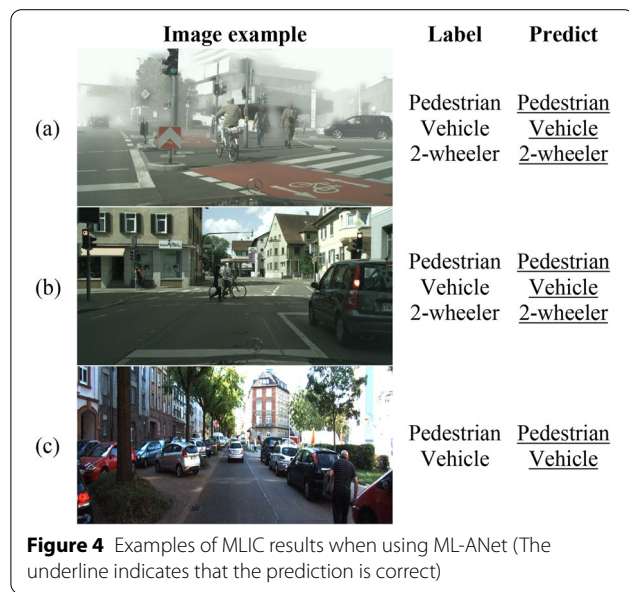
4 Results and Discussion

4.1 Adverse Weather Adaptation

Figure 4(a) and Figure 4(b) show the intuitive examples of ML-ANet for MLIC tasks in hazy and clear weather, respectively. The experimental results of domain shift between clear weather and hazy weather are presented in Table 3. The presented results show that our ML-ANet achieves an average accuracy of 94.83% and 96.85% in transfer tasks of $C \rightarrow F$ and $F \rightarrow C$, respectively, better than the compared transfer learning methods. The classification performance of each object class is also superior to the other compared methods. The results indicate that our proposed ML-ANet could effectively transfer knowledge from a clearly labeled domain to a similar but

different domain with limited or no labels. The advantage of our ML-ANet is probably because it can effectively reduce the distribution discrepancy between the source and target domains caused by weather changes through the adaptation network.

To better quantify the performance of our proposed ML-ANet, the average accuracies of TCA, JDA, BDA, DDC, DAN, and our ML-ANet for transfer tasks $C \rightarrow F$ and $F \rightarrow C$ with respect to epoch numbers are respectively shown in Figure 5(a) and Figure 5(b). Where $C \rightarrow F$ denotes the transfer task from Cityscapes to Foggy Cityscapes, and $F \rightarrow C$ denotes the transfer task from Foggy Cityscapes to Cityscapes. The illustrated results show that: a) TCA has the lowest accuracy because it only adapts to the marginal distribution and does not need iteration. b) The convergence speed and accuracy of ML-ANet is substantially higher than DDC, indicating that single-kernel MMD cannot sufficiently align the probability distribution of the source and target domains. c) The overall change of the ML-ANet curve is higher than DAN, which demonstrates that ML-ANet has better MLIC performance and can further reduce the divergency between the source domain and the target domain. d) ML-ANet can achieve a promising accuracy with a low epoch number.

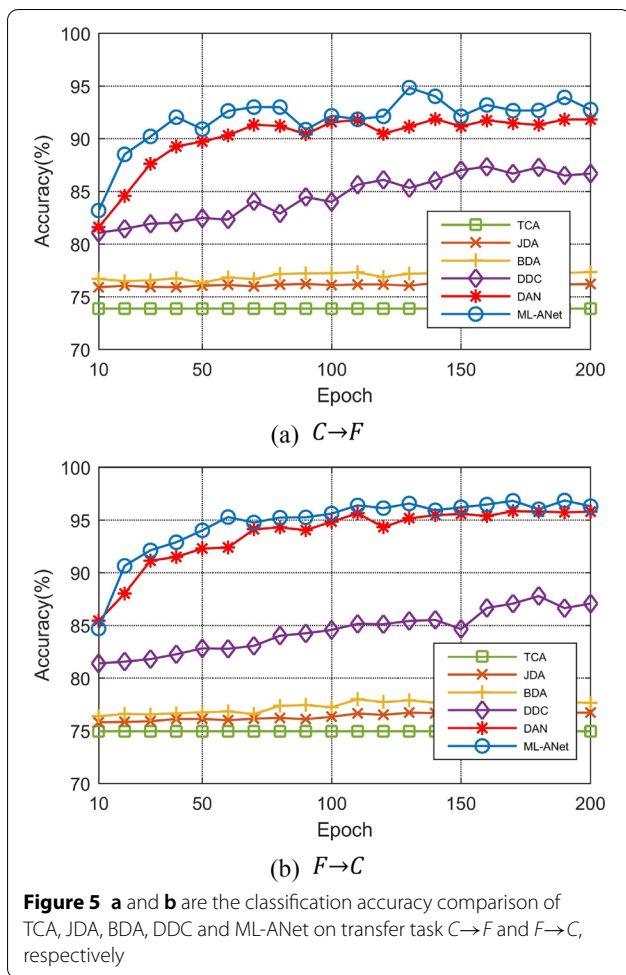


4.2 Cross-Camera Adaptation

The camera mechanisms and underlying settings can also lead to domain-shift, such as substantial differences in visual appearance and image quality. Therefore, cross-camera adaptation is also an important and effective indicator for measuring the quality of transfer learning. Intuitive examples of ML-ANet for MLIC incross-camera adaptation can be found in Figure 4(b) and Figure 4(c). The quantitative experimental results of cross-camera adaptation are shown in Table 4. Specifically, the average MLIC accuracies of ML-ANet are 78.67% and 70.10% in transfer tasks of $C \rightarrow K$ and $K \rightarrow C$, respectively. The numbers are 2.2% and 1.03% higher than the best

Table 3 Comparison of classification accuracies (%) on Cityscapes and Foggy Cityscapes datasets

Method	$C \rightarrow F$				$F \rightarrow C$			
	Pedestrian	Vehicle	2-wheeler	Avg.	Pedestrian	Vehicle	2-wheeler	Avg.
ML-KNN [7]	84.2	95.8	60.6	80.21	89.5	97.6	73.2	86.75
TCA [16]	75.5	94.1	51.9	73.85	76.6	93.0	55.5	74.94
JDA [17]	76.7	93.9	58.3	76.29	77.1	93.4	59.7	76.74
BDA [18]	77.9	93.9	60.3	77.35	78.3	94.0	61.6	77.97
DDC [23]	87.7	97.7	76.6	87.33	87.6	98.1	77.6	87.77
DAN [31]	90.3	99.1	86.2	91.85	95.5	99.3	92.6	95.83
Ours (ML-ANet)	93.1	99.5	91.9	94.83	96.8	99.3	94.5	96.85



comparison method DAN in the same two transfer tasks. The presented results indicate that ML-ANet is superior to the other comparison methods, showing more powerful adaptability.

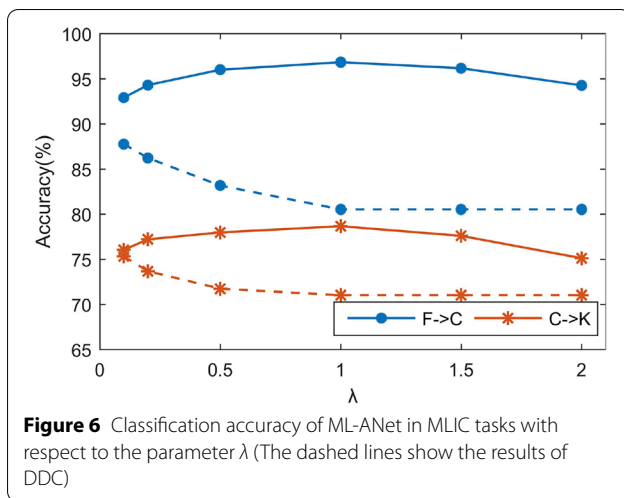
Besides, the pedestrian and vehicle classification accuracies of ML-ANet in task $K \rightarrow C$ are substantially lower than the numbers in task $C \rightarrow K$, which

is reasonable as the number of K samples is relatively smaller compared to C , especially the number of vehicles. However, the classification performance of our ML-ANet on two-wheelers in $K \rightarrow C$ does not show a similar trend, and the classification accuracies of two-wheelers are all lower than the accuracies of pedestrians or vehicles in either $C \rightarrow K$ or $K \rightarrow C$. That's probably because two-wheelers are always with bikes or motorcycles which increase the noise in feature learning, and the adaptability of different transfer tasks is different. The classification accuracies of the examined methods with respect to the number of epochs in transfer tasks of $C \rightarrow K$ and $K \rightarrow C$ show similar trends to the illustrated results in Figure 5, indicating the superiority of ML-ANet in cross-camera adaptation. In summary, the results show that ML-ANet is effective for domain-shift caused by cross-cameras.

The loss weight parameter λ may also influence the performance of ML-ANet. Figure 6 shows the effect of parameter λ on ML-ANet and DDC performance in $C \rightarrow K$ and $F \rightarrow C$ tasks. The other three methods do not include parameter λ , thus only DDC and ML-ANet are compared in Figure 6. Where $F \rightarrow C$ and $C \rightarrow K$ denote the transfer from Foggy Cityscapes to Cityscapes and from Cityscapes to KITTI, respectively. The experimental results show that the classification accuracy of ML-ANet is obviously outperforming DDC for any λ in any of the tasks, supporting the advantage of our proposed ML-ANet. The curves of ML-ANet are bell-shaped with initial rises and following decreases when λ increases. This trend is reasonable because the network focuses more on MK-MMD loss when λ initially increases, resulting in the transferability improvement and increasing accuracy of ML-ANet. However, when λ is too large, the training of the network ignores the classification loss, which causes the accuracy of the network to decrease. The illustrated results in Figure 6 show that the best performance of ML-ANet is achieved when $\lambda = 1.0$ which is the best trade-off for transferability enhancement in ML-ANet.

Table 4 Comparison of classification accuracies (%) on Cityscapes and KITTI datasets

Method	$C \rightarrow K$				$K \rightarrow C$			
	Pedestrian	Vehicle	2-wheeler	Avg.	Pedestrian	Vehicle	2-wheeler	Avg.
ML-KNN [7]	62.9	94.7	54.2	70.58	56.2	74.7	56.2	62.38
TCA [16]	59.5	89.6	50.7	66.61	69.5	71.8	43.4	61.57
JDA [17]	65.7	87.1	52.7	68.49	67.3	74.5	44.2	62.03
BDA [18]	71.3	92.0	48.6	70.61	70.0	74.3	45.6	63.27
DDC [23]	77.1	93.4	55.5	75.30	73.5	80.4	52.1	68.67
DAN [31]	77.4	93.6	58.4	76.47	74.4	81.2	51.6	69.07
Ours (ML-ANet)	81.7	94.9	59.4	78.67	75.2	83.8	51.3	70.10



4.3 Discussion on the Novelities of Our Proposed Method

Different from all the previous methods, the novelties of our proposed method include: (1) The structure of our proposed network is different from the previous ones. Compared with other methods such as DAN, the combination of identity blocks (i.e., the Id_B module in Figure 2), convolutional blocks (i.e., the Conv_B module in Figure 2), and MK-MMD is newly developed to help accelerate the convergence rate of the model and to improve the image classification accuracy and adaptation capability of the network. (2) MK-MMD is innovatively used to simultaneously align the data distribution of multiple labels to enhance the generalization of multi-label image classifiers for ADASs and AVs, which extends the previous work concerning MK-MMD from single-label image classification to multi-label image classification. The results presented above show that our method outperforms the others, both qualitatively and quantitatively on the different urban cross-scenes, which demonstrates that ML-ANet has a better adaptive ability to effectively alleviate domain gap.

Other reasons why our method can utilize unlabeled auxiliary data to improve the generalization of the network comes from: (1) the generated feature maps are mapped to RKHS, and (2) the distribution of data in the source and target domains are aligned during network training. Therefore, our well-trained model can effectively perform MLIC in the source and target domains, which indicates that our ML-ANet has a strong generalization capability for multi-label image classification in different urban scenes.

This paper mainly focuses on the moving objects including pedestrians, vehicles, and two-wheelers. The static objects like traffic sign and traffic light are usually

much smaller than the mentioned moving objects in images [40, 41], therefore challenging the performance of the related methods in the literature. In our future studies, we will focus on developing innovative methods to classify the other objects in various scenarios for traffic safety improvement [42]. Meanwhile, the adaptive object detection and tracking in various weather and illumination scenarios based on the transferred knowledge from daytime dry weather scenarios will also be considered.

5 Conclusions

- (1) To obtain a robust multi-label classifier, a novel and effective method (ML-ANet) is proposed for cross-domain MLIC. The proposed ML-ANet consists of two different sub-networks, ML-Net and A-Net, for multi-label learning and transfer learning, respectively.
- (2) In adverse weather adaptation, ML-ANet achieves an average accuracy of 94.83% and 96.85% in the transfer tasks of $C \rightarrow F$ and $F \rightarrow C$, respectively. The accuracy of ML-ANet could even be better than the compared methods with a low epoch number.
- (3) In cross-camera adaptation, the average accuracy of ML-ANet is 2.2% and 1.03% higher than the best comparison method in the $C \rightarrow K$ and $K \rightarrow C$ transfer tasks, respectively, showing its better adaptability.
- (4) The sensitivity analysis of the loss weight parameter λ show that a good trade-off between the MK-MMD loss and the multi-label classification loss can enhance the feature transferability, and the best performance of ML-ANet is achieved when $\lambda = 1.0$.
- (5) The results from this study demonstrate that our ML-ANet can make ADASs and AVs adaptable to all-around-the-clock illuminations in various weather conditions, and promote the development efficiencies of ADASs and AVs.
- (6) Our future work will focus on the MLIC of other objects and the adaptive object detection and tracking.

Acknowledgements

Not applicable.

Authors' Contributions

GL was in charge of the whole trial; ZJ and GL wrote the original manuscript; XQ reviewed and edited the original manuscript; XQ and DC assisted on the conceptualization and supervised this research work; ZJ, YC, and SL cooperated on the method development, experiment validation, and formal analysis. All authors read and approved the final manuscript.

Authors' Information

Guofa Li, born in 1986, is currently an associate research professor at *Institute of Human Factors and Ergonomics, College of Mechatronics and Control Engineering, Shenzhen University, China* and also a visiting scholar at *Department of Mechanical and Mechatronics Engineering, University of Waterloo, Canada*. He received his Ph.D. degree from *Tsinghua University, China*, in 2016. His research interests include environment perception and decision making for autonomous vehicles.

Zefeng Ji, born in 1996, is currently a master candidate at *Institute of Human Factors and Ergonomics, College of Mechatronics and Control Engineering, Shenzhen University, China*.

Yunlong Chang, born in 1995, is currently an engineer at *Interactive Entertainment Group, Tencent, China*. He received his master degree from *University of Liverpool, UK*, in 2019.

Shen Li, born in 1989, is currently a postdoc at *Traffic Operations and Safety Laboratory, University of Wisconsin – Madison, USA*. He received his Ph.D. degree from *University of Wisconsin – Madison, USA*, in 2019. His research interests include cooperative control method of autonomous connected vehicles in intelligent transportation systems.

Xingda Qu, born in 1978, is currently a professor at *Institute of Human Factors and Ergonomics, College of Mechatronics and Control Engineering, Shenzhen University, China*. He received his PhD degree from *Virginia Tech., USA*, in 2008. His research interests include transportation safety, occupational safety and health, and human computer interaction.

Dongpu Cao, born in 1978, is currently a professor at *Department of Mechanical and Mechatronics Engineering, University of Waterloo, Canada*. He received his PhD degree from *Concordia University, Canada*, in 2008. His research interests include vehicle dynamics, control, and intelligence.

Funding

Supported by Shenzhen Fundamental Research Fund of China (Grant No. JCYJ20190808142613246), National Natural Science Foundation of China (Grant No. 51805332), and Young Elite Scientists Sponsorship Program funded by the China Society of Automotive Engineers.

Competing interests

The authors declare no competing interests.

Author Details

¹Institute of Human Factors and Ergonomics, College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen 518060, China. ²Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada. ³Interactive Entertainment Group, Tencent, Shanghai 200233, China. ⁴Traffic Operations and Safety Laboratory, University of Wisconsin-Madison, Madison, WI 53715, USA.

Received: 1 June 2020 Revised: 22 February 2021 Accepted: 11 August 2021

Published online: 23 August 2021

References

- [1] H Gao, B Cheng, J Wang, et al. Object classification using CNN-based fusion of vision and LIDAR in autonomous vehicle environment. *IEEE Transactions on Industrial Informatics*, 2018, 16(9): 4224-4231.
- [2] G Li, Y Yang, X Qu, et al. A deep learning based image enhancement approach for autonomous driving at night. *Knowledge-Based Systems*, 2021, 213: 106617.
- [3] G Li, Y Yang, T Zhang, et al. Risk assessment based collision avoidance decision-making for autonomous vehicles in multi-scenarios. *Transportation Research Part C: Emerging Technologies*, 2021, 122: 102820.
- [4] G Li, S Li, S Li, et al. Deep reinforcement learning enabled decision-making for autonomous driving at intersections. *Automotive Innovation*, 2020, 3: 374-385.
- [5] Y Chen, W Li, C Sakaridis, et al. Domain adaptive faster R-CNN for object detection in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018: 3339-3348.
- [6] G Li, S E Li, R Zou, et al. Detection of road traffic participants using cost-effective arrayed ultrasonic sensors in low-speed traffic situations. *Mechanical Systems and Signal Processing*, 2019, 132: 535-545.
- [7] S E Li, G Li, J Yu, et al. Kalman filter-based tracking of moving objects using linear ultrasonic sensor array for road vehicles. *Mechanical Systems and Signal Processing*, 2018, 98: 173-189.
- [8] X Zhang, Z Chen, Q M J Wu, et al. Fast semantic segmentation for scene perception. *IEEE Transactions on Industrial Informatics*, 2018, 15(2): 1183-1192.
- [9] L Mou, P Ghamisi, X X Zhu. Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2017, 55(7): 3639-3655.
- [10] Z Yan, W Liu, S. Wen, et al. Multi-label image classification by feature attention network. *IEEE Access*, 2019, 7: 98005-98013.
- [11] M L Zhang, Z H Zhou, ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 2007, 40(7): 2038-2048.
- [12] A Elisseeff, J Weston. A kernel method for multi-labelled classification. *Neural Information Processing Systems*, 2001, 14: 681-687.
- [13] M L Zhang, Z H Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 2013, 26(8): 1819-1837.
- [14] Y Wei, W Xia, M Lin, et al. HCP: A flexible CNN framework for multi-label image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 38(9): 1901-1907.
- [15] J Wang, Y Yang, J Mao, et al. CNN-RNN: A unified framework for multi-label image classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2285-2294.
- [16] J Zhang, Q Wu, C Shen, et al. Multilabel image classification with regional latent semantic dependencies. *IEEE Transactions on Multimedia*, 2018, 20(10): 2801-2813.
- [17] L Song, J Liu, B Qian, et al. A deep multi-modal CNN for multi-instance multi-label image classification. *IEEE Transactions on Image Processing*, 2018, 27(12): 6025-6038.
- [18] F C Heilbron, J C Niebles. Collecting and annotating human activities in web videos. *Proceedings of International Conference on Multimedia Retrieval*, 2014: 377-384.
- [19] G Li, Y Chen, D Cao, et al. Extraction of descriptive driving patterns from driving data using unsupervised algorithms. *Mechanical Systems and Signal Processing*, 2021, 156: 107589.
- [20] S J Pan, Q Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2009, 22(10): 1345-1359.
- [21] S J Pan, I W Tsang, J T Kwok, et al. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 2010, 22(2): 199-210.
- [22] M Long, J Wang, G Ding, et al. Transfer feature learning with joint distribution adaptation. *Proceedings of the IEEE International Conference on Computer Vision*, 2013: 2200-2207.
- [23] J Wang, Y Chen, S Hao, et al. Balanced distribution adaptation for transfer learning. *IEEE International Conference on Data Mining*, 2017: 1129-1134.
- [24] J Wang, Y Chen, H Yu, et al. Easy transfer learning by exploiting intra-domain structures. *IEEE International Conference on Multimedia and Expo*, 2019: 1210-1215.
- [25] J Yosinski, J Clune, Y Bengio, et al. How transferable are features in deep neural networks?. *arXiv preprint arXiv*, 2014: [arxiv:1411.1792](https://arxiv.org/abs/1411.1792).
- [26] N Tajbakhsh, J Y Shin, S R Gurudu, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 2016, 35(5): 1299-1312.
- [27] X Zhang, F X Yu, S F Chang, et al. Deep transfer network: Unsupervised domain adaptation. *arXiv preprint arXiv*, 2015: [arxiv:1503.00591](https://arxiv.org/abs/1503.00591).
- [28] E Tzeng, J Hoffman, N Zhang, et al. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv*, 2014: [arxiv:1412.3474](https://arxiv.org/abs/1412.3474).
- [29] A Gretton, K M Borgwardt, M J Rasch, et al. A kernel two-sample test. *The Journal of Machine Learning Research*, 2012, 13(1): 723-773.
- [30] C Sakaridis, D Dai, L Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 2018, 126(9): 973-992.
- [31] G Li, Y Wang, F Zhu, et al. Drivers' visual scanning behavior at signalized and unsignalized intersections: A naturalistic driving study in China. *Journal of Safety Research*, 2019, 71: 219-229.

- [32] G Li, W Lai, X Sui, et al. Influence of traffic congestion on driver behavior in post-congestion driving. *Accident Analysis and Prevention*, 2020, 141:105508.
- [33] G Li, S E Li, B Cheng, et al. Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities. *Transportation Research Part C: Emerging Technologies*, 2017, 74: 113-125.
- [34] M Cordts, M Omran, S Ramos, et al. The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 3213-3223.
- [35] A Geiger, P Lenz, R Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 3354-3361.
- [36] G Li, Y Yang, X Qu. Deep learning approaches on pedestrian detection in hazy weather. *IEEE Transactions on Industrial Electronics*, 2019, 67(10): 8889-8899.
- [37] Q Wen, Z Luo, R Chen, et al. Deep learning approaches on defect detection in high resolution aerial images of insulators. *Sensors*, 2021, 21: 1033.
- [38] Z Wojna, V Ferrari, S Guadarrama, et al. The devil is in the decoder. *British Machine Vision Conference*, 2017: 1-13.
- [39] M Long, Y Cao, J Wang, et al. Learning transferable features with deep adaptation networks. *International Conference on Machine Learning*, 2015: 97-105.
- [40] G Li, Y Lin, X Qu. An infrared and visible image fusion method based on multi-scale transformation and norm optimization. *Information Fusion*, 2021: <https://doi.org/10.1016/j.inffus.2021.02.008>
- [41] G Li, H Xie, X Qu, et al. Detection of road objects with small appearance in images for autonomous driving in various traffic situations using a deep learning based approach. *IEEE Access*, 2020, 8: 211146-211172.
- [42] G Li, Y Liao, Q Guo, et al. Traffic crash characteristics in Shenzhen, China from 2014 to 2016. *International Journal of Environmental Research and Public Health*, 2021, 18: 1176.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
