# Incremental Cross-Modality Deep Learning for Pedestrian Recognition

Danut Ovidiu Pop, Alexandrina Rogozan, Fawzi Nashashibi, Abdelaziz Bensrhair

# Incremental Cross-Modality Deep Learning for Pedestrian Recognition

Dănuţ Ovidiu Pop[1], Alexandrina Rogozan[2], Fawzi Nashashibi[3] and Abdelaziz Bensrhair[4]

*Abstract*— **In spite of the large amount of existent methods, pedestrian detection is still an open challenge. In recent years, deep learning classification methods combined with multi-modality images within different fusion schemes achieved the best performance. It was proven that late-fusion scheme outperforms both direct and intermediate integration of modalities for pedestrian recognition. Hence, in this paper, we focus on improving the late-fusion scheme for pedestrian classification on the Daimler stereo vision data set. Each image modality among Intensity, Depth and Flow, is classified by an independent Convolution Neural Network (CNN). The CNN outputs are then fused by a Multi-layer Perceptron (MLP) before the recognition decision. We propose different methods based on Cross-Modality deep learning of CNNs: (1) a correlated model where a unique CNN is learned with Intensity, Depth and respectively Flow images for each frame, (2) an incremental model where a CNN is learned with the first modality images frames, then a second CNN, initialized by transfer learning on the first CNN, is learned on the second modality images frames, and finally a third CNN initialized on the second CNN, is learned on the last modality images frames. The experiments show that the incremental cross-modality deep learning of CNNs allows the improvement of classification performances not only for each independent modality classifier, but also for the multi-modality classifier based on late-fusion. Different learning algorithms were also investigated.**

## I. INTRODUCTION

Pedestrian detection is a challenging task of great importance in the domain of object recognition and computer vision. It is a key problem for surveillance, robotics applications and automotive safety [1] where an efficient Advanced Driver Assistance System (ADAS) for pedestrian detection is needed to reduce the number of accidents and fatal injures[1]. These systems usually have multi-modality sensors and/or camera networks to capture the road data, and signal/image processing components to extract pertinent features which are then classified by recognition components.

A study performed by ABI Research published in 2015 shows that Mercedes-Benz, Volvo and BMW dominate the market of car enhancing ADAS systems. As from 2013, BMW cars have been fitted with a Driver Assistance package for Pedestrian Warning, based on an infrared night-vision and a monocular vision cameras. Recently, the Mercedes system has combined stereo vision cameras with long, medium and short-range radars to monitor the area in front of the vehicle. In 2016 Continental company proposed an Advanced Radar Sensor (standard for VW Tiguan) able to detect both objects and pedestrians, at a distance of up to 170 meters. The Nissan company developed a system which detects the environment and surrounding vehicles, such as the road, other vehicles and pedestrians.

These existing ADAS systems still have difficulty distinguishing between human beings and nearby objects, especially in a crowded urban environment where they are not able to detect all partially occluded pedestrians, and they do not work efficiently in extreme weather conditions. Moreover, it is difficult to find an ADAS system that is able to ensure stable, real-time and effective full functionality. We believe it is necessary to improve the classification component of an ADAS system to be able to discriminate between the obstacle type (pedestrian, cyclist, children, old person) in order to adapt the car driver system behavior according to the estimated risk level.

Our work is concerned with the improvement of the classification component of a pedestrian detector. In recent research studies, deep learning neural networks including convolution neural networks (CNNs), like LeNet, AlexNet, GoogLeNet, have usually proved classification performance improvement [2], [3], [4]. Moreover, deep learning classification methods combined with multi-modality images within different fusion schemes achieved remarkable results. The disadvantage for those models is that they require a large amount of annotated data for each modality.

It usually happens that one has not (enough) annotated data in one modality compared with other modalities. The inquiry is whether one modality can be used exclusively (standpoint one) for training the classification model used to recognize pedestrians in another modality or only partially (standpoint two) for improving the training of the classification model in another modality? To our knowledge, this question was not answer yet for the pedestrian recognition task. This paper proposes to answer this brain-teaser through various experiments based on Daimler stereo vision data set.

The paper is organized as follows: Section 2 briefly presents our main contribution and some existing approaches from the literature. Section 3 presents the architecture and methods approach based on Cross-Modality deep learning of CNNs. In Section 4 presents the experiments performance on

[1]Dănuţ Ovidiu Pop is a PhD student at RITS Team, INRIA Paris, 2 Rue Simone IFF, 75012 Paris, France in collaboration with Normandie Univ, INSA Rouen, LITIS, 76000 Rouen, France and Department of Computer Science, Babeş-Bolyai University, 7- 9 Universitatii street, 400084 Cluj-Napoca, Romania. danut-ovidiu.pop@inria.fr

[2] Dr. Alexandrina Rogozan is Associate Professor at Normandie Univ, INSA Rouen, LITIS, 76000 Rouen, France. alexandrina.rogozan@insa-rouen.fr

[3] Dr. Fawzi Nashashibi is the head of RITS Team at INRIA Paris, 2 Rue Simone IFF, 75012 Paris, France. fawzi.nashashibi@inria.fr

[4]Dr. Abdelaziz Bensrhair is Professor at Normandie Univ, INSA Rouen, LITIS, 76000 Rouen, France. abdelaziz.bensrhair@insa-rouen.fr

[1]According to European Commission statistics published in 2016, the number of pedestrians injured in road accidents in 2014 was 1.419.800 and there were 25.900 fatalities

Daimler dataset. Finally, Section 5 presents our conclusions.

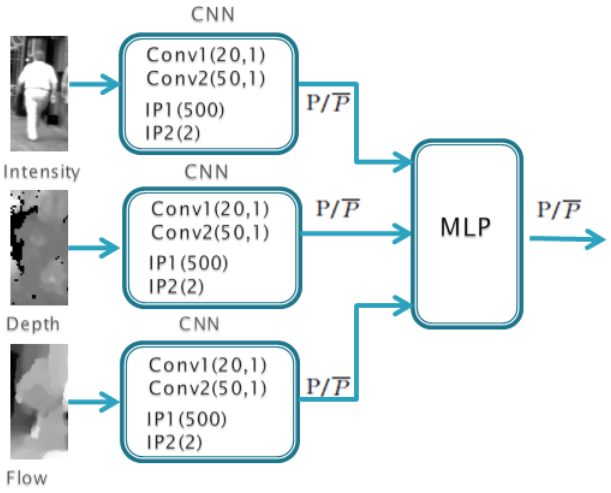



Fig. 1. The correlated cross-modality training architecture

## II. PREVIOUS WORK

In the last decade, the pedestrian detection issue was investigated intensely. Therefore a widely varying detection methods were developed using a combination of features such as Integral Channel Features, Histograms of Oriented Gradients (HOG), Local Binary Patterns (LBP), Scale Invariant Feature Transform (SIFT), among others, followed by a trainable classifier such as Support Vector Machine (SVM), Multilayer Perceptrons (MLP), boosted classifiers or random forests [5], [6]. In [7] is presented a mixture-of-experts framework performed with HOG, LBP features and MLP or linear SVM classifiers. Recently, in [8] was presented a CNN to learn the features with an end-to-end approach. This experiment was focused on the detection of small scale pedestrians on Caltech data set. A combination of three CNNs to detect pedestrians at different scales was proposed on the same monocular vision data set [9]. A cascade Aggregated Channel Features detector is used in [10] to generate candidate pedestrian windows followed by a CNN-based classifier for verification purposes on monocular Caltech and stereo ETH data sets. Two CNN based fusion methods of visible and thermal images on the KAIST multi-spectral pedestrian data set were presented in [11]. The first method combines the information of these modalities at the pixel level (early fusion), the second architecture used separate sub-networks to generate a feature representation for each modality before classification (intermediate fusion). The authors showed that the intermediate fusion outperforms the early fusion.

We compared in [12] the performance of the early fusion and late fusion models on the Daimler stereo vision data set. The early fusion model was built by concatenating three image modalities (intensity, depth and flow) to feed a unique CNN. The late fusion model consists in fusing the outputs of three independent CNNs, trained on intensity, depth and respectively flow images, by an SVM classifier. We showed the early-fusion model is less efficient than the late-fusion one. Moreover, the early fusion is less robust than the late fusion one, since it needs strong image calibration and synchronization. Its training is less effective since it needs for a given image frame an item for each modality and therefore the classifier needs more items to learn the problem. Within the early-fusion model, it is impossible to take advantage of inter-dataset training methods by using modality images from different unimodal and/or multi-modal datasets where all the modalities are not acquired and/or annotated, in order to improve the training by extending the number and the variety of items.

In the literature studies, for the intermediate and late fusion methods, the training was made independently on each modality, exclusively with annotated images acquired from that modality. The aim of this paper is to improve the late-fusion training by using a cross-modality approach. We will prove that the incremental cross-modality is effective for the training of each modality classifier not only with images from that modality, but also with images from other modalities among Intensity, Depth and Flow. A synthetic dataset (Virtual Pedestrian dataset [13]) is used for an initial training, and two different real-world datasets (KITTI Vision Benchmark Suite and Daimler Mono Pedestrian Detection Benchmark) for fine-tuning and evaluation.

To the best of our knowledge any study was made on cross-modality training for pedestrian recognition. In [14], the authors proposed an incremental cross-dataset learning algorithm for the pedestrian detection problem.

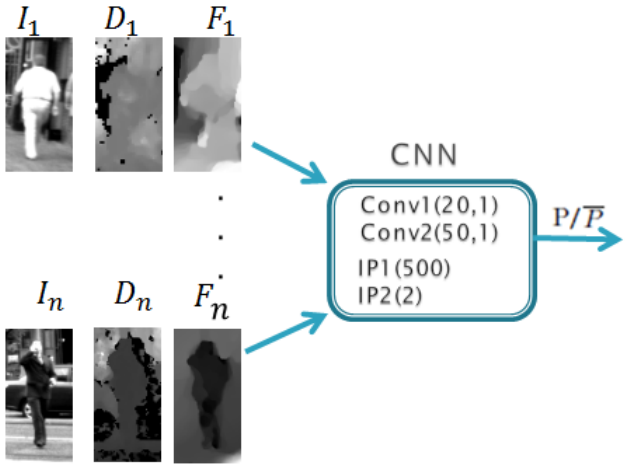



Fig. 2. The general CNN architecture for input data set inserted consecutively, intensity, depth and flow, which trains and validates on the identical multi-modal or unimodal data.

## III. THE PROPOSED ARCHITECTURES

In this paper, we propose fusing of stereo-vision information between three modalities: Intensity (I), Depth (D) and Flow (F). We investigate the late-fusion architecture using three different methods for the training of the CNN-based classifiers: a classical intra-modality approach and two different methods for cross-modality approach.

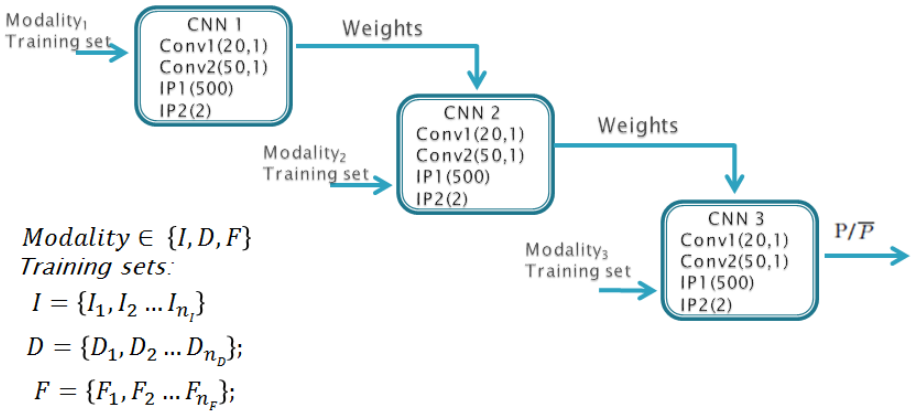



Fig. 3. The incremental cross-modality deep learning architecture.

### A. Our baseline late fusion architecture

We propose a late-fusion architecture (see Fig 1) were a MLP is used to discriminate between pedestrians (P) and non-pedestrians ($\overline{P}$) on the classification results (the class probability estimate of the previous estimated class) of three modality independent CNNs. Each CNN is exclusively trained on the same modality in an independent manner: learned and validated on the same modality.

Each modality CNN is based on the LeNet architecture which consists of 7 layers, except input layer, 2 convolution layers, 2 pooling layers, 2 inner product (IP) layers and one rectified linear unit (ReLU) layer. We use 20 filters with one stride for the first convolution layer followed by 50 filters with one stride for the second one. We use two IP layers with 500 neurons for the first IP layer and respectively 2 neurons for the second IP layer. The final layer returns the final decision of the classifier system, P or $\overline{P}$.

### B. Correlated cross-modality training of CNNs

We propose a correlated cross-modality approach where a unique CNN is learned with the same image frames, but provided in different modalities, Intensity $I_i$, Depth $D_i$ and Flow $F_i$ with i=1,n (see Fig. 2). The CNN model is validated in two different manners: on a multi-modality validation set or on a single modality one. The learning and validation sets are disjointed.

We believe that the disadvantage of the correlated cross-modality training is that it compels one to use a unique CNN model. This is a too strong constraint, if different modalities can ameliorate the learning process with different CNN architectures and/or with different settings (learning algorithms and learning rates).

### C. Incremental cross-modality training

Our experiments show that not all modality CNNs provide the best results with the different architecture and settings. Therefore, we propose an incremental cross-modality training, based on a transfer learning approach (see Fig.3).

A first CNN is learned and validated with the first modality images frames, then a second CNN, initialized by transfer learning on the first CNN, is learned and validated on the second modality images frames, and finally a third CNN initialized on the second CNN, is learned and validated on the last modality images frames. Transfer learning consists in transfer weights information from previous CNN that has already been learned, to a new CNN which follow to be trained.

The advantage of this method is that its architecture is more flexible allowing for adaptive settings for each classifier (i.e. different learning algorithms and rate policies). Learning this model does not require any inter-modality correlated data, neither synchronized modality images. This could be an interesting point if the multi-modality images are various and not acquired with similar sensors/cameras and/or extracted from the same database. The approach can be extended cross-datasets training.

For this approach, the inquiry is whether the order of modality training within the previous model has any importance. We investigate different combinations and we conclude that for the classification in the Intensity modality the optimal order for training is Depth images first, followed by Flow images and finally Intensity images (D,F,I training model of I); for the classification in the Flow modality the optimal order for training is Depth images first, followed by Intensity images and finally Flow images (D,I,F training model of F), and respectively for the classification in the Depth modality the optimal order for training is Intensity images first, followed by Flow images and finally Depth images (I,F,S for training of D) (see Table II).

## IV. EXPERIMENTS AND RESULTS

The training and testing were carried out on Daimler stereo vision images of 48 x 96 px with a 12-pixel border around the pedestrians images extracted from three modalities: Intensity, Depth and optical Flow.

We use 84577 samples for training, 75% of which have been used for learning, 25% for validation and 41834 for testing. The training set contains:

- 52112 samples of pedestrians
- 32465 samples of non pedestrians

The testing set contains:

TABLE I

COMPARISON OF LEARNING ALGORITHMS AND RATE POLICIES ON INTENSITY, DEPTH AND FLOW DATA SET

| Modality Name | Learning rate polics / Algorithm Learning | Accuracy | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | EXP | FIX | INV | POLY | SIG | STEP | MS |
| Intensity | SGD | 95.96 % | 96.07% | 96.01% | 96.09% | 96.01% | 96.20% | 95.78 % |
| | RMSPROP | 95.53% | 61.19% | 95.24% | **96.55%** | 96.42% | 95.91% | 93.37% |
| | ADADELTA | 88.67% | 93.08% | 91.77% | 88.79% | 91.96% | 91.10% | 89.75% |
| | ADAGRAD | 95.02% | 95.41% | 95.83% | 95.49% | 95.46% | 95.87% | 95.02% |
| Depth | SGD | **89.78%** | 61.2% | 89.26% | 89.69 | 88.24% | 88.97% | 61.2% |
| | RMSPROP | 88.64% | 61.17% | 81.99% | 89.10% | 88.66% | 89.22% | 83.54% |
| | ADADELTA | 87.14% | 88.11 | 87.64% | 87.27% | 88.24% | 87.72% | 87.77% |
| | ADAGRAD | 88.77% | 88.81% | 89.44% | 89.25% | 89.44% | 89.09% | 88.71% |
| Flow | SGD | 86.53% | 61.2% | 86.69% | 86.90% | 86.72% | 86.84% | 61.2% |
| | RMSPROP | 86.89% | 61.91% | 80.33% | 85.69% | 87.16% | 86.33% | 86.57% |
| | ADADELTA | 86.56% | **87.34%** | 87.08% | 86.78% | 87.03% | 86.82% | 87.18% |
| | ADAGRAD | 87.22% | 86.46% | 87.11% | 86.17% | 86.59% | 86.68% | 86.97% |

TABLE II

PERFORMANCE OF INCREMENTAL VS CROSS-MODALITY CLASSIFIERS

| Trained on | Validated on | Tested on | TPR | FPR | ACC |
|---|---|---|---|---|---|
| Correlated cross-modality | Intensity | Intensity | 0.972 | 0.0737 | 94.4% |
| Correlated cross-modality | Depth | Depth | 0.9112 | 0.0172 | 86.06 % |
| Correlated cross-modality | Flow | Flow | 0.9115 | 0.152 | 87.38 % |
| Depth+Flow+Intensity | Depth, Flow, Intensity | Intensity | 0.9619 | 0.029 | 96.7 % |
| Intensity+Flow+Depth | Intensity, Flow, Depth | Depth | 0.8764 | 0.095 | 89.39 % |
| Depth+Intensity+Flow | Depth, Intensity, Flow | Flow | 0.9436 | 0.056 | 94.34 % |

- 25608 samples of pedestrians
- 16235 samples of non pedestrians

The experiments are performed in the Caffe deep neural network framework. The performances are measured by the Accuracy (ACC) and using the Receiver Operating Characteristics (ROC) curve created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Withal, the complexity of classification system is also investigated by the area under the curve (AUC).

### A. Benchmark of uni-modal classifiers

We start by comparing for each modality images the classification performances with LeNet architecture with different learning algorithms: Stochastic Gradient Descent (SGD), Adaptive Gradient (ADAGRAD), RMSPROP, ADADELTA and learning rate polices: Fixed (FIX), Exponential (EXP), Step Down (STEP), Polynomial Decay (POLY), Sigmoid (SIG), Multi-Step (MS) and Inverse Decay (INV) (see Table I). It is to be noted that the CNNs were optimized on the training set through 29760 epochs and 0.01 learning rate for both the single modality and the incremental cross-modality models. Each modality classifier is exclusively trained with imagines of its own modality. For the Intensity modality the best performance (ACC = 96.56%) was achieved with the LeNet architecture using the RMSPROP [2] algorithm learning, with POLY rate policy. The best performances are obtained in Depth images with SGD and EXP settings (ACC = 89.78%) and respectively in Flow images ADADELTA and FIX settings (ACC = 87.34%). Therefore, different modalities need different learning algorithms and rate policies for an optimal training.

[2]Tieleman, T. and Hinton, G.,Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural Networks for Machine Learning, 2012

### B. Benchmark of cross-modality training methods

The CNNs were optimized on the training set through 29760 epochs and 0.01 learning rate for both the single modality and the incremental cross-modality models. The CNN belonging to the correlated cross-modality approach needs three more times training epochs (89220 epochs) for the same learning rate.

Since the RMSPROP with POLY settings achieved on Intensity modality the best performance, we decided to use the same settings to train the correlate cross-modality (CCM-CNN). The CNN model is validated following two different approaches on the multi-modality union data set, or single modality ones (see Table II). The second approach gives better results. This correlated cross-modality training outperforms the classical intra-modality training only on Flow testing set. This may be explained by the fact that, with more complex training data, the breadth and depth of the network should be increased. However the complexity would be limited by computing resources, which thus hinders performance (see Table II) [15], [16].

For the training following the incremental cross-modality method, we use for RMSPROP with POLY settings for all three CNNs through 29760 epochs. The results are given in (sea Tab II) are better than those achieved with classical training (see Tab I) for the same settings.

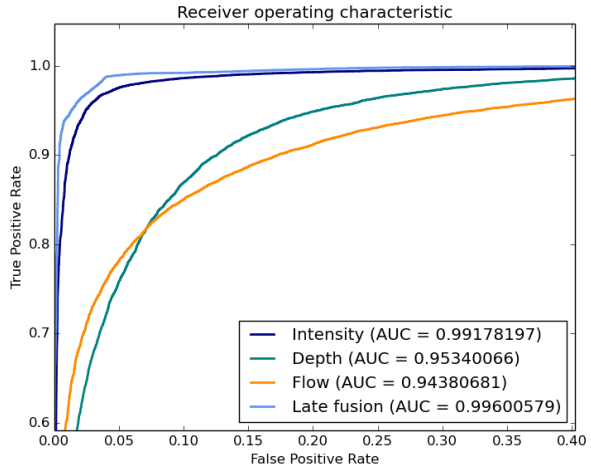
### C. Late-fusion with classical vs cross-modality training

We show in Table III the performance obtained with classical training versus incremental cross-modality. The

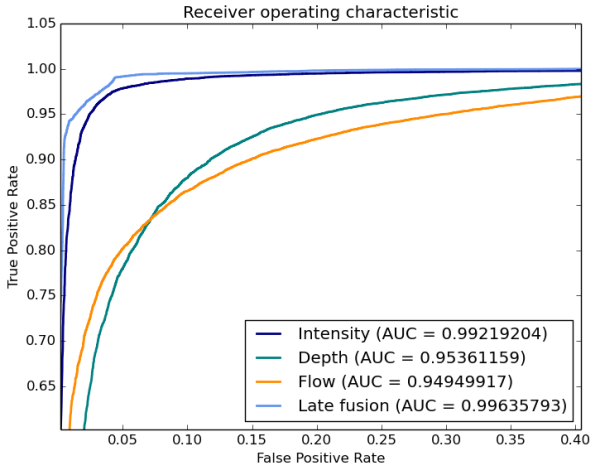
TABLE III

PERFORMANCE WITH LATE FUSION ON DAIMLER TESTING SET

| Late-fusion | TPR | FPR | ACC |
|---|---|---|---|
| classical training | 0.9518 | 0.0109 | 97.46 % |
| incremental cross modality | 0.9534 | 0.0092 | **97.62 %** |




(a) Classical training




(b) Incremental cross-modality

Fig. 4. ROC classification performance on Daimler testing data set.

incremental cross-modality late-fusion solution we proposed outperforms not only all the single modality classifiers but also the classical late-fusion solution. These performances are also shown in the ROC curves (see Fig 4).

## V. CONCLUSIONS

In this paper, we proposed different cross-modality training approaches for late-fusion architecture to improve the pedestrian recognition. The incremental correlated cross-modality approach outperforms the correlated cross-modality one. The incremental method improves the classification performance compared to a classical training of unimodal CNNs through late-fusion schemes on Daimler data set. We believe that the correlated approach is the promising one.

Future work will be concerned with improving that model by using optimal settings for different training modality sets and also by extending the model to cross data-sets training.

## VI. ACKNOWLEDGEMENTS

The research for this paper was financially supported by the Normandy Region and Inria Paris.

## REFERENCES

[1] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann Lecun. Pedestrian detection with unsupervised multi-stage feature learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[2] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. Taking a deeper look at pedestrians. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[3] H. Fukui, T. Yamashita, Y. Yamauchi, H. Fujiyoshi, and H. Murase. Pedestrian detection based on deep convolutional neural network with ensemble inference network. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 223–228, June 2015.

[4] Anelia Angelova, Alex Krizhevsky, and Vincent Vanhoucke. Pedestrian detection with a large-field-of-view deep network. In *IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*, pages 704–711, 2015.

[5] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. *Ten Years of Pedestrian Detection, What Have We Learned?*, pages 613–627. Springer International Publishing, Cham, 2015.

[6] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743–761, April 2012.

[7] M. Enzweiler and D. M. Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. *IEEE Transactions on Image Processing*, 20(10):2967–2979, Oct 2011.

[8] R. Bunel, F. Davoine, and Philippe Xu. Detection of pedestrians at far distance. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2326–2331, May 2016.

[9] M. Eisenbach, D. Seichter, T. Wengefeld, and H. M. Gross. Cooperative multi-scale convolutional neural networks for person detection. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 267–276, July 2016.

[10] Xiaogang Chen, Pengxu Wei, Wei Ke, Qixiang Ye, and Jianbin Jiao. *Pedestrian Detection with Deep Convolutional Neural Network*, pages 354–365. Springer International Publishing, Cham, 2015.

[11] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 509–514, April 2016.

[12] *Fusion of Stereo Vision for Pedestrian Recognition using Convolutional Neural Networks*, 2017.

[13] David Vazquez, Antonio M. Lopez, Javier Marin, Daniel Ponsa, and David Geronimo. Virtual and real world adaptation for pedestrian detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 36(4):797–809, 2014.

[14] C. Karaoguz and A. Gepperth. Incremental learning for bootstrapping object classifier models. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1242–1248, Nov 2016.

[15] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.