Single Shot Multitask Pedestrian Detection and Behavior Prediction

Prateek Agrawal
Volkswagen Group
Innovation Center California
Belmont, California 94002
prateek.agrawal@vw.com

Pratik Prabhanjan Brahma
Volkswagen Group
Innovation Center California
Belmont, California
pratik.brahma@vw.com

Abstract

Detecting and predicting the behavior of pedestrians is extremely crucial for self-driving vehicles to plan and interact with them safely. Although there have been several research works in this area, it is important to have fast and memory efficient models such that it can operate in embedded hardware in these autonomous machines. In this work, we propose a novel architecture using spatial-temporal multi-tasking to do camera based pedestrian detection and intention prediction. Our approach significantly reduces the latency by being able to detect and predict all pedestrians' intention in a single shot manner while also being able to attain better accuracy by sharing features with relevant object level information and interactions.

1 Introduction

A typical autonomous driving modular stack consists of several modules like sensors, perception, localization, prediction, planning, and control. The perception module consists of tasks, like object detection or free space detection, that perceive the environment around the ego vehicle. The prediction module can anticipate the intended behavior and future trajectories of traffic agents. In order to deal safely with Vulnerable Road Users (VRUs) like pedestrians, both of these modules are extremely crucial.

The standard pipeline typically perceives the pedestrians first. This is followed by predicting their behavior or future trajectory by using information like their past motion from corresponding sensor data. We call this as the sequential approach. Pedestrians typically do not follow kinematic models like vehicles and thus it is challenging to have fast and highly accurate predictions in a computationally efficient manner. As an overall system, the sequential approach demands high memory usage and high inference time which may be problematic given the limited computing capability.

In this work, we present a novel approach to solve the problem of pedestrian intention prediction, as defined in [18], by taking a parallel multi-tasking approach between detection and prediction. As shown in Figure 1, we hypothesize that our approach can have manifold advantages: (a) Low latency by executing parts of prediction and perception in parallel, (b) Practical and fixed inference time by detecting and predicting for all pedestrians in a single shot manner, (c) Low parameter and feature memory usage, (d) Multi-tasking leverages correlations to help improving accuracy. In the following sections, we explain the design and compare its performance with a baseline sequential approach on the Pedestrian Intention Estimation (PIE) data set [18].

Machine Learning for Autonomous Driving Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

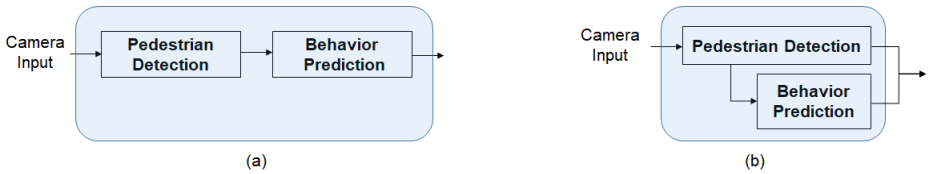


Figure 1: (a) Sequential approach with perception followed by prediction, and (b) proposed auxiliary approach that multi-tasks both followed by a spatial association of pedestrian behavior

2 Related Work

Pedestrian understanding Several recent works [11, 12, 7] have pushed the boundaries in the field of pedestrian detection. Also for most multi-object detection and pixel wise semantic segmentation driving data sets [4, 23], pedestrian is usually a distinct category. Pedestrian understanding however goes beyond that by attempting to detect multiple aspects [2] like pose [5], gesture [19] and actions [3] of human beings and being able to predict the intended behavior and eventually the actual trajectory that the pedestrian is expected to execute in future. Most of the approaches [18, 13] for intention prediction however fall under the sequential approach mentioned in Figure 1(a). Predicting pedestrian behavior is also an interactive problem since each person's future depends on the state of other VRUs, vehicles, traffic lights and other proximal map elements. Although there have been several works [6, 22, 1] towards conducting a social prediction for all interacting agents, most of these again assume that the pedestrians have already been detected and are being tracked.

Multi-task learning Many works [21, 15, 9] have shown the advantage of using multi-task learning for different applications, like Mask R-CNN [8] for instance level segmentation. There have been efforts like [16] to do multi-task pedestrian detection and action recognition but the actual time-to-cross prediction is only done after the bounding boxes are obtained through detection. It is also similar in [17] where multiple prediction sub-modules are multi-tasked but only run after obtaining semantic segmentation output. PnPNet [10] proposes end-to-end learning of perception and prediction but the architecture still executes the trajectory prediction LSTM only after the actual detection network. FaF [14] is another end-to-end solution for detection, tracking and motion forecasting but also places the forecasting prediction model after detection. Thus, the inference time for these methods is typically large. Contrary to these approaches, our proposed methodology leverages low level feature sharing to capture better interactive information and parallel execution of task specific layers for doing detection and behavior prediction of all pedestrians simultaneously.

3 Methodology

The baseline model that we compare with is as suggested in [18]. They use a convolutional LSTM model on top of pre-trained VGG16 features obtained from cropped image regions of pedestrians using ground truth annotations (for a real life system, this input would come from the perception system and hence it is a sequential model) in 15 continuous frames.

Our proposed method contains a multi-task architecture that consists of an object detection network called the primary network and a pedestrian intention network called auxiliary network as shown in Figure 2. For every input image, the auxiliary network receives its inputs from the feature maps calculated at an intermediate L^{th} layer of the primary network and thus operates in parallel to the remainder of the primary network. Given a sequence of images from time T_0 to time T_t, the primary network detects the bounding boxes of all the pedestrians in the images. The auxiliary network uses the information from previous frames and outputs the intention to cross for all the pedestrians in the frame at time T_t. In our implementation, the primary network is a single shot YOLOv2 [20] detector with output shape of $H \times W \times A \times (1 + N_C + 4)$, where A is the number of anchor boxes and N_C defines the number of classes to be detected. We consider $N_C = 4$ with pedestrians, cross-walks, vehicles and traffic lights as the object classes since this can help the feature maps to capture relevant information regarding the spatial-temporal relationships between the scene objects and can be useful for the auxiliary pedestrian intention network. The auxiliary network is a three layered ConvLSTM model. It receives a time sequence of $t = 15$ feature maps from an intermediate

layer $L = 18$ of the YOLOv2 and produces a similar spatial grid output of shape $H \times W \times A \times N_I$, where first three dimensions H, W and A has the same values as the primary network's output and the fourth dimension N_I defines the number of pedestrian intention classes. The raw images of size 1920×1080 are resized to 640×360 before feeding as input to the YOLOv2 and our final layer parameters are $H = 11$ and $W = 20$ with $A = 5$ anchor boxes. $N_I = 2$ indicating whether the pedestrian intends to cross the street or not. We use binary cross entropy as the auxiliary loss. The last step is a constant time spatial mapping between the pedestrian detected from the primary network and the corresponding intentions from auxiliary network. If a pedestrian is detected at {i,j,k} grid in the YOLOv2 output, we assign the softmax output at the same {i,j,k} grid at the auxiliary network as the corresponding classified intention for the detected pedestrian. Thus, our architecture does not have to wait for the detection to be over before performing prediction. During the training process for the auxiliary side, only the loss corresponding to the cells that contain pedestrians are computed and back propagated to train the layers below. The modules can either be trained end-to-end in a multi-task learning manner or just the auxiliary network can be trained separately by using features from a pre-trained object detector. The latter scenario can be particularly useful when the perception network is either trained on a separate and comparatively bigger data set or in real life automotive system integration scenarios where the detection module may be obtained from a different party or supplier. For our experiments in the current paper, we assume an oracle tracking system and thus use the object tracking IDs directly from annotations. However, tracking can also be included in the whole end-to-end system, similar to as it is described in FaF [14].

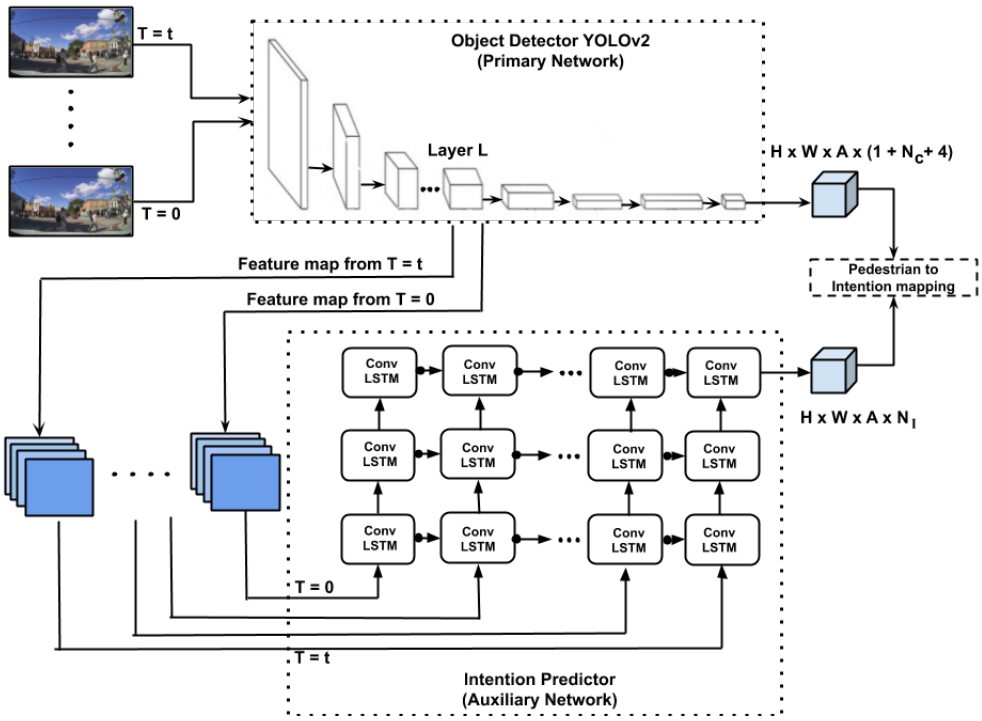


Figure 2: Spatial-temporal multitask network to do detection and intention prediction of all pedestrians simultaneously

4 Experiments

4.1 Dataset

In this paper, we use the PIE data set [18] to train both the detection and prediction models. The pedestrian intention was annotated using Amazon Mechanical Turks where each human subject was asked to observe a highlighted pedestrian in a sequence of consecutive frames and answer whether the

Table 1: Comparing intention prediction results for auxiliary and multitask approaches with baseline sequential approach. The branching is done at the 18^{th} layer of YOLOv2 detector

Method	Pedestrian Height (px)	Accuracy	F1 Score
Sequential	All peds > 0	79%	87%
Auxiliary Training	All peds > 0	81.19%	88.67%
	All peds > 120	82.00%	88.27%
Multitask Training	All peds > 0	83.83%	90.59%
	All peds > 120	84.37%	90.246%

Table 2: Comparing intention predictions results for different values of L

Input Layer	Accuracy	F1 Score
17^{th}	76.79%	83.85%
19^{th}	76.26%	83.87%
20^{th}	75.1%	82.82%

pedestrian wants to cross the street. Intention of a pedestrian is defined as the ultimate objective of the pedestrian and not the actual path the pedestrian executes based on the traffic scenarios. All videos are recorded in HD format (1920 $\times$ 1080 pixels) at 30 frames per second with intention annotated for a total of 1842 pedestrians.

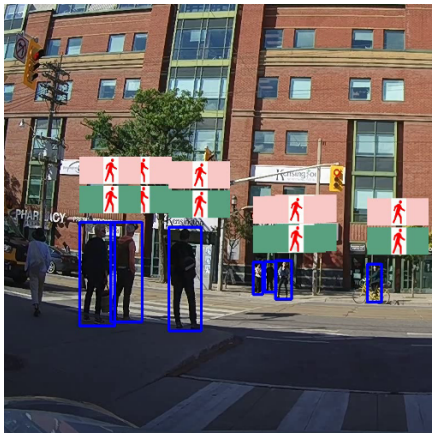
However, we found out that not all the pedestrians that can potentially interact with the ego vehicle are annotated. This posed difficulty for us in training the primary YOLOv2 network because of too many false negatives. Thus, in our current paper, we primarily compare the improvement in the intention prediction task only using our methodology over the sequential approach.

4.2 Quantitative Results

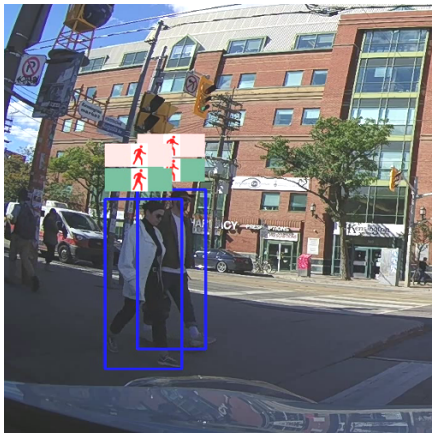
The results shown in this section are on the default test set of the PIE data set. Table 1 summaries the accuracy and F1 score obtained for the baseline as well as the approaches presented in this paper for the 18^{th} layer feature map from YOLOv2 as input to auxiliary model. Using the code and model weights provided by [18], we were able to reproduce the intention prediction accuracy of 79% (and an F1 score of 87%) for the baseline sequential approach as claimed in their paper.

For the auxiliary only training, we first trained the YOLOv2 object detector and subsequently froze all its layers. Given a training image, we pass the image through the frozen object detector, extract features from an intermediate layer and use that as input to the auxiliary intention network. For a sequence of images, we would then compute the loss for training only the layers in the auxiliary network. Keeping all the hyper-parameters same as set by [18] and trying out multiple different intermediate layers from object detector, the auxiliary approach was able to obtain an accuracy of 81.19% and F1 score of 88.67% with no size restriction on the pedestrians. Restricting the pedestrian size to a more humanly perceivable constraint (that is by considering only those pedestrians with size more than 120 pixels in the original high definition whole scene image) increased the accuracy to 82%. We found that the best accuracy is achieved when the 18^{th} layer from YOLOv2 is used as the input to auxiliary network. We did a grid search to fix this hyperparameter and results from some other intermediate layers as input to auxiliary network are shown in Table 2. By L^{th} layer, we mean taking the features generated after the application of batch normalization and activation on the L^{th} convolutional layer of the YOLOv2 architecture.

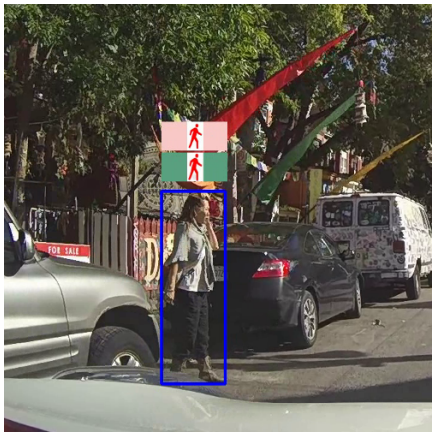
Keeping all the hyper parameters same, we finally did the end-to-end multi-task training of both the detection and prediction networks simultaneously. This helped the common feature extraction layers in both detection and prediction tasks to be generic enough for both the tasks. This gave us a further increase in the accuracy to 83.83% and F1 score to 90.59%. Restricting the size of the pedestrian to a height of greater than 120 pixels only further increased the accuracy to 84.37% and F1 score to 90.246%. We thus see an improvement of 4.83% in the pedestrian intention from the sequential approach. This end-to-end multi-task training also helped to improve the mean average precision (mAP) of the object detection model by 2%.



(a)

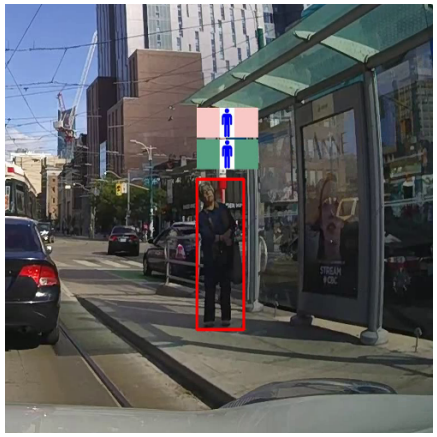


(b)

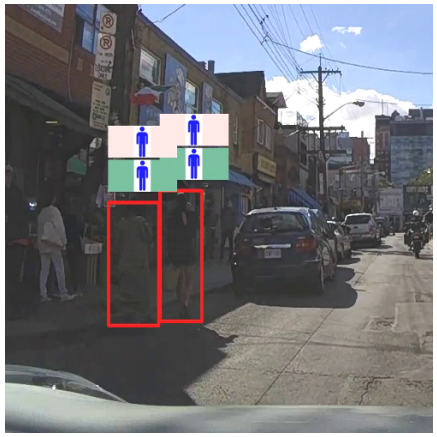


(c)

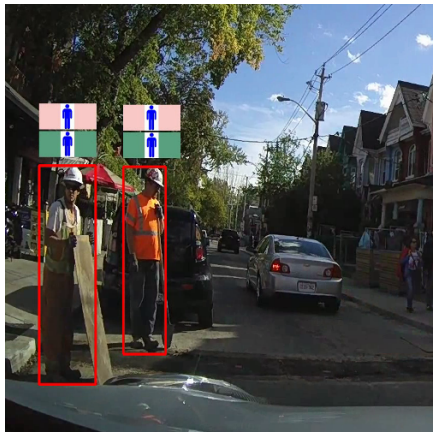
Figure 3: Pedestrian crossing results - Green background icon shows the ground truth intention and pink background icon shows the predicted intention



(a)



(b)



(c)

Figure 4: Pedestrian not crossing- Green background icon shows the ground truth intention and pink background icon shows the predicted intention

Table 3: Comparing memory and latency at the overall system level

Method	Memory (MB)	Time to prediction
Sequential	276.1	137.576ms
Auxiliary/Multi-Tasking	261.5	50.4ms

Other major advantage of the proposed approach is the lower latency and lower parameter memory footprint. As shown in Table 3, the parameter memory of the auxiliary model was about 15 MB lower than the total memory of sequential approach. This can be attributed to the fact that the sequential method uses an extra feature extractor (VGG16) where as the auxiliary method utilizes the features extracted by the object detection network itself. The average inference time for a single image also reduced from 137.5 ms to 50.4 ms. This is because of the single shot inference behaviour of our approach in contrast to running the intention model for individual pedestrians separately in sequential method. Inference time of the sequential approach depends on the number of pedestrians in the image. Increased number of detected pedestrians increases the inference time for sequential approach. On the other hand, the inference time for the auxiliary approach is fixed and does not depend on the number of pedestrians in the image. The numbers reported here are corresponding to the average pedestrians per image in the PIE data set, which is 2.12. This would help for faster detection of the dynamic behaviour of the pedestrians. A more elaborate study on reducing the memory requirement and inference time with better hyper parameter optimization and architecture search is left as future work.

4.3 Qualitative Results

Figure 3 and 4 compare the predicted intention of the pedestrians with the ground truth annotations. In a crowded scene as seen in Figure 3(a), we can see the benefit of such a single shot approach for predicting the intention of all the pedestrians at the same time irrespective of the number of pedestrians in the image. Also, since the pedestrians are never cropped out of the original image, the intention network has access to the features of stimulus objects in the image such as traffic light, cross walk and vehicles. Figure 3(b) and 3(c) show the intention being predicted correctly for pedestrians that are walking and standing in the images respectively. Figure 4(a) shows a pedestrian who is waiting for something and the model correctly predicts the intention of not crossing. Figure 4(b) and (c) show pedestrians walking parallel to the ego vehicle and construction workers with no intention to cross the road respectively. Our model is able to accurately predict their intended behavior.

5 Conclusion

Given the complex dynamics and interactive nature of pedestrian motion patterns, it is extremely critical for autonomous vehicles to predict their intended behavior with maximum accuracy and in the most computationally efficient manner. We proposed a novel way to multitask detection and behavioral prediction directly from sensors data. We could show 2.7x inference speed increase and accuracy improvement of 4.83%. It is worth noting here that the memory and latency improvement obtained by our methodology is orthogonal to other techniques like pruning, quantization and other hardware optimization of neural networks. Such multi-task methodology can be applied onto dealing with any combination of robotic perception and prediction tasks. For example, one can also use such spatial temporal feature sharing for single shot detection as well as prediction of the future trajectories of all vehicles in the scene in order to anticipate possible scenarios like cut ins and sudden braking.

One possible weakness of our approach can be that the compressed features may or may not capture very fine details like hand gesture, eye gaze or head pose that may be needed to predict pedestrian intention whereas sequential approach may have an advantage here since its input is rectangular crops of pedestrians at the original whole image resolution. However, we hope that joint multi-task training can force the common layers to focus and retain such needed features. Although it requires further and extensive experiments to validate the safety and reliability of this methodology, our initial results seem encouraging to continue working in this direction.

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[2] F. Camara, N. Bellotto, S. Cosar, D. Nathanael, M. Althoff, J. Wu, J. Ruenz, A. Dietrich, and C. W. Fox. Pedestrian models for autonomous driving part i: low level models, from sensing to tracking. *arXiv preprint arXiv:2002.11669*, 2020.

[3] L. Chen, N. Ma, P. Wang, J. Li, P. Wang, G. Pang, and X. Shi. Survey of pedestrian action recognition techniques for autonomous driving. *Tsinghua Science and Technology*, 25(4):458–470, 2020.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[5] Z. Fang and A. M. López. Intention recognition of pedestrians and cyclists by 2d pose estimation. *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[6] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

[7] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao. Pedestrian detection: The elephant in the room. *arXiv preprint arXiv:2003.08799*, 2020.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[9] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[10] M. Liang, B. Yang, W. Zeng, Y. Chen, R. Hu, S. Casas, and R. Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2020.

[11] W. Liu, I. Hasan, and S. Liao. Center and scale prediction: A box-free approach for pedestrian and face detection. *arXiv preprint arXiv:1904.02948*, 2019.

[12] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2019.

[13] J. Lorenzo, I. Parra, F. Wirth, C. Stiller, D. F. Llorca, and M. A. Sotelo. Rnn-based pedestrian crossing prediction using activity and pose-related features. *arXiv preprint arXiv:2008.11647*, 2020.

[14] W. Luo, B. Yang, and R. Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018.

[15] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016.

[16] D. O. Pop, A. Rogozan, C. Chatelain, F. Nashashibi, and A. Bensrhair. Multi-task deep learning for pedestrian detection, action recognition and time to cross prediction. *IEEE Access*, 7:149318–149327, 2019.

[17] A. Ranga, F. Giruzzi, J. Bhanushali, E. Wirbel, P. Pérez, T.-H. Vu, and X. Perotton. Vrunet: Multi-task learning model for intent prediction of vulnerable road users. *Electronic Imaging*, 2020(16):109–1, 2020.

[18] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *International Conference on Computer Vision (ICCV)*, 2019.

[19] S. S. Rautaray and A. Agrawal. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial intelligence review*, 43(1):1–54, 2015.

[20] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[21] S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[22] L. A. Thiede and P. P. Brahma. Analyzing the variety loss in the context of probabilistic trajectory prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9954–9963, 2019.

[23] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2020.