Multimodal Late-Fusion for stereo Object Detection

Amaya Dharmasiri
University of Moratuwa
170131R@uom.lk

Ramith Hettiarachchi
University of Moratuwa
ramithuh@gmail.com

Isuru Dissanayake
University of Moratuwa
170146R@uom.lk

Sadeep Jayasumana
Five AI
sadeep@five.ai

Abstract

Most of the advanced driver assistant systems and autonomous driving systems use stereo cameras to obtain depth information. Due to the baseline, camera orientations and different levels of truncation and occlusion of objects with respect to two cameras, object information which isn't available in one image could be available in its complementary stereo image. A less explored subject is the combination of these stereo images to improve the overall 2D object detection ability. In this paper, we investigate different aspects of combining object detections between pairs of stereo images, after applying the same object detector on both images separately, thus improving the precision and recall. We also deduce a theoretical maximum for this improvement (for a particular object detector), based on the correlation of the considered stereo pairs. Experiments were carried out with the KITTI- 2D object detection data set [3].

1. Introduction

Object detection has been a classical computer vision problem for a long time and is still an active research area. Throughout the years, numerous highly efficient approaches have been proposed and used to solve this problem, out of which deep neural networks take a prominent place. The performances of these models have been improving over the years.

With the development of advanced driver assistant systems and autonomous driving cars, the traditional object detection problem has been given a vital importance. The aforementioned object detection algorithms and models have been deployed to identify cars and other vehicles, pedestrians, cyclists and other similar classes of objects that frequently appear in the scenes that are encountered by these vehicles.

Figure 1: A sample image taken from KITTI data set. Top image shows an object detected only in the left stereo image, and the bottom image, another object detected only in the right stereo image(plotted on the left stereo image) Two vantage points for the same scene has resulted in detecting objects that are not clearly visible in its complementary stereo image

These vehicles are equipped with a set of cameras and other sensors to get data about its surrounding. A pair of stereo cameras are used mainly to obtain the depth information of the scene. While traditionally, the object detection and localization algorithms are implemented on the feed of one of these cameras, the data that can be obtained from the second stereo camera is sometimes neglected. But due to its relative positioning and the angle with respect to the first stereo camera, additional object information that is not available in the first camera field due to high truncation and occlusion levels are available in the second camera field.

This additional information can be utilized to increase the precision and recall of the object detection, which holds a vital importance in the context of driver assistant and autonomous driving systems. In this paper, we are exploring the effects of utilizing this second set of data in improving the performance of the object detectors.

The data set used for this investigation is KITTI 2D object detection data set. The sensor setup of the KITTI vision benchmark suite contains two RGB stereo cameras with a baseline of 0.54m. 7480 images each for training and testing is provided in the form of rectified stereo image pairs [3].

For our investigation, we used a faster R-CNN resnet101 [11] [6] model trained on the KITTI data set with a pascal mAP@0.5 of 87 (obtained from the tensorflow detection model zoo [7]). At the initial stage, the investigation was limited to two object classes; cars and pedestrians. This model is separately implemented on the two set of stereo images; left and right, detections are mapped onto the same image plane, combined and evaluated as explained sequentially in the rest of the paper.

2. Related work

2.1. Object detection

Object detection has been a component with vital importance in computer vision for a long time. Over the years, optimized versions of region proposal methods and region-based convolutional neural networks have been developed. The training and testing times for data sets have been reduced drastically with the introduction of region-based convolutional neural networks such as R-CNN, Fast R-CNN[4] and Faster R-CNN[11].
Faster R-CNN, the network used in this paper, eliminates the need for a selective search to get regional proposals. Rather, a trained region proposal network is used which further optimizes it with regards to time. In the model we used, Resnet101 [6] has been used as a feature extractor.

2.2. Combining object detectors

So far, numerous approaches have been proposed to combine the results of different object detectors on a particular image or image set, such that the combined/fused detector performs better on images under different conditions. Fusions based on approaches such as, fuzzy logic [2] [9], Dempster-Shafer theory of evidence [1], correspondence analysis [5] and voting [10] have been proposed. Some of the recent research on combining object detectors such as [13] have been based on Selective cross domain alignment and have surpassed the performance of earlier methods substantially.

In contrast to the above methods, here the same object detector is used on two sets of images, resulting in two sets of detections. And once we map them on the same image plane, the problem simplifies to a fusion/combining problem similar to those cited above.

3. Stereo Vision

Stereo vision basically refers to the use of two or more vantage points of the same scene to extract 3D information. This has a great importance in the contexts of driver assistant and autonomous driving systems as it provides a sense of depth to the surrounding objects. The KITTI benchmark suite has been recorded with a sensor setup including 2 color cameras, 1.4 Megapixels, with a baseline length of 0.54m [3]. The frames thus recorded are rectified, and provided along with the camera calibration data in the object detection data set. In this paper we investigate the use of additional information from the second stereo image to optimize the precision and recall of 2D object detectors.

3.0.1 Perspective projection

In a stereo camera system, the corresponding point on the complementary stereo image to a particular point on one image plane can be related to each other using the projection matrix specific to that image, and the depth to that particular object/point in the camera coordinate system. In this paper, an approach similar to J. Siswantoro et al.[12] was used.

Let us consider a scenario where we need to map a particular point a on the right image plane to its corresponding point on the left image plane. a can be represented in the homogeneous and world coordinate system(A) as ,

$$a = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad and\, A = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

The camera projection matrix P for each image provided in the datset, contains the following information.

$$P = \begin{pmatrix} f & 0 & c_u & k_1 \\ 0 & f & c_v & k_2 \\ 0 & 0 & 1 & k_3 \end{pmatrix}$$

f- focal length
$c_u, c_v - origin\, of\, image\, coordinate\, system\, in\, camera\, coordinates$
$k_1, k_2, k_3 - origin\, of\, camera\, coordinate\, system\, in\, world\, coordinates$
Hence, the homogeneous(image) coordinates of point a can be written in terms of world coordinates (A),

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = P \times \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{1}$$

And the world coordinates (A) in terms of image coordinates (a) can be calculated as,

$$\begin{aligned} X &= ((x - c_u) * Z + k_3 * x - k_1) / f \\ Y &= ((y - c_v) * Z + k_3 * y - k_2) / f \end{aligned} \quad (2)$$

Depth of the point in the world coordinates (Z) can be obtained directly from depth analysis which is explained in section 4

The bounding boxes obtained by running the object detector on right stereo images are represented in terms of the bottom left and top right corners. These two points, were then transformed to world coordinates using equation 2 and right camera projection matrix. Then, the world coordinates are transformed back to image coordinates of the left stereo image using equation 1 and the left camera projection matrix.

4. Depth Analysis

The detections (bounding boxes) obtained by running the object detector on the right stereo images are mapped to their corresponding locations on the left images as described above. During the procedure, the depth of that particular object (bounding box) in the camera coordinate system is required. Two approaches were deployed to obtain the depth information.

4.1. LIDAR

The sensor setup for the KITTI vision benchmark suite is equipped with a laser scanner that spins at 10 frames per second, capturing 100k points per cycle with a vertical resolution of 64. These LIDAR scans are rectified and cropped according to the scene to obtain the depth maps corresponding to the object 2D data set.
These LIDAR scans are sparse; with depth values for around 6% of the total pixels. In order to obtain depth values for the rest of the pixels, a classical image processing approach proposed by [8] was used.

4.2. Stereo

The stereo camera setup of the KITTI vision benchmark enables us to calculate the disparity between two stereo pairs. OpenCV's StereoBinarySGBM was used for the stereo matching and disparity calculation. Then, along with the baseline and focal length information provided with the data set, the pixel-wise depth were calculated according to

$$Depth = \frac{Focal\ length \times baseline}{Disparity} \quad (3)$$

4.3. Combining LIDAR and stereo

LIDAR depth information provided by the data set, are accurate to a considerable extent, but sparse. On the other hand, the stereo depths obtained as mentioned above are dense, but lack in accuracy. Hence to obtain more accurate depth information for all the pixels, a simple combining mechanism was devised. Consider any point(P) on an image. If LIDAR (sparse) data provided in the data set is available for the point P, depth for this point is taken directly from that. But if LIDAR data is not available, a new depth value is computed by a weighted average of the stereo depth values obtained as explained in section 4.2 and the depth values obtained by filling the sparse LIDAR matrix as in [8].

5. Combinations of detections

Once the depths to each object were obtained as explained in section 4.3, the bounding boxes of the objects detected from the images of right stereo camera were mapped onto the corresponding left camera image using the projection matrices and the equations 2,1. After this mapping, it is observed that, very often there are two bounding boxes for the same object. Therefore, to obtain a single bounding box for each object, a bounding box fusion/combination is carried out as described below.

5.1. Bounding Box Pairing

In order to do the fusing, first it is required to pair the corresponding bounding boxes of the considered object. For this, for each image a bounding box overlapping matrix (BBO matrix) is calculated by considering the extent of overlapping of the bounding boxes from left stereo image with the bounding boxes from right stereo image. BBO matrix contains the bounding box overlapping scores (BBO scores) calculated as follows.

Assume there are n number of detections from left stereo image and m number of detections from right stereo image. Then the BBO matrix is a (n*m) matrix with BBO scores. Consider the i^{th} bounding box from left stereo image(BB_i) and j^{th} bounding box from right stereo image(BB_j). Then two areas are calculated. We call them bounding box intersection area 6 and bounding box union area 7.

$$BB_i = ((x_{i0}, y_{i0}), (x_{i1}, y_{i1})) \quad (4)$$

$$BB_j = ((x_{j0}, y_{j0}), (x_{j1}, y_{j1})) \quad (5)$$

$$Bounding\ box\ intersection\ area = Area(BB_i \cap BB_j) \quad (6)$$

$$Bounding\ box\ union\ area = Area(BB_i \cup BB_j) \quad (7)$$

Using the above calculated two areas,bounding box overlapping score for i^{th} bounding box from left stereo image and j^{th} bounding box from right stereo image(BBO_{ij}) is

calculated as follows,

$$BBO_{ij} = \frac{Area(BB_i \cap BB_j)}{Area(BB_i \cup BB_j)} \quad (8)$$

After calculating the BBO matrix as above, the pairing is done by comparing the BBO scores obtained for the bounding boxes. Initially, a cut-off threshold is set for BBO scores such that bounding box pairs below this threshold will be disregarded in pairing and those BBO scores will be replaced by zeros in the BBO matrix. Afterwords, bounding boxes from left stereo image are paired with bounding boxes from right stereo image by considering the non-zero maximum BBO scores in the BBO matrix.

After this pairing, it is observed that there can be two basic types of detections mapped onto the left stereo image; paired detections and unpaired detections. Unpaired detections again consist of two types; unpaired detections of left stereo image(left unpaired detections) and unpaired detections of right stereo image(right unpaired detections).

5.2. Unions

The paired detections, left unpaired detections and right unpaired detections are combined to form the union-combined detection set. Bounding boxes for left unpaired detections and right unpaired detections, are taken as it is. On the other hand, since there are two bounding boxes for the paired detections, a union bounding box is computed using the two detections available as a pair according to equation9

$$BB_i \cup BB_j = ((min(x_{i0}, x_{j0}), min(y_{i0}, y_{j0})), (max(x_{i1}, x_{j1}), max(y_{i1}, y_{j1}))) \quad (9)$$

5.3. Intersections

Only the paired detections are considered in intersection-combined detections. Since there are two detections available as a pair, an intersection bounding box is computed according to the equation 10 to represent that particular object.

$$BB_i \cap BB_j = ((max(x_{i0}, x_{j0}), max(y_{i0}, y_{j0})), (min(x_{i1}, x_{j1}), min(y_{i1}, y_{j1}))) \quad (10)$$

6. Analysis of combined detections

In order to clearly observe the effect of combining the stereo object detections, the evaluations were carried out with the use of the most basic and obvious criteria; precision and recall, defined as;

$$Precision = \frac{True\ positives}{True\ positives + False\ positives}$$

$$Recall = \frac{True\ positives}{True\ positives + False\ negatives}$$

$$Miss\ rate = 1 - recall$$

To decide whether a particular detection (in the form of a bounding box) is a true positive or a false positive, the following methodology was followed.

Initially a pairing procedure similar to section 5.1 is implemented to pair the ground truths with detections. If there are any unpaired detections they are directly taken as false positives. On the other hand, if there are any unpaired ground truths, they are directly taken as false negatives. In the case of paired detections, true positives and false positives are differentiated according to a threshold.

Algorithm .1: Analysis of combined detections.

```
1  begin
2     for (bounding boxes ):
3          if (unpaired):
4               return False positives
5          else:
6               if BBO_ij >= Threshold :
7                    return True positive
8               else:
9                    return False positive
10    for (ground truth boxes):
11         if (unpaired):
12              return False negative
13         else:
14              pass
15 end
```

6.1. Unions

In this combination type, a bounding box is mapped onto the combined image plane only if the relevant object has been detected in either one of the stereo images. In addition to that, when the same object has been detected by both the detectors, the combined bounding boxes is obtained as the rectangle with minimum area that encloses all the pixels which comes under either bounding box from two stereo images according to equation 9.

Hence the intuitive expectation is that the recall of union-combined detections($Recall_U$) will increase.This effect can be quantitatively represented as below.

$$Pr(M_U) = Pr(\ M_L \cap M_R\ ) = \ Pr(\ M_L | M_R\ ) Pr(M_R) \quad (11)$$

In a case where the two set of detections are independent of each other, the above equation can be simplified to

$$Pr(M_U) = Pr(\ M_L) Pr(M_R\ )$$

Hence,

$$Recall_U = 1 - (1 - recall_L)(1 - recall_R) \quad (12)$$

6.2. Intersections

As opposed to the union-combinations, here a bounding box is mapped onto the combined image plane only if the same object has been detected by running the detector on both of the stereo image pairs. The resulting bounding boxes only enclose the pixels that come under both of the aforementioned single camera image bounding boxes according to the equation 10.

Hence the intuitive expectation is that the precision will increase, as the possible false positives are eliminated. The miss rate, and hence the recall can be quantitatively analyzed as below

$$Pr(M_I) = 1 - Pr(\ M_L \cup M_R\ )' = 1 - Pr(M'_L \cap M'_R) \tag{13}$$

If the two sets of detections are independent,

$$Pr(M_I) = 1 - Pr(M'_L)Pr(M'_R)$$

Hence,

$$Recall_I = recall_L \times recall_R \tag{14}$$

But as the two stereo images of the same scene have a high correlation between each other, the independence cannot be assumed in any of the above mentioned scenarios; union-combinations and intersection-combinations. Hence the deviation of the actual experimental results from the theoretical miss rates calculated in equations 4 and 6 (assuming independence of detections) can be used to quantify the correlation of the two sets of stereo detections.

As the correlation between the stereo images depend on factors like the baseline of cameras which are practically constrained, We can further expect that the miss rate obtained by the equation 4 is the minimum achievable miss rate for an object detector with miss rates 1 and 2 on the two seperate stereo image sets.

7. Experimental results

(Refer Tables 1 through 4)

At an IOU threshold of 0.5, a **1.9%** improvement in the number of true positives (for union-combined detections (cars)), a **1.9%** improvement in the recall (for union-combined detections (cars)), and a **1.7%** improvement in precision (for intersection-combined detections (cars)) were achieved, confirming our hypothesis.

As per the assumed independance of the stereo detections, equation 14 expects a recall value of 0.7874 for intersection-combined detections (cars), but the experimental results give a recall of 0.83655.

Again equation 12 expects a recall value of 0.9883 for union-combined detections (cars), but the experimental results give only 0.9788.

These differences in expected and real values give a measure of correlation between the two stereo image sets. In addition to that, 0.9883 can be stated as the theoretical maximum of the recall achievable for this detector, and the data set under combinations of stereo image object detections.

The labels for the KITTI object detection data set are annotated on the left stereo images only. Therefore some of the objects captured by running the detector on right stereo images are not clearly visible in the left stereo images, hence not annotated as ground truths. Due to this reason, the evaluation results for the right stereo image object detection is lesser than that of the left stereo images. For the experiment, we could manually annotate the objects in right stereo images as well, but for the time-being, evaluation is done only with respect to the provided ground truth. Nevertheless, a significant improvement in the number of true positives can be observed (in the union- combined detections).

8. Conclusion

In this paper we have presented a simple yet an effective approach to improve the overall object detection performance by combining the object detections of a pair of stereo image sets. The additional object information present in one stereo image enables the detector to identify objects that are not in obvious sight of its complementary stereo image. In addition to that, we have derived a theoretical maximum (for a specific detector and a stereo image set) for the recall achievable by the aforementioned combining method. Further research could carried out about other possible types of combinations (in addition to unions and intersections).

Table 1: Evaluation on Left Images

	IOU Threshold									
	0.4		0.5		0.6		0.7		0.8	
	Car	Ped	Car	Ped	Car	Ped	Car	Ped	Car	Ped
Precision	**0.9635**	**0.9260**	0.9629	0.9184	0.9612	0.9031	0.9533	0.8585	0.9217	0.7401
Recall	**0.9186**	**0.8623**	0.9180	0.8551	0.9164	0.8409	0.9088	0.7994	0.8787	0.6891
TP	**26401**	**3869**	26385	3837	26338	3773	26121	3587	25255	3092

Table 2: Evaluation on Right Images

	IOU Thresholds									
	0.4		0.5		0.6		0.7		0.8	
	Car	Ped	Car	Ped	Car	Ped	Car	Ped	Car	Ped
Precision	**0.9345**	**0.8067**	0.9295	0.7726	0.9175	0.7241	0.8861	0.6279	0.7760	0.4483
Recall	**0.8623**	**0.7228**	0.8577	0.6922	0.8465	0.6488	0.8176	0.5626	0.7160	0.4016
TP	24783	3243	24651	3106	24331	2911	23500	2524	20580	1802

Table 3: Evaluation on Union Boxes

	IOU Thresholds									
	0.4		0.5		0.6		0.7		0.8	
	Car	Ped	Car	Ped	Car	Ped	Car	Ped	Car	Ped
Precision	0.9206	0.8255	0.9170	0.8122	0.9101	0.7789	0.8944	0.6999	0.8246	0.5376
Recall	**0.9391**	**0.8846**	0.9354	0.8703	0.9283	0.8346	0.9124	0.7499	0.8411	0.5761
TP	**26991**	**3969**	26886	3905	26682	3745	26225	3365	24176	2585

Table 4: Evaluation on Intersection of Boxes

	IOU Thresholds									
	0.4		0.5		0.6		0.7		0.8	
	Car	Ped	Car	Ped	Car	Ped	Car	Ped	Car	Ped
Precision	**0.9816**	**0.9471**	0.9788	0.9136	0.9713	0.8648	0.9511	0.7797	0.8791	0.6081
Recall	0.8389	0.7136	0.8365	0.6884	0.8301	0.6517	0.8128	0.5875	0.7513	0.4582

Figure 2: Objects detected on the left stereo image

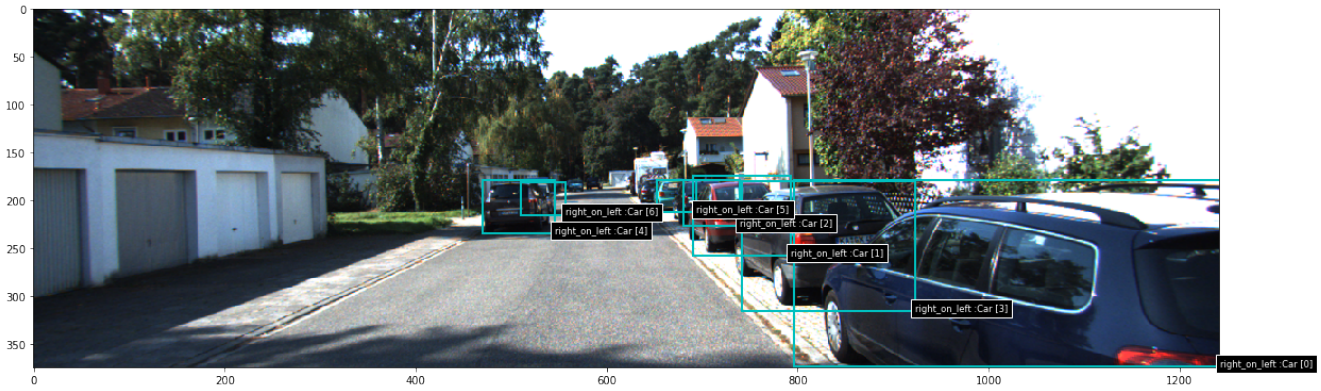

Figure 3: Objects detected on the right stereo image(plotted on the left stereo image)

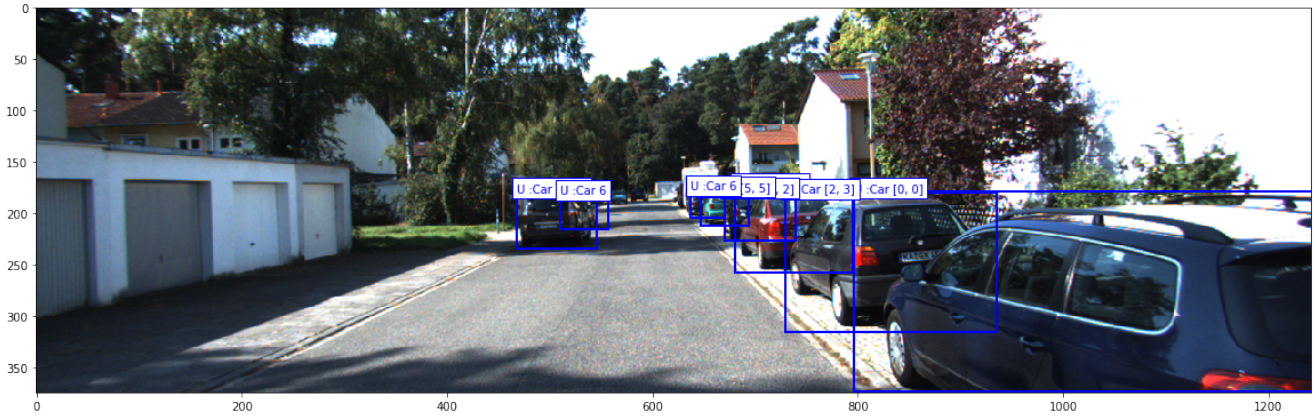

Figure 4: Union detections of right and left stereo images

Figure 5: Intersection detections of right and left stereo images

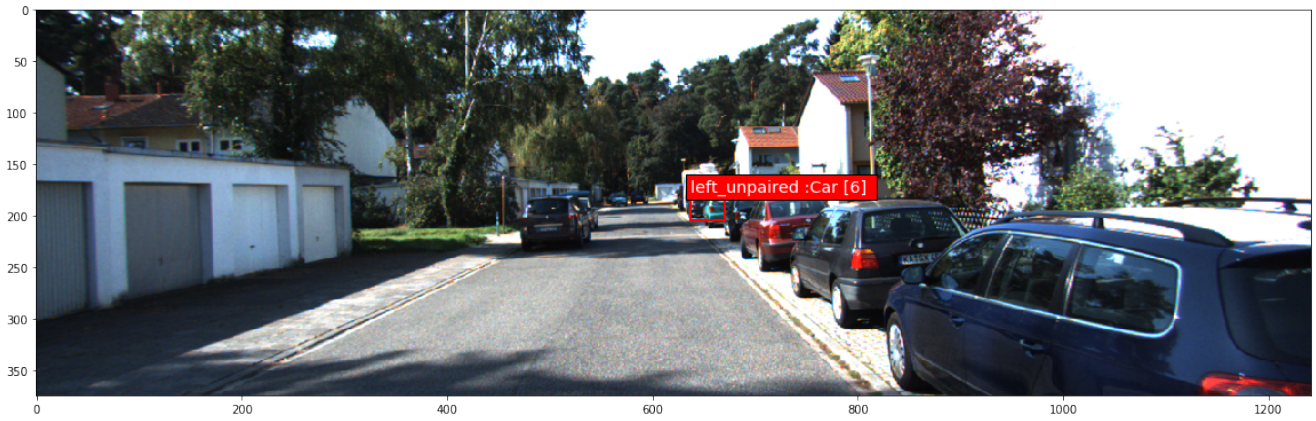

Figure 6: Objects detected only in the left stereo image

Figure 7: Objects detected only in the right stereo image(plotted on the left stereo image)

References

[1] A. Al-Ani and M. Deriche. A new technique for combining multiple classifiers using the dempster-shafer theory of evidence. *Journal of Artificial Intelligence Research*, 17:333–361, 2002.

[2] X. Chen, R. Harrison, and Y.-Q. Zhang. Genetic fuzzy fusion of svm classifiers for biomedical data. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 654–659. IEEE, 2005.

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

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

[5] M. J. Greenacre. *Correspondence analysis*. London: Academic Press, 1984.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017.

[8] J. Ku, A. Harakeh, and S. L. Waslander. In defense of classical image processing: Fast depth completion on the cpu. *arXiv preprint arXiv:1802.00036*, 2018.

[9] L. I. Kuncheva. " fuzzy" versus" nonfuzzy" in combining classifiers designed by boosting. *IEEE Transactions on fuzzy systems*, 11(6):729–741, 2003.

[10] K. T. Leung and D. S. Parker. Empirical comparisons of various voting methods in bagging. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 595–600. ACM, 2003.

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[12] J. Siswantoro, A. S. Prabuwono, and A. Abdullah. Real world coordinate from image coordinate using single calibrated camera based on analytic geometry. In *International Multi-Conference on Artificial Intelligence Technology*, pages 1–11. Springer, 2013.

[13] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin. Adapting object detectors via selective cross-domain alignment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.