

TOWARDS ROBUST AND EFFICIENT EMBODIED  
VISUAL PERCEPTION

ANUPAM SOBTI



AMAR NATH AND SHASHI KHOSLA  
SCHOOL OF INFORMATION TECHNOLOGY  
INDIAN INSTITUTE OF TECHNOLOGY DELHI

2022

# Towards Robust and Efficient Embodied Visual Perception

towards the degree of

**Doctor of Philosophy**

**Anupam Sobti**

2015ANZ8497

under the guidance of

**Prof. M. Balakrishnan**

**Dr. Chetan Arora**



Amar Nath and Shashi Khosla

School of Information Technology

Indian Institute of Technology Delhi



*To my parents  
and teachers*

# Certificate

This is to certify that the thesis titled “**Towards Robust and Efficient Embodied Visual Perception**” being submitted by **Anupam Sobti** for the award of **Doctor of Philosophy** is a record of bonafide work carried out by him under my guidance and supervision in the Amar Nath and Shashi Khosla School of Information Technology, Indian Institute of Technology Delhi. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma unless otherwise stated explicitly.

Certain works included in this thesis involved collaboration with other researchers, which has been explicitly specified/acknowledged in the corresponding chapters and the part done by those collaborators appeared in their respective reports/theses.

**M. Balakrishnan**

Professor

Department of Computer Science and Engg.

Indian Institute of Technology Delhi

New Delhi- 110016

**Chetan Arora**

Associate Professor

Department of Computer Science and Engg.

Indian Institute of Technology Delhi

New Delhi- 110016

## Acknowledgement

I express my sincere gratitude to my supervisor, **Prof M. Balakrishnan** who took the time to understand my motivation, always had the right thing to say at the right time, and provided valuable insights into building life as a researcher, a teacher and a citizen of the world. It was an extremely humbling experience to see the world through his eyes and wisdom, and I would carry this experience all through my life. **Prof. Chetan Arora**, my co-supervisor, has been a constant pillar in pulling me through the tough times, focussing me when I was scattered in my thoughts and helping me see brilliance in simple ideas. He has been instrumental towards the development of my rigor and ambition in research. I would like to thank the Ministry of Electronics, Government of India for awarding me with the Visvesvaraya Fellowship which supported me financially during the PhD.

I would like to thank all the B.Tech./M.Tech. students, interns and project staff with whom I worked on MAVI, as a teaching assistant or on research projects. It has been an enriching experience to work with, mentor and learn from them in various ways. I would also like to thank my colleagues who enriched the atmosphere with thought provoking discussions and inspired me through their disciplined work ethic.

I am ever grateful to my family for supporting my decisions throughout my career and their constant encouragement and understanding through the journey.

Last but not the least, I would like to thank all the administrative and lab staff of the Department of Computer Science and Amar Nath and Shashi Khosla School of Information Technology for their continuous support with various activities.

# Abstract

Computer vision has had tremendous success recently in perception tasks, especially, in cloud applications. An embodied system brings the additional objectives of memory, compute and energy optimization in addition to system accuracy. As opposed to processing in the cloud, an embodied agent has to perform in the wild without access to additional infrastructure. In this thesis, we address four major limitations of visual perception in embodied systems.

**From Images to Videos:** When dealing with image streams, there is often redundancy in the temporal neighborhood since objects are often repeated. Thus, a frame-level analysis leans towards the objects present in a larger number of frames rather than a fair evaluation for all objects in a video. To mitigate this, we propose an evaluation metric that clubs together different appearances of each object from multiple frames. In our work, we show how this results in metrics that are sensitive to bias against any objects in the videos.

**Resource Constrained Object Detection** When object detectors run at different speeds while processing an image stream, a number of frames have to be skipped to maintain real-time operation. Thus, a faster but less accurate detector may be able to process more frames than a slower detector which may be more accurate at the frame level. It is unclear, however, which of these object detectors would be able to detect more objects. This shows the dependence of object detection performance on the speed at which the algorithms are run. In this work, we propose a resource-aware evaluation metric that jointly reasons about speed and accuracy of object detectors for a true evaluation of object detection performance in an embodied system.

**Generalization to Unseen Environments:** Owing to the large number of parameters in deep neural networks, models often over-fit to the data distribution on which the models are trained. Even as perception improves on a certain dataset, the test environment changes continuously for an embodied agent. There is often limited visibility into the test data distribution in a deployed system. Thus, one has to ensure that the perception module in an embodied agent is aware and adaptive of this shift. We propose a method to use sparse data from an alternate modality (LiDAR) in order to adapt a pre-trained depth prediction network to the current context and provide accurate depth even in unseen environments.

**Energy Efficiency with Multimodal Sensing:** Visual perception is an energy intensive task. Even human beings rely on their model of the environment or alternate senses for perception, e.g., using sound as a cue for visual engagement while resting. In this work, we observe a similar phenomenon in a multimodal obstacle detection system. We show how a low power ultrasonic sensor is able to provide a low energy sensing mechanism to trigger visual perception. We take a detailed look at the available runtime modes for an obstacle detection system and perform an accuracy energy tradeoff to demonstrate different possibilities.

## संक्षेप

दृश्य बोध के क्षेत्र में पिछले दिनों वातावरण के अनुभूति कार्यों में काफी सफलता हासिल की गई है। एक अवतीर्ण उपकरण में सटीकता के अलावा मेमोरी, गणना और ऊर्जा संरक्षण के भी उद्देश्य होते हैं। एक अवतीर्ण उपकरण को ये सभी उद्देश्य बिना किसी क्लाउड सर्वर के कार्यान्वित करने होते हैं। इस शोध प्रबंध में हम दृश्य बोध की चार मुख्य बाधाओं का आँकलन करेंगे।

**चित्रों से वीडियो तक:** जब हम चित्रों की धारा में किसी वस्तु को देखते हैं तो उसमें बहुत सी अतिरेकता होती है। आस पास के चित्रों में एक ही वस्तु बार बार प्रस्तुत होती है। इसलिए हर चित्र पे किए गए वस्तु ढूँढने के कार्य का आँकलन सभी वस्तुओं का निष्पक्ष आँकलन न होकर उन वस्तुओं पे केंद्रित रहता है जो ज्यादा चित्रों में दिखाई देते हैं। इस बाधा का निवारण करने के लिए हमने एक नया आँकलन माप का निजात किया है। इस कार्य में हमने अनेक प्रयोगों के माध्यम से यह दिखाया है की हमारे सुझाए माप से हम वस्तु ढूँढने के कार्य का निष्पक्ष आँकलन कर पाते हैं।

**संसाधन विवश वस्तु ढूँढने का कार्य:** जब वस्तु ढूँढने के तरीके अलग अलग गति से किसी वीडियो पर चलते हैं तो वास्तविक समय के संचालन को बनाए रखने के लिए उन्हें बीच के चित्रों को छोड़ कर आगे बढ़ना पड़ता है। इस प्रकार एक तेज मगर कम सटीक तरीका अधिक चित्रों पे चल के अधिक वस्तुओं को ढूँढ सकता है। हालांकि फिर भी यह अस्पष्ट है की इनमें से कौनसा तरीका ज्यादा वस्तुओं को ढूँढ पाएगा। यह वस्तु ढूँढने के तरीकों की गति पर निर्भरता को दर्शाता है। इस कार्य में हम एक संसाधन समावेशी माप का निजात किया है जो गति और वस्तु ढूँढने की क्षमता का साथ में अनुमान लगाता है।

**उनदेखें वातावरण में समान्यकरणीय क्षमता:** अधिक मापदंडों के कारण गहरे तांत्रिक संजाल से चलने वाले अनुभूति कार्य अक्सर अधिकतम फिट हो जाते हैं। इस कारण ये अनदेखे वातावरणों में कार्य नहीं कर पाते। किसी एक वातावरण में अनुभूति सीखने के बावजूद एक अवतीर्ण उपकरण का वातावरण लगातार बदलता रहता है। अक्सर सिस्टम के तैनात होने के बाद भी परीक्षण का माहौल बदलता रहता है। इसलिए यह महत्वपूर्ण है की अनुभूति मापांक इस बदलाव के प्रति जागरूक और अनुरूप है। इस कार्य में हम एक वैकल्पिक साधन (लाइडार) का प्रयोग करके एक पूर्व-प्रशिक्षित संजाल को वर्तमान संदर्भ के प्रति अनुरूपित करने का तरीका प्रस्तुत कर रहे हैं।

**बहुविध संवेदन द्वारा ऊर्जा दक्षिता:** दृश्य बोध एक ऊर्जा गहन कार्य है। मनुष्य भी अपनी अनुभूति के लिए अपनी वतरणीय समझ और वैकल्पिक इंद्रियों पर निर्भर रहता है, जैसे की आराम करते हुए आँखों से ज्यादा ध्वनि पर निर्भर रहना। इस कार्य में हम बहुविध संवेदन में इसी तरह के तथ्य का अनुभव करते हैं। हम एक ऊर्जा संरक्षणीय अल्ट्रासोनिक संवेदक के माध्यम से दृश्य बोध को संचालित करते हैं जिससे ऊर्जा की बचत की जाती है। हम विस्तृत रूप से इस उपकरण की संचालन विधि का परीक्षण करते हैं।



# List of Figures

1.1	The different stages of the development of a smart edge device are demonstrated. In this thesis, we focus on questions centered around the choice of which models are to be deployed and how one can enable self-monitoring and adaptation for neural networks. . . . .	2
1.2	Primary perception tasks in the computer vision community focus on object level attributes and semantics. Major focus of the community has been to get the perception tasks working for independent frames. Pictures courtesy of Detectron2 [33], a popular framework for scene perception tasks. . . . .	4
2.1	Deep learning algorithms for object detection generate hundreds of bounding box proposals around the objects. These proposals are then filtered with confidence and non-maximal suppression thresholds to obtain the final predictions. Image Courtesy [75]. . . . .	8
2.2	The blue boxes represent the prediction made by the algorithm while the red boxes represent the ground truth boxes. The intersection area is the area shaded on the left dog. The union area is the area within both the blue and red boxes. The ratio of the bounding box intersection and union represents the amount of overlap in the predicted and the ground truth box. . . . .	9

- 2.3 Consider a scenario, when two pedestrians A, and B, are walking in front of a mobile robot. Each pedestrian is visible in the camera for 10 frames. Suppose, a detector  $\mathcal{D}_1$  detects each pedestrian A, and B for 5 random frames,  $\mathcal{D}_2$  detects B in all frames and misses A completely, and  $\mathcal{D}_3$  detects A, and B in 5 frames but only in the frames when the objects are nearer (and thus visually larger). Let us assume all the detectors have the same 100% precision. It is easy to see that the recall of all 3 detectors is identical since they miss and detect an equal number of objects. However,  $\mathcal{D}_1$  is most likely to detect all view variations of objects A and B in a video,  $\mathcal{D}_3$  would only work when object size is large, whereas  $\mathcal{D}_2$  is unfair and may even cause the robot to collide with pedestrian A. The example highlights the case where three detectors with very different characteristics (or even having a bias) may have a similar performance in terms of mAP. . . . . 10
- 2.4 In this figure, we show the start (dark rectangle) and end (light rectangle) frame for three objects in video. We also show the centre of bounding box in intermediate frames (red dots). The sets contain bounding boxes of an object in a similar spatial neighborhood (as decided by the unifying criteria), as its initial location in the consecutive video frames. We blend the frames in each set for visualization. . . . . 15
- 2.5 For every object  $O_i$ , the bounding boxes in the frames which contain the object  $i$  are combined into sets. The object bounding box in every new frame (orange frame) is compared against the minimal (first frame) of the last set (green frame from  $b_{min}(S(i, j))$ ). If the location constraint  $\mathcal{U}_l$  is above a threshold, the new bounding box is added to the previous set  $S(i, j)$ . Otherwise, a new set  $S(i, j + 1)$  is instantiated. . . . . 15

- 2.6 The figure shows examples of the biases introduced and behavior of mAP and VmAP on increasing these biases. First row shows the examples of objects less likely to be detected (small/dark/low contrast/fast moving/color) while the second row shows examples of their counterparts which are more likely to be detected. The X axis shows the increasing degree of biases and Y axis shows the behaviour of mAP and VmAP. mAP is fairly insensitive to biases in all the cases, whereas VmAP reduces on increasing the bias. The amount of sensitivity of VmAP varies in each case. Other metrics like mAP normalized with length of set (LNmAP) and mAP of just keyframes (KFmAP) are also found insensitive. Average Delay[58] also increases with increase in bias of the detectors. . . . . 20
- 2.7 The timeline plots for a video (ref. Sec. 2.4.1) for two different detectors are shown on the left side. In the first row, **FGFA** (mAP 78.0) and **MEGA** (mAP 75.1) have similar mAP accuracy. However, FGFA fails to detect the small watercraft in the video. This is captured by the VmAP metric (MEGA - 57.8 Vs FGFA - 44.5). Similarly, in the second row, **CATDET** is able to detect almost all sets of the object without having a single false positive, while **RFCN** has multiple false positives during the video. The similar VmAP (CATDET - 81.8 Vs RFCN - 77.6) despite a vastly different mAP (CATDET - 23.8 Vs RFCN - 61.2) can be observed from the number of sets that are detected. . . 24
- 2.8 The detectors are presented in the order of ground truth rankings (Frame Recall using a Detector + Ideal Tracker). While the Frame Recall is low for some detectors (see FGFA, DETR), the post-tracking performance is quite high. On the other hand, VR follows the post-tracking performance closely. . 27
- 2.9 The sets are smaller for the appearance ( $\mathcal{U}_a$ ) and even smaller for the ( $\mathcal{U}_{at}$ ) criteria. As expected, the set size for the time criteria is the same ( $\mathcal{U}_t$ ). We believe that all criteria could be useful in different scenarios as discussed in Sec. 2.5.1. . . . . 29

2.10	The figure demonstrates how the different VmAP definitions respond to increase in different types of bias. We find that as the % value on number of frames in a set is increased, the behavior of the metric resembles that of mAP. In higher percentage values, a true positive set becomes very difficult to achieve resulting in extremely low scores. We, therefore, recommend the usage of VmAP as defined in the main paper. . . . .	30
2.11	The PR curve shows the effect of temporal NMS across frames. Higher recalls now have better precision due to reduced false positives. . . . .	31
2.12	The figure shows the first and last frame of two of the sets formed by using the location criteria (left) and appearance criteria (right). In the upper row, the appearance criteria declares a new set when the object has come too close. This might be dangerous for perceive and control systems like self driving cars. In the second row, the appearance criteria declares a new set even when the train hasn't moved much. In both the cases, the location criteria gives a better estimation. . . . .	32
3.1	Deviating from a per-frame mAP calculation, we jointly account for latency and accuracy of the object detectors and suggest the best model for the given hardware. . . . .	34
3.2	Entropy is a measure for identifying the activity level in a video. A higher entropy poses a greater challenge for detectors since it has to run faster to be able to detect all objects. . . . .	38
3.3	The figure shows how detectors which are better as per their mAP value are able to detect more objects at all distances. The numbers are calculated assuming an FPR of 15 for all detectors. . . . .	38
3.4	We define row numbers as the pixel distances which are indicative of the distance from the camera. We pick the indicated distances to highlight the area of interest where we see maximum pedestrian activity. . . . .	40

- 3.5 Histogram of the number of frames for which people stay in the video accumulated for all datasets. The high density around low times suggests that the majority of pedestrian are in the field of view for a relatively short time. . . . 44
- 3.6 The accuracy on Bahnhof dataset falls on using a lower frame rate (Blue), while it remains relatively constant for MOT16-02 video (Red) since the entropy is higher in Bahnhof video. Therefore, the application being targeted also determines the applicability of the detector. . . . . 45
- 3.7 Variation in the number of people detected at different FPRs of the detectors in a video sampled at 15 FPS. . . . . 45
- 3.8 The results vary a lot on different platforms. It is clear that the faster detectors perform much better than the more accurate but slower detectors. As we go towards low-resource devices, smaller networks perform even better. The depletion of performance of high-end detectors is evident in all systems without a high end GPU (`faster_rcnn`). The cross-overs in Figures 3.8b,3.8c indicate that depending on criticality of detecting objects from a distance, one detector may be better than the other. Faster detectors do better at smaller distances, however, more accurate detectors (still running at a reasonable speed with respect to entropy of video) work better at a larger distance. . . . 46
- 4.1 The 1st row shows an image from KITTI [94] and its depth prediction using the MonoDepth [34] network trained on the same dataset. The depth is highly accurate. In the 2nd row, we show an image from NuScenes [16] dataset, semantically similar to the 1st row in terms of least cosine similarity criteria in the VGG [84] embedding space. However, [34] doesn't seem to generalize well to this semantically similar, but unseen image. Using our technique, with LiDAR points available only at the test time (as seen overlaying the 2nd row), we adapt the depth prediction neural network at the test time to predict a refined depth image (3rd row). . . . . 52

- 4.2 The video sequence/real-time stream is divided into groups  $G_1 - G_k$ . For each group  $G_i$ , the same pre-trained network is used. For adaptation of the network  $F(\theta(\mathcal{D}), X)$ , the batch  $b_i$  is used as the training data. For every image in the batch  $b_i$ , the depth map is predicted using  $\theta(\mathcal{D})$  (Step 1, shown as 1 in the square box). In Step 2, an L1 loss is calculated between the depth prediction and the available LiDAR points which adapts the network to the sample. Finally, in step 3, the adapted network  $\theta(b_i)$  is used to predict the depth for the entire group  $G_i$ . One can observe the improvements in the predicted depth, especially, in the areas highlighted by a yellow circle. . . . . 54
- 4.3 Top and bottom figures show the % improvement in RMSE with respect to MonoDepth for varying levels of fog and rain respectively. We show improvement over MonoDepth since it is the base network in our technique over which distillation is done. We see that as number of valid depth evidence points increase (for more visibility and less rainfall) the % improvement also increases. 66
- 4.4 Examples from the NuScenes dataset show the input images (with LiDAR data) - row 1, the depth predicted using the pre-trained monodepth network - row 2 and the refined depth after distillation - row 3. Left to right: Starting with a failure case on the left, it sometimes deteriorates far away points where LiDAR is not present. The cross-over on the top of the image becomes visible in the depth image. The closely predicted points due to scattered rain get corrected. The depth of all far away objects gets corrected despite rain. Far away objects are also corrected sometimes despite no LiDAR evidence in the area. . . . . 68
- 4.5 While the raw predictions from KITTI are good by themselves, using test-time distillation further improves aspects of the predicted depth. Closer objects with atypical perspectives get corrected with the help of LiDAR, e.g., the bus in the third column. . . . . 79
- 4.6 Examples from the NuScenes dataset. The depth profile improves for the entire scene. Parts of the scene also deteriorate, e.g., person/round pillar which does not reflect LiDAR laser. . . . . 79

4.7	Examples from the NuScenes Rain dataset images. The distillation also removes the aberrations due to blurred/distorted images which appear due to water in front of the camera. . . . .	80
4.8	The LiDAR based distillation is specially useful at night time when the raw output is very distorted. Test-time distillation recovers over-exposed and under-exposed objects. . . . .	80
4.9	Examples of the depth maps obtained using our distillation framework. The last row shows the effect of scaling using a grid of $16 \times 8$ . As one can see, the predicted depth maps are close to the ground truth despite poor initial estimates. The scaling is necessary for the correct metric depth, which is not evident in the visualization. . . . .	81
4.10	Examples of robustness when the input image has blur . . . . .	82
5.1	An ultrasonic sensor is mounted on top of the camera. Both the sensors are connected through the GPIO pins of the Raspberry Pi. The complete system is powered through a power bank. . . . .	83
5.2	An ultrasonic sensor system is extremely low power, a visible spectrum camera has an intermediate power consumption while an IR camera with active illumination is the most energy consuming. An ideal system should have an effective policy for actual usage of each sensor, to achieve high performance with energy efficiency. . . . .	83
5.3	Execution flow for the software setup. Ultrasonic samples are available in the serial buffer at the rate of 6 samples per second. . . . .	89
5.4	Samples from the day/night sequence. The data contains time stamped ultrasonic samples and image samples with annotated bounding boxes, object number and class labels. The time series information for the ultrasonic sensor data for 1.2 seconds around the timestamp of the image is shown in the plot. We can see the distance of the object reducing as the user walks towards it. .	90
5.5	Exploring the energy-accuracy tradeoff for different design points in day and night environments . . . . .	97

5.6 Histogram of persistence of objects in the two sequences . . . . . 98

# List of Tables

2.1	The table shows mAP, and VmAP values for the same object detector (biased against small objects) when tested on different subsets of a dataset. Notice that depending upon which videos we include or exclude in the dataset, the detector can be made to look arbitrarily good or bad. $f_s(f_l)$ and $O_s(O_l)$ represent number of frames, and number of sets in the small and large category respectively. Refer Sec. 2.4.1 for the detailed discussion. . . . .	23
2.2	The table shows the set recall (VR) and false positives (FP) with the P-R operating point and the VP-VR operating point on videos of hamster class from the Imagenet VID dataset. It is observed that Video Precision/Recall operating point has much lesser false positives while having a similar set recall.	25
2.3	We test the ranks of detectors as determined by the Frame Recall (FR) and Video Recall (VR) at 0.9 frame precision. The ground truth order is determined by using the Frame Recall (FR*) when an ideal tracker is used along with the detector at the given precision. Video Recall (VR) achieves a spearman correlation of 0.97 as compared to 0.93 for Frame Recall (FR) indicating the suitability of VR as a measurement for post-tracking effectiveness. . . . .	26
2.4	The table shows that although the range of values are quite different, there isn't any difference in the ranking of algorithms by VmAP with different set formation strategies. . . . .	28
3.1	Details of sequences used from Multi-object Tracking Challenge [60]. . . . .	39

3.2	Details of Processing Units and the platforms used for our experiments. We have used four configurations: CPU + GPU, High-end CPU, Low-end CPU and an ARM processor for the resource-constrained settings. . . . .	42
3.3	The Tensorflow models from the model zoo for various object detectors used in our experiments. . . . .	43
3.4	Frame Processing Rates (FPRs) of different detectors on a given system. Refer to Table 3.2 for details of the platforms. . . . .	49
4.1	GE - Germany, US - United States of America, SGP - Singapore, JA - Japan, AR - Artificial Chamber and IN - Indoor. MS LiDAR - Multi-Sweep LiDAR. /N - subsampled by N in both X, Y directions. We use datasets based in different contexts with different types of expected noise (weather variations, simulated environments and indoor dataset). . . . .	63
4.2	Generalization ability of different modalities trained on KITTI and tested on various datasets. The first row merely indicates the baseline for all the methods. We perform distillation and scaling over the MonoDepth backbone, thus improving its performance. The numbers in rows 1–7, indicate RMSE wrt the ground truth (lower is better). WM-RMSE is the Weighted Mean RMSE weighted by the number of images in the dataset. Runtimes in the last row are calculated on the NuScenes dataset for an approximate estimate of runtime. * is assigned to datasets which do not have sequences of images. Thus, for this case, Fast mode uses a sorted sequence of images based on file names. Note that RGB methods are reported with median scaling as in [34, 36]. The best performing method is <b>bold</b> and the next-best is <u>underline</u> .	67
4.3	Object-wise analysis of depth improvement showing the effectiveness (RMSE - lower is better) of our method on static/dynamic objects against DeepLiDAR (DLidar: next-best) and MD (base network). The depth prediction performance for both static and dynamic objects improve greatly. . . . .	70

4.4	Streaming mode runs cross-modal distillation as a background process, thus providing zero overhead during inference time. The accuracy is close to the Fast Mode described in the main paper. . . . .	70
4.5	Ablation study for our scaling strategy shows errors (RMSE - lower is better) in scaled depth without the distillation step. In general, Huber Loss performs well across datasets due to lesser focus on outliers. In case of noisy data like the DENSE dataset, median scaling and RANSAC perform well since the number of noisy points are high. Smoothing (subscript $s$ ) as a post-processing step also improves performance. . . . .	71
4.6	Do more iterations help? The accuracy of the data increases with the number of iterations and then saturates. Therefore, we use $e = 10$ as the number of iterations. . . . .	73
4.7	Does a larger batch help? Increasing the size of the tuning batch beyond 7 deteriorates the overall improvement in the group. This suggests that the improvement comes from tuning the network on a small number of images and the adaptation does not converge for a large number of images. This can happen when the optimal for each of the images is significantly different from others. We use $ b  = 1$ in the fast mode as a tradeoff between speed and accuracy. . . . .	73
4.8	When the group size is reduced, the adaptation is done more often. However, the accuracy improvement does not vary with increasing group size due to similar context within a scene. Therefore, we use a large group size $ G  = 500$ since that provides the most benefit in terms of runtime (provided a long enough scene exists). . . . .	74
4.9	LiDAR is often noisy in the presence of dust/rain/fog particles which cause backscattering. Noise sampled from a gaussian distribution from 0 to the indicated magnitude is added to points from the LiDAR scan. We observe that our method is able to provide improvement with respect to the MonoDepth network in varying noise conditions through cross-modal distillation and scaling strategy. . . . .	75

4.10	We introduce gaussian blur in the RGB image and record the performance of our method. We consistently observe improvements over MonoDepth (base network) across increasing levels of blur radius. . . . .	76
4.11	Various weather conditions simulated in VKITTI show competent results with our method. However, as noted in the main paper, our method is not the best performer across methods. We attribute this to the similarity of VKITTI data to the original KITTI data. Therefore, NLSPN and PNCNN also generalize well. . . . .	76
4.12	The table shows how the optimal scaling strategy changes with deterioration in quality of LiDAR for the DENSE dataset. RANSAC performs better in low visibility conditions while Median performs better elsewhere. . . . .	78
5.1	The different components of the static power are shown. The numbers above are from measurements during the idle state. . . . .	93
5.2	The dynamic power of different scenarios measured over 1 minute of operation. This is used to identify the energy per frame which is further used for the power analysis. . . . .	94
5.3	Time taken to capture and process a single frame in one of the configurations	95
5.4	Online validation results for sequences in the same environment captured in different modes . . . . .	99

# Contents

Certificate	4
Abstract	i
<b>1 Introduction</b>	<b>1</b>
1.1 Life cycle of an edge device . . . . .	2
1.1.1 Data setup . . . . .	2
1.1.2 Intelligence setup . . . . .	3
1.1.3 Hardware/Device setup . . . . .	4
1.2 Thesis focus . . . . .	5
<b>2 Bias Sensitive Metrics</b>	<b>7</b>
2.1 Background . . . . .	7
2.2 Problem introduction . . . . .	9
2.3 Proposed metric . . . . .	13
2.3.1 Set formation . . . . .	14
2.3.2 Scoring . . . . .	16
2.4 Experiments . . . . .	18
2.4.1 Fairness against biases . . . . .	19
2.4.2 Operating point . . . . .	23
2.4.3 Post-Tracking assessment . . . . .	25
2.5 Ablation studies . . . . .	26
2.5.1 Unifying criteria ablation . . . . .	27

2.5.2	VmAP definition ablation . . . . .	29
2.6	Discussion . . . . .	29
2.7	Summary . . . . .	31
<b>3</b>	<b>Latency sensitive metrics</b>	<b>33</b>
3.1	Related work . . . . .	35
3.2	Proposed approach . . . . .	36
3.2.1	Dataset . . . . .	37
3.2.2	Entropy: estimating application requirements . . . . .	37
3.2.3	Infinite resource setting: finding applicable algorithms . . . . .	39
3.2.4	Resource-constrained setting: finding apt hardware/algorithm . . . . .	40
3.3	Experimental setup . . . . .	41
3.4	Results . . . . .	42
3.4.1	Entropy comparison . . . . .	43
3.4.2	Infinite Resource Setting (IRS) . . . . .	43
3.4.3	Resource-Constrained Setting (RCS) . . . . .	45
3.5	Summary . . . . .	47
<b>4</b>	<b>Generalization Through Multimodal Sensing</b>	<b>51</b>
4.1	Background . . . . .	51
4.2	Problem introduction . . . . .	52
4.3	Related work . . . . .	55
4.4	Proposed method . . . . .	57
4.4.1	Sample specific knowledge distillation . . . . .	58
4.4.2	Batch mode distillation (fast mode) . . . . .	59
4.4.3	Scaling methodology . . . . .	62
4.5	Experimental setup . . . . .	62
4.5.1	Choice of datasets . . . . .	64
4.6	Experiments and results . . . . .	66
4.7	Ablation studies . . . . .	70
4.7.1	Scaling . . . . .	70

4.7.2	Hyperparameters for fast mode . . . . .	72
4.7.3	Robustness to noise . . . . .	75
4.7.4	Noise in LiDAR samples . . . . .	75
4.7.5	Noisy RGB images . . . . .	76
4.7.6	Scaling strategy in adverse weather . . . . .	77
4.8	Qualitative results . . . . .	77
4.9	Summary . . . . .	77
<b>5</b>	<b>Energy Efficiency Through Multimodal Sensing</b>	<b>83</b>
5.1	Background . . . . .	84
5.2	Problem introduction . . . . .	85
5.3	Related work . . . . .	86
5.4	System description . . . . .	88
5.4.1	Prototype . . . . .	88
5.4.2	Software setup . . . . .	89
5.4.3	Experiments . . . . .	90
5.4.4	Methodology . . . . .	93
5.5	Results . . . . .	95
5.5.1	Occupancy and persistence . . . . .	98
5.5.2	Online validation . . . . .	99
5.6	Summary . . . . .	100
<b>6</b>	<b>Conclusion</b>	<b>101</b>