https://doi.org/10.2352/EI.2022.34.16.AVM-147

FisheyePixPro: Self-supervised Pretraining using Fisheye Images for Semantic Segmentation

***Ramchandra Cheke*[1], *Ganesh Sistu*[2], *Ciarán Eising*[1], *Pepijn van de ven*[1], *Varun Ravi Kumar*[3] *and Senthil Yogamani*[2]**

[1] ***University of Limerick, Ireland***

[2] ***Valeo Vision Systems, Ireland***

[3] ***Valeo DAR Kronach, Germany***

Abstract

Self-supervised learning has been an active area of research in the past few years. Contrastive learning is a type of self-supervised learning method that has achieved a significant performance improvement on image classification task. However, there has been no work done in its application to fisheye images for autonomous driving. In this paper, we propose FisheyePixPro, which is an adaption of pixel level contrastive learning method PixPro [1] for fisheye images. This is the first attempt to pretrain a contrastive learning based model, directly on fisheye images in a self-supervised approach. We evaluate the performance of learned representations on the WoodScape dataset using segmentation task. Our FisheyePixPro model achieves a 65.78 mIoU score, a significant improvement over the PixPro model. This indicates that pre-training a model on fisheye images have a better performance on a downstream task.

INTRODUCTION

Recent advancements in deep learning have acted as a catalyst for achieving human-level performance in various computer vision tasks. Availability of large datasets, development of novel architectures and access to faster GPUs are the key factors in the success of deep learning. One of the main challenges in training a deep neural network in a supervised way is the requirement for a large amount of labelled data, which is costly to generate. Self-supervised learning methods focus on learning a generic visual representation from a large amount of unlabelled images, alleviating the requirement for an annotated dataset. Self-supervised learning can be divided into two major categories: 1) Pretext task and 2) contrastive learning.

In pretext task methods, the labels are generated by defining a pseudo task, with the intuition that the network should learn generic features while solving a pretext task. Examples of such pretext tasks are context prediction [2], image colourisation [3], jigsaw puzzle [4], and rotation prediction [5]. The transfer learning performance of these tasks was limited as the network was unable to learn robust feature representations while solving pretext tasks [6].

Contrastive learning means learning by comparing the input samples. The objective of contrastive learning is to maximise the agreement between ”similar” inputs or ”positive pairs” and also maximise the distance between ”dissimilar” inputs or ”negative pairs” in the embedding space. Two views from a single image can be considered as positive pairs, while two views from different images can be considered as negative pairs. Contrastive learning methods are based on a principle of instance discrimination [7], where each image is considered as a single class, and the aim is to distinguish each class from other classes. In order to classify two different views from the same image as a single class, the need for data augmentation arises. Hence data augmentation proves one of the critical aspect of contrastive learning. Numerous methods [8, 9, 10, 11, 12] have shown promising results on downstream tasks of image classification on the ImageNet-1K dataset using a ResNet-50 backbone which was pre-trained using a contrastive learning framework. However, to the best of our knowledge, no work has been done in leveraging fisheye images to pretrain a model using contrastive learning methods.

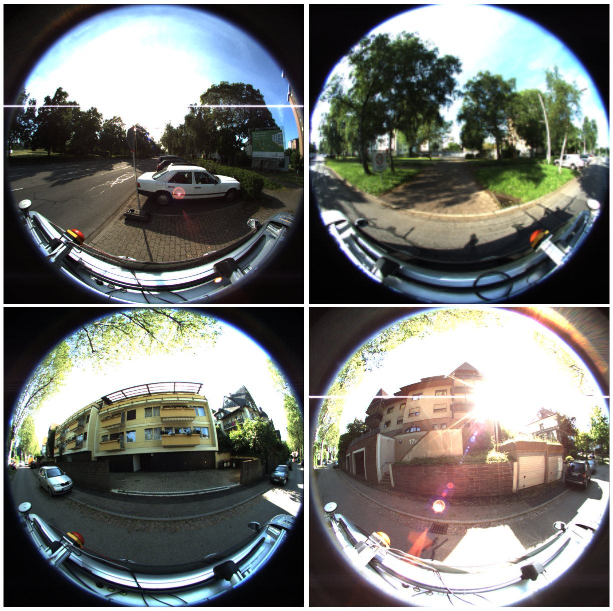
Figure 1: Sample images from KITTI-360 dataset.

Traditional deep learning models offer little performance benefits when applied directly to fisheye photos (e.g. fig 1) due to the large radial distortion in fisheye images. Still fisheye cameras are one of the major components of computer vision systems in autonomous driving cars because only four fisheye cameras are necessary to provide a full 360° coverage around the vehicle. Therefore, it has become popular in near field sensing at low speed [13, 14]. Several experiments have been conducted to enhance the performance of CNN on fisheye dataset by investigating the impact of adversarial attacks[15] on a multi-task visual perception network [16]. The domain of autonomous driving in-

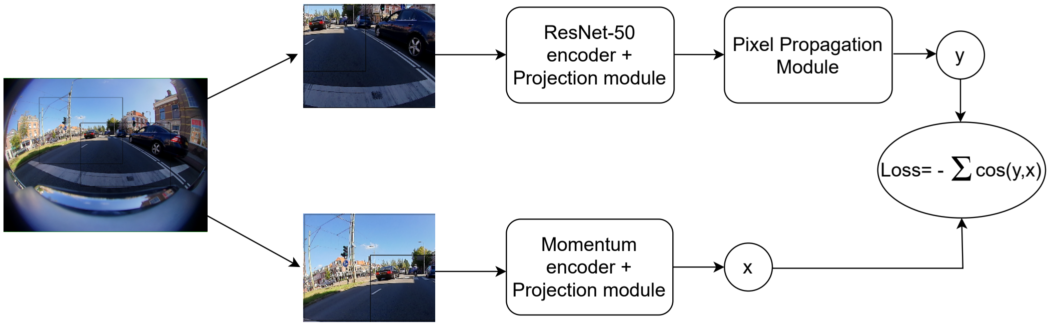
ResNet-50 encoder + Projection module
Pixel Propagation Module
y
Loss= - ∑ cos(y,x)
Momentum encoder + Projection module
x

Figure 2: Self-supervised FisheyePixPro training framework for pretraining the PixPro [1] model using fisheye images.

volves object detection [17, 18, 19], soiling detection [20, 21, 22], semantic segmentation [23], weather classification [24, 25], dynamic object detection [26], depth prediction [27, 28, 29, 30, 31], fusion [32], key-point detection and description [33] and multi-task learning [34, 35, 36]. It also poses many challenges due to the highly dynamic and interactive nature of surrounding objects in the automotive scenarios [37].

Pretext task-based models have lower performance compared to contrastive learning frameworks [10]. As a result, we chose to adopt a contrastive learning framework. However, contrastive learning-based methods such as a Simple framework for Contrastive Learning of visual Representations (SimCLR) requires a very large batch size to maintain the ratio of negative pairs. The Momentum Contrast (MoCo) method [12] addresses this issue by using a momentum encoder and maintaining a queue from previously generated samples that can be utilised as negative pairs. These approaches use the notion of instance discrimination, with the network being pre-trained on the ImageNet dataset. ImageNet dataset typically consist of a single item in a particular image. Thus, two different views from a single image may have some features from the main object in the image. Therefore, instance discrimination methods can be applied to ImageNet dataset. Whereas, fisheye pictures collected for self-driving car purposes are fundamentally different from the ImageNet dataset. These images consist of multiple objects like a bus, car, bike, road, humans, traffic signs etc. in a single image. Due to this, instance discrimination methods like MoCo and SimCLR are not a suitable choice for pretraining models with fisheye images for autonomous driving.

PixPro [1] method is based on pixel level contrastive learning. In this method, each pixel in a given image is considered as a single class and the objective is to differentiate each pixel from other pixels within the same image. The main advantage of using PixPro is that it does not require a large batch size similar to SimCLR. The negatives are selected based on the features obtained from different pixels from the same image. Additionally, the pixel propagation module provides a smoothing effect that removes noise and allows propagation of the features with similar pixels.

In this work, we propose a novel training method FisheyePixPro, which is the first attempt to train a contrastive learning based model, directly on fisheye images in a self-supervised approach. We use PixPro [1] as a base for pretraining on fisheye images. Fisheye images have geometrical distortion and it leads to drop in the performance when ImageNet pretrained model is directly applied on fisheye images. We demonstrated that FisheyePixPro pretrained representations obtained higher score on segmentation task than standard PixPro model.

METHODS

Datasets

We used a subset of the KITTI-360 [38] dataset and a Valeo internal fisheye image dataset for pretraining. The KITTI-360 is a large scale 3D video dataset with 300k images and 3D laser point clouds. The dataset was collected with the help of a station wagon using two fisheye cameras along each side covering 360° view. Sample fisheye images from KITTI-360 dataset are shown in Figure 1. To remove duplicate images, total of 50k images were sampled for pretraining. In addition, we used an internal fisheye dataset from Valeo. These fisheye images are obtained under same conditions as WoodScape [39] and it consists of around 50k unlabelled fisheye images. Therefore, total of 100k images were used for pretraining.

In addition to this, we evaluated the performance on the WoodScape dataset using segmentation task. WoodScape dataset consists of total 10k images with annotations for nine classes. These classes are road, lanes, curbs, person, rider, vehicle, bicycle, motorcycle and traffic sign. From total 10k images only 8215 images are publicly available. Therefore, we randomly choose 7200 images for training a deeplabv3+ model on segmentation task using ResNet-50 encoder and evaluated performance on remaining images as test set.

FisheyePixPro Pretraining for segmentation

PixPro is based on two properties of an image: spatial sensitivity and spatial smoothing. Spatial sensitivity is defined as an ability to differentiate between adjacent pixels. This property is useful in delineation of boundary areas. On the other hand, spatial smoothing operation involves removal of noise or high frequency signals from an image. These two properties are core components of pretext task. The features from the corresponding pixels of the two views taken from the randomly cropped image are encouraged to be consistent. This pixel level pretext task focuses on learning representation from two different views of the same image by minimising the distance between two pixel level representations using cosine similarity loss.

The model architecture is shown in figure 2. It is a siamese architecture with two input branches to process different views from the same input image under different data augmentations. One branch consists of ResNet-50 encoder with projection head and pixel propagation module. Whereas, the other branch consists of only momentum encoder with projection head. A random crop is extracted from given image, then it is resized to 224x224 pixels. Different data augmentations such as random horizontal flip, colour jitter, grayscale, gaussian blur and solarization operation are applied to the input image. An encoder and momentum encoder is used to calculate features from these two extracted patches. The spatial resolution of features map reduces to 7x7. Then each pixel in feature map is mapped to original image space and the distance between every pair of pixels in the two feature maps is calculated according to equation 1

$$A(i,j) = \begin{cases} 1 & \text{if dist(i,j)} \leq \tau \\ 0 & \text{if dist(i,j)} > \tau \end{cases} \tag{1}$$

If the distance between two pixels from different views is less than the threshold τ then those two are considered as positive pairs and if the distance between two pixels from different views is greater than τ then it is considered as negative pairs. Typical value of τ is 0.7.

Pixel propagation module is applied to only one branch of the network after regular encoder. The purpose of this module is to provide smooth representation using self attention mechanism, according to Equation 2.

$$y_i = \sum_{j \in \Omega} (max(cos(x_i, x_j), 0))^{\gamma}.g(x_j) \tag{2}$$

where, cosine function is used to calculate the distance between pair of pixels and γ is control parameter for similarity function. The default value of γ is set to 2. Function denoted as g(·), is a transformation function which composed of batch normalisation and a ReLU layer. Finally, the loss is calculated by equation 3, where y is feature of a pixel from pixel propagation module and x is feature of a pixel from momentum encoder module.

$$loss = -cos(y_i, x'_j) - cos(y_j, x'_i) \tag{3}$$

To pretrain FisheyePixPro, the network was first initialised with PixPro weights as the weights are already available; then further pretrain on fisheye images.

DeepLabv3+

DeepLabv3+[40] incorporates encoder-decoder architecture and is an upgraded version of DeepLabv3 [41]. Deeplabv3+ offers a number of benefits for semantic segmentation tasks, including dense prediction with Atrous convolution [42], memory optimisation with depth-wise separable convolution [43] and multi-scale processing using Atrous Spatial Pyramid Pooling(ASPP) module. The following are some key points in the DeepLabv3+ architecture:

Atrous convolution: Atrous convolution also called as dilated convolution, it allow us to increase the spatial resolution of feature maps. The dilation rate in Atrous convolution determines the distance between consecutive values in the kernel. As a result, multi-scale information is acquired by regulating the dilation rate, boosting the network's generalisation capacity.

	ImageNet		PixPro		FisheyePixPro	
Pretraining	Supervised		Self-supervised		Self-supervised	
Class	IoU	Acc	IoU	Acc	IoU	Acc
void	97.23	**98.59**	97.13	98.29	**97.27**	98.51
road	93.76	96.12	93.64	96.51	**93.91**	**96.32**
lanes	**71.46**	**83.45**	69.92	82.47	70.00	83.26
curbs	**53.30**	81.25	50.05	**84.81**	52.54	83.05
person	55.29	**79.88**	52.25	77.02	**55.63**	78.64
rider	**54.92**	76.46	52.03	73.84	53.90	**76.75**
vehicle	88.28	93.11	87.91	92.86	**88.56**	**93.64**
bicycle	48.35	**72.85**	46.45	72.45	**48.47**	71.81
motorcycle	**60.14**	**80.25**	56.74	70.46	59.04	77.14
traffic sign	**39.26**	**65.22**	35.76	58.06	38.45	61.26

Table 1: Class-wise IoU score and accuracy on validation dataset. FisheyePixPro pretraining performs better than PixPro pretraining in a self-supervised setting.

Depth-wise separable convolution: Depth-wise separable convolution operation splits a regular convolution into two components as depth-wise convolution followed by a point-wise convolution. The depth-wise convolution conducts a spatial convolution for each input channel individually, whereas the point-wise convolution is used to combine the depth-wise convolution's output. This innovative solution not only significantly reduces calculation complexity but also enhances performance.

Atrous Spatial Pyramid Pooling: The size of the same object varies according to its position in front of the camera. To deal with different sizes of the same object, several studies have been proposed to extract features at multiple scales [44] [42]. DeepLabv3+ uses Atrous Spatial Pyramid Pooling (ASPP) with different atrous rates of 6,12 and 18 to process the convolution neural network output.

Network backbone: In this work, we follow [1, 12, 10] and use ResNet-50 [45] backbone for pretraining.

EXPERIMENTS

We investigated whether pretraining a FisheyePixPro model leads to better representation learning than a regular PixPro model. To evaluate this hypothesis, we adopt state-of-the-art Deeplabv3+ model with ResNet-50 backbone on WoodScape dataset. These experiments were carried out using a PyTorch [46] based implementation. For PixPro model, the ResNet-50 encoder is initialised with the weights provided by [1] and ImageNet model uses weights from ImageNet pretrained weights. All training images were resized to 640x480 pixels. These models were trained for 100k iterations using Nvidia V100, 16GB GPU with batch size of 6. We used SGD optimiser with initial learning rate=0.01, momentum=0.9, weight decay=0.0005. We adopted poly learning rate scheduling scheme with power=0.9, minimum learning rate=0.0001. To overcome the problem of class imbalance, we also used weighted categorical cross-entropy.

Table 1 shows the detail view of class-wise IoU score for PixPro, FisheyePixPro and ImageNet pretrained model on the WoodScape dataset for segmentation task. It can be seen that our FisheyePixPro model outperforms PixPro model on fisheye

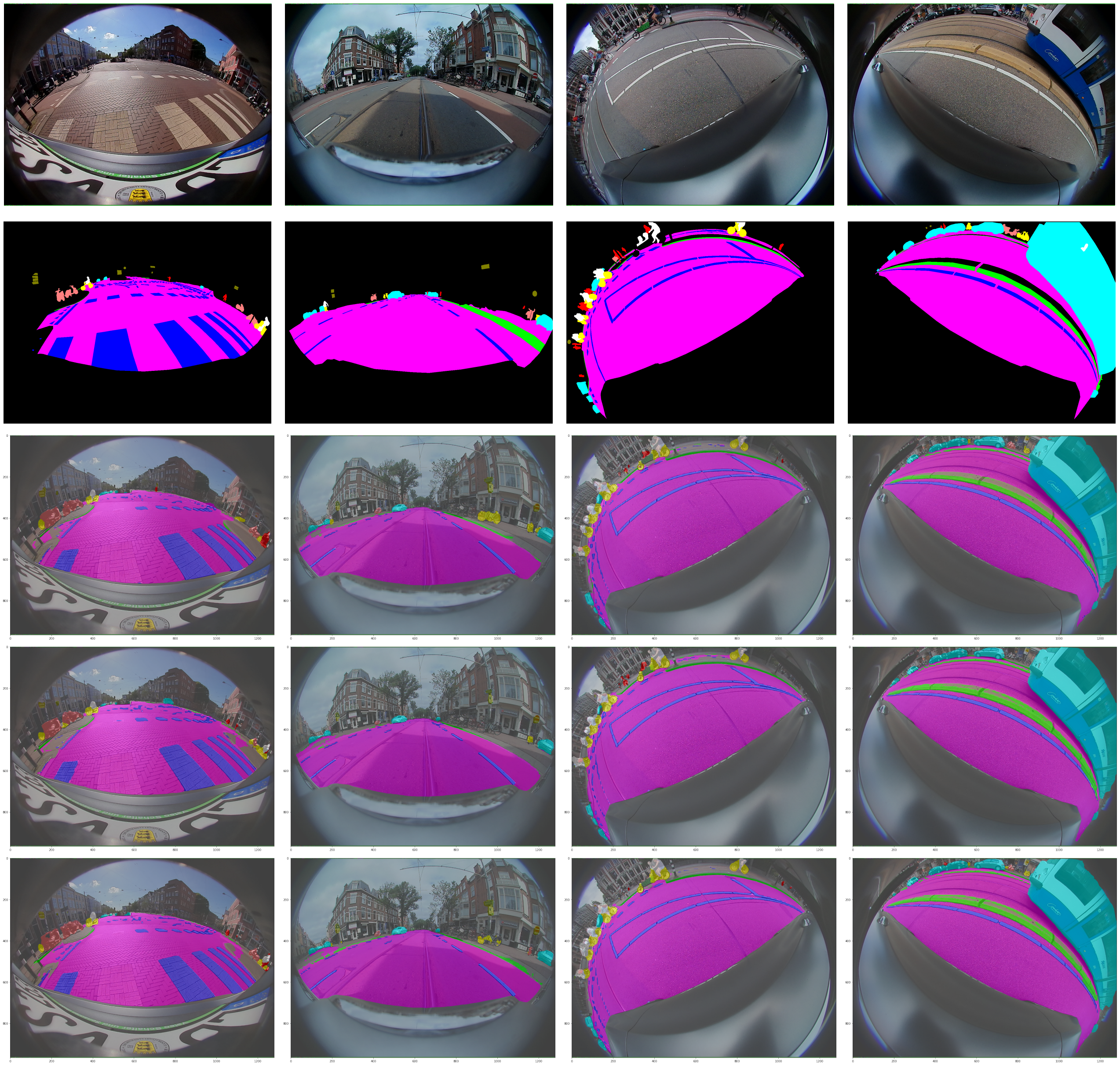

Figure 3: Qualitative results on downstream task of segmentation on WoodScape dataset. Rows represent fisheye images, ground truth, results from ImageNet supervised pretrained model (baseline), PixPro and our FisheyePixPro respectively. The PixPro and FisheyePixPro models were trained using unlabelled dataset. While ImageNet pretrained model was trained on 1.2M images with corresponding classification labels.

dataset, whereas it achieves comparable performance with supervised ImageNet pretrained model. Kindly note that supervised ImageNet pretrained model was pretrained on ImageNet dataset which requires 1.2M images with its classification label. However, FisheyePixPro model was pretrained without using labels.

We have also provided the visual results on validation set in fig 3. Table 2 compares the mean intersection over union (mIou) score and average accuracy on WoodScape segmentation task for our FisheyePixPro, PixPro and ImageNet pretrained model. Our FisheyePixPro method achieves significantly higher mIoU score as compared to normal PixPro. On the other hand, supervised ImageNet pretrained model achieves the highest score. This re-

	Pretraining	mIoU	aAcc
ImageNet	Supervised	**66.2**	97.20
PixPro	Self-supervised	64.19	97.06
FisheyePixPro	Self-supervised	65.78	**97.21**

Table 2: Comparing proposed FisheyePixPro pretraining with PixPro and supervised pretraining on ImageNet dataset. The results are evaluated on WoodScape dataset for semantic segmentation.

sults demonstrates that our FisheyePixPro pretrained model helps to learn better representations as compared to PixPro model.

CONCLUSION

In this work, we build a successful pretraining framework using pixel level pretext task of contrastive learning for fisheye images. We demonstrated that our FisheyePixPro method achieved better feature representation and transfer performance using fisheye images for dense prediction. According to our knowledge, this is the first attempt to pretrain a contrastive learning based model directly on fisheye images. Our findings show that there is potential to define a pixel level pretext task for fisheye images that can alleviate the effect of non-linear distortion and learn generic visual representations.

ACKNOWLEDGEMENTS

”This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6049”. ”The authors wish to acknowledge the Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support”. The authors would like to thank Valeo Vision Systems for providing the opportunity to work on the research project.

References

[1] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, “Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning,” *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16679–16688, 2021. 1, 2, 3

[2] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1422–1430, 2015. 1

[3] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 649–666, Springer International Publishing, 2016. 1

[4] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 69–84, Springer International Publishing, 2016. 1

[5] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” *ArXiv*, vol. abs/1803.07728, 2018. 1

[6] I. Misra and L. v. d. Maaten, “Self-supervised learning of pretext-invariant representations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020. 1

[7] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, “Discriminative unsupervised feature learning with exemplar convolutional neural networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1734–1747, 2015. 1

[8] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018. 1

[9] P. Bachman, R. D. Hjelm, and W. Buchwalter, “Learning representations by maximizing mutual information across views,” in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019. 1

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, PMLR, 13–18 Jul 2020. 1, 2, 3

[11] I. Misra and L. van der Maaten, “Self-supervised learning of pretext-invariant representations,” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6706–6716, 2020. 1

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. 1, 2, 3

[13] C. Eising, J. Horgan, and S. Yogamani, “Near-field perception for low-speed vehicle automation using surround-view fisheye cameras,” *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–18, 2021. 1

[14] L. Gallagher, V. R. Kumar, S. Yogamani, and J. B. McDonald, “A hybrid sparse-dense monocular slam system for autonomous driving,” in *Proc. of ECMR*, pp. 1–8, IEEE, 2021. 1

[15] I. Sobh, A. Hamed, V. Ravi Kumar, and S. Yogamani, “Adversarial attacks on multi-task visual perception for autonomous driving,” *Journal of Imaging Science and Technology*, vol. 65, no. 6, pp. 60408–1, 2021. 1

[16] V. Ravi Kumar, S. Yogamani, H. Rashed, G. Sitsu, C. Witt, I. Leang, S. Milz, and P. Mäder, “Omnidet: Surround view cameras based multi-task visual perception network for autonomous driving,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2830–2837, 2021. 1

[17] H. Rashed, E. Mohamed, G. Sistu, V. Ravi Kumar, C. Eising, A. El-Sallab, and S. Yogamani, “Generalized Object Detection on Fisheye Cameras for Autonomous Driving: Dataset, Representations and Baseline,” in *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 2272–2280, 2021. 2

[18] A. Dahal, V. R. Kumar, S. Yogamani, and C. Eising, “An online learning system for wireless charging alignment using surround-view fisheye cameras,” *arXiv preprint arXiv:2105.12763*, 2021. 2

[19] R. Hazem, E. Mohamed, V. R. K. Sistu, Ganesh and, C. Eising, A. El-Sallab, and S. Yogamani, “FisheyeYOLO: Object Detection on Fisheye Cameras for Autonomous Driving,” *Machine Learning for Autonomous Driving NeurIPS 2020 Virtual Workshop*, 2020. 2

[20] M. Uricar, G. Sistu, H. Rashed, A. Vobecky, V. Ravi Kumar, P. Krizek, F. Burger, and S. Yogamani, “Let’s get dirty: Gan based data augmentation for camera lens soiling detection in autonomous driving,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 766–775, 2021. 2

[21] A. Das, P. Křížek, G. Sistu, F. Bürger, S. Madasamy, M. Uřičář, V. Ravi Kumar, and S. Yogamani, “TiledSoilingNet: Tile-level Soiling Detection on Automotive Surround-view Cameras Using Coverage Metric,” in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6, IEEE, 2020. 2

[22] M. Uricár, J. Ulicny, G. Sistu, H. Rashed, P. Krizek, D. Hurych, A. Vobecky, and S. Yogamani, “Desoiling dataset: Restoring soiled areas on automotive fisheye cameras,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Work-*

shops, pp. 0–0, 2019. 2

[23] A. Dahal, E. Golab, R. Garlapati, V. Ravi Kumar, and S. Yogamani, "RoadEdgeNet: Road Edge Detection System Using Surround View Camera Images," in *Electronic Imaging*, 2021. 2

[24] M. M. Dhananjaya, V. R. Kumar, and S. Yogamani, "Weather and light level classification for autonomous driving: Dataset, baseline and active learning," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 2816–2821, 2021. 2

[25] L. Yahiaoui, M. Uřičář, A. Das, and S. Yogamani, "Let the sunshine in: Sun glare detection on automotive surround-view cameras," *Electronic Imaging*, vol. 2020, no. 16, pp. 80–1, 2020. 2

[26] M. Yahiaoui, H. Rashed, L. Mariotti, G. Sistu, I. Clancy, L. Yahiaoui, V. R. Kumar, and S. K. Yogamani, "Fisheyemodnet: Moving object detection on surround-view cameras for autonomous driving," *ArXiv*, vol. abs/1908.11789, 2019. 2

[27] V. R. Kumar, S. A. Hiremath, M. Bach, S. Milz, C. Witt, C. Pinard, S. Yogamani, and P. Mäder, "Fisheyedistancenet: Self-supervised scale-aware distance estimation using monocular fisheye camera for autonomous driving," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 574–581, 2020. 2

[28] R. K. Varun, S. Yogamani, S. Milz, and P. Mäder, "FisheyeDistanceNet++: Self-Supervised Fisheye Distance Estimation with Self-Attention, Robust Loss Function and Camera View Generalization," in *Electronic Imaging*, 2021. 2

[29] V. Ravi Kumar, M. Klingner, S. Yogamani, S. Milz, T. Fingscheidt, and P. Mader, "Syndistnet: Self-supervised monocular fisheye camera distance estimation synergized with semantic segmentation for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 61–71, 2021. 2

[30] R. K. Varun, M. Klingner, S. Yogamani, M. Bach, S. Milz, T. Fingscheidt, and P. Mäder, "SVDistNet: Self-supervised near-field distance estimation on surround view fisheye cameras," *IEEE Transactions on Intelligent Transportation Systems*, 2021. 2

[31] R. K. Varun, S. Yogamani, M. Bach, C. Witt, S. Milz, and P. Mäder, "UnRectDepthNet: Self-Supervised Monocular Depth Estimation using a Generic Framework for Handling Common Camera Distortion Models," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 2020. 2

[32] K. El Madawi, H. Rashed, A. El Sallab, O. Nasr, H. Kamel, and S. Yogamani, "Rgb and lidar fusion based 3d semantic segmentation for autonomous driving," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 7–12, IEEE, 2019. 2

[33] A. Konrad, C. Eising, G. Sistu, J. B. McDonald, R. C. Villing, and S. K. Yogamani, "Fisheyesuperpoint: Keypoint detection and description network for fisheye images," *ArXiv*, vol. abs/2103.00191, 2021. 2

[34] G. Sistu, I. Leang, and S. Yogamani, "Real-time joint object detection and semantic segmentation network for automated driving," *NeurIPS 2018 Workshop on Machine Learning on the Phone and other Consumer Devices*, 2019. 2

[35] P. Maddu, W. Doherty, G. Sistu, I. Leang, M. Uřičář, S. Chennupati, H. Rashed, J. Horgan, C. Hughes, and S. Yogamani, "Fisheyemultinet: Real-time multi-task learning architecture for surround-view automated parking system.," in *Proceedings of the Irish Machine Vision and Image Processing Conference (IMVIP)*, 2019. 2

[36] I. Leang, G. Sistu, F. Bürger, A. Bursuc, and S. Yogamani, "Dynamic task weighting methods for multi-task networks in autonomous driving systems," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–8, IEEE, 2020. 2

[37] S. Houben, S. Abrecht, M. Akila, A. Bär, *et al.*, "Inspect, Understand, Overcome: A Survey of Practical Methods for AI Safety," *CoRR*, vol. abs/2104.14235, 2021. 2

[38] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment — a modern synthesis," in *Vision Algorithms: Theory and Practice* (B. Triggs, A. Zisserman, and R. Szeliski, eds.), (Berlin, Heidelberg), pp. 298–372, Springer Berlin Heidelberg, 2000. 2

[39] S. K. Yogamani, C. Hughes, J. Horgan, G. Sistu, P. Varley, *et al.*, "Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9307–9317, 2019. 2

[40] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018. 3

[41] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017. 3

[42] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017. 3

[43] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017. 3

[44] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017. 3

[45] 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*, pp. 770–778, 2016. 3

[46] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark." `https://github.com/open-mmlab/mmsegmentation`, 2020. 3