Attending Multiple Visual Tasks for Own Failure Detection

Jiaxu Xing*[1], Boyang Sun*[2], Hermann Blum[2], Roland Siegwart[3], and Cesar Cadena[4]

***Abstract*— Autonomous robots deal with unexpected scenarios in real environments. Given input images, various visual perception tasks can be performed, e.g., semantic segmentation, depth estimation, and normal estimation. These different tasks provide rich information for the whole robotic perception system. All tasks have their own characteristics while sharing some latent correlations. However, some of the task predictions may suffer from the unreliability of dealing with complex scenes and anomalies. We propose an attention-based failure detection approach by exploiting the correlations among multiple tasks. The proposed framework infers task failures by evaluating the individual prediction, across multiple visual perception tasks for different regions in an image. The formulation of the evaluations is based on an attention network supervised by multi-task uncertainty estimation and their corresponding prediction errors. Our proposed framework generates more accurate estimations of the prediction error for the different task's predictions.**

I. INTRODUCTION

Autonomous agents utilize the information from various learning-based visual perception predictions. Existing works have shown good performance on cases where the deployment environment has similar distribution to the training set [1]. However, many state-of-the-art deep learning approaches still face the lack of ability in dealing with open and unconstrained world [2]–[4], and will produce failures, especially in unseen environments [5]. Thus, a method to detect prediction failures of various robotics visual perception tasks is crucial for safe robotic deployments. With higher introspection capabilities, autonomous robots will be more controllable in safety-critical scenarios.

This work focuses on identifying failure predictions of various robotics perception tasks by exploiting the latent correlations among them. Those correlations have been recently used to improve tasks performance [6,7]. Our basic idea is to exploit the complementary information from multiple tasks to improve the introspection capability of the perception system on every single task based on an attention mechanism. Our failure detection model has a *unified* structure that attends the encoded multi-task feature maps with the *expressive power* to perform failure prediction for different tasks.

Existing research recognizes uncertainty as a common measurement of the multi-task prediction's confidence [5]. The uncertainties are ideally correlated to the corresponding task prediction errors, which measure the reliability of the predictions. General uncertainty estimation methods are based on a single task, e.g., softmax entropy from semantic segmentation [8]. However, the quality of the uncertainty estimation is limited by several factors, such as environmental conditions (e.g., clean/foggy weather), anomalies, and more [9].

* Authors contributed equally to this work. Jiaxu Xing is with the Robotics and Perception Group, University of Zurich. Boyang Sun and Hermann Blum are with Computer Vision and Geometry Group, ETH Zurich, Roland Siegwart is with Autonomous Systems Lab, ETH Zurich, and Cesar Cadena is with Robotics Systems Lab, ETH Zurich.

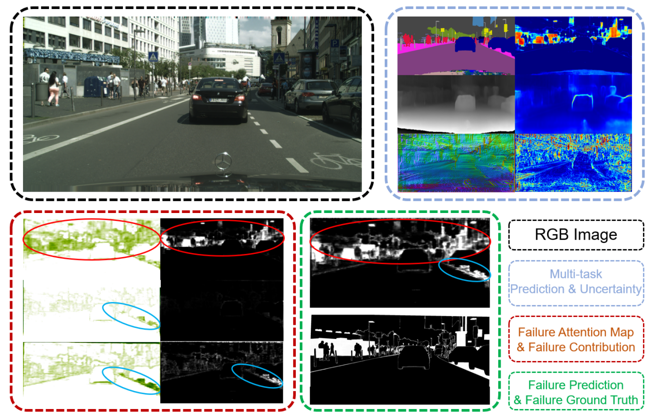

Fig. 1: **Example Result of Our Approach.** Our method captures the high prediction uncertainty regions of a single task using multiple visual tasks. The result maintains the useful uncertainty estimation from the original task (highlighted areas in red circle). Moreover, beneficial from the multi-task setup, our approach captures the relevant information from other tasks (highlighted areas in blue circle) to compensate the missed failure regions.

To investigate the model robustness, we train our model on the Cityscapes [10] dataset and test it on several datasets with different distribution characteristics [11,12]. We evaluate the model for different tasks and compare it against several existing methods. We show that our approach outperforms all other failure prediction approaches. Moreover, our framework is flexible to the number and types of tasks with different task prediction & uncertainty estimation methods.

In summary, the contributions of this work are i) The first work to exploit the multiple visual tasks setup for detecting failures in their prediction in deployment. ii) A novel framework with an attention mechanism over the multiple visual tasks being deployed to extract the complementary information in their uncertainty estimates for failure detection. iii) A thorough evaluation of design components and their influence in open-world scenarios.

II. RELATED WORK

Failure Detection via Uncertainty Estimation

Uncertainty estimation has a close relation with failure detection and introspection. The uncertainty of the prediction

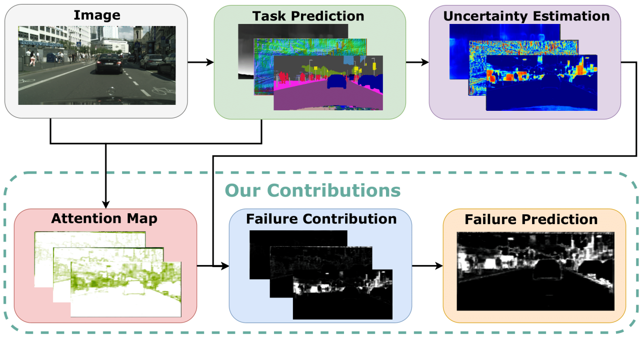

Fig. 2: **Visual Tasks Failure Detection Framework.** Given an image and its multiple task predictions, our approach computes the attention maps to weigh the multiple task uncertainty estimations. This weighted sum of attention and uncertainty maps is our failure prediction for a chosen task.

results reflects the level of its confidence. And the intuition follows that a low-confidence prediction is likely a failure. Therefore, uncertainty estimation could be regarded as a reference to failures prediction. A conventional way to calculate the uncertainty is to directly analyze the distribution of the model prediction such as the *softmax entropy* [13], or *softmax distance* [14] used in classification models. Besides, *image flipping* [2] investigates the model results' difference in dealing with the original and flipped image. Finally, *Bayesian estimation* [15]–[17] estimates the uncertainty by sampling multiple models, e.g. *Monte-Carlo dropout* [5,18] captures the uncertainty by randomly dropping the connection between different layers.

Failure Detection via Learning-based methods

Nowadays, the neural network has become a possible option for the failure detection task. Most of the failure detection methods in the visual perception area focus on detecting semantic segmentation miss-classification. These methods can be roughly divided into two categories. One group directly trains detectors with failure cases [19]–[21]. The other group uses re-synthesis methods [9,20,22,23] that rebuild the image from semantic prediction, and capture the anomalies by comparing the rebuilt image and the original one.

Learning from Multiple Tasks

Prior works have already acknowledged the relation among different perception tasks [24,25]. This latent correlated structure among visual tasks has been exposed in the work of *Taskonomy* [6]. The utilization of cross-task relations also lies in the area of domain adaptation [26,27], transfer learning [7], and multi-task learning [28]–[30]. More specifically, recent attention has focused on using cross-task supervised learning to improve the performance of a single task, such as to improve depth prediction under the supervision of semantic understanding [31,32].

III. Method

A. Multi-task Element Generation

Our approach admits any number of visual perception tasks that provide per-pixel predictions and their per-pixel uncertainties. Different chosen uncertainty estimation methods will certainly influence the final output, and thus we evaluate their influence in Section IV.

B. Attention Network Model

We denote the original image as $\mathcal{I}$. The predictions of all n tasks are denoted as $\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_n$. The uncertainty estimations of them are denoted as $\mathcal{U}_1$, $\mathcal{U}_2$, $\cdots$, $\mathcal{U}_n$, respectively. The architecture of our attention network is shown in Figure 3. The model first encodes the original images $\mathcal{I}$ and its task predictions $\{\mathcal{T}_i\}, i \in \{1, 2 \ldots, n\}$. We chose to encode the image $\mathcal{I}$ with the first several layers of ResNet50 [33] into a 256-channel feature with a 128×128 size. On the other hand, the predictions $\{\mathcal{T}_i\}$ are encoded by part of MobileNetV2 [34]. Each of them is encoded into a 24-channel feature map of size 128×128. All tasks prediction, $\{\mathcal{T}_i\}$, share the same encoding structure. The encoded feature maps are denoted as $C_{\mathcal{I}}$, $\{C_i\}$, correspondingly.

After the encoding process, the encoded features are concatenated along the channel dimension. The resulting feature map, C_{cat} is then forwarded into a neural network with four convolutional layers. In these convolutional layers, the output of the second and third layers will be concatenated together and used as the input to the last layer. In this case, the output layer has one channel for each task. Given a predefined patch size p, two pooling layers are added after the second and the third convolutional layers to resize each output channel into $(256/p) \times (256/p)$, which is the resolution of our attention maps. An extra nearest neighbour rescaling layer is added here to rescale each channel to the size of the uncertainty maps. The rescaled feature maps in each channel are the final attention map generated by our attention network. We denote them as $\mathcal{W}_1$, $\mathcal{W}_2$, $\cdots$, $\mathcal{W}_n$, for each task, respectively. The final failure prediction $\mathcal{E}$. for a single task is generated by calculating the weighted sum of all tasks' uncertainty estimates, with the attention maps as the weights.

C. Training Procedure

As introduced in the last subsection, the attention maps are predicted by our attention network, $\{\mathcal{W}_i\} = f_\theta(\mathcal{I}, \{\mathcal{T}_i\})$. We set a single error approximation loss function to learn the network parameters θ. When training the network to learn a certain task failure, we compute the pixel-wise prediction error $\epsilon_{\{\cdot\}}$ for this task.

IV. Experiments

A. Experimental Setup

1) Model Implementation: Inspired by task networks graph in *taskonomy* [6], we decided to choose tasks from three visual tasks in total in our different experiments: semantic segmentation, depth estimation, and normal estimation. We add later in the experiments a fourth task, instance segmentation. For each task, we implemented publicly

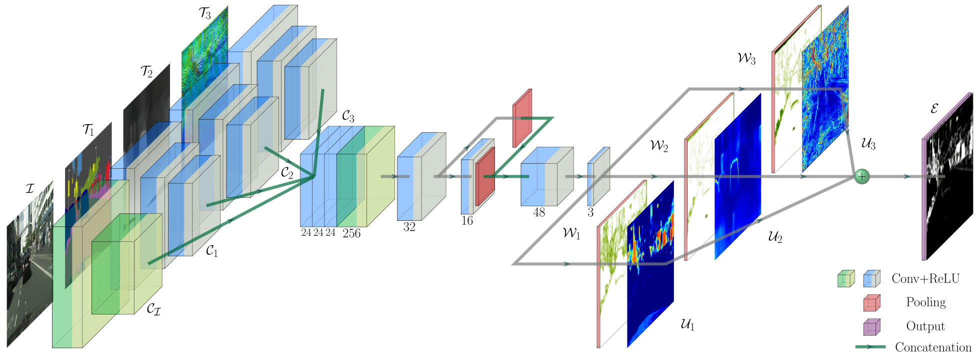

Fig. 3: **Example of the Model Architecture**: This example model uses three different tasks: semantic segmentation, depth, and normal estimation. For the uncertainties $\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3$ and prediction errors $\mathcal{E}$, we resize them to 256×256. Using the predefined attention patch size p, the output attention from the model $\mathcal{W}_1, \mathcal{W}_2, \mathcal{W}_3$ will have the size $(256/p) \times (256/p)$. Then the output would be equally upscaled by a factor of p so that the attention maps' sizes are 256×256. Now the computed attention maps have the same size as the uncertainties, then element-wise multiplication can be performed.

Task	**Prediction Method**	**Uncertainty Estimation Methods**
Semantic Segmentation	SDC Net [4]	Softmax Entropy [8] Softmax Distance [14] Synboost* [9] MC Dropout [35]
Depth Estimation	Monodepth V2 [3]	Bayesian Estimation [17] MC Dropout [35] Self Learning* [18]
Normal Estimation	VNL [36]	Flipping* [2]
Instance Segmentation	EfficientPS [37]	ROI Softmax Uncertainty* [38]

TABLE I: The selected task prediction methods and the uncertainty estimation methods for all different tasks in the experiments. * indicates the method used by default in the experiments unless otherwise mentioned.

Task Entries			**Semantic**				**Depth**
S	D	N	ZNCC↑	AP-Err↑	AP-Suc↑	FPR95↓	ZNCC↑
✓	✓	✓	**0.649**	**0.590**	0.987	0.280	**0.646**
✓	✓		0.609	0.545	**0.990**	**0.278**	0.489
✓			0.494	0.413	0.978	0.570	-
	✓		-	-		-	0.483

TABLE II: Multiple Tasks Experiments: Comparison among multiple different task entries for both semantic segmentation's and depth estimation's failure prediction.

available task prediction methods and uncertainty estimation methods. All evaluated methods are shown in Table I.

2) Dataset: Training was performed on Cityscapes training dataset [10], including 2975 driving scenes images with fine semantic annotations and disparity ground truth. We produced our dataset by applying the methods shown in the Table I. The training set is then composed by set in form of $\{\mathcal{I}, \mathcal{T}_S, \mathcal{T}_D, \mathcal{T}_N, \mathcal{U}_S, \mathcal{U}_D, \mathcal{U}_N, \epsilon_{\{\cdot\}}\}$, where S, D, N denote semantic segmentation, depth estimation, and normal estimation, tasks, respectively, and $\epsilon_{\{\cdot\}}$ correspond to the prediction error of the chosen task to predict its failure.

To test our model's performance, pre-processing of the various test datasets is also required. Here we performed the same pipeline as mentioned in the subsection IV-A.2 for Cityscapes validation set [10], Foggy Cityscapes validation set [39], Wilddash [12] and Dark Zurich dataset [11]. Wildash provides a dataset and benchmark for challenging driving scenarios under real-world conditions, it contains scenarios from very diverse environments, locations, and weather conditions. Dark Zurich is a dataset designed for semantic uncertainty-aware model evaluations. It contains driving scenes images captured at night time, twilight and day time. Here we only use the night time images for our evaluation. The purpose of testing on these extra two datasets is to validate the model robustness when dealing with the *challenging unseen scenarios*.

3) Metrics: We choose the zero-mean normalized cross-correlation (ZNCC) in our experiments as a measurement of how close the predicted failure is to the ground-truth failure. This metric is seamlessly applicable to classification and regression tasks.

In addition, for the classification task of semantic segmentation, previous works on failure detection have used several metrics for evaluation [9,21,23]. Thus we also report: i) AUPR-Error: the area under the Precision-Recall (AUPR) curve ii) AUPR-Success: the AUPR metrics treating correct prediction as the positive class. iii) FPR95: the false positive rate at 95% true positive rate. As for regression tasks, such as depth estimation, we are not aware of any previous work focusing on evaluating the failure prediction model.

B. Comparisons

Are multiple tasks beneficial for single task failure detection? In these experiments, we evaluate two aspects. The first one is the influence of increasing the number of tasks as inputs to our failure detection. And the second one, we evaluate our failure detection on different main tasks: a classification task (semantic segmentation) and a regression task (depth estimation).

We start in our network structure by only having the input of a single task (the task for which the failure is being detected), see last 2 rows of Table II, on the Cityscapes original validation set. This is equivalent to learning failures from single task knowledge and thus can be compared to use the uncertainty input as a proxy to the failure.

Method	Patch	Cityscapes Original		Cityscapes Foggy	
		ZNCC↑	AP-Err↑	ZNCC↑	AP-Err↑
Ours with Synboost	1	**0.649**	**0.590**	**0.560**	**0.518**
	16	0.489	0.400	0.516	0.466
SynBoost	-	0.450	0.387	0.506	0.480
Ours with Soft. Ent.	1	**0.681**	**0.618**	**0.628**	**0.565**
	16	0.568	0.447	0.593	0.497
Soft. Ent.	-	0.572	0.444	0.619	0.520
Ours with MC Dropout	1	**0.576**	**0.506**	**0.430**	**0.408**
	16	0.358	0.274	0.327	0.307
MC Dropout	-	0.249	0.218	0.163	0.236
Ours with Soft. Dis	1	**0.668**	**0.593**	**0.600**	**0.532**
	16	0.540	0.409	0.574	0.479
Soft. Dis.	-	0.527	0.408	0.569	0.489

TABLE III: Effect of Changing the Semantic Uncertainty Input: Our models vs. selected uncertainty inputs for semantic estimation's failure prediction.

Then, we continue by adding a second task (depth or semantics, as appropriate), and a third one (normal estimation), see first row in Table II.

From this experiment we have evidence that, indeed, a multi-task setup improves failure detection of one task, and, this is true for both semantic segmentation and depth estimation. This is a confirmation of our hypothesis that our framework leverages the latent correlations among tasks to improve the introspection capabilities of the every single one of them.

How dependant is our failure detection on the uncertainty estimate input? This experiment investigates whether our conclusions have been biased to the uncertainty input used. For the semantic segmentation task we evaluate three more uncertainty inputs, and for the depth estimation another two (see Table I). For each specific uncertainty method, we select two of our models with different patch sizes (1 & 16).

The results of this investigation can be seen in Tables III and IV. Our framework is consistently outperforming the uncertainty input for both tasks failure detections, within both original and foggy image set from Cityscapes.

How is the failure detection performing when adding one more task? Finally, the question comes to our failure detection based on a multi-task setup is fixed to the already chosen tasks in the previous experiments. For that reason, we add the extra task of instance segmentation (IS), with its corresponding uncertainty estimate, see last row of Table I.

The results of this experiment, shown in Table V, indicate that the addition of an extra task continues to be beneficial for the different tasks failure detections. However, we observe that the contribution is higher for the depth estimation failure detection than for the semantic segmentation. We believe this is due to less extra information provided by the instance segmentation with respect to the semantic segmentation task. While differentiating among different instances of the same class is highly informative for the depth estimation task.

Method	Patch	Original	Foggy
		ZNCC↑	ZNCC↑
Ours with Self Learning	1	**0.646**	**0.570**
	16	0.569	0.529
Self Learning	-	0.248	0.255
Ours with Bayesian	1	**0.757**	**0.645**
	16	0.684	0.584
Bayesian	-	0.091	0.074
Ours with MC Dropout	1	**0.648**	**0.516**
	16	0.575	0.446
MC Dropout	-	0.076	0.075

TABLE IV: Effect of Changing the Depth Uncertainty Inputs: Our models vs. selected uncertainty methods for depth estimation's failure prediction.

Patch	Task Entries				Semantic		Depth
	S	D	N	IS	ZNCC↑	AP-Err↑	ZNCC↑
1	✓	✓	✓	✓	0.641	0.585	**0.655**
	✓	✓	✓		**0.649**	**0.590**	0.646
16	✓	✓	✓	✓	**0.493**	**0.403**	**0.581**
	✓	✓	✓		0.489	0.400	0.529

TABLE V: Adding an Extra Task: Comparison among multiple different task entries for both semantic segmentation's and depth estimation's failure prediction.

Method	Wilddash		Dark Zurich	
	ZNCC↑	AP-Err↑	ZNCC↑	AP-Err↑
Ours with Synboost	**0.412**	**0.595**	**0.238**	**0.775**
SynBoost	0.323	0.584	0.199	0.726
Ours with Soft. Ent.	**0.520**	**0.630**	0.544	**0.867**
Soft. Ent.	0.508	0.625	**0.578**	0.830
Ours with MC Dropout	**0.321**	**0.537**	**0.101**	**0.734**
MC Dropout	0.139	0.428	-0.128	0.678
Ours with Soft. Dis	**0.511**	**0.628**	0.497	**0.861**
Soft. Dis.	0.478	0.619	**0.502**	0.797

TABLE VI: Generalization: Our models vs. selected uncertainty inputs for semantic segmentation's failure prediction on two other datasets: Wilddash and Dark Zurich.

How is our failure detection generalizing to scenarios with larger distribution mismatch? Here, we use our default models trained with the Cityscapes dataset, and deployed them on Wilddash and Dark Zurich (night) datasets, for the tasks of failure detection of the semantic segmentation. Results can be seen in Table VI. Additionally, we include the evaluation of different uncertainty inputs as they are quite dependent on the distribution mismatch between the test and training set. We can conclude that the improvements brought by our framework generalize to more complex scenarios, invariant to the uncertainty estimate input.

V. Conclusion

We propose a framework to detect visual task prediction failures. We leverage the information from multiple visual tasks simultaneously being deployed, and build a learning-based attention neural network to perform a weighted sum of task uncertainties to approximate the task prediction failure. Our approach is more accurate in detecting semantic and depth prediction errors, compared with various uncertainty estimation methods. Additionally, our thorough experimental evaluation also proves its ability to further improve the performance by increasing the attention map resolution, as well as by including in extra correlated visual tasks. Finally, we observe that the multi-task setup allows for better generalization to environments with a larger distribution mismatch to that of the training set.

REFERENCES

[1] S. Bulusu, B. Kailkhura, B. Li, P. K. Varshney, and D. Song, "Anomalous example detection in deep learning: A survey," *IEEE Access*, vol. 8, pp. 132 330–132 347, 2020.

[2] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[3] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3827–3837.

[4] Y. Zhu, K. Sapra, F. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8848–8857, 2019.

[5] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, University of Cambridge, 2016.

[6] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3712–3722.

[7] A. R. Zamir, A. Sax, N. Cheerla, R. Suri, Z. Cao, J. Malik, and L. J. Guibas, "Robust learning through cross-task consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[9] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-wise anomaly detection in complex driving scenes," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[10] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] C. Sakaridis, D. Dai, and L. Van Gool, "Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[12] O. Zendel, K. Honauer, M. Murschitz, D. Steininger, and G. F. Dominguez, "Wilddash - creating hazard-aware benchmarks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[13] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Neural Information Processing Systems (NeuIPS)*, 2017.

[14] M. Rottmann, P. Colling, T. Hack, R. Chan, F. Hüger, P. Schlicht, and H. Gottschalk, "Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities," 07 2020, pp. 1–9.

[15] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Neural Information Processing Systems (NeuIPS)*, 2017.

[16] E. Ilg, O. Cicek, S. Galesso, A. Klein, O. Makansi, F. Hutter, and T. Brox, "Uncertainty estimates and multi-hypotheses networks for optical flow," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 652–667.

[17] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 2012, vol. 118.

[18] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "On the uncertainty of self-supervised monocular depth estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[19] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *arXiv preprint arXiv:1812.04606*, 2018.

[20] C. B. Kuhn, M. Hofbauer, Z. Xu, G. Petrovic, and E. Steinbach, "Pixel-wise failure prediction for semantic video segmentation," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 614–618.

[21] Q. Marufur Rahman, N. Sünderhauf, P. Corke, and F. Dayoub, "Fsnet: A failure detection framework for semantic segmentation," *arXiv e-prints*, pp. arXiv–2108, 2021.

[22] K. Lis, K. Nakka, P. Fua, and M. Salzmann, "Detecting the unexpected via image resynthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2152–2161.

[23] D. Haldimann, H. Blum, R. Siegwart, and C. Cadena, "This is not what i imagined: Error detection for semantic segmentation through visual dissimilarity," *arXiv preprint arXiv:1909.00676*, 2019.

[24] S. Ben-David and R. Borbely, "A notion of task relatedness yielding provable multiple-task learning guarantees," *Machine Learning*, vol. 73, pp. 273–287, 12 2008.

[25] I. Kokkinos, "Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5454–5463.

[26] Y. Aytar and A. Zisserman, "Tabula rasa: Model transfer for object category detection," in *2011 international conference on computer vision*. IEEE, 2011, pp. 2252–2259.

[27] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE signal processing magazine*, vol. 32, no. 3, pp. 53–69, 2015.

[28] X. Wang, K. He, and A. Gupta, "Transitive invariance for self-supervised visual representation learning," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1329–1338.

[29] A. Pentina and C. H. Lampert, "Multi-task learning with labeled and unlabeled tasks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2807–2816.

[30] 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*, 2016, pp. 3994–4003.

[31] V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon, "Semantically-guided representation learning for self-supervised monocular depth," *International Conference on Learning Representations*, 2020.

[32] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[35] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *Proceedings of The 33rd International Conference on Machine Learning*, 06 2015.

[36] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5683–5692.

[37] R. Mohan and A. Valada, "Efficientps: Efficient panoptic segmentation," *International Journal of Computer Vision (IJCV)*, vol. 129, pp. 1551 – 1579, 2021.

[38] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[39] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, no. 9, pp. 973–992, Sep 2018. [Online]. Available: https://doi.org/10.1007/s11263-018-1072-8