Hyperdimensional Feature Fusion for Out-of-Distribution Detection: Supplementary Material

Samuel Wilson
Queensland University of Technology
2 George St, Brisbane, QLD 4000, Australia
s84.wilson@hdr.qut.edu.au

Tobias Fischer
Queensland University of Technology
2 George St, Brisbane, QLD 4000, Australia
tobias.fischer@qut.edu.au

Niko Sünderhauf
Queensland University of Technology
2 George St, Brisbane, QLD 4000, Australia
niko.suenderhauf@qut.edu.au

Feras Dayoub
University of Adelaide
North Terrace, Adelaide, SA 5005, Australia
feras.dayoub@adelaide.edu.au

1. Additional Experiments

As described in the main submission, we complete our evaluation by incorporating the CIFAR100 as in-distribution setting as well as additional metrics.

1.1. CIFAR100 Results

To complete the AUROC evaluation discussed in the main paper, we find the corresponding results for the OOD detection setting with CIFAR100 as the in-distribution dataset in Table 1 and Table 2.

Table 1 provides the OOD detection results for the **statistics** stream but does not include the Energy-based model [7] due to no published results existing for the CIFAR100 setting. The results from Table 1 reinforce the findings from the main submission, with the gap between the top performers (HDFF and Gram) growing significantly between the other methods, averaging out to ≈10% AUROC difference. Comparative to these shifts in scores, the difference between HDFF and Gram remains small with HDFF taking significantly less computational time to attain its respective results.

Table 2 displays the additional OOD detection results for the **training** stream but does not contain NMD [1] due to the absence of published results in the CIFAR100 setting. Similarly to the results from the main submission, we see that HDFF in combination with other state-of-the-art OOD detectors increases performance across the majority of benchmarks. Specifically, we see that HDFF-1DS outperforms the Spectral Discrepancy Detector [12] in two of the four comparative benchmarks despite HDFF requiring 50x less computation to achieve these results. We note that in this CIFAR100 setting, Table 2 shows that HDFF-MLP is weaker at the SVHN and CIFAR10 OOD datasets.

Statistics Stream - CIFAR100

OOD Dataset	HDFF (Ours)	HDFF-Ens (Ours)	Gram [11]	MSP [3]	ML [2]
iSun	95.2	*95.8*	**98.8**	82.5	85.5
TINc	93.1	*93.8*	**98.2**	83.5	86.3
TINr	95.4	*96.0*	**98.5**	81.6	84.3
LSUNc	91.7	*92.5*	**96.0**	83.9	86.5
LSUNr	94.5	*95.3*	**99.3**	82.7	85.5
SVHN	*99.2*	**99.4**	99.0	86.7	90.0
MNIST	*99.8*	*99.8*	**99.9**	82.4	84.6
KMNIST	99.5	*99.6*	**99.99**	86.6	87.5
FMNIST	*98.4*	*98.4*	**99.4**	91.0	93.3
DTD	92.9	*93.5*	**97.5**	78.1	79.7
CIFAR10	65.7	68.2	74.2	*80.9*	**81.5**
Average	93.2	*93.8*	**96.4**	83.6	85.9

Table 1: OOD detection results for the against the methods contained belonging to the **statistics** stream with CIFAR100 as the in-distribution dataset. Comparison metric is AUROC, higher is better. **Best** results are shown in **blue and bold**, *second* best results are shown in *green and italics*. The ensemble in *HDFF-Ens* always consists of 5 models. Due to an absence of published data, Energy-based model [7] is not included. In the far-OOD settings, HDFF and Gram achieve significant (≈10%) improvements in AUROC over the competing methods.

Due to the high compatibility of HDFF with the MLP on other ODO benchmarks, it is unclear if this drop is due to the MLP or the HDFF representation without comparative results from other representations. Future work on auxiliary OOD detection networks may investigate the sensitivity of these auxiliary networks to their input data, determining which representations are most effective for individual OOD datasets.

Training Stream - CIFAR100

OOD Dataset	HDFF-MLP (Ours)	HDFF-1DS (Ours)	Spectral [12]	DDU [9]	MOOD [6]
iSun	**99.9**	*94.4*	-	-	77.8
TINc	**99.4**	*93.5*	88.6	83.13*	-
TINr	**99.8**	*94.0*	93.7	83.13*	-
LSUNc	*93.9*	90.5	93.8	-	**96.8**
LSUNr	**99.96**	94.9	*95.7*	-	77.6
SVHN	54.1	**92.8**	-	*87.53*	85.9
MNIST	**99.8**	*96.9*	-	-	91.3
KMNIST	**99.6**	*98.4*	-	-	97.2
FMNIST	**99.7**	97.8	-	-	*99.1*
DTD	**91.0**	*86.4*	-	-	71.7
CIFAR10	*44.9*	**77.7**	-	-	-
Average	89.3	*92.5*	**93.0**	84.6	87.2

Table 2: OOD detection results for the against the methods contained belonging to the **training** stream with CIFAR100 as the in-distribution dataset. Comparison metric is AUROC, higher is better. **Best** results are shown in **blue and bold**, *second* best results are shown in *green and italics*. Due to an absence of published data, NMD [1] is not included. When HDFF is combined with other state-of-the-art detectors it consistently provides state-of-the-art performance across many benchmarks.

Statistics Stream - FPR95

ID Dataset	OOD Dataset	HDFF (Ours)	HDFF-Ens (Ours)	Gram [11]	MSP [3]	ML [2]
CIFAR10	iSun	*2.7*	2.8	**0.6**	21.8	10.5
	TINc	6.8	*6.6*	**2.7**	28.4	16.2
	TINr	*2.7*	*2.7*	**1.0**	29.9	17.0
	LSUNc	*20.7*	20.8	**8.8**	25.7	14.3
	LSUNr	2.1	*1.8*	**0.4**	21.4	10.3
	SVHN	2.7	**2.3**	*2.6*	25.3	14.3
	MNIST	*0.01*	**0**	**0**	56.6	43.9
	KMNIST	1	*0.4*	**0**	42.4	31.4
	FMNIST	*1.2*	**0.3**	**0.3**	37.2	25.1
	DTD	27.3	*26.4*	**8.7**	46.7	39.3
	CIFAR100	83.6	82.8	68.1	*52.8*	**47.8**
CIFAR100	iSun	25.3	*22.9*	**6.3**	72.4	68.9
	TINc	35.2	*33.1*	**10.0**	67.5	63.4
	TINr	23.9	*21.3*	**6.7**	72.0	68.7
	LSUNc	39.1	*37.3*	**23.3**	70.1	66.7
	LSUNr	29.8	*26.0*	**5.9**	72.1	68.3
	SVHN	*3.1*	**2.4**	4.1	66.1	59.2
	MNIST	**0**	**0**	**0**	83.3	*82.5*
	KMNIST	0.04	*0.01*	**0**	70.2	70.3
	FMNIST	*6.7*	*6.7*	**1.7**	49.0	40.9
	DTD	31.9	*30.5*	**15.1**	79.3	78
	CIFAR10	92.2	91.0	86.8	*77.2*	**76.4**
Average		19.9	*19.0*	**11.5**	53.1	46.1

Table 3: OOD detection results for the against the methods contained belonging to the **statistics** stream. Comparison metric is FPR@95, lower is better. **Best** results are shown in **blue and bold**, *second* best results are shown in *green and italics*. Due to an absence of published data, Energy-based model [7] is not included. HDFF and Gram consistently produce state-of-the-art level performance, with a significant margin between them and the other statistical baselines.

1.2. Additional Metrics

Tables 3 and 4 complete the **statistics** stream evaluation from our main submission against the current state-of-the-art under the FPR95 and Detection Error metrics respectively with both CIFAR in-distribution datasets. Note that neither Table 3 nor Table 4 contain results for the Energy-based Model [7] due to published results being absent for these metrics. Across the board, Tables 3 and 4 reflect the results from the AUROC evaluation in the main submission with HDFF and Gram [11] outperforming the other comparison methods. Consistent with the results from the main submission, we see that the performance gap between HDFF and Gram is relatively minor, by comparison to the gaps of other methods, with HDFF having a significant lower computational cost than Gram. Overall, the original evaluation against the AUROC metric provides an accurate assessment of the relative performance of the compared methods, with Tables 3 and 4 reinforcing the findings from the original submission.

The results for the **training** stream are reported in Tables 5 and 6 for the FPR95 and Detection Error metrics respectively. Note that neither table contains results for the DDU [9] OOD detector and Table 6 additionally does not contain MOOD [6]; the absence of both is due to a lack of published results. The results from Tables 5 and 6 repeat the same patterns as both the main submission and the additional results described in Table 2. In general, HDFF-MLP provides the strongest performance across the majority of the benchmarks, with HDFF-1DS also boasting large performance benefits over the original Spectral Discrepancy Detector [12] in the CIFAR10 setting, with smaller benefits in the CIFAR100 setting. In summary, the results from Tables 5 and 6 help strengthen the claims main in the main submission by broadening the standard of good performance.

2. Ablations

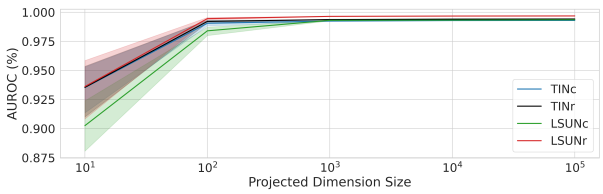
2.1. Projection Dimensionality

Figure 1 visualises the influence of the number of dimensions in the hyperdimensional space on the OOD detection performance. Specifically, we plot the mean and 95% confidence interval over 10 initialisations of the random projection matrices at each order of magnitude for a single 1D Subspaces [12] trained model. The general trend seen in both figures is that as the number of dimensions increases up to 10^3, the mean performance of the HDFF detector increases. Similarly, we see that the bounds of 95% confidence all but effectively disappear when considering spaces of 10^3 dimensions or greater. The results of this ablation show that standard choices of hyperdimensional space in the range of $10^3 - 10^4$ will produce reasonable results, consistent with conventions used in HDC literature [10, 8, 4].

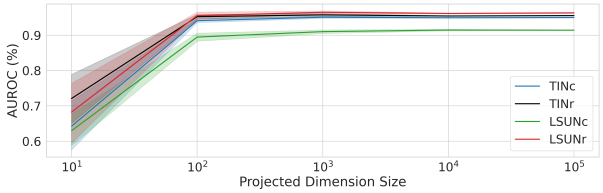
Statistics Stream - Detection Error

ID Dataset	OOD Dataset	HDFF (Ours)	HDFF-Ens (Ours)	Gram [11]	MSP [3]	ML [2]
CIFAR10	iSun	*3.7*	3.9	**1.8**	8.2	6.9
	TINc	5.5	*5.4*	**3.5**	9.4	8.3
	TINr	*3.8*	*3.8*	**2.1**	9.8	8.6
	LSUNc	9.7	9.7	**6.7**	9.0	*7.9*
	LSUNr	*3.3*	3.4	**1.2**	8.0	6.7
	SVHN	3.7	**3.5**	*3.6*	7.8	7.2
	MNIST	1.5	*1.1*	**0.1**	16	15.4
	KMNIST	2.9	*2.6*	**0.3**	12.3	12.0
	FMNIST	3.1	*2.5*	**1.6**	11.9	10.7
	DTD	12.8	*12.7*	**6.7**	15.3	15.6
	CIFAR100	29.9	30.0	28.0	**17.4**	*18.0*
CIFAR100	iSun	11.8	*11.0*	**5.0**	25.2	22.0
	TINc	14.6	*13.6*	**6.4**	24.5	21.2
	TINr	11.4	*10.6*	**5.4**	26.0	23.0
	LSUNc	16.9	*15.7*	**11.2**	23.7	20.6
	LSUNr	12.6	*11.4*	**4.9**	24.9	21.8
	SVHN	*3.9*	**3.5**	4.3	20.7	16.7
	MNIST	0.7	**0.5**	*0.6*	23.1	20.6
	KMNIST	*1.6*	*1.6*	**0.3**	20.0	18.8
	FMNIST	5.0	*4.9*	**3.0**	16.9	13.8
	DTD	15.3	*14.4*	**8.2**	28.0	26.6
	CIFAR10	37.1	35.1	32.8	*25.9*	**25.3**
Average		9.6	*9.1*	**6.3**	17.5	15.8

Table 4: OOD detection results for the against the methods contained belonging to the **statistics** stream. Comparison metric is Detection Error, lower is better. **Best** results are shown in **blue and bold**, *second* best results are shown in *green and italics*. Due to an absence of published data, Energy-based model [7] is not included.



(a) CIFAR10 as In-Distribution



(b) CIFAR100 as In-Distribution

Figure 1: Ablation of the required size of projected HD space to achieve best OOD detection performance. Individual plot lines show mean and 95% confidence interval over 10 independent initialisations of projection matrices on the same model trained with the 1D Subspaces [12] methodology. Features should be projected into a HD space in the range of $10^3 - 10^4$ dimensions to remove the potential for variation in OOD detection performance.

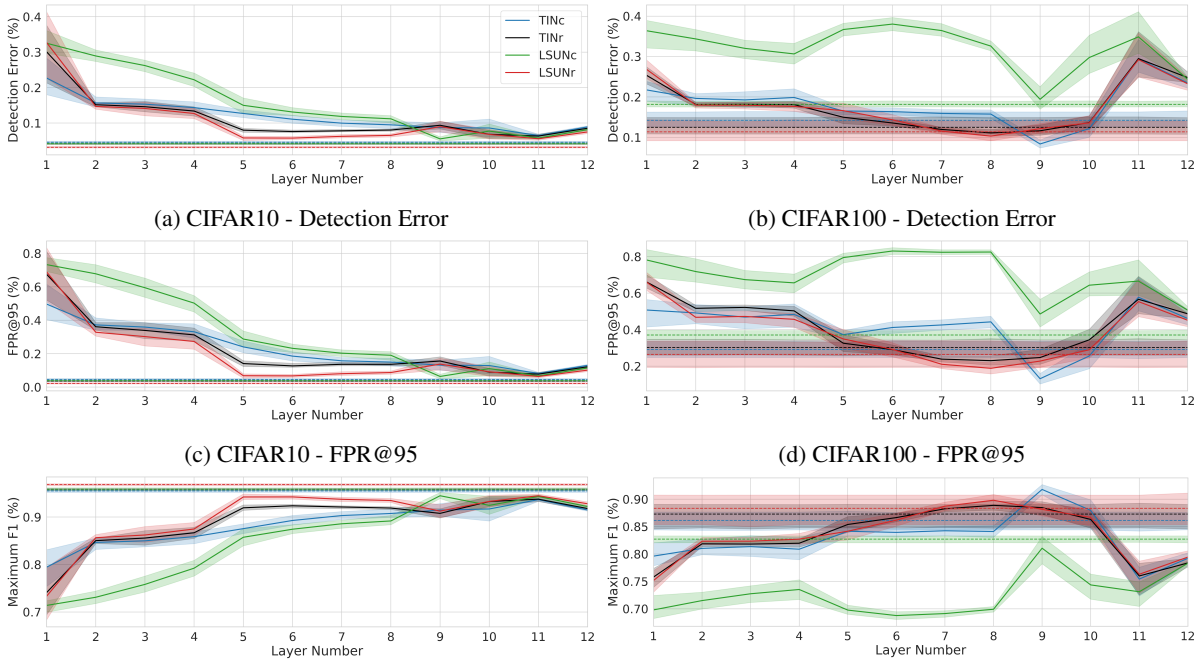
Training Stream - FPR95

ID Dataset	OOD Dataset	HDFF (MLP)	HDFF (1DS)	Spectral [12]	NMD [1]	MOOD [6]
CIFAR10	iSun	**0.01**	0.5	-	*0.3*	38.8
	TINc	**0.4**	*1.7*	9.0	3.9	-
	TINr	**0.1**	*1.3*	7.6	-	-
	LSUNc	9.6	3.3	**2.8**	6.1	*3.2*
	LSUNr	**0**	*0.4*	3.4	-	36.2
	SVHN	66.4	*5.2*	-	**2.3**	17.2
	MNIST	**0.2**	2.8	-	-	*0.4*
	KMNIST	5.0	*2.5*	-	-	**0.3**
	FMNIST	*0.3*	4.1	-	-	**0.1**
	DTD	18.9	*14.8*	-	**6.0**	56.0
	CIFAR100	85.8	*42.6*	-	**36.2**	-
CIFAR100	iSun	**0.2**	*23.6*	-	-	81.5
	TINc	**2.4**	*28.8*	41.7	-	-
	TINr	**0.7**	*24.3*	47.2	-	-
	LSUNc	*25.1*	48.2	50.2	-	**17.0**
	LSUNr	**0.1**	*21.6*	43.0	-	81.2
	SVHN	96.0	**26.0**	-	-	*63.7*
	MNIST	**0**	*11.7*	-	-	57.7
	KMNIST	**0.4**	*6.7*	-	-	16.6
	FMNIST	**0.01**	8.3	-	-	*4.6*
	DTD	**31.1**	*50.9*	-	-	86.8
	CIFAR10	*97.1*	**87.6**	-	-	-
Average		20.0	*19.0*	25.6	**9.1**	35.1

Table 5: OOD detection results for the against the methods contained belonging to the **training** stream. Comparison metric is FPR@95, lower is better. **Best** results are shown in **blue and bold**, *second* best results are shown in *green and italics*. Due to an absence of published data, DDU [9] is not included. Consistent with the findings from the main submission, HDFF when combined with recent state-of-the-art OOD detectors produces improved results, setting a new state-of-the-art.

2.2. Additional Layer Ablations

Figure 2 provides additional ablations of individual layer performance with respect to the other standard metrics for the benchmark defined in [5]; those being FPR@95, Detection Error and Maximum F1 Score. We make use of the same 12 BasicBlock hooks as in the main submission. Across both settings we see that the cropped datasets nearly perfectly mirror each other, indicating that there is a set of features that HDFF is consistently strong at detecting. In the CIFAR10 setting we see that across all metrics, there is no individual layer that performs at or above the level of the the fusion of feature maps. This trend is violated by the later layers, 7 through to 9 in the CIFAR100 setting. In particular, we see that the metrics that measure binarised performance, F1 and FPR@95, appear to have the largest increase above the fusion of features, usually in only a single layer. Overall, we make the same observation as in the main submission; there is no individual layer that performs well across all OOD datasets, assuming we already have the prior knowledge of how well each layer performs at this task. Future work into weighting individual layers based on predicted performance may be an important consideration.



(a) CIFAR10 - Detection Error

(b) CIFAR100 - Detection Error

(c) CIFAR10 - FPR@95

(d) CIFAR100 - FPR@95

(e) CIFAR10 - Maximum F1

(f) CIFAR100 - Maximum F1

Figure 2: Layer-wise ablation across multiple metrics for the 1D Subspaces trained model. We plot mean and 95% confidence interval for each datasets with the dotted lines corresponding to the performance of the fusion of feature maps across the network. For the CIFAR10 setting in (a), (c) and (e) it is clear that the fusion of feature maps from across the network provide the best performance. For the CIFAR100 setting in (b), (d) and (f) we see that violations of this characteristic are present in the later layers of 7, 8 and 9. Although some individual layers violate this characteristic, we note that no individual layer performs well across all of the OOD datasets.

2.3. Preprocessing

In order to apply our hyperdimensional feature fusion to the feature maps of a DNN we first implemented a preprocessing pipeline that made use of a global pooling operation and mean-centering. Table 7 ablates the choice of pooling components on a subset of datasets with the AUROC metric, showing their influence to the overall results described in the main paper.

Table 7 shows that in most cases max pooling increases performance in the CIFAR10 settings and average pooling provides better results on the whole for the CIFAR100 setting. We additionally note that the performance gaps between max and average pooling are relatively minor, with only the LSUN datasets in the CIFAR100 setting having a discrepancy greater than 1%. Overall, both max and average pooling are approximately balanced with no clear better choice.

2.4. Inter-Sample Similarity

The concept of the angular distance between class descriptor vectors and image descriptors as visual similarity between the sample and the input extends to comparisons between pairs of image descriptor vectors. By comparing the descriptor vectors between two images, HDFF provides a quantitative metric that evaluates how visually similar the two input samples are. Since the angles between vectors is bounded between $[0, 90]$ degree we can assign semantic meaning to these bounds with 0 corresponding to exactly matching images and 90 corresponding to images with no similarity at all. Figure 3 demonstrates a few examples of this, visualising the angular difference between images sampled from the CIFAR10 ID set.

References

[1] Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and H.T. Kung. Neural mean discrepancy for efficient out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19217–19227, 2022.

[2] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019.

[3] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural net-

Figure 3: Visualisation of similarity between CIFAR10 ID samples using a 1D Subspaces trained WideResNet. Lines with angles between images corresponds to the angular distance between each images respective image descriptor vector. Comparisons are only made between direct neighbours to avoid clutter.

Training Stream - Detection Error					
ID Dataset	OOD Dataset	HDFF-MLP (Ours)	HDFF-1DS (Ours)	Spectral [12]	NMD [1]
CIFAR10	iSun	**0.5**	1.9	-	*1.4*
	TINc	**1.6**	*2.9*	6.8	4.4
	TINr	**0.8**	*2.5*	6.2	-
	LSUNc	6.6	*4.0*	**3.7**	5.5
	LSUNr	**0.3**	*1.8*	3.8	-
	SVHN	22.3	*4.9*	-	**3.4**
	MNIST	**2.1**	*3.8*	-	-
	KMNIST	*4.7*	**3.7**	-	-
	FMNIST	**2.1**	*4.5*	-	-
	DTD	10.6	*8.7*	-	**5.4**
	CIFAR100	35.2	**15.6**	-	*16.6*
CIFAR100	iSun	**1.1**	*14.0*	-	-
	TINc	**3.3**	*15.2*	18.9	-
	TINr	**1.7**	14.3	*14.2*	-
	LSUNc	**13.2**	18.0	*13.9*	-
	LSUNr	**0.8**	13.1	*11.3*	-
	SVHN	*46.4*	**15.4**	-	-
	MNIST	**0.6**	*7.3*	-	-
	KMNIST	**2.0**	*4.9*	-	-
	FMNIST	**1.3**	*5.7*	-	-
	DTD	**16.4**	*23.6*	-	-
	CIFAR10	*49.9*	**26.6**	-	-
Average		10.2	*9.7*	9.9	**6.1**

Table 6: OOD detection results for the against the methods contained belonging to the **training** stream. Comparison metric is Detection Error, lower is better. **Best** results are shown in **blue and bold**, *second* best results are shown in *green and italics*. Due to an absence of published data, DDU [9] and MOOD [6] are not included.

ID set	OOD set	Avg Pool	Max Pool
CIFAR10	TINc	98.0	**98.3**
	TINr	98.8	**99.2**
	LSUNc	**96.9**	96.2
	LSUNr	99.1	**99.2**
CIFAR100	TINc	**93.7**	93.1
	TINr	**96.1**	95.4
	LSUNc	87.7	**91.7**
	LSUNr	**95.8**	94.5

Table 7: Ablation of our feature preprocessing in terms of the AUROC metric in the single-inference pass setting with the standard cross-entropy trained WideResNet model. **Best** result per ID + OOD dataset combination is coloured. Differing pooling operations provide better results dependant on the specific ID/OOD dataset configuration with no clear winner between the two.

works. In *Proceedings of the International Conference on Learning Representations*, 2017.

[4] Alejandro Hernández-Cano, Cheng Zhuo, Xunzhao Yin, and Mohsen Imani. Real-time and robust hyperdimensional classification. In *Great Lakes Symposium on VLSI*, page 397–402, 2021.

[5] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the International Conference on Learning Representations*, 2018.

[6] Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15313–15323, 2021.

[7] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

[8] Justin Morris, Yilun Hao, Roshan Fernando, Mohsen Imani, Baris Aksanli, and Tajana Rosing. Locality-based encoder and model quantization for efficient hyper-dimensional computing. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2021. to appear.

[9] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deep deterministic uncertainty: A simple baseline. *arXiv preprint arXiv:2102.11582*, 2022.

[10] Peer Neubert, Stefan Schubert, Kenny Schlegel, and Peter Protzel. Vector semantic representations as descriptors for visual place recognition. In *Robotics: Science and Systems*, 2021.

[11] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8491–8501. PMLR, 2020.

[12] Alireza Zaeemzadeh, Niccolo Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9452–9461, 2021.