Improving Visual Place Recognition Based Robot Navigation Through Verification of Localization Estimates: Supplementary

Owen Claxton, Connor Malone, Helen Carson, Jason J. Ford, Gabe Bolton, Iman Shames, Michael Milford *Senior Member, IEEE*

We provide the following more detailed results and analysis of the proposed verification system with respect to verification performance and standard VPR performance.

First, we would like to present several tables and figures which evaluate the baseline and verification systems with respect to 'classification' type metrics. That is, these metrics assess the system's ability to accurately classify whether a proposed VPR position estimate is 'in-tolerance' or 'out-of-tolerance' of the true position. In the following tables, we highlight the best performing verification method using **bold** if it is better than both the baseline and other verification method, and using an underline if it is the best verification method but not better than the baseline.

We use Accuracy, Precision, and Recall as defined below for these tables:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Where, TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

For these results, we assume the 'baseline' system has no way of rejecting VPR position estimates and therefore considers all position estimates to be 'in-tolerance'. For our method, we use the single-query verification method presented in Section III.D of the manuscript.

In the case of binary classification, the Precision and Recall values change depending on which class is considered the 'Positive'. That is, the Precision and Recall can be calculated from the perspective of classifying which queries are 'in-tolerance' ***or*** which queries are 'out-of-tolerance'. We provide results for the 'out-of-tolerance' case in Table 2, and results for the 'in-tolerance' case in Table 3.

Table 1: 'Classification' accuracy for verification systems.

		Office, Normal			Office, Adverse			Campus, Normal			Campus, Adverse		
		Baseline	Verification Method		Baseline	Verification Method		Baseline	Verification Method		Baseline	Verification Method	
		VPR	SVM	Ours	VPR	SVM	Ours	VPR	SVM	Ours	VPR	SVM	Ours
AP-GeM	Accuracy	92%	49%	77.8%	72.2%	49.3%	61.7%	94.2%	30.4%	76.6%	67.4%	40.1%	63.8%
NetVLAD	Accuracy	92.1%	40.4%	77.9%	70.6%	44.6%	65%	93.5%	28.9%	66.9%	45.4%	53.7%	**66%**
SALAD	Accuracy	96%	35.7%	74.2%	90%	33.3%	65.1%	95.7%	40.9%	84%	94.8%	39.3%	63.3%

For the above classification accuracy results, it can be seen that our proposed verification method is always better than the previously state-of-the-art SVM method but often is still theoretically worse than the baseline of 'accepting all position estimates'. However, this is an artifact of the datasets and the individual VPR method's performance on them. This indicates that the VPR methods are able to provide accurate position estimates for $> 90\%$ of queries a lot of the time. This table in particular shows that our method is able to correctly reject 'bad' VPR position estimates without incorrectly rejecting as many 'good' position estimates compared to the SVM method.

Table 2 shows that our method is generally able to correctly predict which VPR position estimates are 'bad' for a greater number of queries (higher recall) and without rejecting as many 'good' position estimates (higher precision). With the exception being the SALAD VPR method where the SVM achieves better recall but worse precision than our method.

Table 3 shows that our method generally has a higher precision, if not a comparable one, to the SVM method. However, it also shows that at this precision, our method is always able to achieve a higher recall. Practically, this means that our method is able to perform the verification of VPR position estimates to a higher or similar level of precision without rejecting as many 'good' position

Table 2: 'Classification' precision and recall from the perspective of classifying which queries are 'out-of-tolerance'. That is, out-of-tolerance = Positive

		Office, Normal			Office, Adverse			Campus, Normal			Campus, Adverse		
		Baseline	**Verification Method**		**Baseline**	**Verification Method**		**Baseline**	**Verification Method**		**Baseline**	**Verification Method**	
		VPR	**SVM**	**Ours**	**VPR**	**SVM**	**Ours**	**VPR**	**SVM**	**Ours**	**VPR**	**SVM**	**Ours**
AP-GeM	Precision	-	11.2%	**24.8%**	-	33.9%	**41.7%**	-	6.8%	**7.4%**	-	32.4%	**47.1%**
	Recall	0%	77.5%	**87.6%**	0%	86.3%	**94.3%**	0%	**86.6%**	26.2%	0%	77.1%	**90.1%**
NetVLAD	Precision	-	10.3%	**23.7%**	-	29%	**43.3%**	-	7.4%	**11.3%**	-	55.7%	**63.4%**
	Recall	0%	**85.2%**	80.7%	0%	60.7%	**60.9%**	0%	**86.3%**	59.8%	0%	73.8%	**89.1%**
SALAD	Precision	-	5.7%	**6.8%**	-	12%	**20.3%**	-	4.7%	**5.1%**	-	6%	**7.1%**
	Recall	0%	**95.6%**	42.2%	0%	**88.6%**	84.7%	0%	**65.7%**	15.3%	0%	**73%**	49.5%

Table 3: 'Classification' precision and recall from the perspective of classifying which queries are 'in-tolerance'. That is, in-tolerance = Positive.

		Office, Normal			Office, Adverse			Campus, Normal			Campus, Adverse		
		Baseline	**Verification Method**		**Baseline**	**Verification Method**		**Baseline**	**Verification Method**		**Baseline**	**Verification Method**	
		VPR	**SVM**	**Ours**	**VPR**	**SVM**	**Ours**	**VPR**	**SVM**	**Ours**	**VPR**	**SVM**	**Ours**
AP-GeM	Precision	92%	96%	**98.6%**	72.2%	86.9%	**95.7%**	94.2%	**97%**	94.6%	67.4%	66.7%	**91.4%**
	Recall	100%	46.5%	76.9%	100%	35%	49.2%	100%	27%	79.7%	100%	22.1%	51.1%
NetVLAD	Precision	92.1%	96.6%	**97.9%**	70.6%	69.8%	**80.3%**	93.5%	**96.3%**	96%	45.4%	48.4%	**74.6%**
	Recall	100%	36.5%	77.6%	100%	37.9%	66.7%	100%	24.9%	67.3%	100%	29.5%	38.3%
SALAD	Precision	96%	**99.4%**	96.9%	90%	95.5%	**97.4%**	95.7%	**96.3%**	95.8%	94.8%	**96.2%**	95.8%
	Recall	100%	33.2%	75.5%	100%	27.1%	62.9%	100%	39.8%	87.1%	100%	37.4%	64.1%

estimates. It is worth noting that these performance metrics are nearly equivalent to the precision and recall that would be computed for VPR performance. The precision values will be the same, however, the recall is calculated slightly differently in the case of VPR (as shown later). As a result, the baseline system always is shown as having 100% recall in this table of results.

In addition to these tables, we provide the raw TP, TN, FP, and FN values to provide further analysis of how our system performs with respect to both the baseline and SVM methods. The following figures are for the case where the verification system is classifying which queries are 'in-tolerance', i.e. in-tolerance = Positive:

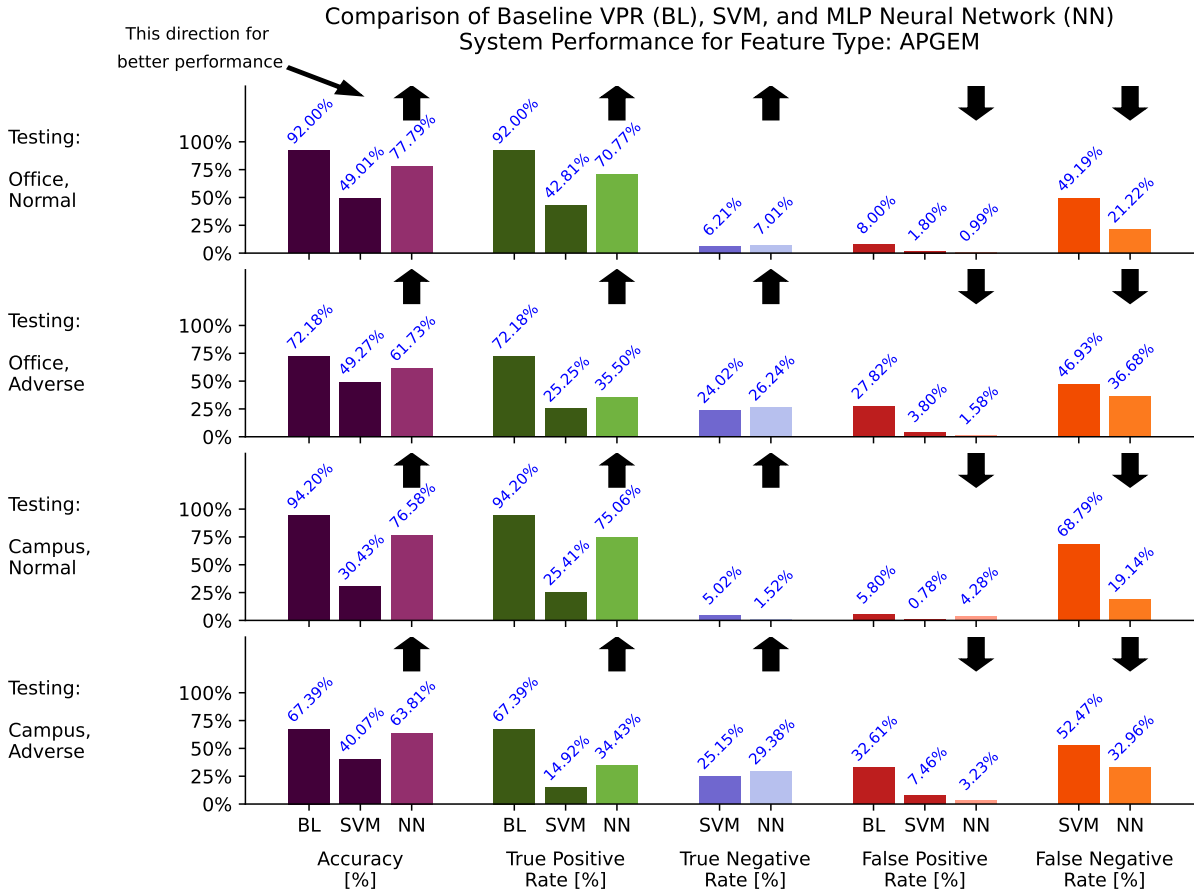

Figure 1: Results for 'classification' performance with **AP-GeM** as the VPR descriptor. The bar plots show the accuracy (far left), and the separate confusion matrix values (True Positives, True Negatives, False Positives, False Negatives).

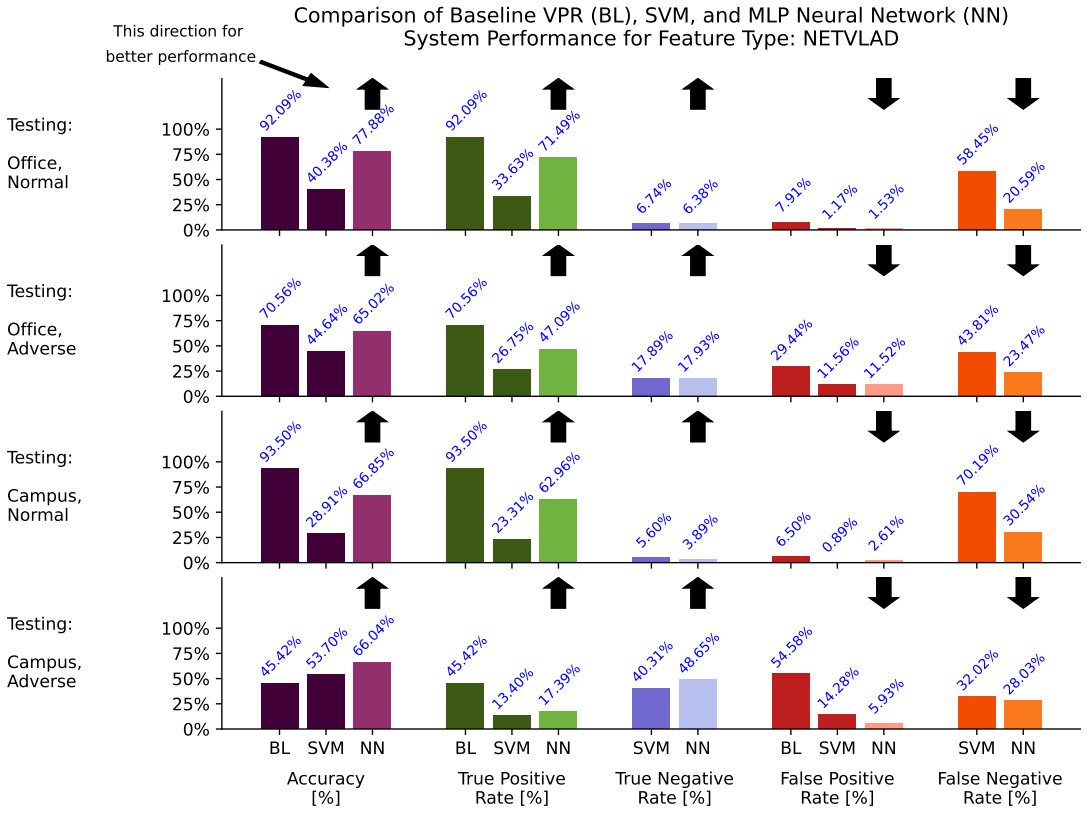

Figure 2: Results for 'classification' performance with **NetVLAD** as the VPR descriptor. The bar plots show the accuracy (far left), and the separate confusion matrix values (True Positives, True Negatives, False Positives, False Negatives).

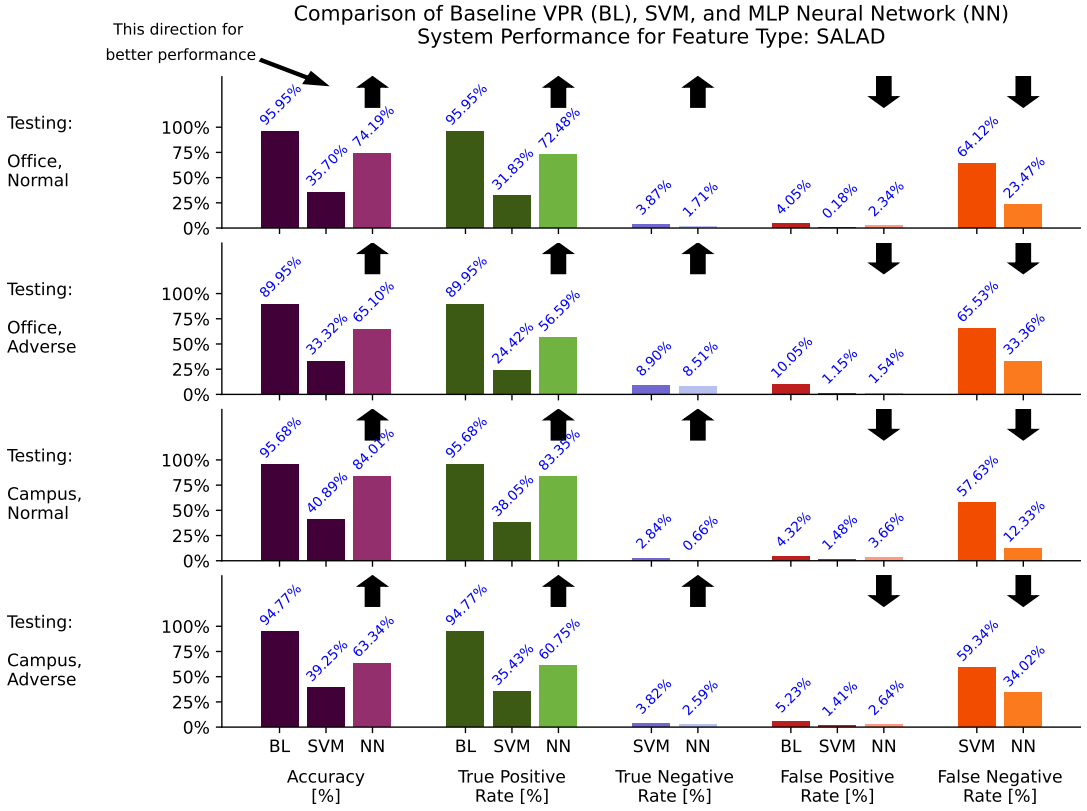

Figure 3: Results for 'classification' performance with **SALAD** as the VPR descriptor. The bar plots show the accuracy (far left), and the separate confusion matrix values (True Positives, True Negatives, False Positives, False Negatives).

For completeness, we also include here the classification Accuracy, Precision, and Recall values for the additional datasets now included in the manuscript. These datasets are the Oxford RobotCar dataset, using an Overcast set for query data and a Sunny set for reference data, and the Nordland dataset, using the Summer set as query data andthe Winter set as reference data. For both the SVM and our MLP verification methods, we use the models which were trained on data from the QCR training datasets, therefore making both of these new datasets completely unseen.

We once again provide the values from the perspective of classifying which queries are 'in-tolerance', i.e. in-tolerance = Positive. At this stage, we provide results for the AP-GeM and NetVLAD VPR descriptors, however, we plan to also add results for the SALAD VPR descriptor once we are able to extract features for the respective datasets.

Table 4: 'Classification' accuracy for verification of VPR position estimates across the new VPR datasets.

		Nordland			**RobotCar**		
		Baseline	**Verification Method**		**Baseline**	**Verification Method**	
		VPR	**SVM**	**Ours**	**VPR**	**SVM**	**Ours**
AP-GeM	Accuracy	3.4%	7.6%	**57.0%**	65.1%	53.0%	**68.2%**
NetVLAD	Accuracy	8.6%	8.3%	**69.8%**	77.0%	54.9%	**77.1%**

Interestingly, for both of these datasets, Table 4 shows that VPR with verification of position estimates using our method is able to achieve the highest accuracy in all cases. This is a good indicator that these datasets are harder for the baseline VPR systems than the QCR datasets. In the case of Oxford RobotCar, the SVM method is worse than the baseline for both VPR descriptors.

Table 5: 'Classification' precision and recall for verification of VPR position estimates across the new VPR datasets from the perspective of classifying which queries are 'in-tolerance'. That is, in-tolerance = Positive.

		Nordland			**RobotCar**		
		Baseline	**Verification Method**		**Baseline**	**Verification Method**	
		VPR	**SVM**	**Ours**	**VPR**	**SVM**	**Ours**
AP-GeM	Precision	3.4%	3.4%	**5.5%**	65.1%	63.2%	**75.5%**
	Recall	100.0%	96.8%	72.3%	100.0%	66.2%	75.7%
NetVLAD	Precision	8.6%	7.0%	**13.6%**	77.0%	71.4%	**83.1%**
	Recall	100.0%	79.3%	46.8%	100.0%	69.3%	88.1%

Table 5 shows that our method in the case of the Oxford RobotCar is able to somewhat generalize and provide a higher precision verification system with a smaller drop in recall compared to the SVM method. In the case of Nordland, the SVM method has a much higher recall, however, this is likely a result of accepting the majority of VPR position estimates whether they are correct or incorrect. This assumption is reflected in the very low accuracy of the SVM method. By observing the table it can be seen that both verification methods and the baseline perform relatively poorly on the Nordland dataset. Not only is this dataset considered a difficult dataset for VPR, but given that it captures footage from a train through the countryside, the data is from a significantly different environment to the one which the SVM and MLP verification methods were trained on. Incorporating similar data into the training process for these methods should help increase performance.

In addition to these results focusing on the 'classification' ability of the different verification methods for VPR position estimates, we would like to provide some VPR-focused results. As mentioned above, the precision for VPR will be equal to the 'classification' precision in the case where it is evaluated from the perspective of classifying which queries are 'in-tolerance', i.e. in-tolerance = Positive. However, the recall in the case of VPR is typically calculated differently. We use the definition for recall as set in [1]:

$$Recall = \frac{TP}{GTP} \tag{4}$$

Where GTP is the number of queries where it is possible for VPR to retrieve a correct match from the reference dataset. That is, all query locations have been visited before in the reference traverse. Therefore, for all of the datasets discussed here (QCR, Nordland and Oxford RobotCar), this would simply become:

$$Recall = \frac{TP}{\#\ of\ Queries} \tag{5}$$

Table 6: VPR-based precision and recall values for all datasets and VPR descriptors

		AP-GeM			NetVLAD			SALAD		
		Baseline	Verification Method		Baseline	Verification Method		Baseline	Verification Method	
		VPR	SVM	Ours	VPR	SVM	Ours	VPR	SVM	Ours
Office, Norm	Precision	92.00%	95.97%	**98.62%**	92.09%	96.64%	**97.91%**	95.95%	**99.44%**	96.88%
	Recall	92.00%	42.81%	70.77%	92.09%	33.63%	71.49%	95.95%	31.83%	72.48%
Office, Adv	Precision	72.18%	86.92%	**95.73%**	70.56%	69.83%	**80.35%**	89.95%	95.50%	**97.35%**
	Recall	72.18%	25.25%	35.50%	70.56%	26.75%	47.09%	89.95%	24.42%	56.59%
Campus, Norm	Precision	94.20%	**97.02%**	94.61%	93.50%	**96.32%**	96.02%	95.68%	**96.26%**	95.80%
	Recall	94.20%	25.41%	75.06%	93.50%	23.31%	62.96%	95.68%	38.05%	83.35%
Campus, Adv	Precision	67.39%	66.67%	**91.42%**	45.42%	48.41%	**74.56%**	94.77%	**96.17%**	95.83%
	Recall	67.39%	14.92%	34.43%	45.42%	13.40%	17.39%	94.77%	35.43%	60.75%
Nordland	Precision	3.4%	3.4%	**5.5%**	8.6%	7.0%	**13.6%**	-%	-%	-%
	Recall	3.41%	3.30%	2.46%	8.59%	6.8%	4.02%	-%	-%	-%
RobotCar	Precision	65.1%	63.2%	**75.5%**	77.0%	71.4%	**83.1%**	-%	-%	-%
	Recall	65.1%	43.1%	49.2%	77.0%	53.4%	67.9%	-%	-%	-%

Table 6 shows that in general our MLP verification method maintains a significantly higher recall compared to the SVM method with higher or comparable precision. This demonstrates that our MLP verification method is able to correctly identify a higher or comparable number of 'bad' VPR position estimates without rejecting as many of the 'good' VPR position estimates compared to the SVM method. In the table, the precision and recall is equivalent for the baseline performance because we assess the case where the baseline method accepts all VPR position estimates.

To provide some further context for these values, we plot precision-recall curves for the baseline systems and include where the respective verification methods fit on the plot for the new datasets, Nordland and Oxford RobotCar. For the precision-recall curves below the baseline performance uses VPR match rejection based on match distances. The curve is computed by varying the threshold required to either accept or reject a VPR position estimate.

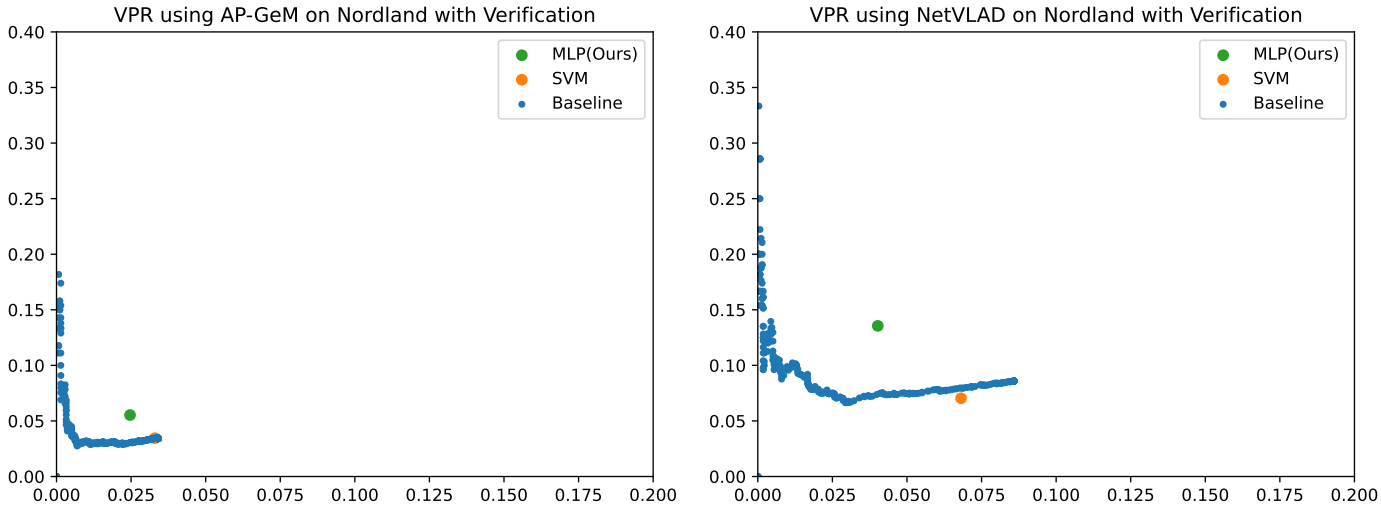

Figure 4: Precision and recall curves for VPR performance across the Nordland dataset. **a)** using AP-GeM as the baseline. **b)** using NetVLAD as the baseline.

Figure 4 demonstrates that even when trained on data from a completely different environment, our MLP verification method is able to maintain either a much higher precision at the same recall, ***or***, a much higher recall at the same precision. Compared to the SVM method which generally falls along the performance curve of the baseline system. It is important to note that the SVM method still has a potential advantage over the baseline system as there is no additional requirement to select a threshold specifically for this dataset.

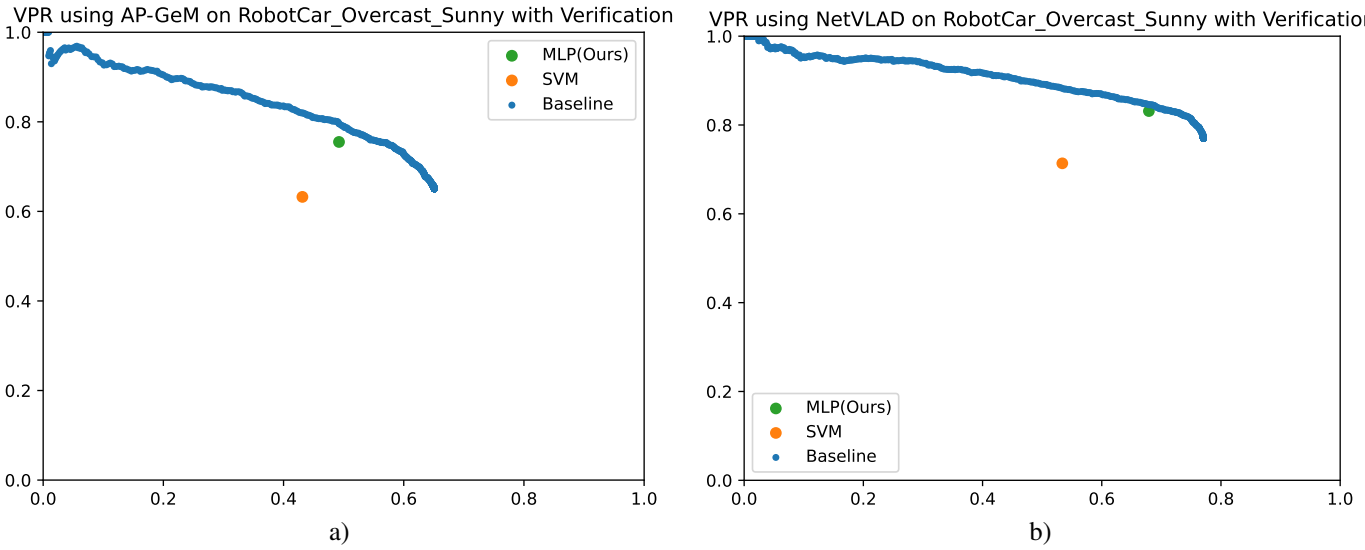

Figure 5: Precision and recall curves for VPR performance across the Oxford RobotCar dataset. **a)** using AP-GeM as the baseline. **b)** using NetVLAD as the baseline.

Figure 5 shows that VPR performance using the verification methods is not as strong compared to the baseline method for the Oxford RobotCar dataset. However, our MLP method clearly outperforms the previous SVM method and is, therefore, more robust when operating in environments that differ from the training dataset. Our MLP method performs comparably to operating points along the baseline curve, however, without the need to manually select a threshold for VPR match rejection, still offers an advantage over the baseline method.

References

[1] S. Schubert, P. Neubert, S. Garg, M. Milford, and T. Fischer, "Visual place recognition: A tutorial," *IEEE Robotics & Automation Magazine*, pp. 2–16, 2023.