You’re your own best teacher: A Self-Supervised Learning Approach For Expressive Representations

Johan Vik Mathisen

June 4, 2024

TODO: The spacing and figure placement is under development.

TODO: Have not yet polished formulations and language.

In this thesis, we focus on two main objectives, which relates back to the research questions. Firstly, in Stage 1, we aim to determine whether NC-VQVAE can learn more expressive representations compared to VQVAE. Specifically, we investigate whether NC-VQVAE can achieve reconstruction performance on par with VQVAE while simultaneously enhancing downstream classification. In Stage 2, our interest lies in examining the impact of NC-VQVAE on prior learning and time series generation.

Our evaluation process begins with assessing the tokenization models, focusing on their reconstruction capability and performance in downstream classification tasks. Subsequently, we then evaluate the performance of the generative models using metrics such as IS (Inception Score), FID (Fréchet Inception Distance), and CAS (Classification Accuracy Score). Additionally, visual inspections are conducted to provide further insights into the models' performance.

0.1 Stage 1

In this section we present the results of the tokenization model, in particular the reconstruction loss and the downstream classification accuracy. We address research question 1 and 2, if the proposed NC-VQVAE is able to reconstruct on par with the naive VQVAE and if the learned latent representations are more expressive, in the sense that they simultaneously improve the downstream classification accuracy. We see that some configuration of NC-VQVAE is the top performer on the majority of datasets for both metrics, and provides significant increase in probe accuracy.

0.1.1 Reconstruction

We present top 1 and mean reconstruction loss across the four runs in table **??** and table **??** respectively.

Mean validation reconstruction error

Dataset	Baseline	SSL Method					
	Regular	Barlow Twins			VIbCReg		
	None	Warp	Slice	Gauss	Warp	Slice	Gauss
FordA	0.217	0.127	0.134	**0.108**	0.173	0.169	0.203
ElectricDevices	**0.041**	0.067	0.044	0.049	0.105	0.042	0.049
StarLightCurves	**0.032**	0.042	0.069	0.071	0.052	0.050	0.068
Wafer	0.044	0.037	0.048	0.049	**0.035**	0.042	0.039
ECG5000	**0.048**	0.083	0.170	0.104	0.093	0.205	0.064
TwoPatterns	0.197	0.201	**0.184**	0.230	0.214	0.186	0.207
UWaveGestureLibraryAll	0.190	**0.172**	0.190	0.245	0.189	0.178	0.237
FordB	0.150	0.115	0.122	0.123	**0.114**	0.121	0.142
ShapesAll	**0.045**	0.056	0.066	0.102	0.064	0.069	0.073
SonyAIBORobotSurface1	0.402	0.509	0.494	0.491	**0.360**	0.363	0.418
SonyAIBORobotSurface2	0.623	0.622	0.618	0.640	0.487	**0.454**	0.589
Symbols	0.110	0.143	0.134	0.173	0.078	**0.067**	0.105
Mallat	0.066	0.081	0.091	0.096	0.066	0.067	**0.060**

Table 1: Mean validation reconstruction error across all 13 datasets. Results are averaged over four runs.

Top 1 validation reconstruction error

Dataset	Baseline	SSL Method					
	Regular	Barlow Twins			VIbCReg		
	None	Warp	Slice	Gauss	Warp	Slice	Gauss
FordA	0.158	0.108	0.111	**0.087**	0.130	0.134	0.113
ElectricDevices	0.036	0.060	0.034	0.043	0.092	**0.031**	0.045
StarLightCurves	**0.026**	0.037	0.057	0.055	0.043	0.048	0.065
Wafer	0.038	0.031	0.045	0.043	**0.027**	0.031	0.038
ECG5000	**0.044**	0.069	0.156	0.084	0.080	0.181	0.056
TwoPatterns	0.181	0.184	**0.169**	0.208	0.200	0.172	0.185
UWaveGestureLibraryAll	0.159	**0.145**	0.167	0.201	0.155	0.169	0.233
FordB	0.117	0.094	0.090	0.103	**0.082**	0.094	0.102
ShapesAll	**0.035**	0.043	0.046	0.092	0.061	0.063	0.067
SonyAIBORobotSurface1	0.381	0.473	0.472	0.465	0.329	**0.328**	0.408
SonyAIBORobotSurface2	0.513	0.577	0.536	0.588	0.444	**0.414**	0.470
Symbols	0.088	0.111	0.122	0.150	0.062	**0.059**	0.090
Mallat	0.061	0.075	0.076	0.088	0.059	0.059	**0.057**

Table 2: Top 1 validation reconstruction error across all 13 datasets. Lowest value of the four runs for each model is selected.

We observe that NC-VQVAE reconstructs on par with the baseline model, and that some configuration outperforms the naive VQVAE on mean reconstruction loss for 9 out of 13 datasets.

In figure **??** observe that the difference in reconstruction loss is small for most datasets, both across SSL methods and augmentations. Nevertheless we observe that VIbCReg generally performs slightly better than Barlow Twins, except for FordA. Additionally the use of gaussian augmentation introduces less of a regularizing effect compared to the two other sets of augmentations, with the exception of Slice and Shuffle on ECG5000. These results show that the introduction of a non contrastive loss does not hurt the reconstruction capabilities, compared to naive VQVAE.

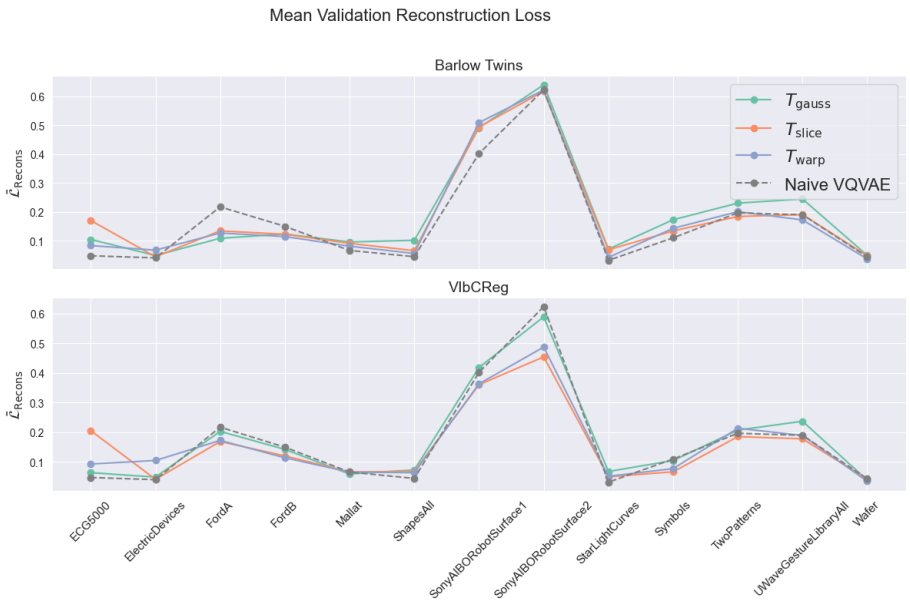

Figure 1: Mean validation reconstruction loss for the two models, compared to naive VQVAE

Regarding the effect of by the reconstruction loss of augmented branch on validation reconstruction, a small initial experiment was conducted. The results show that the validation reconstruction loss was quite robust to the particular value of augmentation reconstruction weight, indicating a minor role played, as seen in Figure **??** and **??**.

By investigating the development of the validation reconstruction loss during training, we have observed that right configuration for NC-VQVAE can act as a regularizer. In Figure **??** we see the development on FordA.

0.1.2 Classification

We present the mean and max downstream classification accuracy in table **??** and **??** respectively.

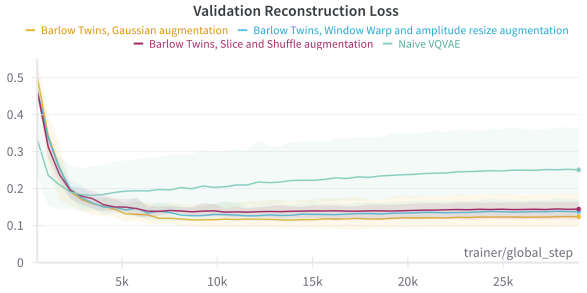

Figure 2: Development of the validation reconstruction loss for Barlow Twins and naive VQVAE on FordA during training. Averaged across all four runs.

Mean linear probe accuracy

Dataset	Baseline		SSL Method											
	Regular		Barlow Twins						VIbCReg					
	None		Warp		Slice		Gauss		Warp		Slice		Gauss	
	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
FordA	0.70	0.74	0.83	0.84	**0.91**	**0.89**	0.80	0.83	0.80	0.74	0.87	0.86	0.76	0.78
ElectricDevices	0.35	0.41	0.35	**0.44**	0.38	0.41	**0.40**	0.42	0.33	0.38	0.36	0.39	0.39	0.43
StarLightCurves	0.87	0.89	0.93	0.93	**0.94**	**0.94**	0.88	0.88	0.92	**0.94**	0.91	0.93	0.89	0.89
Wafer	0.93	0.89	0.96	**0.94**	0.96	**0.94**	0.96	0.93	**0.97**	0.94	0.96	0.92	**0.97**	0.92
ECG5000	0.80	0.83	0.85	0.81	**0.88**	0.84	0.86	**0.84**	0.86	0.82	**0.88**	**0.84**	0.84	0.82
TwoPatterns	0.34	0.53	**0.69**	**0.91**	0.66	0.82	0.47	0.71	0.64	0.90	0.68	0.80	0.55	0.72
UWaveGestureLibraryAll	0.31	0.40	**0.62**	0.70	0.56	0.63	0.40	0.54	**0.62**	**0.73**	0.55	0.66	0.44	0.55
FordB	0.58	0.60	0.64	0.67	0.74	0.76	0.64	0.68	0.63	0.64	0.70	0.70	0.61	0.64
ShapesAll	0.29	0.30	0.49	0.55	0.53	**0.60**	0.40	0.48	0.48	0.56	**0.54**	**0.60**	0.40	0.46
SonyAIBORobotSurface1	0.56	0.68	0.54	0.70	**0.61**	**0.74**	0.53	0.70	0.48	**0.74**	0.58	0.71	0.54	0.69
SonyAIBORobotSurface2	**0.81**	**0.86**	0.77	0.79	0.80	0.80	0.80	0.81	0.77	0.85	0.80	0.85	0.80	0.85
Symbols	0.50	0.60	**0.59**	0.60	0.50	**0.66**	**0.59**	**0.66**	0.45	0.61	0.42	0.62	0.43	0.63
Mallat	0.63	0.77	0.72	0.81	0.76	0.83	0.68	0.78	**0.79**	**0.87**	0.77	0.85	0.69	**0.86**

Table 3: Summary of mean linear probe accuracy by SSL Method and Augmentation. Average across 4 seeds. Best result for KNN and SVM are highlighted in bold.

Top 1 linear probe accuracy

Dataset	Baseline		SSL Method											
	Regular		Barlow Twins						VIbCReg					
	None		Warp		Slice		Gauss		Warp		Slice		Gauss	
	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN	SVM
FordA	0.75	0.78	0.84	0.88	**0.93**	**0.92**	0.85	0.87	0.81	0.77	0.88	0.90	0.86	0.85
ElectricDevices	0.35	0.43	0.36	0.45	0.39	0.43	**0.45**	**0.46**	0.34	0.42	0.39	0.42	0.42	0.45
StarLightCurves	0.89	0.91	0.94	0.95	**0.96**	**0.96**	0.90	0.91	0.95	0.95	0.93	0.95	0.90	0.90
Wafer	0.94	0.89	**0.97**	**0.95**	**0.97**	**0.95**	**0.97**	0.93	**0.97**	**0.95**	**0.97**	**0.95**	**0.97**	0.94
ECG5000	0.83	0.84	0.88	0.86	**0.90**	**0.88**	**0.90**	**0.88**	0.88	0.85	0.89	0.86	0.86	0.85
TwoPatterns	0.37	0.62	**0.75**	**0.96**	0.68	0.85	0.55	0.75	0.70	0.92	0.71	0.81	0.63	0.76
UWaveGestureLibraryAll	0.34	0.43	**0.67**	0.74	0.60	0.67	0.43	0.54	**0.67**	**0.76**	0.58	0.67	0.48	0.58
FordB	0.60	0.63	0.67	0.71	**0.76**	**0.80**	0.69	0.74	0.67	0.65	0.74	0.77	0.63	0.68
ShapesAll	0.33	0.34	0.53	0.59	**0.59**	**0.65**	0.44	0.50	0.50	0.56	0.57	0.63	0.44	0.48
SonyAIBORobotSurface1	0.67	**0.80**	0.61	0.77	0.76	**0.80**	0.60	0.74	0.51	0.79	0.63	0.75	0.63	0.75
SonyAIBORobotSurface2	**0.84**	**0.89**	0.80	0.86	0.82	0.84	0.83	0.82	0.81	0.88	0.81	0.88	0.83	0.87
Symbols	0.56	0.66	0.65	0.69	0.55	0.73	**0.64**	**0.71**	0.51	0.65	0.45	0.67	0.46	0.69
Mallat	0.54	0.88	0.57	0.87	0.74	0.89	0.66	0.80	0.74	**0.92**	0.72	0.88	0.62	**0.90**

Table 4: Summary of max linear probe accuracy by SSL Method and Augmentation. Maximum value across 4 seeds. Best result for KNN and SVM are highlighted in bold.

We observe a significant improvement in probe accuracy with NC-VQVAE, compared to naive VQVAE. Some configuration is best on 12 out of 13 datasets, while the one where our model falls short, the difference is one percent for both SVM and KNN. The largest differences are seen on FordA, FordB, Mallat, ShapesAll, TwoPatterns and UWaveGestureLibraryAll.

In Figure **??** we observe that even though the specific choice of augmentation has a large impact, on the majority of datasets, all choices result in significantly improved probe accuracy. We note that both SSL methods produce similar probe accuracies for a given augmentation, highlighting the importance of selecting suitable augmentations.

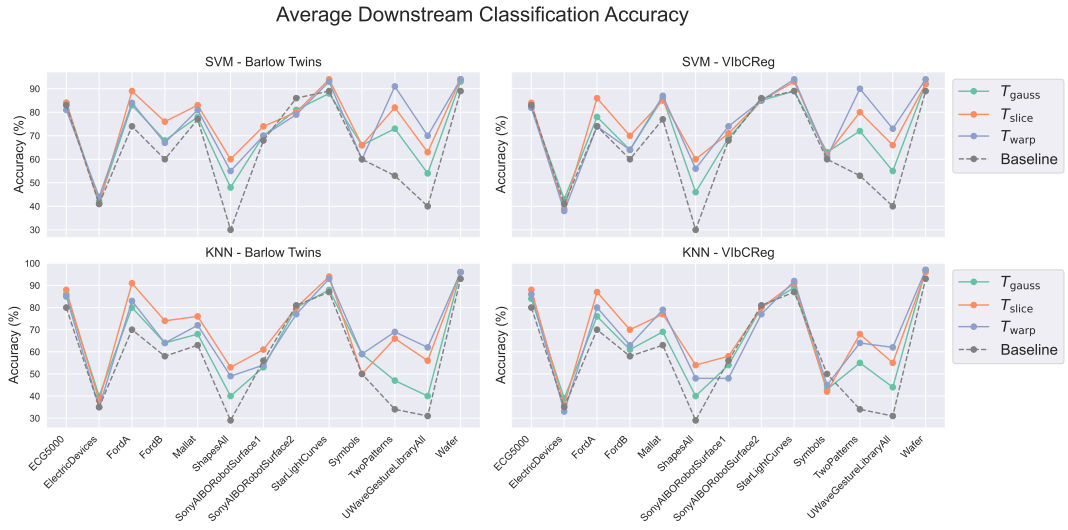

Figure 3: Mean probe accuracies.

We observe that Slice and Shuffle, and Window Warp and Amplitude Resize result in the most dramatic increases in accuracy, while Gaussian noise consistently results in less drastic improvement. We hypothesize that, since Slice and Warp often result in augmented views that deviate quite a lot from the original view, the SSL loss pushes the representations in different directions, which again might result in a better utilization of the latent space. In Figure **??** and **??** we see the effect of NC-VQVAE on the discrete latent representations of FordA and TwoPatterns. From these visualizations it is evident that representations learned using NC-VQVAE are more structured than those of the naive VQVAE. Similar samples, typically with the same label, are clustered closer together in latent space. This further indicate that the SSL loss introduces information regarding shape and semantics into the latent representations.

To summarize the results from stage 1, NC-VQVAE is able to reconstruct on par

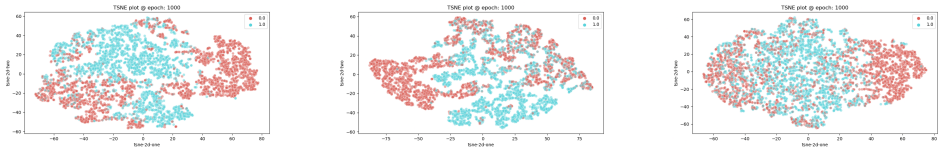

Figure 4: TSNE plots of FordA. Barlow (left) and VIbCReg (center) with Slice and Shuffle, naive VQVAE (right). Best performing model in terms of KNN accuracy is chosen.

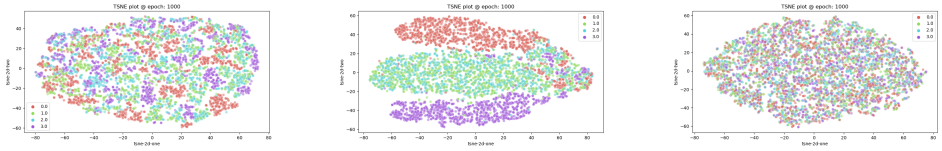

Figure 5: TSNE plot of discrete latent representations from VIbCReg with Slice and Shuffle (left), Barlow Twins with Window Warp and Amplitude Resize (center) and naive VQVAE (right). Dataset is TwoPatterns. The latent space is significantly more structured with NC-VQVAE.

with naive VQVAE, and in some cases improve the reconstruction loss, while significantly improving the probe accuracy for most datasets. To address research question 1, we conclude that the representations learned using NC-VQVAE are more expressive compared to the naive VQVAE. As the representations separates classes more effectively, they could encode more class specific information. Since we are not sacrificing reconstruction quality, this could be beneficial for prior learning. To address research question 2, the specific augmentation used plays a prominent role in the results, and we observe that warp and slice typically leads to better performance than gaussian, especially in terms of probe accuracy. We additionally note that there are large differences across datasets, which support the the hypothesis that the optimal choice of augmentations is highly dataset dependent.

0.1.3 Losses

We investigate some trends in the development of different loss terms in this section. To summarize, using VIbCReg results in more easily minimizable losses compared to Barlow Twins, and the Gaussian augmentation results in significantly easier minimization of the SSL loss as well as reducing the VQ loss.

In figure **??** we observe the typical pattern of the SSL loss during training.

We see that the Gaussian augmentation results in a SSL loss which is easier to minimize, which might be attributed to the fact that it affects the samples in a more predictable way. We too see that the VIbCReg loss decreases more rapidly than the Barlow Twins loss. Both observations are seen across datasets.

Previously in figure **??** we have seen that the gaussian augmentation often

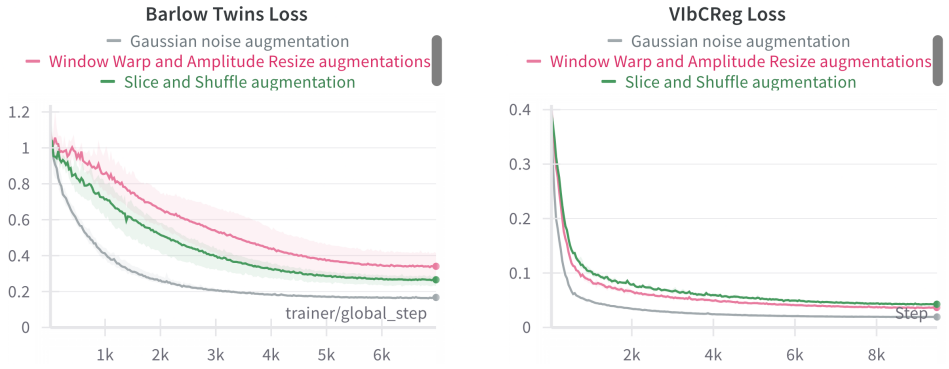

Figure 6: SSL loss during training on UWaveGestureLibraryAll. Averaged across four runs.

resulted in lower probe accuracy than the other two. In figure **??** we see that, on the datasets with a significant increase in probe accuracy, the augmentations that result in a more challenging SSL loss typically has higher downstream classification accuracy. We also see that for a specific augmentation, the pattern is rather flat, indicating that the particular augmentation plays the most prominent role in probe accuracy. The SSL loss for a specific augmentation varies very little compared to probe accuracy.

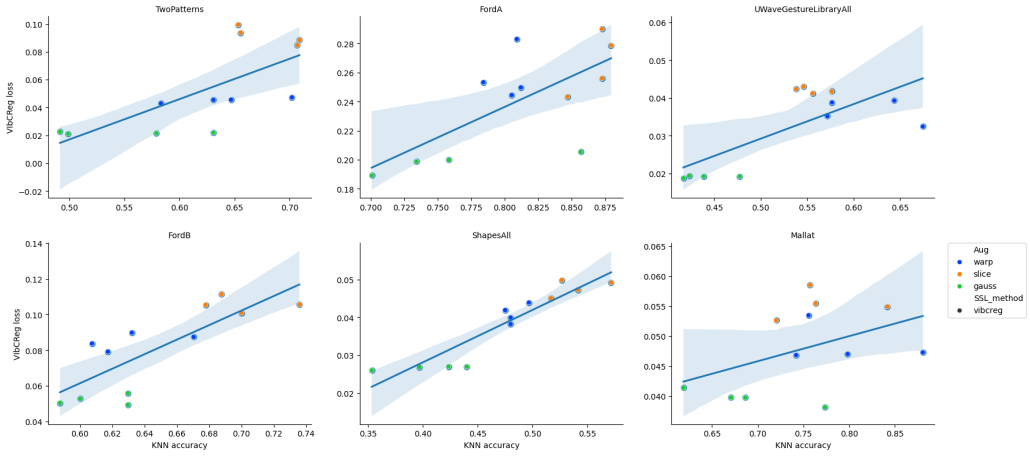

Figure 7: KNN accuracy plotted against VIbCReg loss. Each point correspond to a single run of the model. Similar tendency is shown for Barlow Twins.

The training reconstruction losses are heavily minimized, both across models and augmentations. The only consistent noticeable difference is the augmented reconstruction loss, where models using Slice and Shuffle have a slightly higher loss. The differences in VQ loss for the different models is mainly due to the code-

book, where we too observe that VIbCReg minimizes more effectively than Barlow Twins, and again that gaussian augmentations results in the hardest minimization followed by Window Warp and then Slice and Shuffle.

Both VIbCReg and Barlow Twins with Gaussian augmentation routinely perform on par with naive VQVAE in terms of VQ loss during training. The minimization of the codebook loss indicates that the encoder is properly aligned with the discrete latent codes. We hypothesize that when the SSL loss is not properly minimized, the encoder must adjust its weights more throughout training which keeps the encoder outputs and the discrete codes from aligning completely.

0.2 Stage 2

The generative quality of or models are evaluated according to FID, IS and CAS. All results will be presented in this section. For datasets with very few samples, or very few per class, the generative scores must be taken with a grain of salt. Both the classifier, and the evaluation metrics is dependent on a certain number of samples to be considered reliable. We rather look more closely on the visual inspection for these.

0.2.1 FID and IS

We present the top 1 and mean score across the four runs for both FID and IS in table **??** and **??**. From the tables we see that our model produces better IS score for 12 out of 13 datates, and better FID for 10 out of 13.

Top 1 FID and IS

Dataset	**Baseline**		**SSL Method**											
	Regular		Barlow Twins						VIbCReg					
	None		Warp		Slice		Gauss		Warp		Slice		Gauss	
	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS ↑	FID↓	IS↑
FordA	2.59	1.30	1.93	**1.51**	2.13	1.48	1.80	**1.51**	2.83	1.38	2.50	1.43	**1.66**	1.41
ElectricDevices	12.05	3.97	11.82	4.20	**8.91**	4.07	9.89	3.86	12.38	**4.23**	11.08	3.94	13.96	3.71
StarLightCurves	**0.74**	1.99	0.89	**2.43**	1.50	2.36	0.75	2.39	0.92	2.39	0.85	**2.40**	0.79	2.26
Wafer	5.27	**1.39**	3.31	1.29	3.82	**1.26**	2.77	1.35	3.33	1.29	3.60	1.30	**2.52**	1.34
ECG5000	1.56	2.01	2.43	2.02	2.27	2.00	2.15	2.02	2.15	**2.03**	2.21	2.00	**1.52**	2.02
TwoPatterns	3.63	2.47	3.59	2.65	2.74	2.73	**2.24**	2.70	3.45	2.64	2.90	2.70	**2.19**	**2.77**
UWaveGestureLibraryAll	8.16	2.24	6.45	2.94	**6.26**	**3.13**	7.31	2.79	6.52	2.99	6.33	3.06	7.09	2.79
FordB	2.92	1.52	2.10	1.52	2.44	1.61	1.93	**1.67**	1.76	1.65	2.12	1.64	**1.66**	1.52
ShapesAll	**21.35**	4.32	35.89	**5.22**	29.61	5.16	27.91	4.83	30.03	4.95	31.59	4.92	27.20	4.94
SonyAIBORobotSurface1	18.21	1.27	26.20	1.32	28.90	1.28	21.63	1.32	21.98	1.36	25.20	1.38	**15.73**	**1.55**
SonyAIBORobotSurface2	3.85	1.69	2.50	1.82	3.34	1.79	**0.82**	1.82	2.61	1.81	2.75	1.83	1.24	**1.84**
Symbols	8.50	2.43	5.86	3.20	7.39	2.82	**4.25**	**3.50**	6.78	3.39	7.21	3.23	8.21	3.30
Mallat	**1.31**	3.41	2.01	3.67	2.24	3.72	1.85	3.66	1.87	3.34	2.30	3.05	**1.31**	**3.92**

Table 5: Summary of FID and IS scores by SSL Method and Augmentation. Best achieved results are highlighted in bold

In figure **??** we get a better overview of the results, and observe that both Barlow Twins and VIbCReg produces better samples than the naive VQVAE in

Mean FID and IS

Dataset	Baseline		SSL Method											
	Regular		Barlow Twins						VIbCReg					
	None		Warp		Slice		Gauss		Warp		Slice		Gauss	
	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS ↑	FID↓	IS↑
FordA	5.15	1.16	2.59	1.41	2.36	**1.45**	**2.28**	**1.45**	3.01	1.34	2.90	1.41	3.73	1.29
ElectricDevices	13.48	3.75	16.51	3.95	**10.20**	3.93	11.54	3.75	13.99	**4.17**	11.82	3.85	15.20	3.55
StarLightCurves	**1.01**	1.93	1.29	2.35	1.91	2.32	1.08	2.25	1.07	2.35	1.19	**2.36**	1.05	2.22
Wafer	5.72	**1.33**	3.70	1.25	4.20	1.24	2.85	1.31	3.67	1.26	3.86	1.26	**2.84**	1.31
ECG5000	**1.62**	1.94	2.61	**2.00**	2.56	1.98	2.47	**2.00**	2.60	1.99	2.39	**2.00**	1.76	1.99
TwoPatterns	4.04	2.41	4.00	2.54	2.96	2.66	**2.44**	**2.67**	4.05	2.56	3.15	2.66	2.62	**2.67**
UWaveGestureLibraryAll	8.48	2.13	6.77	2.86	6.64	2.96	7.35	2.73	6.80	2.91	**6.49**	**2.99**	7.34	2.72
FordB	4.05	1.28	2.66	1.48	3.49	1.50	2.88	**1.52**	**2.49**	1.48	3.07	1.51	3.04	1.31
ShapesAll	**27.64**	4.22	38.22	5.07	32.54	**5.04**	32.25	4.56	36.59	4.72	35.79	4.76	31.56	4.71
SonyAIBORobotSurface1	23.71	1.20	30.65	1.22	31.97	1.21	25.29	1.28	26.11	1.32	28.20	1.32	**18.61**	**1.44**
SonyAIBORobotSurface2	5.42	1.62	3.35	1.77	4.41	1.74	**1.78**	**1.81**	4.43	1.74	3.32	1.79	2.36	1.79
Symbols	13.62	1.99	9.78	2.92	9.78	2.67	8.61	3.14	8.84	3.20	9.74	3.03	**8.58**	**3.24**
Mallat	2.09	3.01	2.54	3.29	3.68	2.94	2.12	3.53	2.11	3.18	2.40	2.96	**1.65**	**3.72**

Table 6: Summary of FID and IS scores by SSL Method and Augmentation. Best mean achieved FID and IS are highlighted in bold

terms of FID and IS. Additionally we see that the use of gaussian augmentation results in the largest improvements for most datasets. The high IS scores indicate that NC-VQVAE captures the conditional distributions better than naive VQVAE in many datasets. This will be explored further in section **??**. The improved FID scores indicates that the synthetic samples more closely resemble the test data. The moderate decrease in FID, compared to the increase in IS, could indicate that the generated samples does not generalize too well to the test data. The discrete latent representations from NC-VQVAE provides more information regarding the classes, as we saw from the improved downstream classification accuracy in stage 1. This additional class specific information seems to assist the prior learning in capturing class conditional distributions.

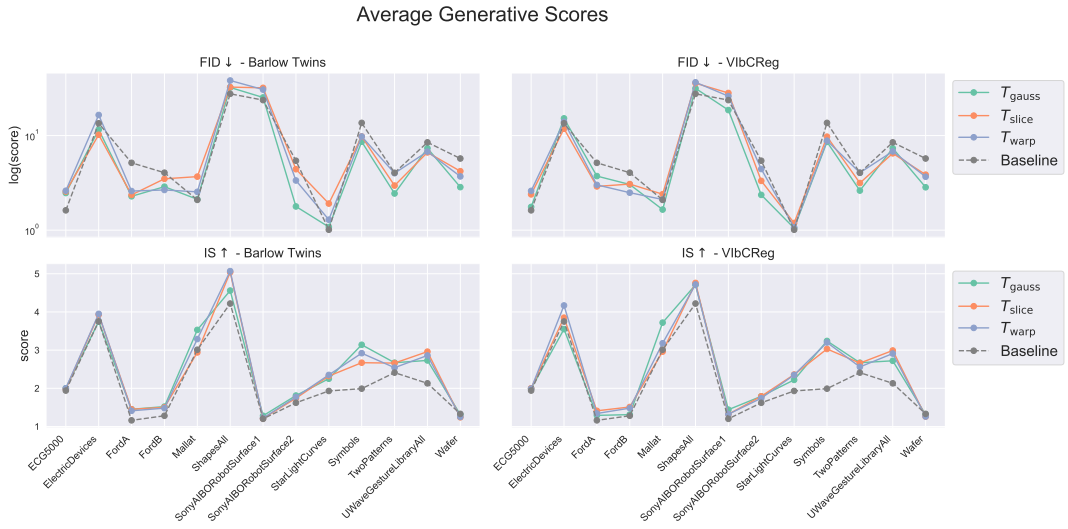

Figure 8: Mean FID and IS scores for Barlow Twins and VIbCReg VQVAE. FID is plotted on a log scale because of the large difference in values across datasets.

It is worth mentioning that the FID and IS score is calculated using the SupervisedFCN, which is also trained on the UCR Archive. Thus, the FID and IS scores could have a bias toward samples that mimic the training data.

0.2.2 CAS

We present the mean CAS for all models across datasets in table **??**.

Mean CAS

Dataset	**Baseline**	**SSL Method**					
	Regular	Barlow Twins			VIbCReg		
	None	Warp	Slice	Gauss	Warp	Slice	Gauss
FordA	0.864	0.884	**0.902**	0.878	0.864	0.895	0.870
ElectricDevices	0.614	0.588	0.607	0.599	**0.618**	0.610	0.594
StarLightCurves	0.960	0.953	0.955	**0.965**	0.962	0.954	0.964
Wafer	0.976	0.977	0.978	0.968	0.979	0.976	**0.984**
ECG5000	0.866	0.881	0.863	0.880	0.877	0.892	**0.910**
TwoPatterns	0.808	0.770	0.788	**0.847**	0.715	0.781	0.846
UWaveGestureLibraryAll	0.333	0.300	0.367	0.313	0.360	**0.401**	0.383
FordB	0.725	0.748	**0.756**	0.741	0.750	0.738	0.750
ShapesAll	0.361	0.344	0.329	**0.420**	0.379	0.367	0.404
SonyAIBORobotSurface1	0.975	0.933	0.957	0.979	0.982	0.976	**0.985**
SonyAIBORobotSurface2	0.929	0.956	0.951	0.969	0.960	**0.970**	0.964
Symbols	0.956	0.929	0.930	0.930	0.969	**0.974**	0.963
Mallat	0.471	0.642	0.563	0.661	0.827	0.876	**0.908**

Table 7: Mean CAS score across datasets. Results averaged across four runs.

We see that some configuration of NC-VQVAE outperforms the naive VQVAE on all datasets, as well as VIbCReg with gaussian augmentation outperforming the baseline on 12 out of 13, where the one dataset where it falls short its within one percent. In general we observe that NC-VQVAE performs well across all datasets, and in particular with gaussian augmentation. The dataset where we see the most dramatic increase is Mallat, with an improvement of 0.437. This particular case will be investigated in section **??**.

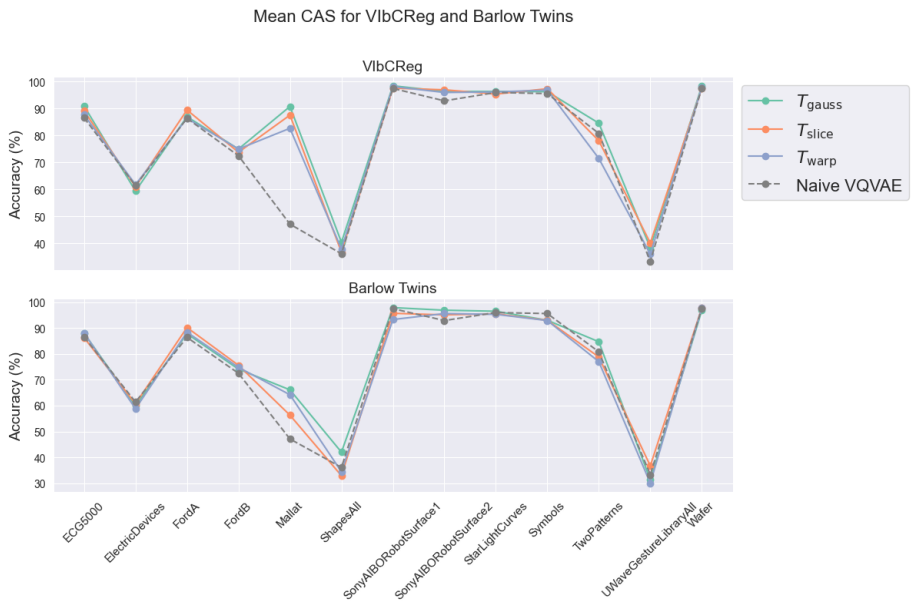

Figure 9: Mean CAS across all datasets.

0.2.3 Prior loss

TODO: Under construction

Naive VQVAE outperforms NC-VQVAE in terms of validation prior loss across datasets. There is the occasional dataset where a model with gaussian augmentation performs equally well. The minimization of the validation prior loss does therefore not correspond to improved FID and IS in general.

0.2.4 The influence of stage 1 on stage 2

We further attempt to address research question 3 and 4, how expressive representations influence synthetic sample quality and the role of augmentations. For this we inspect relationship between probe accuracy and FID and IS. We consider only the datasets where NC-VQVAE provided a prominet increase in probe accuracy, which are FordA, FordB, Mallat, ShapesAll, TwoPatterns and UWaveGestureLibraryAll.

In Figure **??** and **??**, we show scatterplots of KNN against FID and IS with the least square regression line. The corresponding plots with SVM accuracy show similar trends. From Figure **??** we see a trend, with higher probe accuracy correlating with higher IS. Upon closer inspection, we see a pattern of the prominent effect of augmentations. For each specific augmentation, the correlation between KNN and IS is close to 0. It seems to be that augmentations that result in higher KNN accuracy tends to high IS scores, though the pattern is not consistent across all datasets. From Figure **??** we see that the specific augmentation is a better indicator of FID score than the KNN accuracy. Both figures too gives an indication of the model performance sensitivity to initialization, especially in terms of probe accuracy.

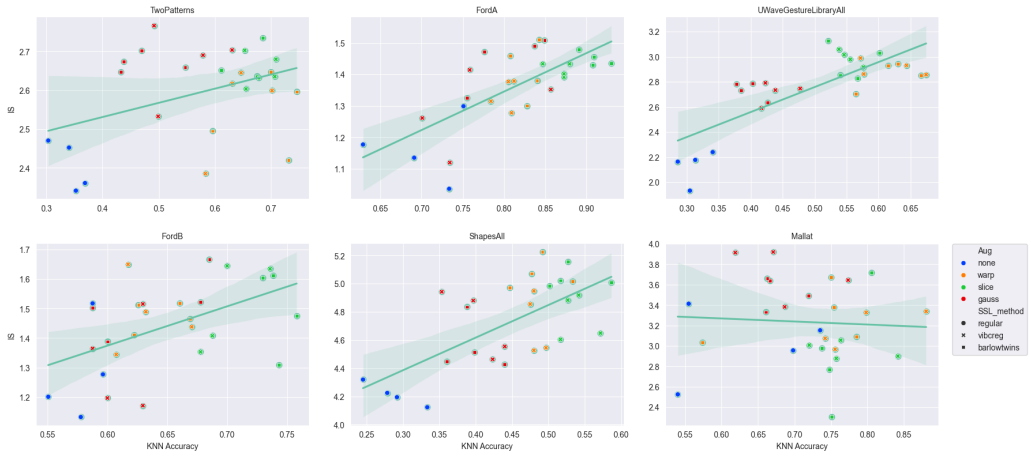

Figure 10: KNN plotted against Inception Score on the subset of datasets with significant improvement in probe accuracy.

0.2.5 Visual inspection

In the following we present generated samples from naive VQVAE and NC-VQVAE for some selected dataset. The ground truth, both test and train, are additionally included in order to better make sense of the IS, FID and CAS scores.

Some datasets, such as FordA and B, are poorly suited for this type of visual inspection, as illustrated previously in Figure **??**. As a result, the selection of datasets is primarily based on how well they lend themselves to this type of presentation. For each figure in the following sections, 50 samples are generated from each model. For the ground truth, we plot a subset of 50 randomly selected samples, or the entire set if the dataset contain less than 50 samples. For the datasets with complex distribution, or many classes, it is very difficult to visually asses the unconditional distribution. Thus, we mainly provide class conditional samples. We

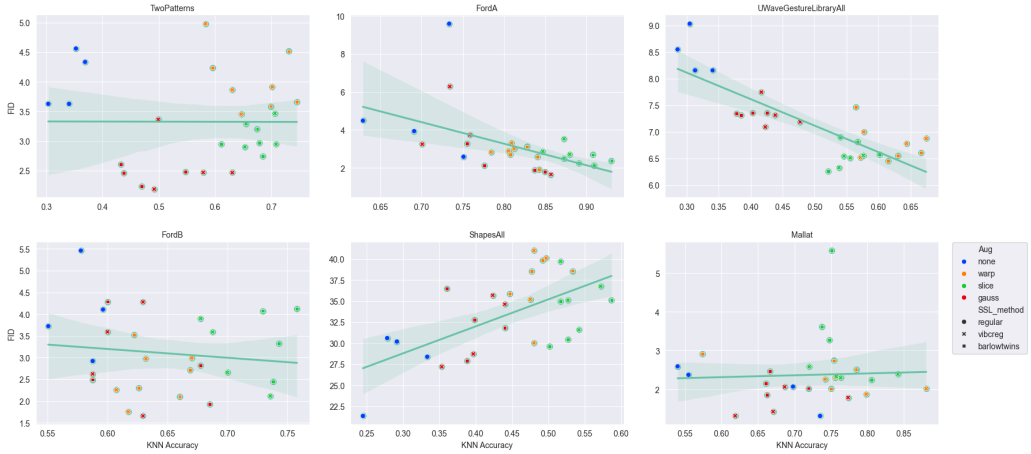

Figure 11: KNN plotted against Fréchet Inception Distance on the subset of datasets with significant improvement in probe accuracy.

surprisingly only observe minor differences in the generated samples form NC-VQVAE trained with different augmentations.

Typically naive VQVAE has more trouble with capturing the global consistency of the samples when samples are scarce and diverse, as seen on ShapesAll and Symbols. In contrast, our method will tend to overfit in these cases. The overfitting issue is most prominent in the class conditional distributions. This is likely because some classes only have 2-5 sample in certain datasets.

ECG5000

In Figure **??** we present generated samples from naive VQVAE and NC-VQVAE trained with Window Warp and Amplitude Resize augmentations on ECG5000.

We see some evidence that VIbCReg maintains more variability than Barlow Twins, while both has good mode coverage. In class 4, where the training data only consists of 2 samples, both Barlow Twins and VIbCReg catches the pattern, while producing some variation which resemble the training samples. The naive VQVAE samples does not capture this distribution, but when compared to the test data, it is more similar. This could explain the minor increase in CAS for NC-VQVAE. Looking at the unconditional sample, it is not evident why naive VQVAE performs better in terms of FID score than NC-VQVAE.

Mallat

Mallat is a simulated dataset, where the classes have very little variability and training and test distribution are almost indistinguishable, except for sample size.

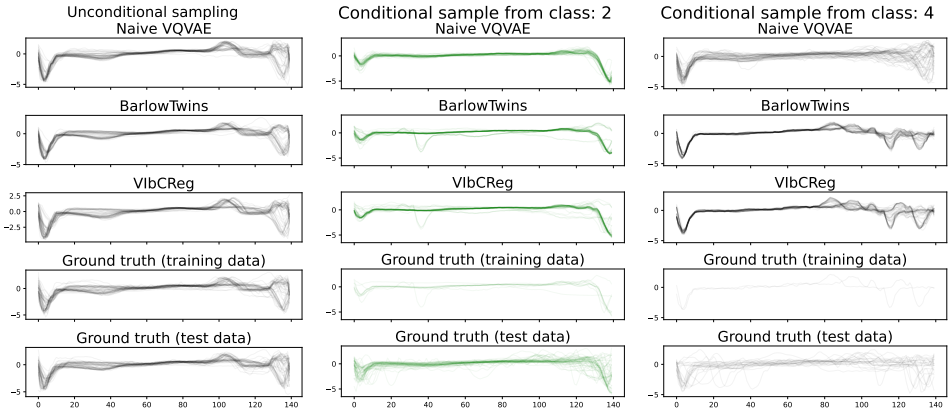

Figure 12: Dataset: ECG5000. Barlow and VIbCReg both trained with window warp and amplitude resize augmentations. 50 samples from each model.

We observe that VIbCReg is superior in capturing the variability, compared to Barlow Twins and naive VQVAE. This is most evident in the first 300 timesteps of class 5 in Figure **??**. Looking as class 7, we see Barlow Twins completely collapsing, essentially producing the same sample over and over.

These figures explain the significant increase in CAS seen in Figure **??**, particularly for VIbCReg. It too explains why VIbCReg with gaussian augmentation both increases IS and reduces FID.

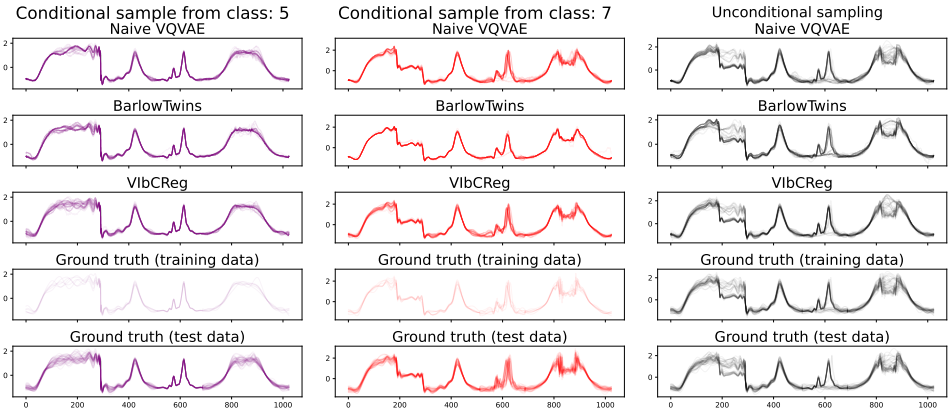

Figure 13: Class conditional distribution for some selected classes of Mallat, in addition to unconditional samples. Barlow and VIbCReg both trained with gaussian augmentation.

By inspecting the PCA plots of both data space and the discrete latent representations of samples from Mallat, compared to synthetic samples form VIbCReg and Barlow Twins in Figure **??**, we see a clear case of representation collapse for

Barlow Twins. We hypothesize that the variance term in VIbCReg assists in maintaining variability in the representations. Making it more effective in avoiding this type of collapse.

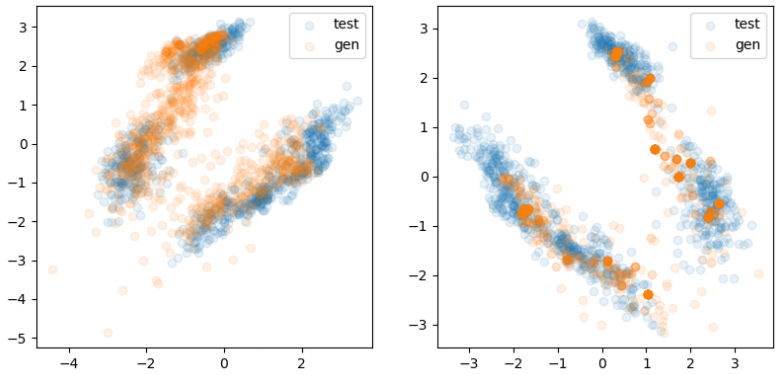

Figure 14: PCA of discrete latent representation from Mallat. Both VIbCReg (left) and Barlow Twins (right) are trained with gaussian augmentation.

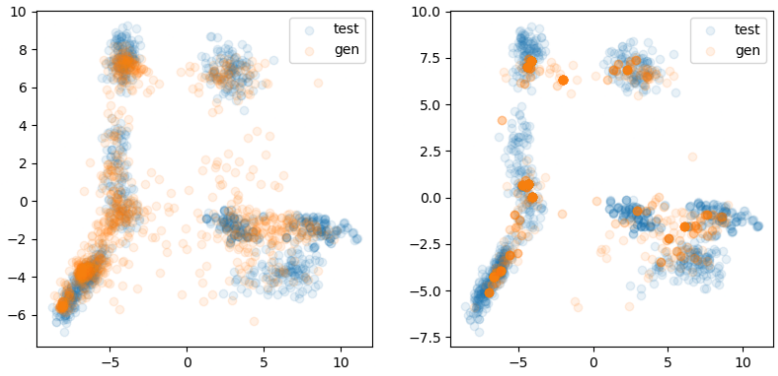

Figure 15: PCA of generated time series from Mallat. Both VIbCReg (left) and Barlow Twins (right) are trained with gaussian augmentation.

Symbols

The Symbols dataset consists of several distinct, but simple, patterns. The dataset is very small, where each class there is less than 5 training samples.

In Figure **??**, particularly the unconditional sample, we see that naive VQVAE does not capture the entire underlying distribution, some classes are not represented/not recognizable, while global consistency for the sinusoids are poor, particularly towards the end.

In class 3, we observe that both Barlow Twins and VIbCReg mimic the training data to a large degree,VIbCReg to a slightly higher degree than Barlow Twins. At first glance the naive VQVAE looks to produce the most desirable distribution, but upon closer inspection we see an excessive amount of noise and lack of consistency.

The IS on Symbols, for both VIbCReg and BarlowTwins, is substantially higher than naive VQVAE. While this is not surprising after inspecting the samples, it exposes an issue with the IS metric. It fails to take intraclass diversity into account, and is therefore oblivious to overfitting.

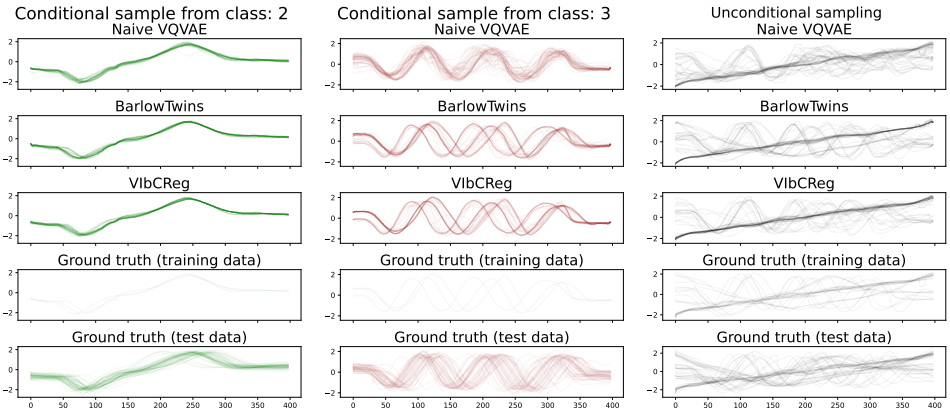

Figure 16: Class conditional distribution for some selected classes of Symbols. Barlow and VIbCReg both trained with gaussian augmentation.

ShapesALL

The dataset ShapesAll consists of 60 classes, with 10 samples within each class. Each class has distinct patterns, with varying complexity.

In Figure **??**, we observe clearly that naive VQVAE struggles with capturing the global consistency of the samples. We too observe that Barlow Twins mimic the training slightly more closely than VIbCReg, which provides some insight as to why Barlow Twins improves CAS by about 10 percent compared to naive VQVAE. Both Barlow Twins and VIbCReg improve IS, but fail to improve FID. As we have seen in the other datasets with few samples for a specific class, NC-VQVAE has a tendency to overfit.

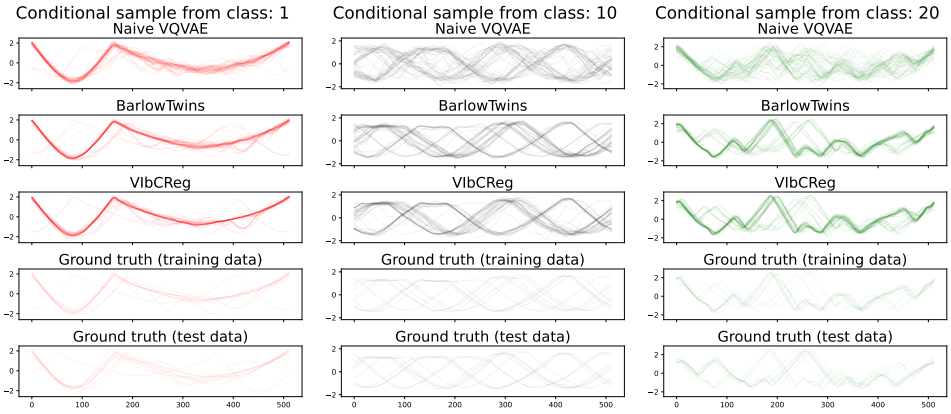

Figure 17: Class conditional distribution for some selected classes of ShapesAll. Barlow and VIbCReg both trained with gaussian augmentation.

UWaveGestureLibraryAll

The dataset UWaveGestureLibraryAll contains time series with distinct discontinuities and sharp changes in modularity. As noted in [**TimeVQVAE**], such datasets are challenging to model.

In Figure **??** a selected subset of classes are illustrated. We observe upon close inspection that VIbCReg maintains variability in the samples to a greater degree than Barlow Twins, as well at slightly better capturing the "dead spots" following the discontinuities.

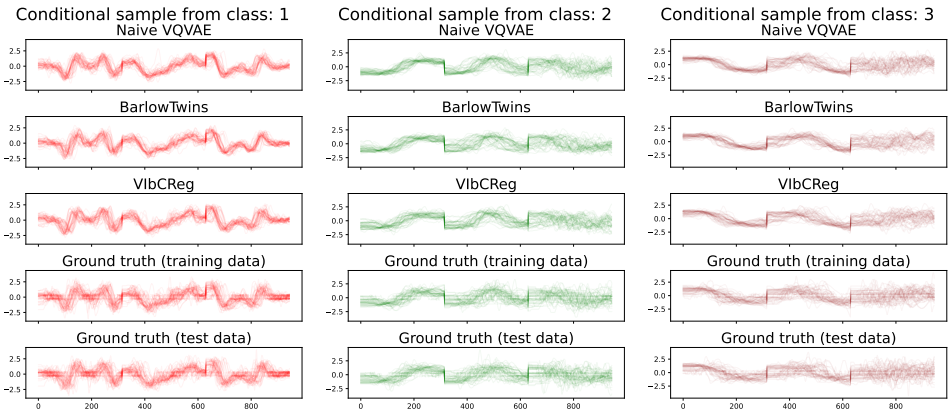

Figure 18: Class conditional distribution for some selected classes of UWaveGestureLibraryAll. Barlow and VIbCReg both trained with window warp and amplitude resize augmentations.

0.2.6 Augmentation Reconstruction Weight

TODO: Should this be included at all?

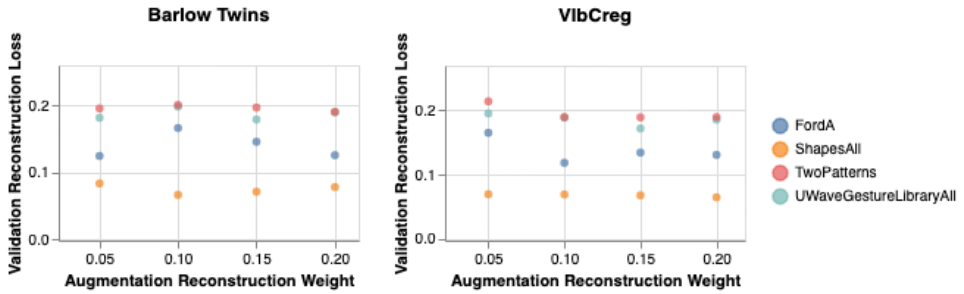

Figure 19: Augmentation: Window Warp and Amplitude Resize. Averaged across 2 runs. Trained for 250 epochs

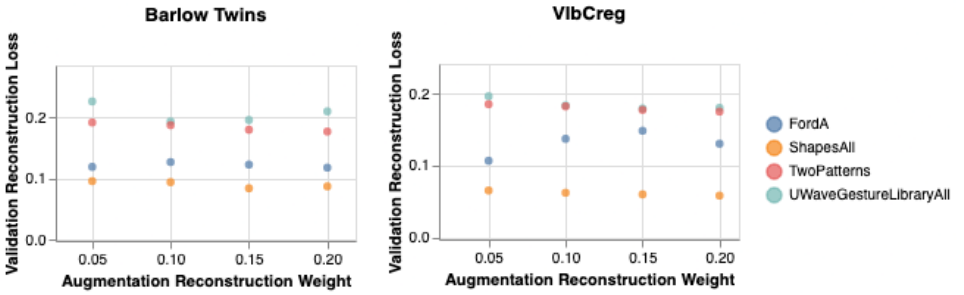

Figure 20: Augmentation: Slice and Shuffle. Averaged across 2 runs. Trained for 250 epochs

0.3 Discussion

0.3.1 Differences in Barlow Twins and VIbCReg

VIbCReg seems to keep the variability in the conditional distribution a bit better than Barlow Twins. We hypothesize that it is due to the variance term present in VIbCReg, and wonder whether increasing its weight might assist in producing more diverse samples.

0.3.2 Overfitting problem

We have observed on multiple occasions that when the sample size is small and the patterns in the data are simple NC-VQVAE has a tendency to overfit, and memorize the training data to a large degree. We have not employed any form of dropout or other regularization techniques when training on small datasets.

0.3.3 Temporal vs frequency influence of augmentations

We assessed and chose our different sets of augmentations based on their effect on the temporal representation on the time series. As we model the time frequency domain, future work should more thoroughly investigate the effects of augmentation on spectrograms. Additionally, all models considered compress the input only along temporal axis in the encoder, which in a sens puts more emphasis in the frequency components rather than the exact temporal structure.

The gaussian augmentation introduces HF component, though its influence in the time domain is visually clear, its affect on the spectrograms is minor, as the LF components typically has much larger magnitude. Window Warp and Amplitude Resize changes LF, often changing the exact location of dominant frequencies on the time axis, but not their order. The effect of Slice and Shuffle is variable, but has a tendency to create sharp discontinuities, which in many cases is a significant HF component.