Impact of View Generation Strategies on Self-Supervised Dual-View Representation Learning for 12-Lead ECG

Shady Ali[1*]

[1*]College of Science and Engineering, University of Minnesota, Minneapolis, MN, USA.

Corresponding author(s). E-mail(s): ahme0971@umn.edu;

Abstract

The scarcity of high-quality annotated data and significant inter-observer variability in clinical interpretation present major challenges for supervised ECG diagnosis. This work investigates Self-Supervised Learning (SSL) as an alternative by evaluating the impact of different view-generation strategies on a novel Dual-View Transformer architecture. We compared three distinct augmentation paradigms on the PTB-XL dataset: (1) Stochastic Augmentations, (2) Interleaved Subsampling, and (3) Non-Overlapping Temporal Partitioning. Our results demonstrate that Non-Overlapping Temporal Partitioning provides superior diagnostic features, achieving **72%** linear probing accuracy and better embedding space separation compared to noise-based methods. Furthermore, attention map analysis reveals that Random Lead Masking (RLM) induces the model to prioritize inter-lead spatial relationships, which effectively captures spatial context but potentially underutilizes temporal information. These findings highlight that the choice of view generation is a critical determinant of representation quality in ECG foundation models and suggest that segmentation-based strategies are essential for capturing clinically relevant features.

Keywords: ECG, Foundation Models, Augmentation, Contrastive Learning

1 Introduction

Cardiovascular disease is and has been a leading cause of death in the United States since as early as 1950 [1]. Heart disease can also have multiple risk factors, such as

The full source code is available at the project's GitHub repository.
All runs' logs are available in TensorBoard logs

smoking, high cholesterol, blood pressure, and glucose control, with almost half of the U.S. population having one of these factors [2].

Electrocardiogram (ECG) is a simple test to evaluate the heart's performance by recording the electrical patterns from the heart over a period of time. These ECG signals can then be used for diagnosis, risk-stratification management decision-making, and therapy response assessment [3].

ECGs are, however, hard and complex to interpret, and they require high expertise and pattern recognition ability to interpret and classify correctly [3]. The requirement for high pattern recognition ability can lead this process to be tackled by machine learning, where a model can be trained on already classified ECG signals to be able to predict ones with no label.

While supervised Deep Learning has shown promise, it is fundamentally constrained by the scarcity of high-quality annotated data. Manual ECG interpretation is labor-intensive and subject to significant inter-observer variability. For instance, the PTB-XL dataset contains numerous records with ambiguous or conflicting annotations, reflecting the inherent uncertainty in expert diagnosis and limiting the reliability of purely supervised objectives [4].

Consequently, Self-Supervised Learning (SSL) has emerged as a promising alternative, focusing on learning signal representations directly from intrinsic patterns rather than explicit labels. By pre-training on large-scale unlabeled datasets, SSL produces "foundation models" that generate versatile embeddings. These representations can facilitate a wide range of downstream tasks, including clustering, classification, and generation, drastically reducing the dependency on extensive annotated datasets.

One of the commonly used self-supervised techniques is contrastive learning, where a sample of signals are treated as similar (positive) to each other, leaving the rest as negatives, to then try to minimize the distance between the representation of these positive samples while maximizing that distance from the negatives. Defining which samples are positive or similar in a self-supervised setting is tricky since the data is not labeled. Standard practice involves generating positive pairs via data augmentation and treating the augmentation(s) as positive(s), hence treating all other samples as negatives, as in 1.

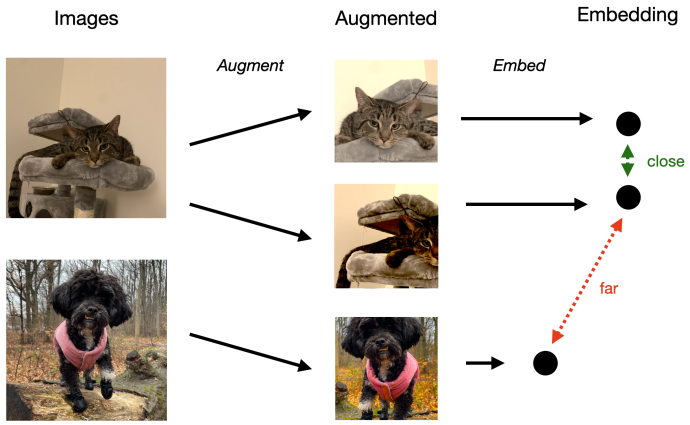


Fig. 1: Self-supervised contrastive learning between images, per [5].

Recent literature demonstrates that the representation quality and training of these models is highly sensitive to the quality and the augmentation technique used. Some works adopted augmentation techniques from image/audio-based models [6, 7]. Another works proposed a family of different types of augmentations under the same type, which they proved effective for contrastive training [8]. There have been generally multiple papers that succeeded in contrastive training. However, there has not been much of a discussion or analysis on why some types of signal augmentations work effectively while some others do not.

The following work tries to discuss and analyze different types of augmentations and their effect on contrastive training outcomes, along with proposing a new dual-view transformer model architecture for ECG foundation models.

Our contributions can be summed up as follows:

1. A novel Dual-View Transformer architecture that explicitly models temporal and channel dependencies independently.
2. A systematic evaluation of three view-generation strategies for ECGs: (1) Stochastic Augmentations, (2) Interleaved Subsampling, and (3) Non-Overlapping Temporal Partitioning.
3. Empirical evidence demonstrating that Random Lead Masking (RLM) combined with temporal partitioning significantly outperforms noise-based augmentations for downstream diagnostic tasks.

This work compared 3 types of augmentations: (1) sampling and applying random augmentations sequentially from a pool of augmentations, (2) signal non-overlapping segmentation and Random Lead Masking (RLM) [7], and (3) signal interleaving segmentation and Random Lead Masking (original contribution). The analysis was mainly done based on the clustering tendency in the representation space (embedding

space) and downstream classification like linear probing and fine-tuning with few-shot learning.

Three models were trained on each augmentation and preprocessing technique and then evaluated with PTB-XL [4] as the training and evaluation dataset.

While absolute performance was constrained by dataset scale and compute resources, the relative performance gap between augmentation strategies provides evidence that temporal partitioning captures superior diagnostic features. Attention map analysis also finds that RLM induces the models to rely heavily on the time dimension across all channels (channel tokens) while ignoring the channel dimension.

This work was fairly limited by both time and physical constraints, and there is still a need to see how the findings consistently scale with the size of the data and the model, which are relatively small in the experiments done here. This is also done mainly based on the Transformer architecture [9], so claims cannot be made on other types like pure Convolutional Neural Networks (CNNs).

2 Background

3 Related Works

There have been multiple attempts to train machine learning models on ECG signals through both supervised and unsupervised setups; we describe and provide background on some of them below.

3.1 Supervised learning and ECGs

The application of deep learning to 12-lead ECG diagnosis has been extensively studied, with CNNs serving as the primary backbone for most state-of-the-art systems.

CNN Baselines: 1D-CNNs and ResNet architectures are currently considered the gold standard for supervised ECG classification. [10] proposes two different fully convolutional or recurrent architectures on the PTB dataset to predict myocardial infection. Their method operates directly on the ECG data without prior preprocessing, and they also investigate the interpretability of their results and the model's performance. They found that their fully convolutional architecture outperforms both the previous state-of-the-art and also the recurrent architecture, even though one would expect that the recurrent networks would be more suitable for time-series data. They used multiple attribution methods to see which parts of the signals affect the model the most. The network's decision was strongly influenced by features established in cardiology, such as ST-segment elevation, pathological Q-waves, and T-wave inversions. High attribution scores were consistently found in these regions across different heartbeats, even when baseline shifts were present.

3.2 Unsupervised learning and ECGs

There have been multiple works aiming to mainly propose new pretraining methods to allow differing models to produce good signals' representations, where they try to exploit different invariances in the signals or specific loss objectives. [8] introduce

the CLOCS contrastive learning methods family, where these methods exploit both temporal and spatial information in the ECG signals. They proposed three types of augmentations:

- Contrastive Multi-segment Coding (CMSC): Segment the signal into multiple segments over the time dimension, so for a signal of length S seconds, it can be segmented into V segments each of length $\frac{S}{V}$ seconds.
- Contrastive Multi-lead Coding (CMLC): Since each lead can be interpreted as a different representation or projection of the heart's electrical signals, an ECG signal can be segmented by the lead instead. This results in a positive pair of S lengths as the original signal and K leads in each, which is less than the total lead count in the original signal.
- Contrastive Multi-segment Multi-lead Coding (CMSMLC): It exploits both temporal and spatial invariances based on the earlier two techniques.

Additionally, they also treat samples from the same patient as positive pairs, leveraging spatiotemporal patient data. They compared their pretraining approach to that of other established supervised and unsupervised state-of-the-art approaches through multiple datasets, where their approach demonstrated better performance in downstream tasks.

[7] proposes a novel pretraining method that captures both local and global context. Their approach proposes doing multi-objective pretraining by encoding the signals with a convolutional block to produce local representations, which are then masked and quantized through codebooks. They try to maximize the cosine similarity between these local representations, then feed them into a transformer block to produce global representations, which also try to maximize the cosine similarity between the positive pair and minimize it between negatives. They build on the CMSC approach by adding Random Lead Masking (RLM) in order to help the model learn more generalized representations and to make it "agnostic" to signals with missing leads. RLM typically works by just masking random leads, with all having a probability 0.5 of being masked.

Their proposed method outperforms other state-of-the-art ECG pretraining methods for cardiac arrhythmia classification and patient identification as downstream tasks, which they chose specifically because "classification and identification require local and global contextualized representations, respectively."

Most recent works proposed hybrid architectures that mostly build on both CNNs and Transformers. Most approaches use the CNN blocks as a latent space encoder for the signals, then the transformer block to produce the final representation. The ECG-FM model from [6] proposes a foundation model pretrained on both masked reconstruction and contrastive objectives. They build on the previous two papers as well as other works like Wav2Vec [11] by following CMSC and RLM and the masked prediction in Wav2Vec. Pretraining is done on a dataset of 1.5 million 12-lead ECGs. ECG-FM is built on a multi-layer CNN and then a BERT-like transformer encoder, where the latent representations are masked and quantized, then loss is computed between the quantized representation and the final representation from the transformer, and then also between the positive pairs of the final representations to compute the global contrastive loss. They concluded that their model can outperform strong

task-specific models on downstream tasks in the small-to-medium-scale data regime but may not provide significant value downstream given sufficiently large task-specific datasets.

These approaches provide very competitive models and methods for pretraining, but they do not, however, provide deeper analysis on why these augmentations/objectives work or what actually happens inside the model given each one, which is what this work tries to tackle and explore.

4 Methods

We describe the dataset used, different types of preprocessing and augmentations done, and the model architecture. We then describe the training process and approach used.

4.1 Dataset

We used PTB-XL as the main training and evaluation dataset [4]. The dataset has 21799 clinical 12-lead ECGs from 18869 patients of 10 second length. The raw waveform data was annotated by up to two cardiologists, who assigned potentially multiple ECG statements to each record. There are two levels of classification: superclasses, which are (1) Normal ECG (NORM), (2) Myocardial Infarction (MI), (3) ST/T Change (STTC), (4) Conduction Disturbance (CD), and (5) Hypertrophy (HYP), which is our main classification objective when assessing clustering tendency and downstream tasks.

The patient demographics are mainly 52% males and 48% females. Their ages are distributed between 0 and 95 years, with the median being 62 and the interquartile range being 22. The dataset collects not only different co-occurring pathologies but also a large population of healthy control samples.

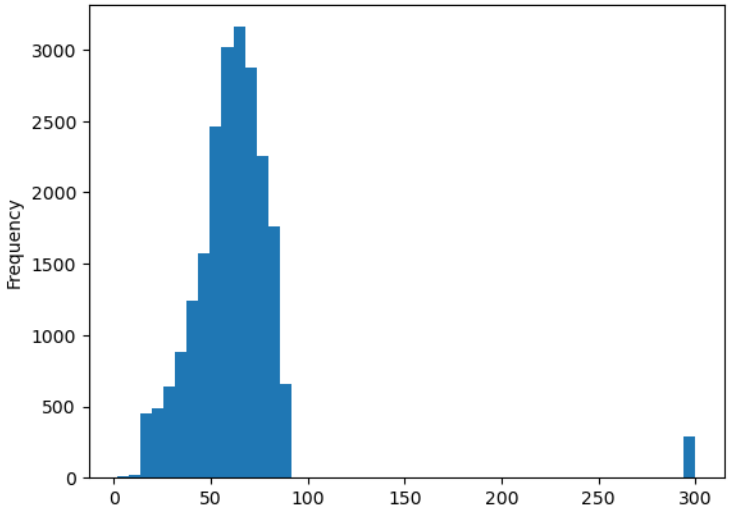


Fig. 2: Age distribution in PTB-XL. Age counts above 89 are shifted to 300+ in accordance to *HIPAA*.

Table 1: Distribution of ECG Superclasses: a single ECG can hold multiple classifications at once.

#Records	Superclass
9514	NORM
5469	MI
5235	STTC
4898	CD
2649	HYP

Two main reasons were taken into account to use this dataset:

- Having multiple classifications for a single ECG, which validates the hypothesis of the difficulty of annotation for ECGs and would help identify if unsupervised approaches can actually help.
- The dataset size is relatively small for pretraining a foundation model, but it was hypothesized to suffice for initial analysis and study on a similarly small model.

4.2 Preprocessing

There were over 5 different types of augmentations and preprocessing tried in this work, but here we focus on 3 types, two of which are based on RLM and CMSC.

We used the signals sampled at a $500Hz$ rate of 5000 time steps ($T = 5000$) for segmentation-based approaches and sampling at $100Hz$ or ($T = 1000$) timesteps for the stochastic augmentations due to memory constraints. And used all 12 channels/leads ($C = 12$), and signals are normalized per sample and per channel before any augmentation.

4.2.1 Stochastic Augmentations

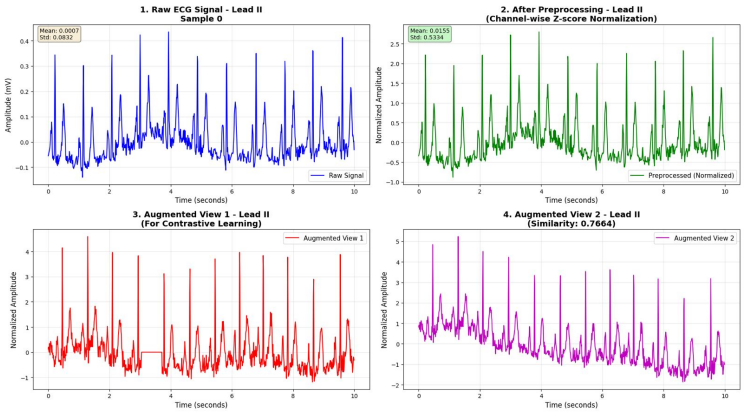


Fig. 3: Visualization of the raw signal, preprocessed, and then two different augmentations for it.

This approach is quite simple, where we create a pool of different augmentations to apply, each typically with a probability of 0.5. Some of the augmentations are

- Random Time Warp.
- Time Crop.
- Bandpass Filter.
- Gaussian Noise.
- Amplitude Scale.

Then we apply these augmentations sequentially on each signal twice in order to create two different views of the raw signal. This also keeps both views at the full length of the original signal, which in our case was 10 seconds or 5000 time steps. We also validated that the augmentations stay consistent with the original signal in the frequency domain by a Fast Fourier Transform (FFT).

We applied these different augmentations through the ecgmentations library [12].

4.2.2 CMSC + RLM: Non-Overlapping Segments

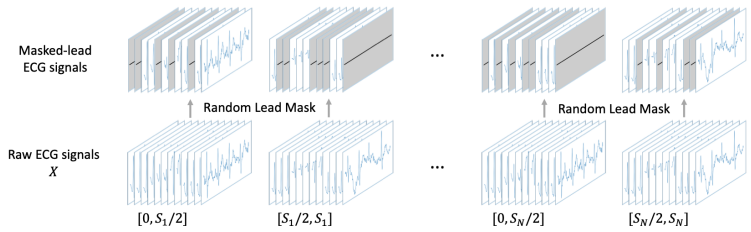


Fig. 4: Splitting signals into two segments and masking leads, per [7]

This approach follows the earlier works exactly [6, 7], where temporal invariance is exploited by having each signal split into two non-overlapping segments, each of length ($\frac{T}{2} = 2500$) time steps. We then randomly mask the channels for each segment; each channel has a 0.5 probability of being masked.

4.2.3 CMSC + RLM: Interleaving Segments

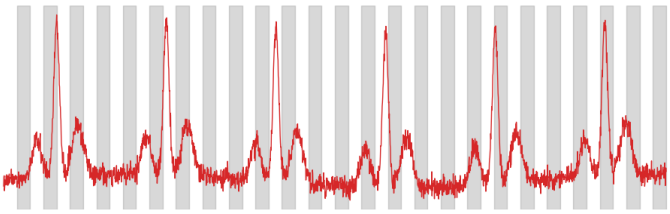


Fig. 5: Interleaving segmentation of a signal

We propose this approach based on the original non-overlapping segmentation, but the intuition here is that instead of having each non-overlapping segment that does not have a full view of the global context of the original signal, we can still keep this information by interleaving the time steps into both segments. So each segment takes non-corresponding time steps, either even-indexed steps or odd ones, like in Figure **??**. This way we can keep both the local and global context information in both segments while also having differences between them, with the addition of lead masking for the contrastive objective.

4.3 Architecture

We propose a Dual-View Transformer-based model, where the input signal would be processed through two different streams: a time stream and a channel stream. Given a signal of shape (T, C), we assume, along the time dimension, that we have C channel

tokens, each with an embedding vector of dimension T. In the time stream, we assume that we have T time tokens, with each having a C dimensional embedding vector. The assumption is that both dimensions hold contextual information, which, if treated like word tokens, can have that information exchanged between them through attention blocks.

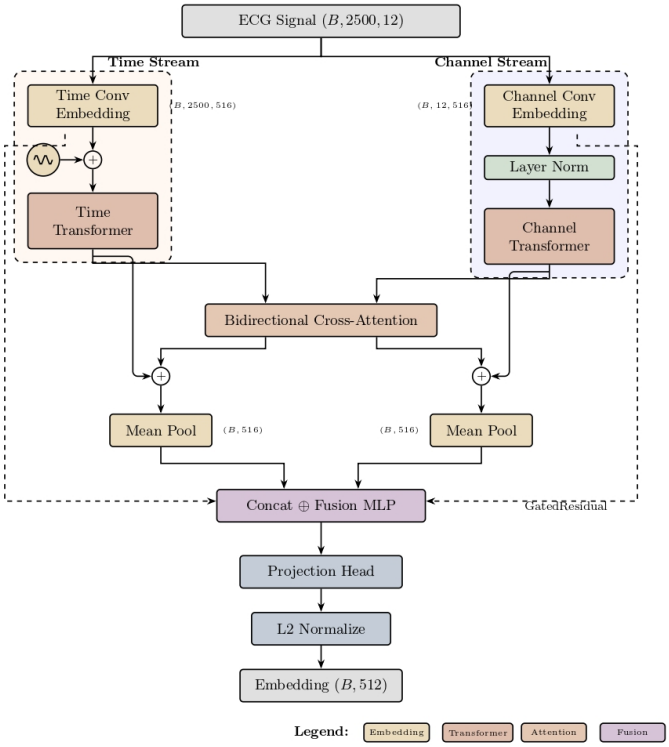


Fig. 6: The Dual-View Transformer Architecture

The convolutional blocks in both streams are mostly used to up/down sample the embeddings of both token types, so the time tokens would have an initial embeddings dimension of 12, which needs to be up-sampled. And the channel tokens have a dimension of 5000 or 2500 (depending on the preprocessing used), which also needs to be down-sampled into a lower dimension.

Then each stream has a self-attention block to allow tokens of the same type to attend to each other. The output tokens from the self-attention are then passed to a bidirectional cross-attention block to allow the time and channel tokens to exchange information back and forth. The model then just pools each type and fuses them together through an FFN to produce the final representation of the signal.

Another benefit of this dual-view architecture is that it allows interpretability of the attention maps for both time and channel dimensions, separately and jointly.

Positional encodings are also added to the time tokens to account for the logical ordering between the time steps, which isn't added to the channel tokens as they do not have such ordering.

4.4 Training

For the training pipeline, Adam was used as the optimizer with a final learning rate (LR) of 3×10^{-5}, with a linear warm-up of 20 iterations starting with an LR of 0. An LR scheduler was also used, specifically a one-cycle scheduler, which raises the LR gradually and then lowers it across steps over 1 cycle only. This scheduler should produce faster convergence for training [13]. Flash Attention was also used instead of vanilla Attention to decrease the memory requirements from $O(N^2)$ to $O(N)$ during training.

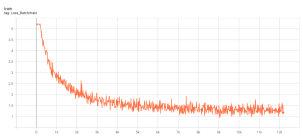
We also added some residual connection between the streams and the fusion head, as the training faced multiple vanishing/exploding gradients. Since we had a limited hardware setup, we used gradient accumulation to simulate larger batch sizes, which is important for contrastive learning. The effective batch size was 1024 samples. The hardware used was $2x$ Nvidia DGX Spark (GB10), with $128GB$ VRAM. The training setup was not distributed across both GPUs due to the need to run multiple models in parallel and other time constraints.

Training was done for 10 hours on average for each approach, and the model had a total of 21.5 million parameters on average, although different sizes were tested.

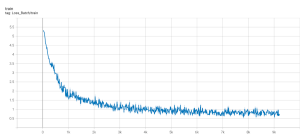
5 Results

We present the different results from the model trained on the 3 different preprocessing methods. Since the training process of different approaches was done iteratively, some figures were created and logged only after the stochastic augmentations experiments were already done running, hence why they could be unavailable for it.

(a) Loss on training and validation for the stochastic augmentations.



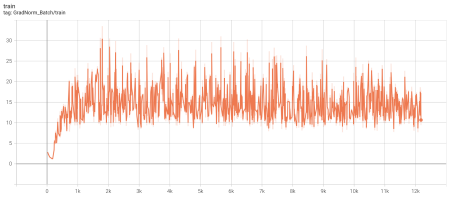
(b) Loss on training for non-overlapping segmentation.



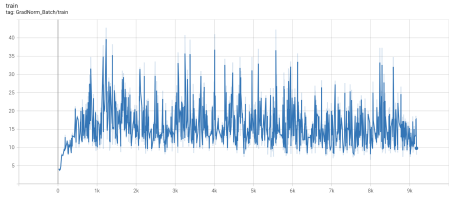
(c) Loss on training for interleaving segmentation.

Fig. 7: Loss curves on all three approaches.

The loss during training did not decrease as much for the random augmentations as it did for the segmentation-based ones, with interleaving segmentation having the lowest loss during training.



(a) Gradient norm for non-overlapping segmentation.



(b) Gradient norm for interleaving segmentation.

Fig. 8: Gradient norm on segmentation-based approaches.

5.1 Embedding Space Separation and Clustering Tendency.

The clustering tendency was measured based on 3 main metrics:

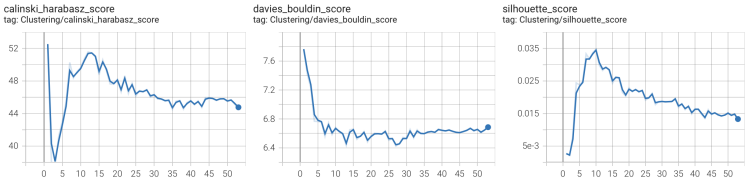
- The silhouette score, which measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It ranges from -1 (misclassified/-clustered) to +1 (perfect clustering).
- Davies-Bouldin score, which is calculated as the average similarity measure of each cluster with the cluster most similar to it, and it makes no assumption about the clusters' shape, unlike the silhouette score. The minimum score is zero, with lower values indicating better clustering.
- Calinski-Harabasz score, which is defined as the ratio of the sum of between-cluster dispersion and of within-cluster dispersion. It ranges from $[0, \infty]$ with higher being better.

Table 2: Best Reported Clustering Evaluation Scores

Method	**Silhouette**	**Davies-Bouldin**	**Calinski-Harabasz**
Stochastic Augmentations	-0.0168	7.8849	9.89
Non-Overlapping Segmentation	0.0327	6.289	37.55
Interleaving Segmentation	0.0350	6.424	50.92

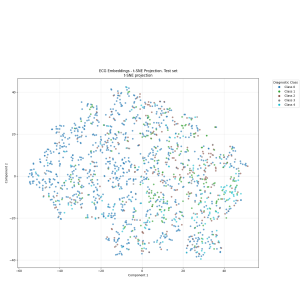
From the best reported values across metrics for all approaches, it can be seen that the random augmentations performance was the worst, with the other two being objectively higher and the interleaving segmentation being the best.

(a) Non-overlapping segmentation clustering tendency along training.

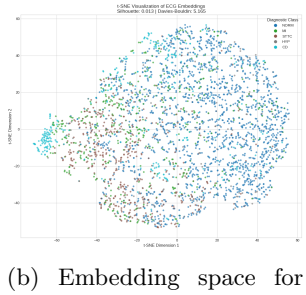


(b) Interleaving segmentation clustering tendency along training.

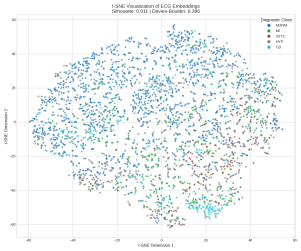
Fig. 9: Clustering tendency metrics along training.



(a) Embedding space for random augmentations.



(b) Embedding space for non-overlapping segmentation.



(c) Embedding space for interleaving segmentation.

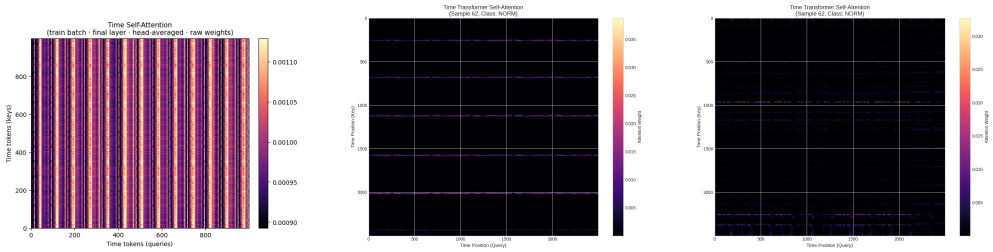
Fig. 10: Embedding space landscape projected with T-SNE.

From Figure 10, it can be seen that, while the three approaches provide a modest separation between the classes and overlap in some cases, the segmentation-based ones have a slightly better separation than the random augmentations.

5.2 Attention Maps

The attention maps for both self- and cross-attention blocks were extracted for a sample with **NORM** classification, which is also seen in Figure 11.

Fig. 11: NORM-classified signal on which the attention maps were made on.



(a) Random augmentations's self-attention.

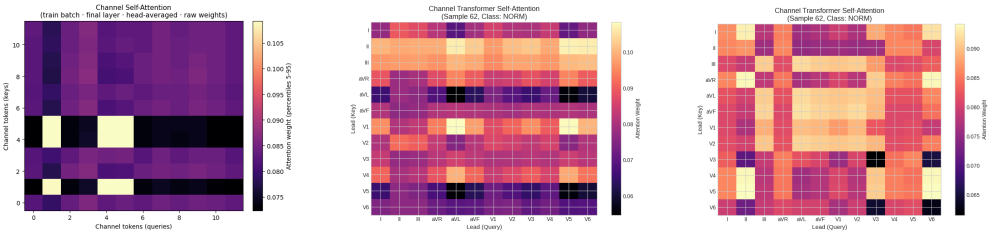
(b) Non-overlapping segmentation's self-attention.

(c) Interleaving segmentation's self-attention.

Fig. 12: Self-Attention maps over time tokens.

Figure 12 illustrates the self-attention weights in the time stream, which seem to be too low in both segmentation-based approaches while being visibly high in the random augmentations approach.

Figure 13 however, shows much higher activation for the channel tokens in the segmentation-based approaches. The random augmentations had median weights for most channels, though it seems that all channels were not attending to channels 1, 4, and 5 (II, aVL, aVF) except themselves.



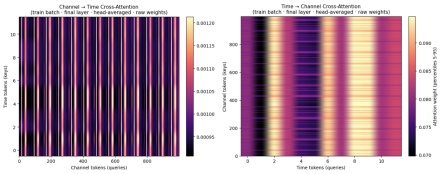
(a) Random augmentations's self-attention.

(b) Non-overlapping segmentation's self-attention.

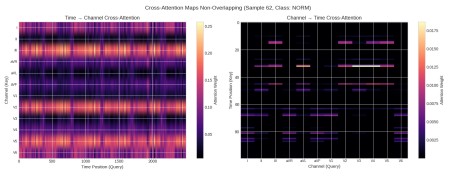
(c) Interleaving segmentation's self-attention.

Fig. 13: Self-Attention maps over channel tokens.

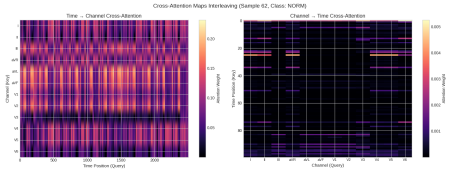
The cross-attention maps in Figure 14 shows a much more interesting story for the segmentation-based approaches, where the channel tokens in both of them attend to the time tokens extremely weakly and selectively. The random augmentations, on the other hand, had no problem for both tokens attending to each other.



(a) Cross-Attention for random augmentations. The x- and y-axes are flipped here.



(b) Cross-Attention for non-overlapping segmentation.



(c) Cross-Attention for interleaving segmentation.

Fig. 14: The bidirectional Cross-Attention weight maps for all approaches.

5.3 Downstream Classification

We tested the three approaches in downstream classification, mainly through linear probing, where we extracted the embeddings of the training set and trained one linear layer on these embeddings in a supervised setting. Supervised fine-tuning was also tried on two settings, by tuning the model on 10% of the training examples or by tuning the model on only 10 examples (few-shot learning). They did not, however, produce significant results or results different from linear probing.

Table 3: Comparison of Model Performance (F1-Score) on downstream classification on all 5 diagnosis classes. Best score in bold.

Class / Metric	**Sequential Augmentations**	**Interleaved Segments**	**Non-Overlapping Segments**
Normal	0.78	0.82	**0.83**
Myocardial Infarction	0.07	0.38	**0.42**
ST/T Change	0.33	0.57	**0.63**
Hypertrophy	**0.11**	0.00	0.01
Conduction Disturbance	0.53	0.60	**0.66**
Overall Accuracy	0.63	0.70	**0.72**
Macro F1-Score	0.36	0.47	**0.51**

In downstream classification with linear probing, it was found that the non-overlapping segmentation delivers a higher F1 score than the other two. Both

segmentation-based approaches have better overall F1 scores than the random one, except for hypertrophy prediction.

6 Discussion

Most results point to the direction that segmentation-based augmentation is better for learning representations for ECG signals. But the general performance, separation in embedding space, and downstream classification have been generally modest across all three approaches.

For the clustering and embedding space, it seemed that all three models' performance would degrade over the training steps, even if the loss for both training and validation kept decreasing. This can be seen by the peaks of the Silhouette and Calinski-Harabasz scores, where they typically peak halfway through training (after 20-30 epochs) and then keep degrading without rising again. To check that further, we made a couple of 1000 epoch runs to see if these scores would rise again, which was not the case. This decrease in the scores was also noticeable through embedding space visualizations, though the general separation between classes did not change much after meeting the peak of the scores.

For the **attention map**, it seems that the interleaving segmentation self-attention on time tokens attended only to the signal's peaks across all channels. Unlike the random augmentations, which had more uniform weights across all attention blocks, which can be attributed to the model learning noise given the bad performance it had against the other segmentation approaches, which have much less attention weights across all boards.

The low weights in all time-token maps are worth noting, as it seems that they are correlated to the random lead masking. What RLM basically does is zeroing out a specific subset of dimensions across all of the time tokens' embeddings, which are then up-sampled into a higher dimension through the CNN block. This process seems to cancel the information that could be gained from the upsampling of the original 12-dimensional embeddings (leads count) across all tokens, as the tokens now have < 12 dimensions of information each.

The model seems to notice this early on and then put most of the focus and learning on the channel stream and tokens, which can be seen by how active the attention is to them in both self- and cross-attention. This shift effectively kills the time stream. This was also confirmed by checking the gradient norm and flow for each stream, where we found the channel stream's gradient flow to be one order of magnitude higher than that of the time stream.

When this phenomenon was noticed during training, we tried to force the model to pass information from the time tokens and improve the gradient flow in two ways:

1. Creating a channel token drop-out and gating the channel stream residual connections.
2. Introducing an additional auxiliary loss on both the channel and time streams.

Both approaches, individually or combined, improved the gradient flow by a bit but not significantly.

Downstream classification did not have much information to uncover, as the performance across all approaches was modest, both in comparison to the existing work, which had state-of-the-art performance (90%+ accuracy/other metrics), and to the ResNet-18 baseline, which performed similarly or slightly worse than these approaches. This can also highlight that the generally poor performance is, for the bigger part, data-based, not only model- or approach-based.

7 Limitations and Future Works

Several limitations were faced during this work, mainly time and compute constraints. DGX Sparks have very good virtual memory but also a really slow memory bandwidth compared to other GPUs such as A100s or H100s, which made the training process much slower. Multiple runs were needed in order to get the training pipeline up and running. The data scale was also an issue, as it seems that both the quality and the size of the dataset had a big contribution to the results across all approaches and tasks. It also seemed that the results were parameter-sensitive, as differing the number of attention heads, hidden layers, and transformer blocks can make significant changes in the training outcomes.

Results or claims cannot be confidently made about the interleaving segmentation, as it seems to be on par or even better, in some instances, than the non-overlapping segmentation. It would be interesting to see how it will scale with more and better data and bigger models as well.

It seems that the dual-view architecture is also promising, as each stream does learn different information about the signals. Applying the lead masking after upsampling the time tokens could help the time stream issue, as it can allow better information representation to reach the attention blocks to work on while still helping the model to be lead-agnostic.

So, generally, scaling up the experiments in terms of both data and model size and qualities can help make more confident claims about the augmentations' effectiveness.

8 Conclusion

This project aimed to investigate the impact of various view-generation strategies on self-supervised representation learning for 12-lead ECG signals, specifically analyzing why certain augmentations outperform others. We proposed a novel Dual-View Transformer architecture that independently models temporal and channel dependencies to create robust "foundational models" from unlabeled data. Our evaluation focused on three strategies: stochastic augmentations, non-overlapping temporal partitioning, and a new interleaved subsampling approach. Results indicated that segmentation-based augmentations, particularly non-overlapping segments, consistently provided superior diagnostic features and better embedding space separation compared to noise-based stochastic methods. Analysis of attention maps revealed that Random Lead Masking (RLM) causes the model to prioritize the channel stream, effectively capturing spatial relationships across leads while potentially underutilizing the time stream.

In a clinical setting, this approach could improve care by providing reliable automated diagnostic support even when datasets are small or labels are uncertain.

By utilizing self-supervised foundational models, healthcare providers could achieve high-accuracy classifications for conditions like myocardial infarction or conduction disturbance with minimal human annotation. To bring this project to practical clinical practice, future work must address the current data-scale limitations by scaling the training to much larger, more diverse datasets. Additionally, refining the architecture to prevent "time-stream killing," perhaps by applying masking after upsampling, is essential to ensuring the model fully leverages both temporal and spatial cardiac information.

References

[1] Heron, M., Anderson, R.N.: Changes in the leading cause of death: Recent patterns in heart disease and cancer mortality. NCHS data brief **254** (2016)

[2] Tsao, C.W., Aday, A.W., Almarzooq, Z.I., Anderson, C.A.M., Arora, P., Avery, C.L., Baker-Smith, C.M., Beaton, A.Z., Boehme, A.K., Buxton, A.E., Commodore-Mensah, Y., Elkind, M.S.V., Evenson, K.R., Eze-Nliam, C., Fugar, S., Generoso, G., Heard, D.G., Hiremath, S., Ho, J.E., Kalani, R., Kazi, D.S., Ko, D., Levine, D.A., Liu, J., Ma, J., Magnani, J.W., Michos, E.D., Mussolino, M.E., Navaneethan, S.D., Parikh, N.I., Poudel, R., Rezk-Hanna, M., Roth, G.A., Shah, N.S., St-Onge, M.-P., Thacker, E.L., Virani, S.S., Voeks, J.H., Wang, N.-Y., Wong, N.D., Wong, S.S., Yaffe, K., Martin, S.S., Epidemiology, Committee, P.S., Subcommittee, S.S.: Heart disease and stroke statistics—2023 update: A report from the american heart association. Circulation **147**(8), 93–621 (2023) https://doi.org/10.1161/CIR.0000000000001123 https://www.ahajournals.org/doi/pdf/10.1161/CIR.0000000000001123

[3] Breen, C.J., Kelly, G.P., Kernohan, W.G.: Ecg interpretation skill acquisition: A review of learning, teaching and assessment. Journal of Electrocardiology **73**, 125–128 (2022) https://doi.org/10.1016/j.jelectrocard.2019.03.010

[4] Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T.: Ptb-xl, a large publicly available electrocardiography dataset. Scientific data **7**(1), 1–15 (2020)

[5] Witter, R.T.: Contrastive Learning. https://www.rtealwitter.com/deeplearning2023/contrastive.html. CSCI 1051: Deep Learning, Winter 2023. Accessed: 2024-05-21 (2023)

[6] McKeen, K., Masood, S., Toma, A., Rubin, B., Wang, B.: Ecg-fm: An open electrocardiogram foundation model. JAMIA open **8**(5), 122 (2025)

[7] Oh, J., Chung, H., Kwon, J.-m., Hong, D.-g., Choi, E.: Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In: Conference on Health, Inference, and Learning, pp. 338–353 (2022). PMLR

[8] Kiyasseh, D., Zhu, T., Clifton, D.A.: Clocs: Contrastive learning of cardiac signals across space, time, and patients. In: International Conference on Machine Learning, pp. 5606–5615 (2021). PMLR

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[10] Strodthoff, N., Strodthoff, C.: Detecting and interpreting myocardial infarction using fully convolutional neural networks. Physiological measurement **40**(1), 015001 (2019)

[11] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)

[12] Epifanov, R.: Ecgmentations. GitHub (2023)

[13] Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial Intelligence and Machine Learning for Multi-domain Operations Applications, vol. 11006, pp. 369–386 (2019). SPIE