Imperial College London

BACKGROUND REPORT

DEPARTMENT OF COMPUTING

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

Multi-Modal Data Fusion Through Contrastive Learning in Geoscience

Author:
Nathan Bailey

Supervisor:
Dr Sibo Cheng

Submitted in partial fulfillment of the requirements for the MSc Computing - Artificial Intelligence and Machine Learning of Imperial College London

Contents

Chapter 1

Introduction

Self-Supervised learning (SSL) is a common technique primarily used within computer vision to generate meaningful and discriminative representations of unlabelled data invariant to transformations (i.e. strong generalisation capabilities) [1]. One goal of SSL is to utilise data where labels are sparse, and leverage the inherent structures in the data to form useful representations, which can be used to improve the performance of downstream tasks on labelled data [2, 3].

Contrastive learning, a particular type of self-supervised learning, aims to learn data representations by contrasting negative and positive data samples. In this way, within the resulting representation space, it pulls positive samples close and pushes negative samples away, creating structure within the latent space [4]. SimCLR [5], for example, creates a positive pair of data by augmenting a data sample twice, and uses an NT-Xent loss function to encourage the representations of positive pairs to be close whilst pushing away negative samples. Through this method, the model learns high-quality embeddings of data that can be used for downstream tasks [4].

Whilst primarily used in computer vision (CV) and natural language processing (NLP), contrastive learning has also been used for time-series data. For example, Yue et al [6] take the representations of the same timestamp as positives in TS2Vec and use a joint temporal and instance contrastive loss to align positive samples in the latent space.

Contrastive learning lends itself naturally to geoscience data, particularly weather analysis datasets, since these are often collected with no labels. Weather analysis datasets are time-series data, whereby measurements of various data variables are collected across latitude and longitude at varying degrees of resolution [7].

Downstream tasks can be performed using weather analysis datasets. A common task is numerical weather prediction (NWP), which traditionally is performed by simulating the atmosphere by solving physical equations [7, 8]. Machine learning methods offer an alternative to this. For example, Bi et al propose a 3D transformer-based neural network for weather forecasting [9]. Whilst downstream tasks such as weather forecasting can be performed on the raw data, due to the fine horizontal resolution at which this data is collected, the samples live in high dimensions, increasing the challenge of which downstream tasks can be performed, due to the curse of dimensionality. For example, the common reanalysis dataset ERA5 has a lower-bound resolution of 64x32, reaching a maximum of 1440×721. Bi et al [9] use some level of compression, but still operate in high dimensions.

Moreover, weather analysis datasets often consist of multiple data variables (called modes). ERA5 consists of 62 data variables, some of which span up to 137 pressure levels [10]. This substantially increases the dimensionality of the data, making downstream performance on the raw data samples substantially harder.

The majority of current self-supervised learning methods only focus on a single mode of data, for example, images or text. The presence of multiple data modes introduces a further challenge within contrastive learning and, therefore, within weather analysis datasets on how to best merge these modes to create a stable and structured representation to improve downstream performance. Whilst contrastive learning for weather analysis datasets has been explored in current work, an extensive investigation into this topic has not been carried out.

Therefore, contrastive learning offers a way to create significantly smaller and structured representations of weather data, reducing the dimensionality to increase downstream performance.

Finally, current self-supervised methods on geoscience data do not consider the fact that incoming data may be sparse and not all samples will be accessible. Commonly, satellite-borne sensors produce sparse observations of the Earth's atmosphere, presenting challenges when using the data for certain applications [11]. Having a robust framework that can align sparse data to full observations within the latent space would, in theory, enable little to no performance change for resulting downstream tasks.

Therefore, this project aims to explore contrastive learning for geoscience data, developing a scalable, robust framework for multimodal data fusion to improve the predictive accuracy for downstream tasks. We aim to create a structured latent space, invariant to the sparsity of data, to enable downstream tasks to be performed even with sparse observations.

Chapter 2

Literature Review

The literature review outlines current research on contrastive learning in computer vision and time-series data. We also summarise the current ongoing contrastive learning research within geoscience and detail the gaps this project aims to fill.

2.1 Current Contrastive Learning Methods

2.1.1 Computer Vision

SimCLR

The first major proposal within computer vision contrastive learning was SimCLR [5]. SimCLR takes a batch of data samples and, for each, creates a positive pair (i, j) by augmenting the data sample twice. These are used to create intermediate representations h_i, h_j and projections z_i, z_j. Within the batch, SimCLR optimises the network using the NT-Xent loss, which is shown in equation 2.1. This maximises the cosine similarity (denoted by sim()) between normalised positive projections and minimises the cosine similarity between normalised negative projections. This is computed across all positive pairs (both (i, j) and (j, i)) in a mini-batch and averaged across all $2N$ pairs. SimCLR showed that using the representations formed from this contrastive learning approach in supervised learning in a simple linear classifier matched the performance of a ResNet-50 approach on labelled data.

$$\ell_{i,j} = -\log \frac{\exp\left(\text{sim}(z_i, z_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp\left(\text{sim}(z_i, z_k)/\tau\right)} \tag{2.1}$$

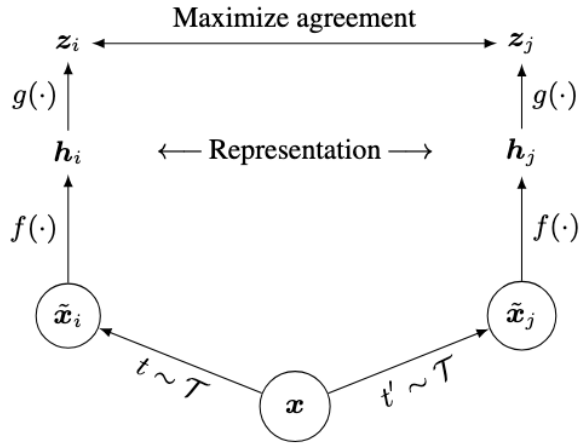

Figure 2.1: SimCLR Model

Due to the setup of SimCLR, the quality of the representations and, therefore, the downstream performance are dependent on the batch size and the quality of the data augmentations used. A larger batch size increases the number of negative samples and makes the separation task harder, increasing the quality of the representations. A stronger data augmentation yields a similar result, forcing the model to produce higher-quality representations, improving downstream performance.

Another consideration is the temperature hyperparameter τ introduced in the loss function; a smaller value, below 1, increases the similarity between samples. This increases the similarity of the sample to negatives, encouraging the model to push negatives further away from the sample, creating more diverse representations. Tuning the temperature hyperparameter is key, and Chen et el demonstrate that a tuned τ can help the model learn from hard negatives, that is, negatives that are close in similarity to the data point. They find that a temperature parameter of 0.1 yields the best performance.

Supervised Contrastive Learning

Supervised Contrastive Learning (SupCon) [12] follows a similar approach to SimCLR, but is extended to leverage existing labels that may be present. In contrast to SimCLR, where there is only one positive sample j for a given chosen anchor sample i, labels are used to define positive samples as data points with the same label. This enables multiple positive samples for a given anchor, as shown in the updated loss function in equation 2.2, where each sample has an arbitrary number of positives p.

$$\mathcal{L}_{\text{out}}^{\text{sup}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_a / \tau\right)} \tag{2.2}$$

SupCon uses the same framework as SimCLR, augmenting each sample twice and using representations and projections. They show that using labelled data with contrastive learning can lead to an increase in top-1 accuracy for ImageNet compared to SimCLR.

Barlow Twins

As mentioned above, existing contrastive methods such as SimCLR are dependent on larger batch sizes to increase downstream performance. Barlow Twins [13] proposes a novel approach to create diverse feature representations that do not depend on batch size by utilising the cross-correlation matrix between representations.

As in SimCLR, and shown in figure 2.2, Barlow Twins data augmentations are applied to a sample to create a positive pair.

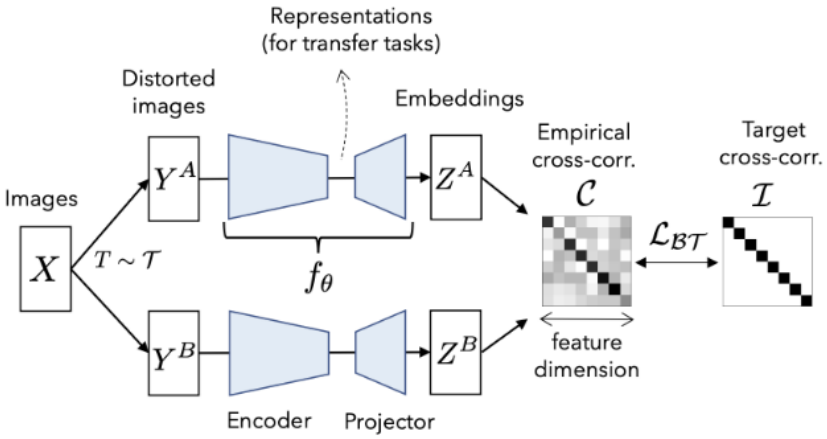

Figure 2.2: Barlow Twins Model

The projected embeddings for a batch of data are centred along the batch dimension, and the loss in equation 2.3 is computed.

$$\mathcal{L}_{\mathcal{BT}} = \underbrace{\sum_i (1 - c_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} c_{ij}^2}_{\text{redundancy reduction term}} \qquad c_{ij} = \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b (z_{b,i}^A)^2} \sqrt{\sum_b (z_{b,j}^B)^2}} \tag{2.3}$$

The invariance term equates the diagonal elements of the cross-correlation matrix to 1, which makes the embedding invariant to the augmentations applied, by ensuring that the same dimensions of the augmented projections are the same. The redundancy reduction term attempts to equate the non-diagonal terms to 0, to encourage diversity within the dimension of the projections. In this way, the projections of the positive pairs are aligned, but collapse is prevented by encouraging non-redundancy. This involves no negative samples and therefore means that Barlow Twins is more invariant to batch size than SimCLR. Barlow Twins works well with batches as small as 256, which is highlighted in figure 2.3.

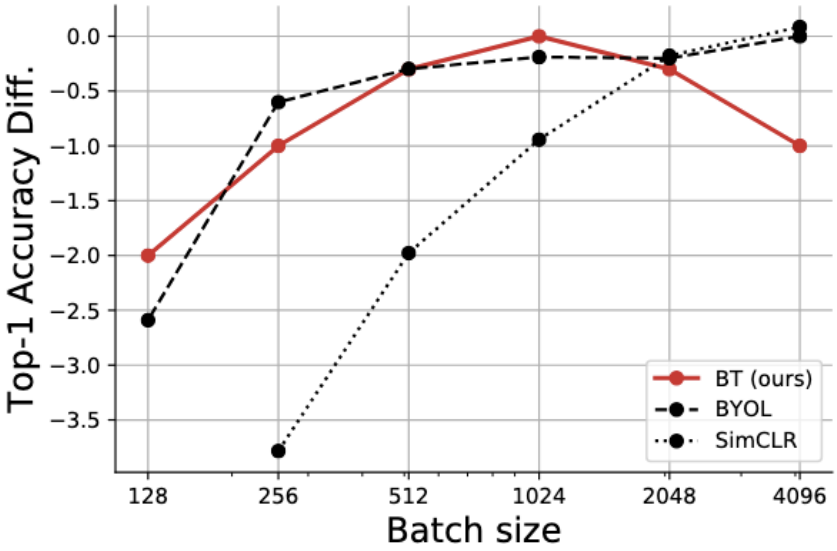

Figure 2.3: Barlow Twins Batch Size vs SimCLR

2.1.2 Time-Series Data

In addition to computer vision, contrastive learning has also been used extensively within time-series data. Eldele et al [14] propose a time-series contrastive model called Temporal and Contextual Contrasting (TS-TCC), as shown in figure 2.4. Two contrastive losses are used; the first contextual contrasting term uses positive pairs that are created by augmenting a single timestep. The model also creates a context vector from t timesteps, which is then used to predict the next k timesteps. This is aligned to the context vector through the second temporal contrastive loss term. In this way, the model learns discriminative features from the contextual contrasting term and temporal features from the temporal contrastive loss.

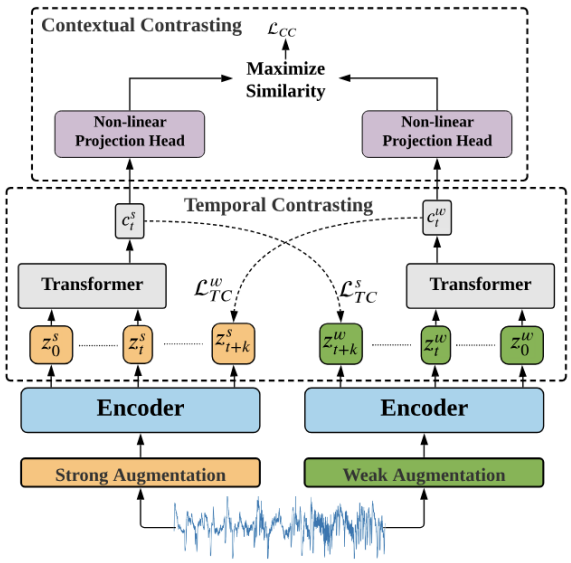

Figure 2.4: Contextual and Temporal Contrastive Loss [14]

One important consideration with time-series data compared to image data is that timesteps close to

the current selected timestep should lie closer to the latent space than timesteps further away. This data is likely to be more similar to the current timestep than to those further away. Lee et al [15] consider this by introducing soft contrastive learning. The samples are weighted according to their similarity in non-embedding space, which is precomputed to save time whilst training the network. This similarity is then incorporated in the contrastive loss.

This concept is shown in equations 2.4 - 2.6, where a positive pair is formed by $(t, t+T)$. As the timestep increases, the resulting weight becomes smaller, points that are closer to the current timestep are assigned more weight and deemed more positive than points farther away. This also has the benefit that for a given sample, it forms a soft positive pair with every other sample.

$$w_T(t, t') = 2 \cdot \sigma(-\tau_T \cdot |t - t'|) \tag{2.4}$$

$$p_T(i, (t, t')) = \frac{\exp(r_{i,t} \circ r_{i,t'})}{\sum_{\substack{s=1 \\ s \neq t}}^{2T} \exp(r_{i,t} \circ r_{i,s})} \tag{2.5}$$

$$\ell_T^{(i,t)} = -\log p_T(i, (t, t+T)) - \sum_{\substack{s=1 \\ s \neq \{t, t+T\}}}^{2T} w_T(t, s \bmod T) \cdot \log p_T(i, (t, s)). \tag{2.6}$$

Finally, Bach et al [16] present a dynamic contrastive method which samples both adjacent timesteps as positive samples and optimises them in the same loss term as shown in figure 2.5. In this way, there are 2 SimCLR losses, as shown in equation 2.7, combined to form the loss term. They also introduce a masking term to encourage masked projections of timesteps to be close to the projection of the unmasked timesteps through MSE loss. This enables feature invariance and stabilises the training process. Finally, they introduce an additional margin term to increase the distance between features in the projection space. They show that the addition of the masking term is critical to ensuring diverse feature representations. The addition of both the masking and the margin term increases unsupervised clustering performance compared to the state of the art, whilst the introduction of adjacent samples increases downstream classification.

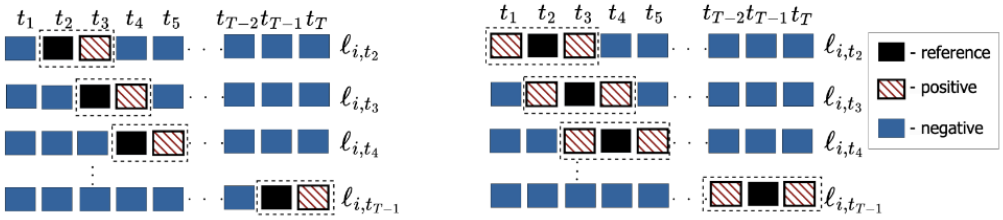

Figure 2.5: Adjacent Positive Samples

$$\ell(i,t) = -\log \frac{\exp\left(\frac{\text{sim}(z_{i,t}, z_{i,t-1})}{\tau}\right) + \exp\left(\frac{\text{sim}(z_{i,t}, z_{i,t+1})}{\tau}\right)}{\sum_{k=1}^{T} \mathbb{1}_{[k \neq t, t-1, t+1]} \exp\left(\frac{\text{sim}(z_{i,t}, z_{i,k+1})}{\tau}\right) + \sum_{l=1}^{T} \mathbb{1}_{[l \neq t, t-1]} \exp\left(\frac{\text{sim}(z_{i,t-1}, z_{i,l})}{\tau}\right)} \tag{2.7}$$

2.1.3 Multi-Modal Fusion

The contrastive learning approaches presented above focus on single modes of data, e.g. images. However, it is possible that data can be collected from multiple different types of sensors, e.g. LIDAR,

video, infrared, etc [17]. In this case, to perform contrastive learning, it is needed to fuse the modes of data. As explained by Tang et al and shown in figure 2.6, there are 3 ways to do this. Early fusion fuses the modes of data before passing them through the model, while late fusion passes each mode separately through the model and then fuses the resulting representations. Early fusion can be better at capturing inter-modal relationships and performs best when the relationship between modes is correlated. Late fusion is better suited when overfitting is an issue, but requires a larger model and therefore may require a trade-off with the batch size. If modes are uncorrelated, late fusion can perform better than early fusion. Finally, a hybrid approach involves combining the two fusion mechanisms, which can be more flexible but adds additional complexity when training the model [17].

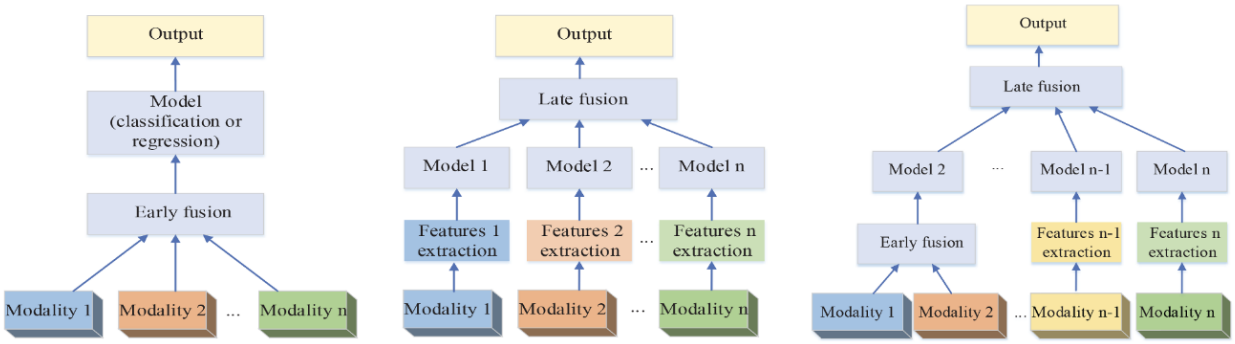

Figure 2.6: Fusion Methods

Current research is sparse on modal fusion methods; recent work focuses mainly on situations with 2 modalities, with attention mechanisms being the main focus. Dufumier et al [18] use late-fusion modal attention to create a shared representation for 2 modes of data after passing them through modality-specific encoders. The representations from each encoder are passed through token adapters to serve as tokens in a self-attention block, which creates the final representation.

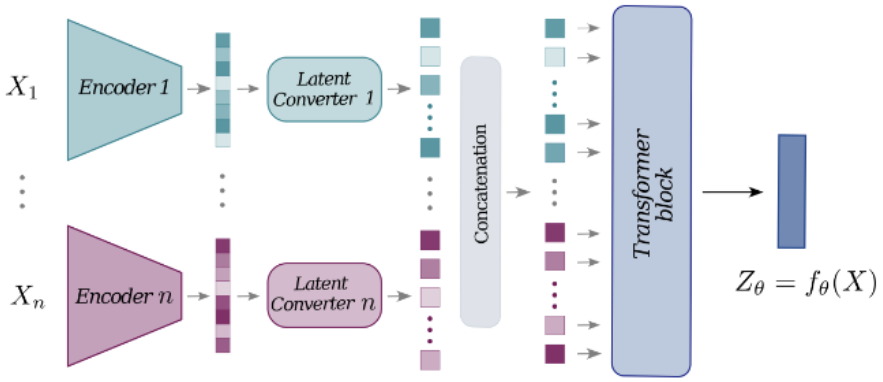

Figure 2.7: Self-Attention Fusion Mechanism [18]

Saioni et al [19] perform a similar approach with text and image data. Instead of using self-attention, they perform cross-modal attention, whereby the representations for one mode act as the values and keys for the other mode. They combine this with skip connections to form the final representation.

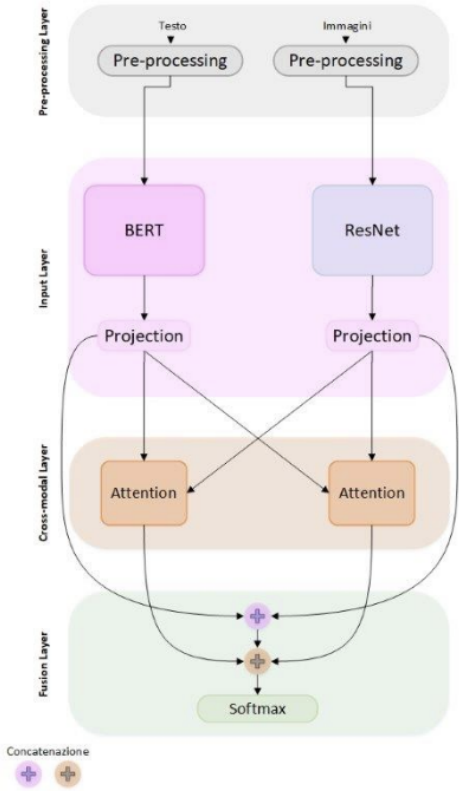

Figure 2.8: Cross Modal Attention

One downside of this approach is that it does not scale for more than 2 modes unless certain modal interactions are missed out. For example, if there were 3 modes, to model all modal interactions $3 * (3 - 1) = 5$ cross-attention models would be needed. Bahaduri et al [20] extend this cross-modal attention mechanism to 4 modes, but choose only one mode to act as keys and values for each mode.

One possible solution to this could extend on Guo et al's [21] work that groups modes based on the modal distances in TSNE. They use grouped mode representations to form a positive pair and perform contrastive learning on the other samples in the same way as SimCLR.

2.2 Contrastive Learning Methods in Geoscience

Whilst contrastive learning methods within geoscience and especially the ERA5 dataset are underexplored, some existing work exists, which is summarised below.

Wang et al [22] use the ERA5 dataset in a SimCLR-style model to build latent representations to classify weather systems in latent space. Positive pairs are created by sampling a data point and then randomly sampling a nearby data point with a monotonically decreasing probability distribution. These 2 samples are augmented using resizing, random cropping and a 5x5 mean filter. A shared ResNet-18 encoder is used with early fusion. The representations are evaluated on the classification of weather systems, showing that the contrastive learning approach outperforms baseline methods such as k-means and self-organising maps.

Wang et al also use a contrastive learning method on a range of datasets, including ERA5 data in the downstream task of tropical cyclone prediction [23]. They opt for a hybrid fusion approach using a separate encoder for each modality, but also complementing the representations with a shared

encoder. All representations are concatenated before being used in the downstream task. Instead of optimising a general contrastive loss function, they optimise the model directly on the downstream task of tropical cyclone prediction. Whilst this tailors the model to the specific downstream task, it is unlikely to build generic representations that could be used for a multitude of tasks.

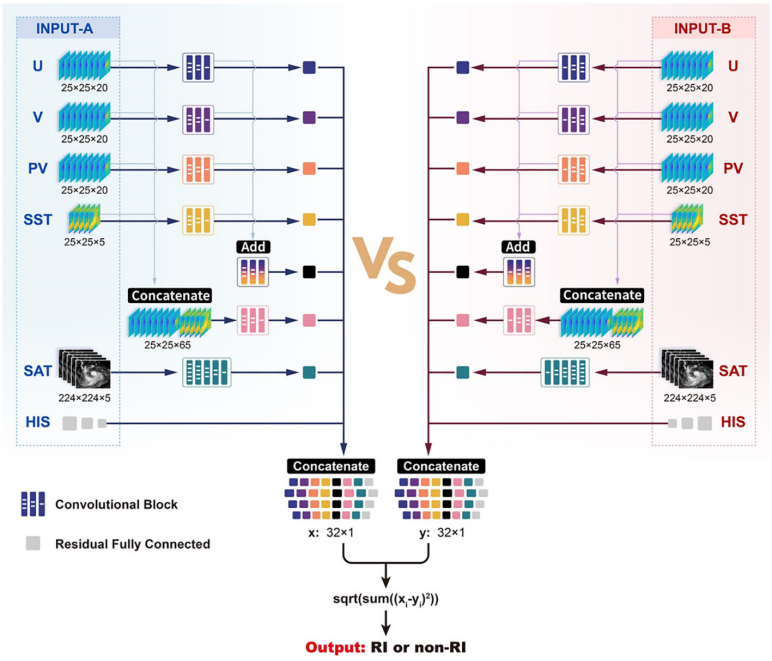

Figure 2.9: Contrastive Learning for Cyclone Prediction

Ballard et al opt to use contrastive learning in the task of climate model bias correction [24]. This refers to a set of techniques used to adjust the output of climate models so they better match observed historical data. Their model uses a super-resolution contrastive unpaired translation GAN to perform bias correction and spatial resolution enhancement. This model introduces an InfoNCE style contrastive loss that operates on patch-level features from output images. It samples patches of the output image and ensures that the samples are similar to corresponding patches from the input. At the same time, it discourages patches from being too similar to other patches of the input image.

Shi et al [25] use a multi-group multi-attention (MGMA) encoder-decoder model to perform atmospheric variable prediction (AVP) over time on ERA5 data. Whilst this is built on a self-supervised learning idea and their model compresses to a latent space, neither of these ideas are used in downstream tasks. The model performs the spatiotemporal prediction in the original space, and the latent space is not explored for downstream tasks.

Finally, Han et al create a compressed version of ERA5, called CRA5 [26]. A transformer-based dual variational autoencoder (VAE) is used to compress ERA5 data to a latent space. The goal of this paper was to create a compressed version of ERA5, such to widen the availability of weather data. The full ERA5 dataset can reach up to 226TB, which limits the availability due to the large storage overhead. The downstream task of weather forecasting was used to evaluate the quality of the compressed data, showing that the compression had a negligible impact on training a forecasting model. Whilst this method does not include contrastive learning, it shows that building a latent space for weather datasets can yield as good performance on downstream tasks as using the full dimensionality.

2.3 Current Methods Tackling Sparsity in Geoscience

Existing work focusing on sparse observations in geoscience is limited, and to the best of our knowledge, current literature only focuses on training autoencoder-style approaches to reconstruct masked observations.

For example, Goh et al propose MAESSTRO [27] and Agabin et al propose a similar model [11], both of which use masked autoencoders to reconstruct sea surface temperature (SST) data using the LLC4320 dataset. Masked autoencoders were originally presented by He et al [28] and involve masking out patches of an image before encoding them, and training a decoder to reconstruct the masked patches. Similar to contrastive methods outlined above, this enables the encoder to form a robust representation of the data, which can then be used in downstream tasks such as classification or semantic segmentation.

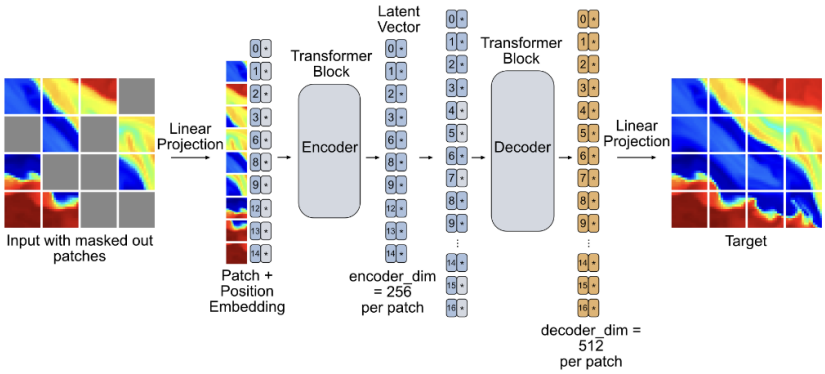

Figure 2.10: MAE Approach for SST data [11]

The MAE models used by both Goh and Agabin focus on developing a model that can reconstruct sparse samples. This is useful as it mitigates against issues that arise if sparse climate data is sampled. However, it does not evaluate the approach for downstream latent space tasks, only for reconstruction.

Vandal et al [29] extend this MAE approach to multi-modal data. They create EarthNet, a multi-modal masked autoencoder model to reconstruct sparse data for multiple modalities. EarthNet is trained as an alternative to data assimilation, which merges observations to produce an optimal initial state of the atmosphere. EarthNet takes in multiple sparse modes, encodes them to tokens, which are then used within an attention mechanism to form representations for the mode-specific decoders to reconstruct the modes. Like Goh and Agabin, this approach is not evaluated for the downstream task of weather forecasting in the latent space.

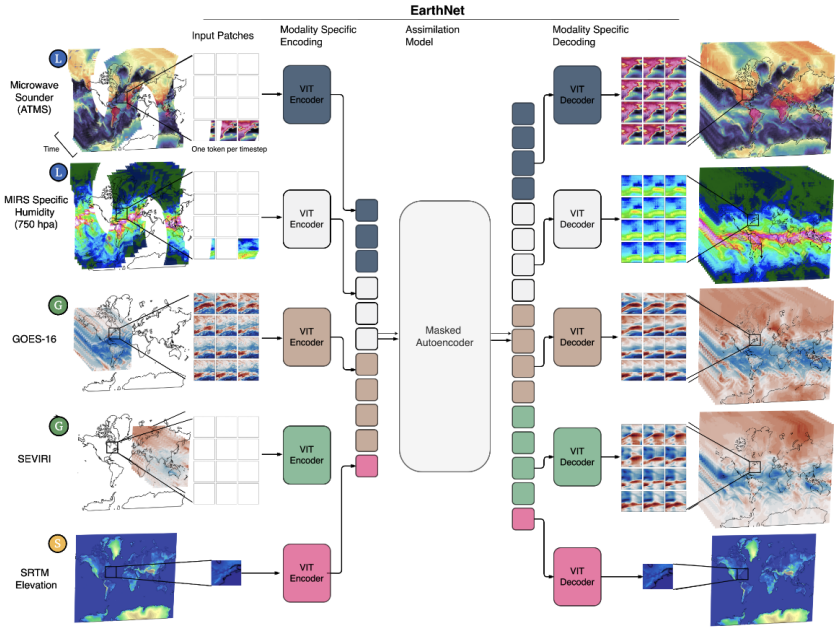

Figure 2.11: Multi-Model MAE - EarthNet

2.4 Decoder Based Contrastive Learning

Whilst contrastive learning has been key to accelerating performance in downstream tasks, there has been little research on how best to train a decoder to map latent representations to the original space. This is likely because the majority of downstream tasks in the field of computer vision do not need the resulting data to be decoded; for example, classification tasks can take place fully in the latent space. However, for the downstream task of weather forecasting, resultant latent space predictions must be decoded back to the original dimensionality to be usable.

Kadeethum et al [30] investigate this and create a unified framework for a Barlow Twins model in the domain of flow and transport phenomena (see figure 2.12). They propose a joint training procedure, where the encoder is trained with contrastive learning with an outer batch, and then jointly trained with the decoder on the inner batch formed from the outer batch. The outer batch is set to 512, whilst the inner batch is set to 32. The decoder is trained to reconstruct the non-augmented data sample separately, whilst augmented inputs are used in the contrastive loss as before. They find that by employing Barlow Twins self-supervised learning, the model maximises the information content within the latent space, leading to more robust and generalizable representations.

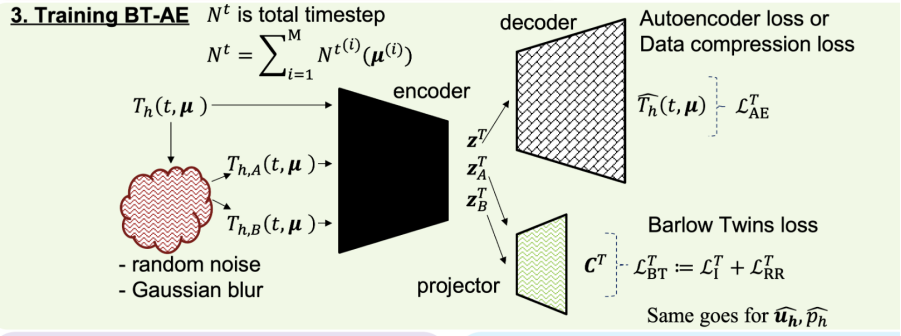

Figure 2.12: Barlow Twins Decoder

Kadeethum et al [31] also extend upon this idea by using an ensemble of encoders, where the average of the representations is used before being passed through the encoder. The same training process is used, but they find that combining multiple encoders in an ensemble model can improve the prediction accuracy and can account for the imbalance in the data in multiphase flow problems.

Similar works [32, 33, 34] also investigate how to train a decoder with a contrastive learning encoder. Quetin et al [32] train a contrastive encoder along with a decoder (figure 2.13). The following joint loss function is used: $L = \alpha \times L_{enc} + (1 - \alpha) \times L_{dec}$ where alpha is slowly decayed over the training epochs. They find that training the encoder and decoder in this joint fashion helps to enhance the representational power of the encoder and improve downstream performance compared to training the decoder on the frozen representations from the encoder.

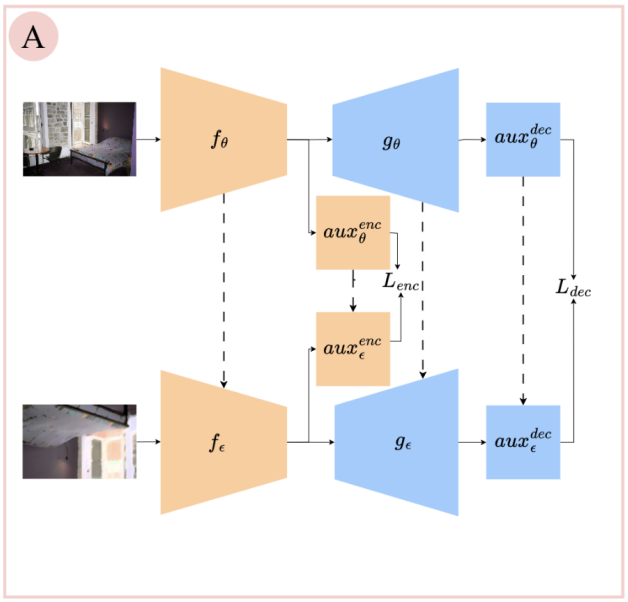

Figure 2.13: Contrastive Decoder Framework from [32]

Yao et al [33] attempt to create a contrastive masked autoencoder model optimising the following joint loss: $L = L_{con} + \alpha \times L_{dec}$. As in SimCLR, they augment the image twice to create a positive pair and then mask both samples to create a total of 4 samples. Contrastive loss (L_{con}) is optimised over both the masked and non-masked pairs separately, with a decoder then trained jointly to reconstruct the masked patches of the augmented images. They find that this contrastive MAE approach leads to

better performance than a plain MAE or common contrastive-only approaches such as MoCo [35].

Finally, Dippel et al [34] extend the SimCLR framework with a decoder to add a self-reconstruction loss on masked inputs, in a similar approach to [33]. The framework masks the input, and the decoder attempts to reconstruct the masked input based on the encoded representations. Like Yao et al [33], the decoder is trained to reconstruct the augmented image, which contrasts with [30], whereby the decoder is trained to reconstruct the non-augmented data point.

In addition to the encoder-decoder structure, they also present an attention-weighted pooling to replace the average global pooling typically used in SimCLR. They argue that this global operation discards salient features. They instead introduce attention-weighted pooling that computes attention weights of the feature map, which is used in global average pooling: $a(\mathbf{x}) = g(A_\phi \odot \phi) \odot \frac{1}{g(A_\phi)}$. Where A_ϕ are the attention weights and ϕ are the feature maps. They show that both attention pooling and decoder training produce improved vector representations for images and outperform regular SimCLR on classification datasets.

2.5 Summary and Research Area

From the literature review presented, contrastive learning is widely used to leverage the underlying structure of data and build robust representations to improve downstream performance. Whilst applied within geoscience and weather prediction, there exists scope to explore this further.

Wang et al [22] build a SimCLR model to classify weather systems in latent space, showing the strength of contrastive learning on weather datasets compared to traditional approaches. They also exhaustively evaluate various augmentation techniques on ERA5 data, showing conclusively that resizing and smoothing produce the best results. However, they only consider the classification of weather systems by using clustering or a linear classifier on the latent representations. The task of weather forecasting and decoding back from the latent representations is not considered. Moreover, this task is only performed on a subset of the latitude and longitude of the data, and only early fusion is considered.

Wang et al [23] consider contrastive learning on ERA5 data using a different hybrid fusion method. However, they optimise the model directly on the chosen downstream task of tropical cyclone prediction, rather than using a contrastive learning function. So, we cannot evaluate the method for creating robust representations using contrastive learning that could apply to many tasks.

Han et al [26] create a VAE model to compress the ERA5 dataset to a latent space. This is done to create a smaller dataset to widen the access to weather datasets, which can reach up to 226TB of storage. They show that this method can achieve comparable forecasting performance on the latent data compared to the full dimensionality. Whilst this is an important first step, they do not consider the benefit that contrastive learning methods could bring. Furthermore, they do not decode the resulting latent predictions back to the original dimensionality, which is important for use in real-world applications. They simply show that using a compressed latent representation of ERA5 data does not downgrade the forecasting ability.

All of these methods consider the reanalysis dataset as-is and do not account for the fact that weather data is often sparse or noisy. Existing methods tackling sparsity within weather datasets [27, 11, 29] only focus on creating models that can reconstruct the sparse data by training masked autoencoders. While this is an important area of focus within weather datasets, none of these methods consider creating robust representations of sparse/noisy data within the latent space for downstream tasks.

These approaches do not thoroughly explore the various multimodal fusion approaches that are available, which we highlighted in section 2.1.3. More importantly, they do not consider that latent repre-

sentations will need to be decoded back to the original dimensionality for analysis.

Therefore, we can conclude that contrastive learning within geoscience applications such as weather forecasting has not been widely applied or explored. This project will focus on exploring this area more thoroughly, creating an end-to-end model that fuses multiple modalities to create robust latent representations. We will incorporate a decoder model within our system that can accurately decode the latent representations, to enable analysis to occur after downstream tasks have been performed. Finally, we will also incorporate sparse and noisy data to create a robust latent representation that is more applicable to the data found in the real world.

Chapter 3

Experimentation

3.1 Dataset

The main dataset to be used in this project will be the WeatherBench 2 Dataset. Specifically, the ERA5 dataset has been chosen to perform initial trials and experiments. ERA5 is the fifth generation European Center for Medium-Range Weather Forecasting (ECMWF) reanalysis dataset and contains time-series data from 1959 onwards. A reanalysis dataset combines observations and model data to create a complete dataset [36]. It therefore provides consistent data on which to perform our experiments. Moreover, it has also been used within the early work on contrastive learning within geoscience [22, 23]. ERA5 consists of 62 data variables (modes) that span the complete longitude and latitude of the globe [10]. Specific variables such as humidity are also measured at multiple pressure levels. The maximum being 13 levels. In addition, ERA5 is measured at various resolutions, allowing the initial work to use lower-dimensional data (64x32) before moving to higher-resolution data (721x1440) [10].

For our initial experiments, we use the following 5 variables: '2m_temperature', 'geopotential', 'u_component_of_wind, 'v_component_of_wind', 'specific_humidity'. All but temperature have associated pressure variables; for ease, we select pressure at level 0 for out initial experiments. We chose to start our experiments with the lowest dimensional data (64x32), which spans the full range of latitude and longitude. This gives data of size $X \in R^{B\times5\times64\times32}$, where each variable is stacked as a channel. This means the total dimensionality of a single data sample is 10240. We show all modes of a randomly selected data sample in figure A.1 appendix A.

ERA5 data is collected at intervals of 6 hours, we resample the data such that we have daily values by averaging across the 4 samples per day. All available values are used from 1959-2022, dedicating 80% of the samples for training and 20% for validation.

We pre-process the data by computing the mean and standard deviation for each data variable across all training samples and standardise the data: $z_i^c = \frac{x_i^c - \mu}{\sigma}$

3.2 Autoencoder

To produce a baseline to compare the contrastive methods against, we first train a simple deterministic autoencoder. We use a ResNet-18 model as the encoder and build a simple decoder to reconstruct the original data. The autoencoder is trained for 300 epochs, using the Adam optimiser [37] and a learning rate of $1e^{-3}$.

We show the training loss curve in figure 3.1 and an example of reconstructed data across all modes

in figure A.2 appendix A. From the curve and example, we can see that the model generally can reconstruct the overall shape of the data well, but misses out on finer details.

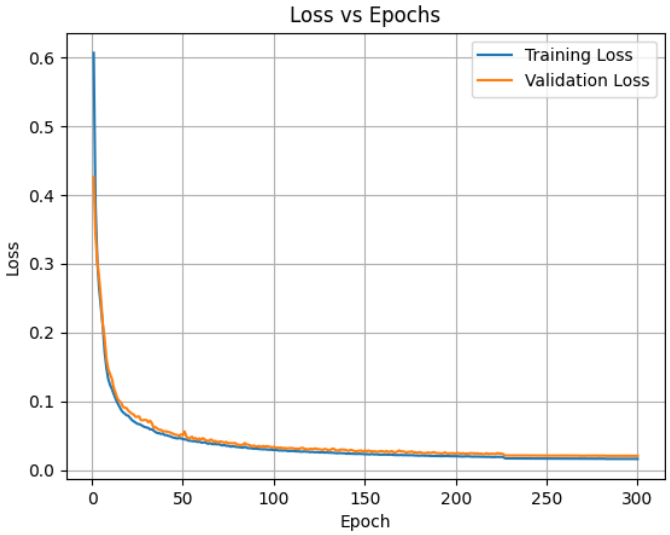

Figure 3.1: Autoencoder Loss Curve

We also show the structure of the embeddings for 20 random validation samples in figure 3.2. We can see that the samples exhibit no structure in the latent space.

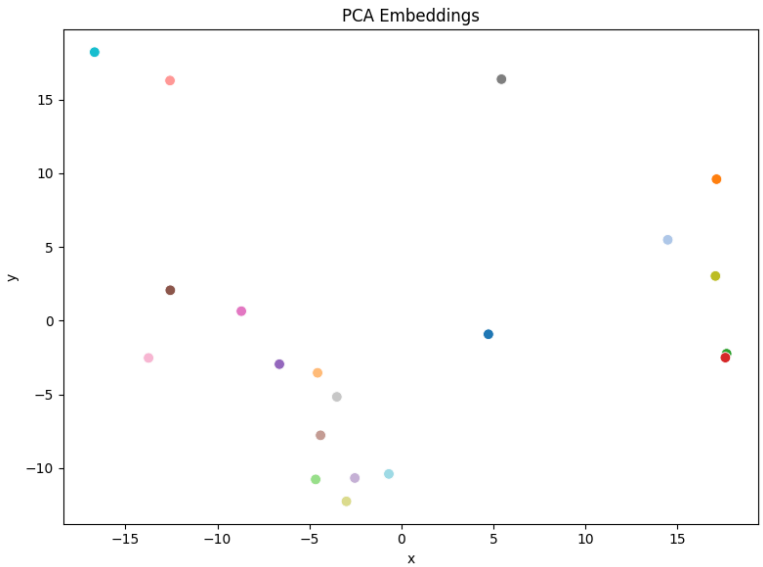

Figure 3.2: Autoencoder PCA Embeddings

3.3 Contrastive Learning Methods

In addition to the autoencoder approach, we also evaluate 3 common self-supervised learning methods: Barlow Twins, SimCLR and Supervised Contrastive Learning. For all methods, we use a simple ResNet-18 encoder and sweep various combinations of hyperparameters to find the best fit for our

data. As the loss value often depends on the individual hyperparameters chosen, we instead evaluate each run by comparing the cosine similarity of the 1000-D embeddings of positive and negative pairs. We also look at the mean variance of the embeddings as well as a TSNE plot of a random batch of validation data to see how well positive pairs cluster in the latent space. We opt to focus our evaluation on the 1000-D embeddings, rather than the projections, as these are used in the downstream task, so the quality is of higher importance than the projections. Results for each method are summarised in table 3.1.

We make no change to the ResNet-18 network, except altering the first layer to accept 5 input channels to conform to our data. The output size is kept as is, giving a latent space of 1000, we chose a projection space of 128. This gives us a total compression of around 10x. A batch size of 128 was used for all trials.

To create positive pairs, we follow the augmentation procedure used by [22]. We sample a random timestep and augment it twice, by resizing to size 160x80 and randomly cropping to 144x72. Each channel in the resulting cropped image is then smoothed with a 5x5 average filter.

We use the Adam optimiser [37], with no weight decay and train each network for 100 epochs.

3.3.1 Barlow Twins

As per the original work [13], we design the projection network such that it uses 3 linear layers, each followed by batch normalisation. The final output used in the loss function is followed by a batch normalisation layer to give the embeddings a mean of 0 over the batch.

For hyperparameters, we tune learning rate and λ, choosing a learning rate of $1e^{-3}$ and λ values of 0.001, 0.01 and 0.1. We find the model that produces the best separation, both quantitatively and qualitatively, has a λ of 0.01. This model produces positive embeddings that have a mean cosine similarity of 0.989, negative embeddings that have a mean cosine similarity of 0.354 and overall embeddings that have a mean variance of 45.6. An example PCA plot of the embeddings is shown in figure 3.3

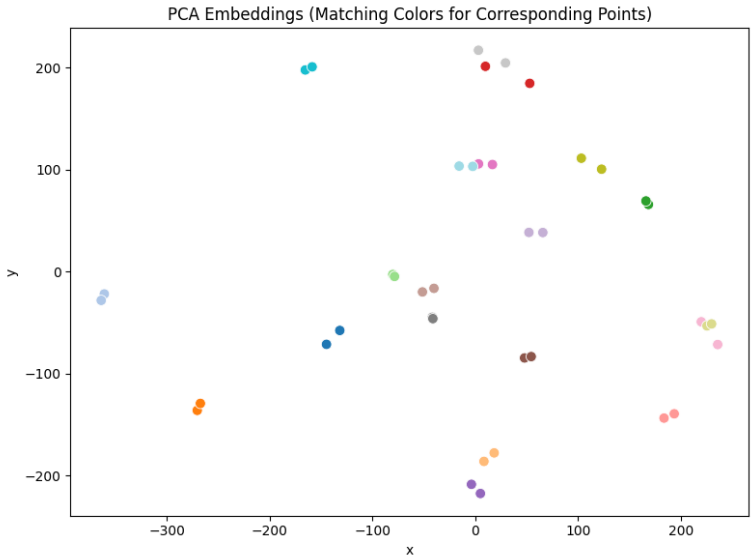

Figure 3.3: Barlow Twins PCA Embeddings

3.3.2 SimCLR

For the SimCLR model, we use the same projection network as in Barlow Twins. We sweep temperature values in the following list, using a learning rate of $1e^{-4}$:
$[0.01, 0.03, 0.05, 0.07, 0.08, 0.1, 0.3]$, finding the best temperature value to be 0.3. This produces positive embeddings that have a mean cosine similarity of 0.999, negative embeddings that have a mean cosine similarity of 0.30 and overall embeddings that have a mean variance of 0.68. An example PCA plot of the embeddings is shown in figure 3.4

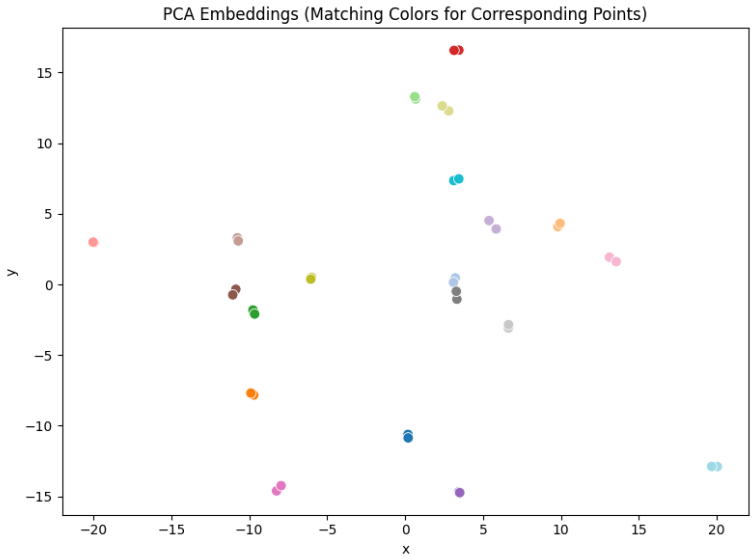

Figure 3.4: SimCLR PCA Embeddings

3.3.3 Supervised Contrastive Learning

As outlined in section 2, supervised contrastive learning utilises labels to increase the amount of positive samples for a given anchor point. To enable this for the ERA5 dataset, we assign pseudo-season labels (winter, spring, summer, autumn) for each sample depending on the month it falls into.

We use the same projection network as in SimCLR and sweep the same hyperparameters, finding the best temperature value to be 0.1. This produces positive embeddings that have a mean cosine similarity of 0.931, negative embeddings that have a mean cosine similarity of 0.23 and overall embeddings that have a mean variance of 44.24. An example PCA plot of the embeddings is shown in figure 3.5.

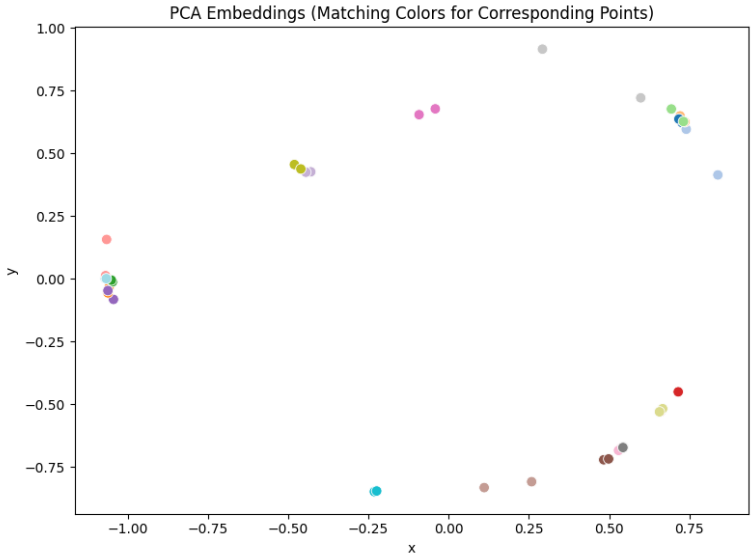

Figure 3.5: Supervised Contrastive Learning PCA Embeddings

Table 3.1: Cosine Similarity and Variance Results

Model	Mean Positive Cosine Similarity	Mean Negative Cosine Similarity	Mean Variance of Embeddings
Barlow Twins	0.989	0.354	45.6
SimCLR	0.999	0.30	0.68
Sup Con	0.931	0.23	44.24

3.3.4 Conclusions

From the evaluated contrastive learning approaches, we determine that Barlow Twins produces the best results. SimCLR produces good separation between positive and negative pairs. Supervised contrastive learning produces good separation between pairs and high variance of embeddings, but as seen from the PCA plots, it tends to produce embeddings that collapse together in latent space. This is expected because supervised contrastive learning has multiple positives for each sample. Since the goal of this project is to create a structured latent space, where ideally we foresee nearby timesteps to be closer to one another, we can deem that supervised contrastive learning is not as suitable for our task as SimCLR or Barlow Twins. Barlow Twins produces embeddings with the highest variance and still good separation. As mentioned in section 2, Barlow Twins is not as sensitive to the batch size and works without explicit negative samples. Henceforth, we will use a Barlow Twins model in our future experiments.

Chapter 4

Project Plan

Following on from section 3, we outline the following project plan. We include steps from the initial experiments for completeness.

1. Implement a range of contrastive learning methods. Barlow Twins, Supervised Contrastive Learning and SimCLR have been chosen. These will be compared to a standard deterministic autoencoder to contrast the additional structure that the contrastive methods produce. **To be completed by April**

2. Following on from step 1, contrastive learning methods will be explored with additional sparsity within the augmentations. **To be completed by May**

    - As in the first step, these methods will be compared against an autoencoder baseline to establish the benefit of the structured latent space that contrastive learning produces over an autoencoder.
    - Unlike step 1, where we evaluated these methods using the quality of the embeddings, these methods will be evaluated using the downstream of weather forecasting.
    - To facilitate the downstream task, a sequence-to-sequence LSTM model will be created to operate in the latent space, focusing on predicting the next timestep, based on the previous 30 timesteps.

3. To create a full comparison with the autoencoder approach, we will next focus on training a decoder on the latent space created by the contrastive learning methods. As discussed in the literature review, this is an active area of research and not as trivial as training a decoder for an autoencoder-style model. **To be completed by May**

4. Once a baseline contrastive encoder-decoder style model has been created, work will be carried out to investigate how to make the model more robust to sparsity and noise in the input. This will enable the model to cope with more variations in the input, making the framework and resulting latent space more robust to changing and varied input data. As of present, the following ideas will be explored. **To be completed by June**

    - Create random mask ratios, to enable the model to align varying masked data to full data.
    - Adding spatially correlated noise to simulate noisy sensor data expected in real-world applications.
    - Incorporating ideas and methodology from contrastive learning for time-series data to create a better-structured latent space to enable better downstream performance.

5. In addition to the downstream task of weather forecasting, additional downstream tasks should be explored, such as data assimilation, autoregressive weather prediction and conditional latent diffusion. **To be completed by July**

6. The penultimate step will investigate if the models trained on lower-dimensional ERA5 data can be naturally extended to higher-resolution data. This is important to test that the model created is scalable to higher-resolution data and robust to various resolution types. Enabling the latent model to work at a higher resolution also enables it to be used in practical scenarios, where forecasting at higher resolutions is desired. **To be completed by July**

7. Finally, additional multi-modal fusion methods will be investigated to create a more robust latent space that better fuses the modes. This area is relatively underexplored within current research, especially for multiple modes and presents an opportunity to explore novelty within this field. **To be completed by August**

Appendix A

Additional Experimentation Figures

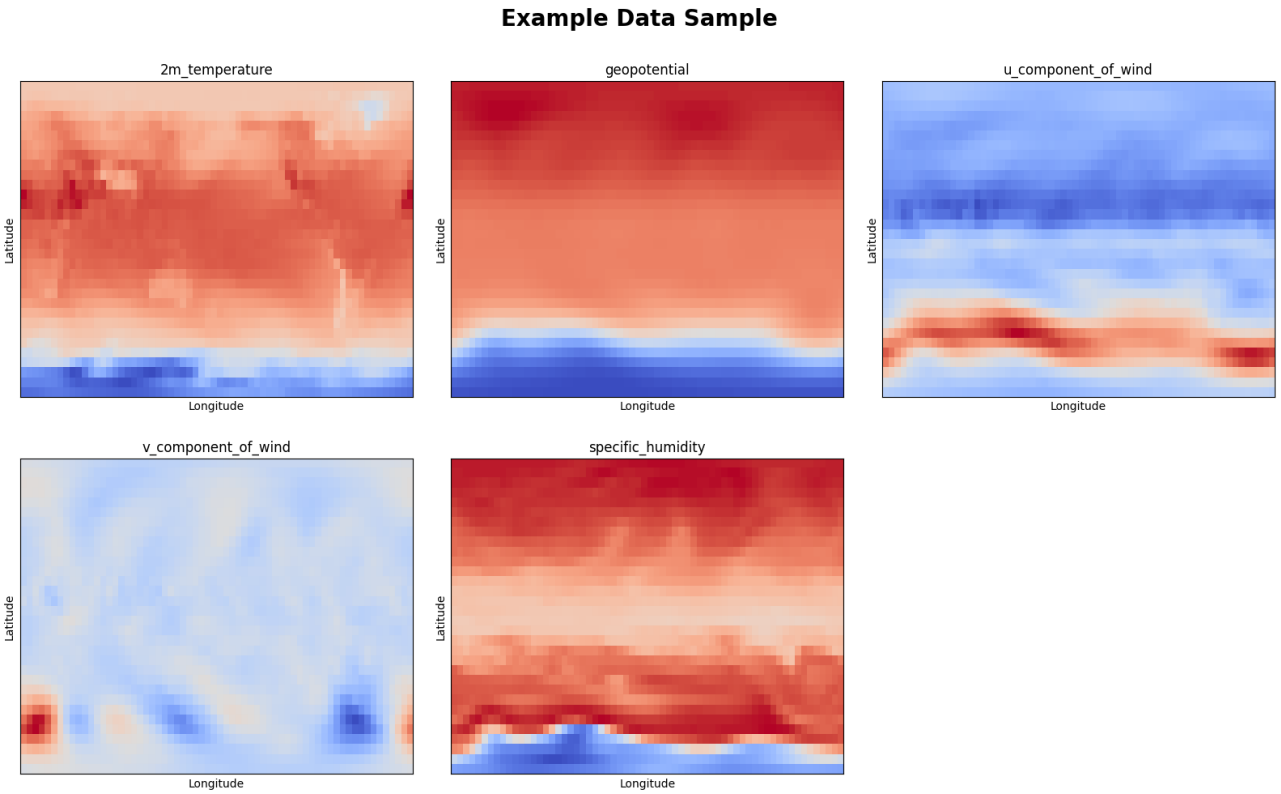

Figure A.1: Example ERA5 64x32 Data Samples

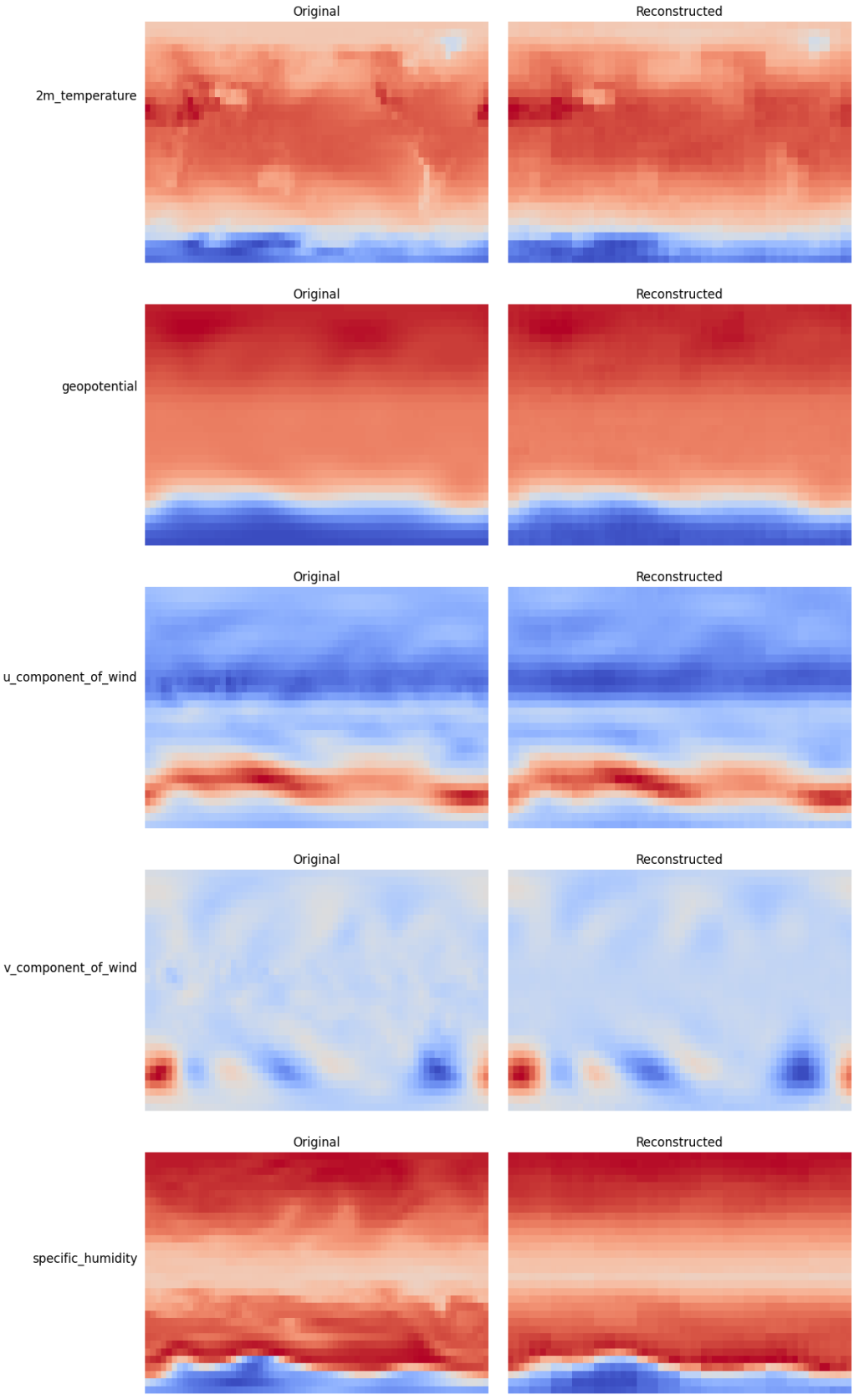

Figure A.2: Autoencoder Reconstruction Results

Bibliography

[1] IBMClimate Data Store. *What is self-supervised learning?* [Online]. Available from: https://www.ibm.com/think/topics/self-supervised-learning;. [Accessed: 2025-04-29]. pages 2

[2] Gui J, Chen T, Zhang J, Cao Q, Sun Z, Luo H, et al. A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. *IEEE Trans Pattern Anal Mach Intell*. 2024;46(12):9052–9071. Available from: https://doi.org/10.1109/TPAMI.2024.3415112. pages 2

[3] Finextra. *Self-supervised Learning The future of Artificial Intelligence* [Online]. Available from: https://www.finextra.com/blogposting/26343/self-supervised-learning-the-future-of-artificial-intelligence;. [Accessed: 2025-04-29]. pages 2

[4] Jaiswal A, Babu AR, Zadeh MZ, Banerjee D, Makedon F. A Survey on Contrastive Self-Supervised Learning. *Technologies*. 2021;9(1). Available from: https://www.mdpi.com/2227-7080/9/1/2. pages 2

[5] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org; 2020. . pages 2, 4

[6] Yue Z, Wang Y, Duan J, Yang T, Huang C, Tong Y, et al. TS2Vec: Towards Universal Representation of Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2022 Jun;36(8):8980-7. Available from: https://ojs.aaai.org/index.php/AAAI/article/view/20881. pages 2

[7] Dexter. *Leveraging 500TB of weather forecasts: A paradigm shift to array-native infrastructure* [Online]. Available from: https://dexterenergy.ai/news/leveraging-500tb-of-weather-forecasts-a-paradigm-shift-to-array-native-infrastructure/;. [Accessed: 2025-04-29]. pages 2

[8] Price I, Sanchez-Gonzalez A, Alet F, Andersson TR, El-Kadi A, Masters D, et al. Probabilistic weather forecasting with machine learning. *Nature*. 2025;637(8044):84-90. Available from: https://www.nature.com/articles/s41586-024-08252-9. pages 2

[9] Bi K, Xie L, Zhang H, Chen X, Gu X, Tian Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*. 2023;619(7970):533-8. pages 2

[10] WeatherBench 2. *WeatherBench 2 Data Guide* [Online]. Available from: https://weatherbench2.readthedocs.io/en/latest/data-guide.html;. [Accessed: 2025-04-28]. pages 2, 17

[11] Agabin A, Prochaska JX, Cornillon PC, Buckingham CE. Mitigating Masked Pixels in a Climate-Critical Ocean Dataset. *Remote Sensing*. 2024;16(13). Available from: https://www.mdpi.com/2072-4292/16/13/2439. pages 3, 12, 15

[12] Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Red Hook, NY, USA: Curran Associates Inc.; 2020. . pages 5

[13] Zbontar J, Jing L, Misra I, LeCun Y, Deny S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In: Meila M, Zhang T, editors. *Proceedings of the 38th International Conference on Machine Learning*. vol. 139 of Proceedings of Machine Learning Research. PMLR; 2021. p. 12310-20. Available from: https://proceedings.mlr.press/v139/zbontar21a.html. pages 5, 19

[14] Zhao W, Fan L. Time-series representation learning via Time-Frequency Fusion Contrasting. *Frontiers in Artificial Intelligence*. 2024;Volume 7 - 2024. Available from: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1414352. pages 7

[15] Lee S, Park T, Lee K. Soft contrastive learning for time series arXiv [Preprint]; 2024 Version 3. Available from: https://arxiv.org/abs/2312.16424. pages 8

[16] Shamba AK, Bach K, Taylor G. Dynamic Contrastive Learning for Time Series Representation arXiv [Preprint]; 2024 Version 1. Available from: https://arxiv.org/abs/2410.15416. pages 8

[17] Tang Q, Liang J, Zhu F. A comparative review on multi-modal sensors fusion based on deep learning. *Signal Processing*. 2023;213:109165. Available from: https://www.sciencedirect.com/science/article/pii/S0165168423002396. pages 9

[18] Dufumier B, Castillo-Navarro J, Tuia D, Thiran JP. What to align in multimodal contrastive learning? In: *International Conference on Learning Representations*; 2025. Available from: https://arxiv.org/abs/2409.07402. pages 9

[19] Guo H, Xu X, Wu H, Liu B, Xia J, Cheng Y, et al. Multi-scale and multi-modal contrastive learning network for biomedical time series. *Biomedical Signal Processing and Control*. 2025;106:107697. Available from: https://www.sciencedirect.com/science/article/pii/S1746809425002083. pages 9

[20] Bahaduri B, Ming Z, Feng F, Mokraou A. Multimodal Transformer Using Cross-Channel attention for Object Detection in Remote Sensing Images arXiv [Preprint]; 2024 Version 3. Available from: https://arxiv.org/abs/2310.13876. pages 10

[21] Saioni M, Giannone C. Multimodal Attention Is All You Need. In: Dell'Orletta F, Lenci A, Montemagni S, Sprugnoli R, editors. *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*. Pisa, Italy: CEUR Workshop Proceedings; 2024. p. 873-9. Available from: https://aclanthology.org/2024.clicit-1.94/. pages 10

[22] Wang L, Li Q, Lv Q. Self-Supervised Classification of Weather Systems Based on Spatiotemporal Contrastive Learning. *Geophysical Research Letters*. 2022;49(15):e2022GL099131. E2022GL099131 2022GL099131. Available from: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022GL099131. pages 10, 15, 17, 19

[23] Wang C, Yang N, Li X. Advancing forecasting capabilities: A contrastive learning model for forecasting tropical cyclone rapid intensification. *Proceedings of the National Academy of Sciences*. 2025;122(4):e2415501122. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.2415501122. pages 10, 15, 17

[24] Ballard T. Contrastive Learning for Climate Model Bias Correction and Super-Resolution. In: *AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*; 2022. Available from: https://www.climatechange.ai/papers/aaaifss2022/10. pages 11

[25] Shi Z, Zheng H, Dong J. Spatiotemporal self-supervised predictive learning for atmospheric variable prediction via multi-group multi-attention. *Knowledge-Based Systems*. 2024;300:112090. Available from: https://www.sciencedirect.com/science/article/pii/S095070512400724X. pages 11

[26] Han T, Chen Z, Guo S, Xu W, Bai L. CRA5: Extreme Compression of ERA5 for Portable Global Climate and Weather Research via an Efficient Variational Transformer arXiv [preprint]; 2024 Version 2. Available from: https://arxiv.org/abs/2405.03376. pages 11, 15

[27] Goh E, Yepremyan AR, Wang J, Wilson B. MAESSTRO: Masked Autoencoders for Sea Surface Temperature Reconstruction under Occlusion. *EGUsphere*. 2023;2023:1-20. Available from: https://egusphere.copernicus.org/preprints/2023/egusphere-2023-1385/. pages 12, 15

[28] He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked Autoencoders Are Scalable Vision Learners. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022. p. 15979-88. pages 12

[29] Vandal TJ, Duffy K, McDuff D, Nachmany Y, Hartshorn C. Global atmospheric data assimilation with multi-modal masked autoencoders arXiv [preprint]; 2024 Version 1. Available from: https://arxiv.org/abs/2407.11696. pages 12, 15

[30] Kadeethum T, Ballarin F, O'Malley D, Choi Y, Bouklas N, Yoon H. Reduced order modeling for flow and transport problems with Barlow Twins self-supervised learning. *Sci Rep*. 2022 Nov;12(1):20654. pages 13, 15

[31] Kadeethum T, Silva VLS, Salinas P, Pain CC, Yoon H. Boosting Barlow Twins Reduced Order Modeling for Machine Learning-Based Surrogate Models in Multiphase Flow Problems. *Water Resources Research*. 2024;60(10). Available from: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023WR035778. pages 14

[32] Quetin S, Ghosh T, Maleki F. Should we pre-train a decoder in contrastive learning for dense prediction tasks? arXiv [preprint]; 2025 Version 1. Available from: https://arxiv.org/abs/2503.17526. pages 14

[33] Dippel J, Vogler S, Höhne J. Towards Fine-grained Visual Representations by Combining Contrastive Learning with Image Reconstruction and Attention-weighted Pooling. *CoRR*. 2021. Available from: https://arxiv.org/abs/2104.04323. pages 14, 15

[34] Yao Y, Desai N, Palaniswami M. Masked Contrastive Representation Learning arXiv [preprint]; 2022 Version 1. Available from: https://arxiv.org/abs/2211.06012. pages 14, 15

[35] He K, Fan H, Wu Y, Xie S, Girshick R. Momentum Contrast for Unsupervised Visual Representation Learning. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2020. p. 9726-35. pages 15

[36] Climate Data Store. *ERA5 hourly data on single levels from 1940 to present* [Online]. Available from: `https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview`;. [Accessed: 2025-04-28]. pages 17

[37] Kingma D, Ba J. *Adam: A Method for Stochastic Optimization*. International Conference on Learning Representations. 2014 12. pages 17, 19