# CONTRASTIVE SENSOR TRANSFORMER FOR PREDICTIVE MAINTENANCE OF INDUSTRIAL ASSETS

*Zaharah Bukhsh*

School of Industrial Engineering, Eindhoven University of Technology, Eindhoven, Netherlands

## ABSTRACT

Reliable fault detection of industrial assets is crucial for the success of smart industry. Standard (supervised) data-driven approaches for fault detection perform poorly due to lack of semantic annotations of condition monitoring data, and novel fault types introduced at the operational time. We propose *Contrastive Sensor Transformer* (CST) , a novel approach for learning useful representations for robust fault identification without using task-specific labels[1]. We explore sensor transformations for pre-training in a self-supervised contrastive manner, where the similarity between the original signal instance and its augmented version is maximized. We demonstrate that the powerful transformer architecture applied to condition monitoring data learns highly useful embedding that perform exceptionally well for fault detection in low labeled data regimes and for the identification of novel fault types. Our approach obtains an average of 75% accuracy on the considered bearing benchmark datasets while using less than 2% of the labeled instances.

***Index Terms***— fault detection, predictive maintenance, self-supervised learning, contrastive loss, transformer

## 1. INTRODUCTION

Unplanned downtime of industrial assets is costly, leading to inadequate service levels, loss of resources, and cascading impact on supply chains. Prognostic health management (PHM) of complex industrial systems is crucial to transition towards data-driven *predictive maintenance* solutions in the context of smart industry [1]. The industrial systems are subject to multiple critical faults that must be detected timely, (fault detection) and distinguished into respective types (fault diagnostic) for an appropriate maintenance intervention. For PHM, sensors are used to monitor the condition of machinery continuously. The availability of condition monitoring data triggered the development of several data-driven fault diagnostic methods [2, 3, 4]. However, most of the proposed approaches assume the abundance of semantically labeled data for supervised learning, as also noted in [5]. Additionally, these methods are sensitive to distribution shifts and perform poorly when encountered with novel fault types and varying severity levels.

Industrial assets function under complex operating conditions in real-world settings, making them susceptible to novel and critical faults. Even though a large amount of condition monitoring data is collected, the manual annotation of data is costly, error-prone and labour intensive [6]. Therefore, the huge amount of data remain under-utilized due to the unavailability of semantic labels or lack of sufficient samples belonging to critical fault types resulting in *low labeled data regime* further on referred to as *low-data regime* for simplicity. The supervised learning approaches towards fault diagnostic perform inadequately under such limitations. On the contrary, other semi-supervised methods such as anomaly detection and clustering either use only labeled or unlabeled data to detect deviating patterns or segment novel types of faults [7, 8, 9]. Likewise, few prior approaches that do account for critical fault types by effectively using the condition monitoring data, fail to address the problem of fault detection in extremely low-data regimes or learning from limited labeled instances [5, 10].

Self-supervised learning provide an effective avenue to learn semantic features from the raw unlabeled input data [11]. The general-purpose representations can be used to address data distribution problem using clustering and to solve the end-task having extremely fewer labeled examples. In this work, we propose Contrastive Sensor Transformer (CST), a novel approach for robust fault detection using raw condition monitoring data. The CST is based on a contrastive learning framework for pre-training a sensor transformer model using raw data without the need for explicit semantic labels. Instead, CST draws a supervisory signal from the data itself by leveraging signal transformations [12] and instance discrimination contrastive objective [11]. Combined with powerful transformer architecture and contrastive learning, CST learns broadly useful representations from unlabeled data that can be leveraged for several related downstream tasks.

We demonstrate the effectiveness of learned representations on standard downstream tasks of fault diagnostic under low-data regimes and identification of novel types of faults. We perform the evaluation on bearing benchmark datasets provided by Case Western Reserve University (CWRU) [13]

[1]Our code is available at https://github.com/Zaharah/Contrastive-Sensor-Transformer

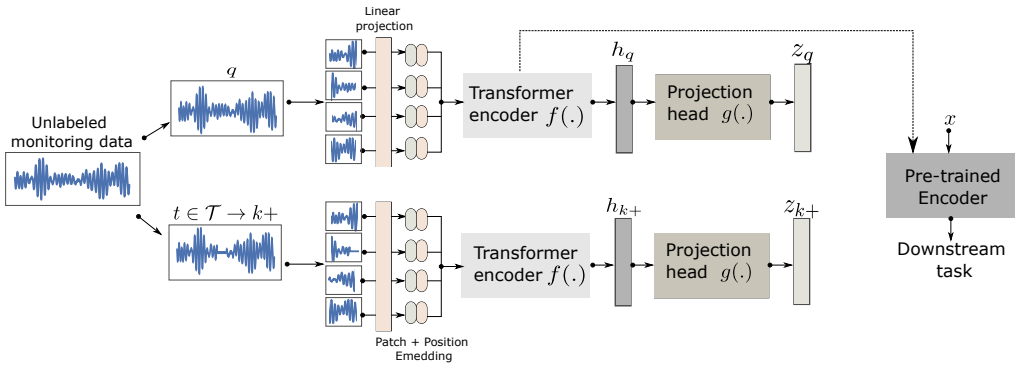


**Fig. 1**: **Illustration of Contrastive Sensor Transformer.** A single transformation $t \in \mathcal{T}$ is applied to input signal to prepare a pair of correlated views as input $q$ and $k+$. The input splitted into multiple fixed-size patches, to form the learnable patch and positional embedding and serve as input to our sensor transformer encoder $f(.)$. The encoder is trained with contrastive loss in an end-to-end manner to maximise similarity between related pairs and minimizing it for rest of the examples in a batch. The pre-trained encoder is then fine-tuned to solve specific downstream task.

and Paderborn University [14]. Our approach outperforms the baselines and achieves above 85% and 65% accuracy with only 2% of labeled CWRU and 0.14% of labeled KAT examples, respectively. Similarly, CST achieves higher homogeneity and completeness score for detection of novel fault types compared to baselines.

## 2. METHOD

*Contrastive sensor transformer* (CST) learns general-purpose representations using the transformer architecture and contrastive loss function. The contrastive learning framework maximizes the agreement between the semantic representations of an input signal segment and its augmented (or transformed) version. The CST is pre-trained using only unlabeled condition monitoring data, which is then fine-tuned for specific downstream tasks. Here, we leverage the CST for downstream tasks of fault detection and novel fault types identification. Our method of learning useful representations is generic and can be used to solve several related downstream tasks, such as remaining useful life estimation.

The CST leverages the recent advancements of training deep neural networks by solving pre-text tasks, especially, in audio [12], computer vision [11, 15], and speech [16] and powerful transformer architecture for NLP [17]. An overview of CST is depicted in Figure 1. We transform an input signal segment using transformations to create similar views for contrastive learning from unlabeled data. We explore nine different transformations, including scaling, permutation, time-wrapping, magnitude wrapping, shifting, random noise, flipping, negation, and masking and found them to be highly effective in learning from accelerometer signal [12]. Here, we consider the input segment as anchor or query $q$ and its perturbed version as positive key $t \in \mathcal{T} \rightarrow k+$ , while all other examples called targets $\mathbb{K}$ in a batch are considered negative as $\mathbb{K} \setminus \{k+\}$.

Instead of dominant RNN and LSTM models, we propose to use transformer architecture as a neural network-based encoder $f(.)$ for diagnostic of faults. Due to its self-attention mechanism, transformer architecture can better capture the complex dynamics of input sensor data compared to traditional sequence models. Typically, the transformer uses a $1D$ sequence of embedding as input for NLP tasks. To use transformers for continuous monitoring data, we split an instance into multiple fixed-sized patches as $x \in \mathbb{R}^{N \times (S.C)}$, where $S$ is signal size, $C$ is channel size, and $N = S/p$ is a number of patches computed based on signal size and chosen patch size as a same strategy is found to be effective in learning representations from images [18]. The patches are mapped to trainable linear projections with a constant latent vector size $D$, forming patch embedding. The positional embedding is added to the patch embedding to retain positional information among segments. The combination of these serves as an input for encoder $f(.)$. The encoder transforms the input $x$ into latent representation $h = f(x) \in \mathbb{R}^d$.

A small network $g$ maps $h$ to $z = g(h)$, where a bilinear similarity comparison is performed between query and targets [19]. We randomly sample a batch of examples to form targets and compute the contrastive loss function as follows:

$$\mathcal{L} = -log \frac{\exp\left(q^T W k+\right)}{\sum_{i=0}^{K-1} \exp q^T W k_i}$$

where $W$ are the learnable bilinear parameters. The contrastive loss function aims to maximize the similarity between the related pairs, i.e., $q$ and $k+$, and minimize it between the unrelated pairs, i.e., $q$ and $\mathbb{K} \setminus \{k+\}$. In other words, for the training of our model, all the positives are used as negatives for other anchors in the batch. After self-supervised training, the projection head $g(.)$ is ignored, and learned representations $h$ are used and fine-tuned for downstream tasks. In the

following section, we evaluate the usefulness of learned embedding on downstream tasks.

## 3. EXPERIMENTS

### 3.1. Datasets and Tasks

We evaluate the learning capabilities of our approach on open benchmark bearing datasets, i.e., CWRU [13] and KAT [14]. Both of these datasets have been used widely for different fault diagnostic and detection tasks [20]. However, to the best of our knowledge, these datasets are not evaluated for the cases, where the semantically labeled data is scarce, and novel fault types are likely to be introduced at an operational time. We found only one study that provides a setup for CWRU to detect novel types of faults [10]. Table 1 provides an overview of classes in datasets for specific fault severity and fault type. The accelerometer sensor collected the CWRU vibration data at locations near to and far-off the motor bearing. The data is recorded with the sampling frequency of $48$kHz and divided into sequence lengths of $512$ points having two channels based on the measurement location. The KAT dataset provides high-resolution data consisting of six healthy and $26$ damaged bearings. The data is collected at a sampling rate of 64kHz, which is further segmented into fixed-length windows of 1200 points.

**Table 1**: **Classes statistics in the considered datasets with respect to fault severity & type.** Fault severity refers to the damage level applied on the bearings. Except for healthy state (H), the fault types present the specific damage location such as ball fault (B), inner race fault (IR), & outer race fault (OR).

| **Type of classes in CWRU dataset.** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Class** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Fault severity (µm) | - | 7 | 14 | 21 | 7 | 14 | 21 | 7 | 14 | 21 |
| Fault type | H | B | IR | OR | B | IR | OR | B | IR | OR |
| **Type of classes in KAT dataset.** | | | | | | | | | | |
| **Class** | **0** | **1** | **2** | **3** | **4** | | | | | |
| Fault severity (mm) | - | $\leq 2$ | $\leq 2$ | $>2$ | $>2$ | | | | | |
| Fault type | H | IR | OR | IR | OR | | | | | |

We evaluate the usefulness of self-supervised learned embedding on two downstream tasks. We performed fault detection using only 5 to 100 labeled samples per fault class in the first task. The second task evaluates the robustness of embedding for novel types of fault introduced at the operational time. We performed stratified sampling to split the dataset into 70% training and 30% of the test set. The `CST` learns the generic representations using the training set without annotated labels.

### 3.2. Model Architecture and Implementation details

The `CST` takes a non-overlapping segments of signal $x \in \mathbb{R}^{N\times(S.C)}$ as an input (see Method section). The transformer encoder consists of $4$ multi-headed projection heads followed by $64$ units of MLP blocks. We apply global max-pooling followed by Layer normalization [21] and *tanh* activation to obtain the embedding for pre-training network. The embedding are further passed to the contrastive model having a Bilinear similarity [19] to compute the loss. The model is trained with negative log-likelihood loss and ADAM optimizer with a learning rate of $0.0001$ for $100$ epochs.

For baselines, we consider Convolutional Autoencoder (Conv-AE) and Convolutional Fully Supervised (Conv-FS) models. The encoder part of Conv-AE consists of four 1D convolutional layers with increasing filters from $16$ to $128$, having kernel size of $7$, stride $2$, and ReLU as a non-linear activation function. Each of the encoder layers are followed by a dropout layer with $0.1$ rate. The decoder consists of 1D Transposed convolutional layers with decreasing units from $128$ to $16$. Similar to an encoder, each layer is followed by a dropout layer with a $0.1$ rate. The Conv-AE is pre-trained with an ADAM optimizer having a learning rate of $0.001$ and mean squared error as loss function. The pre-training is performed for $100$ epochs with a batch size of $24$. For the downstream task, we keep the encoder and applied global max-pooling followed by a single dense classification layer. The Conv-FS model utilizes the encoder architecture of Conv-AE followed by global max-pooling and a dense layer to directly learn to solve considered task.

### 3.3. Results

**Faults diagnostic under low-data regimes**

Figure 2 reports the test accuracy of `CST` and other baselines for the fault diagnostic with limited amount of labeled data. The pre-trained models are fine-tuned using fewer labeled samples from each class. Our approach consistently outperforms the baselines with a significant margin specifically for the KAT dataset. It achieves above 65% accuracy using only 100 samples per class, forming 500 samples in total only. For the CWRU dataset, `CST` achieves above 70% accuracy when using only ten samples per class, constituting 100 labeled samples only. The performance of `CST` is comparable to baselines, when using 20 and 50 labeled samples per class. We also study the impact of different transformations applied to an input signal to create its perturbed version ($k+$) for the self-supervised pre-training. For each downstream training setting, we eliminate a single transformation type as follows: $\mathbb{T}\backslash t = \big\{x : x \in \mathbb{T}, \sim(x \in \{t\})\big\}$, performed pre-training of a `CST` model, and compared the impact on test accuracy given in Table 2. The difference in performance is notable when the training data size is extremely small, consisting of only 5 to 10 samples per class. However, the difference among various test accuracy is not prominent when considering the higher number of samples per class. We found that eliminating random noise, i.e.$\mathbb{T} \setminus t_{noise}$ results in the highest increase in overall test accuracy.

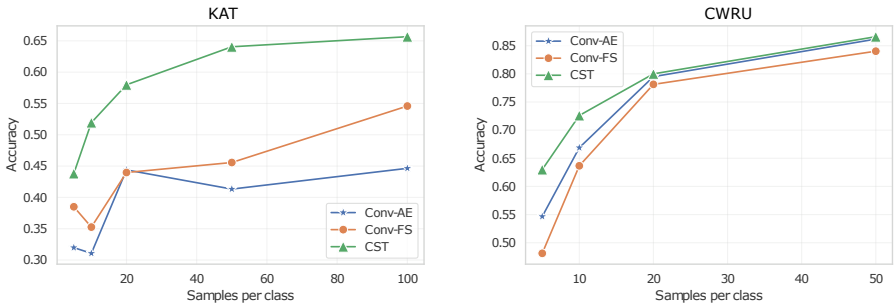


**Fig. 2**: Performance of CST and other baselines for fault diagnostic using fewer labeled samples from each fault class.

**Table 2**: Impact of transformations on test accuracy for CWRU with different number of training samples.

| Samples/class | $\mathbb{T} \setminus t_{scale}$ | $\mathbb{T} \setminus t_{permute}$ | $\mathbb{T} \setminus t_{time}$ | $\mathbb{T} \setminus t_{magnitude}$ | $\mathbb{T} \setminus t_{shift}$ | $\mathbb{T} \setminus t_{noise}$ | $\mathbb{T} \setminus t_{flip}$ | $\mathbb{T} \setminus t_{negate}$ | $\mathbb{T} \setminus t_{cutout}$ | $\mathbb{T}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | **0.687** | 0.578 | 0.664 | 0.674 | 0.594 | 0.637 | 0.584 | 0.677 | 0.659 | 0.630 |
| 10 | 0.754 | 0.714 | 0.721 | 0.708 | 0.690 | 0.730 | 0.685 | 0.751 | **0.774** | 0.726 |
| 20 | 0.791 | 0.783 | 0.762 | 0.760 | 0.781 | **0.815** | 0.770 | 0.784 | 0.797 | 0.800 |
| 50 | 0.845 | 0.841 | 0.853 | 0.856 | 0.862 | **0.875** | 0.859 | 0.852 | 0.858 | 0.866 |
| 100 | 0.881 | 0.875 | 0.848 | **0.896** | **0.891** | **0.891** | 0.867 | 0.874 | 0.882 | 0.836 |

**Detection of novel fault types**

We study the ability of CST to distinguish novel fault types that are introduced at operational times. We follow the experimental setup and evaluation measures introduced in [10]. We create subsets of training data depending on the fault types and fault severity. The training subsets are created as follows: $\mathbb{D} \setminus D_{ft} = \{y : y \in \mathbb{D}, \sim(y \in \{ft\})\}$, where $ft$ is fault type and $y$ is class label. The subset of training data contains all fault types except for the one considered novel for evaluation purposes. For example, $\mathbb{D} \setminus D_B$ consist of samples from all fault types except for the samples belonging to ball fault. We apply a K-means clustering algorithm on top of the embedding extracted from the pre-trained CST classifier. The similar approach was followed for baselines. We report adjusted mutual information (AMI), homogeneity, and completeness to measure the non-linear similarities among the clusters. Table 3 presents the result of unsupervised clustering using the Kmeans algorithm. The method aims to find cluster centroid with the minimum variance within classes. We found that for both datasets, the CST representations performed predominately well compared to baselines. The higher homogeneity and completeness score warrants that the data points belonging to one class are assigned to the same cluster. The CST obtain higher AMI scores which shows that the clustering has a large number of groups.

## 4. CONCLUSIONS

We propose *Contrastive Sensor Transformer* to address prognostic and health management tasks for effective predictive maintenance of industrial assets. Our approach achieves remarkable performance for fault diagnostic on considered datasets using less than 2% of labeled examples only. Similarly, the proposed method performs remarkably well to detect novel types of fault introduced at inference time. We conclude that the contrastive learning combined with transformer architecture provides a promising avenue for health prognostic of assets for label-efficient learning, fault detection, and remaining useful life estimation.

**Table 3**: **Unsupervised clustering to detect novel types of fault introduced at operational time.**

| **Model** | **Adjusted mutual information** | **Homogeneity** | **Completeness** |
|---|---|---|---|
| Result on **KAT** test set when training was performed on $\mathbb{D} \setminus D_{IR}$ | | | |
| Conv-FS | 0.095 | 0.064 | 0.184 |
| Conv-AE | 0.097 | 0.065 | 0.187 |
| CST | **0.544** | **0.452** | **0.684** |
| Result on **KAT** test set when training was performed on $\mathbb{D} \setminus D_{OR}$ | | | |
| Conv-FS | 0.156 | 0.126 | 0.205 |
| Conv-AE | 0.141 | 0.114 | 0.188 |
| CST | **0.283** | **0.238** | **0.350** |
| Result on **CWRU** test set when training was performed on $\mathbb{D} \setminus D_B$ | | | |
| Conv-FS | 0.764 | 0.698 | 0.844 |
| Conv-AE | 0.754 | 0.690 | 0.834 |
| CST | **0.772** | **0.707** | **0.851** |
| Result on **CWRU** test set when training was performed on $\mathbb{D} \setminus D_{IR}$ | | | |
| Conv-FS | 0.749 | 0.677 | 0.838 |
| Conv-AE | 0.682 | 0.616 | 0.764 |
| CST | **0.788** | **0.715** | **0.877** |
| Result on **CWRU** test set when training was performed on $\mathbb{D} \setminus D_{OR}$ | | | |
| Conv-FS | 0.667 | 0.587 | 0.772 |
| Conv-AE | 0.684 | 0.608 | 0.782 |
| CST | **0.718** | **0.658** | **0.792** |

## 5. REFERENCES


[1] Insun Shin, Junmin Lee, Jun Young Lee, Kyusung Jung, Daeil Kwon, Byeng D Youn, Hyun Soo Jang, and Joo-Ho Choi, "A framework for prognostics and health management applications toward smart manufacturing systems," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 5, no. 4, pp. 535–554, 2018.

[2] Fei Shen, Chao Chen, Ruqiang Yan, and Robert X Gao, "Bearing fault diagnosis based on svd feature extraction and transfer learning classification," in *2015 Prognostics and System Health Management Conference (PHM)*. IEEE, 2015, pp. 1–6.

[3] Olivier Janssens, Viktor Slavkovikj, Bram Vervisch, Kurt Stockman, Mia Loccufier, Steven Verstockt, Rik Van de Walle, and Sofie Van Hoecke, "Convolutional neural network based fault detection for rotating machinery," *Journal of Sound and Vibration*, vol. 377, pp. 331–345, 2016.

[4] Jinhao Lei, Chao Liu, and Dongxiang Jiang, "Fault diagnosis of wind turbine based on long short-term memory networks," *Renewable energy*, vol. 133, pp. 422–432, 2019.

[5] Manuel Arias Chao, Bryan T Adey, and Olga Fink, "Implicit supervision for fault detection and segmentation of emerging fault types with deep variational autoencoders," *Neurocomputing*, vol. 454, pp. 324–338, 2021.

[6] Jin Yuan and Xuemei Liu, "Semi-supervised learning and condition fusion for fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 38, no. 2, pp. 615–627, 2013.

[7] Roozbeh Razavi-Far, Ehsan Hallaji, Maryam Farajzadeh-Zanjani, and Mehrdad Saif, "A semi-supervised diagnostic framework based on the surface estimation of faulty distributions," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1277–1286, 2018.

[8] Gabriel Michau and Olga Fink, "Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer," *Knowledge-Based Systems*, vol. 216, pp. 106816, 2021.

[9] Changgen Li, Liang Guo, Hongli Gao, and Yi Li, "Similarity-measured isolation forest: Anomaly detection method for machine monitoring data," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[10] Katharina Rombach, Gabriel Michau, and Olga Fink, "Contrastive learning for fault detection and diagnostics in the context of changing operating conditions and novel fault types," *Sensors*, vol. 21, no. 10, 2021.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[12] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien, "Multi-task self-supervised learning for human activity detection," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, pp. 1–30, 2019.

[13] Wade A Smith and Robert B Randall, "Rolling element bearing diagnostics using the case western reserve university data: A benchmark study," *Mechanical systems and signal processing*, vol. 64, pp. 100–131, 2015.

[14] Christian Lessmeier, James Kuria Kimotho, Detmar Zimmer, and Walter Sextro, "Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification," in *PHM Society European Conference*, 2016, vol. 3.

[15] Michael Laskin, Aravind Srinivas, and Pieter Abbeel, "Curl: Contrastive unsupervised representations for reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5639–5650.

[16] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[20] Dhiraj Neupane and Jongwon Seok, "Bearing fault detection and diagnosis using case western reserve university dataset with deep learning approaches: A review," *IEEE Access*, vol. 8, pp. 93155–93178, 2020.

[21] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.