Superlet-MAE: Self-Supervised Masked Autoencoding for Sleep Staging Using Single-Channel EEG

Jeong-Yun Cha
Department of Brain and Cognitive Engineering
Interdisciplinary Program in Precision Public Health
Korea University
Seoul, Republic of Korea
chajy1212@korea.ac.kr

Choel-Hui Lee
Department of Brain and Cognitive Engineering
Interdisciplinary Program in Precision Public Health
Korea University
Seoul, Republic of Korea
dlcjfgmlnasa28@korea.ac.kr

Hakseung Kim
Department of Brain and Cognitive Engineering
Korea University
Seoul, Republic of Korea
mkhsm@korea.ac.kr

Dong-Joo Kim*
Department of Brain and Cognitive Engineering
Korea University
Seoul, Republic of Korea
dongjookim@korea.ac.kr

***Abstract*—Sleep stage classification is essential for diagnosing sleep disorders, but traditional polysomnography (PSG) is time-consuming and labor-intensive and subject to inter-rater variability, limiting its reliability. We propose a self-supervised learning (SSL) framework that integrates the Superlet Transform (SLT), offering high-resolution time-frequency analysis, with a Masked Autoencoder (MAE) architecture. Single-channel EEG signals (i.e., Fpz-Cz) from the Sleep-EDF dataset were transformed into Superlet scalograms and used for masked reconstruction pre-training. Our method is the first to combine Superlet with MAE for EEG representation learning, enabling robust feature extraction from unlabeled data. Experimental results show that the proposed approach outperforms conventional transforms such as Short-Time Fourier Transform (STFT) and Continuous Wavelet Transform (CWT), achieving state-of-the-art performance in sleep staging. These findings highlight the potential of Superlet-based SSL for scalable and accurate sleep analysis. The source code is available at https://github.com/chajy1212/Superlet-MAE**

***Index Terms*—superlet, masked autoencoder, self-supervised learning, electroencephalography, automatic sleep staging**

I. Introduction

Sleep is an essential biological process for both mental and physical recovery, as well as for maintaining physiological homeostasis [1]. Despite its importance, a large portion of the population suffers from sleep disorders [2]. Polysomnography (PSG) is widely used in clinical practice for accurate diagnosis and assessment [3]. PSG records a wide range of physiological signals during sleep, including electroencephalography (EEG), electrooculography, electromyography, and electrocardiography. These signals are manually annotated into sleep stages by experts, following guidelines such as those provided by the American Academy of Sleep Medicine (AASM) [4]. However, manual scoring is labor-intensive, time-consuming, and subject to inter-rater variability, which can compromise the reliability of the diagnosis [5].

To address these limitations, many studies have explored deep learning-based approaches for automatic sleep stage classification [6]–[8]. In particular, recent research has increasingly focused on using only single-channel EEG signals to enhance usability and reduce hardware complexity [8]–[10]. Notable models such as DeepSleepNet [8], AttnSleep [9], and SleepExpertNet [10] employ various architectures—including convolutional neural networks, long short-term memory networks, and Transformers—to learn temporal and frequency-related features from EEG signals. However, since these models are based on supervised learning, they require large-scale labeled datasets and often suffer from poor generalization to unseen data [5].

In response to these challenges, self-supervised learning (SSL) has emerged as a promising alternative [11]. SSL enables the learning of meaningful representations without the need for labeled data and is typically divided into two major paradigms: contrastive learning [12], [13] and masked prediction [14]. This approach is particularly well-suited for applications like sleep EEG analysis, where large amounts of data can be collected with relative ease, but labeling is costly and labor-intensive. Indeed, several studies have explored contrastive learning on single-channel EEG data and demonstrated its potential [15]–[17]. However, the performance of these methods is often unstable due to their sensitivity to EEG data augmentation strategies and backbone network architectures. Moreover, studies utilizing masked prediction in this domain remain relatively scarce [14], [18].

In this study, we propose the first SSL framework that integrates Superlet [19]—a method capable of high-resolution time-frequency representation (TFR) of EEG signals—with the Masked Autoencoder (MAE) architecture. While prior studies have primarily employed traditional transforms such as the Short-Time Fourier Transform (STFT) [20] or the Continuous Wavelet Transform (CWT) [21], Superlet achieves superior time and frequency resolution simultaneously by adaptively combining wavelets of varying orders. By using Superlet scalograms as input representations to MAE, our approach

enables more fine-grained and generalizable representation learning from unlabeled sleep EEG data.

II. Methods

In this study, we propose a SSL framework that integrates Superlet scalograms with a MAE architecture to effectively capture the multi-scale temporal and frequency dynamics of single-epoch EEG signals. Our objective is to enable the model to learn meaningful representations of sleep EEG by capturing its complex time-frequency patterns in a fine-grained and generalizable manner. This section describes the dataset used, the Superlet transformation method, the MAE architecture, and the evaluation procedure in detail.

A. Dataset

This study utilizes the Sleep Cassette (SC) subset of the Sleep-EDF Expanded dataset (Sleep-EDFX) [22], provided by PhysioNet. The SC subset includes PSG recordings from 153 healthy adult subjects, whose ages range from 25 to 101 years. Although the PSG dataset contains a wide range of physiological signals—including EEG Fpz-Cz and EEG Pz-Oz channels, horizontal electrooculography, and submental electromyography—this study focuses exclusively on the EEG Fpz-Cz channel.

The EEG signals were sampled at 100 Hz and segmented into 30-second epochs based on manual annotations provided by expert sleep scorers. The original annotations followed the Rechtschaffen and Kales sleep scoring manual and included eight sleep stages: Wake, N1, N2, N3, N4, REM, UNKNOWN (?), and MOVEMENT. Following the American AASM guidelines [4], the N3 and N4 stages were merged into a single 'N3' category, and epochs labeled as UNKNOWN and MOVEMENT were excluded. As a result, the final dataset used in this study consists of five sleep stages: Wake, N1, N2, N3, and REM.

B. Time-Frequency Representation with Superlet

To generate TFR of EEG signals, we employed the Superlet Transform (SLT) [19]. Superlet adaptively combines wavelets of varying orders depending on frequency, offering significantly higher resolution in both the time and frequency domains compared to traditional methods such as STFT [20] and CWT [21], and recent biomedical applications that leverage Superlet for noise-sensitive signal analysis [23]. This approach is particularly effective in mitigating resolution degradation at higher frequency bands. By increasing the number of Morlet wavelets [24], Superlet compensates for the limitations of single wavelets and alleviates the inherent trade-off between temporal and spectral resolution. The core component of SLT is the complex-valued Morlet wavelet, which is defined as:

$$\psi_{f,c}(t) = \frac{1}{B_c\sqrt{2\pi}} e^{-\frac{t^2}{2B_c^2}} e^{j2\pi ft}$$

where f is the central frequency (Hz), c is the number of cycles in the wavelet, and B_c is the standard deviation in the time domain, defined as:

$$B_c = \frac{c}{k_{sd} f}$$

Here, k_{sd} is a scaling factor, typically set to 5, which modulates the time spread of the wavelet based on its frequency and cycle count.

SLT generates a high-resolution time-frequency scalogram by applying wavelets of various orders to the input signal and combining their complex-valued responses using a geometric mean. This results in a more accurate and stable representation of EEG dynamics across multiple frequency scales.

In our study, the base number of cycles was set to 3, and the wavelet order varied from 1 to 30. A bandpass filter ranging from 0 to 40 Hz was applied to the input EEG signals to suppress irrelevant noise while preserving physiologically meaningful frequency bands. The resulting time-frequency scalogram was log-scaled, normalized, and resized into a single-channel image of size (30, 100), which was then used as input to the MAE model.

C. Masked Autoencoder Architecture

In this study, we adopt a MAE [14] architecture built upon the Vision Transformer (ViT) [25] backbone. The pretraining process is carried out in a self-supervised manner. The input data is a single-channel 2D scalogram of size (1, 30, 100), which is divided into non-overlapping patches of size (5, 5), resulting in a total of 120 patches. Each patch is flattened and passed through a linear projection layer to obtain a 256 dim latent vector. A fixed two-dimensional sine-cosine positional encoding is added to each patch embedding to retain spatial information.

Among the 120 patches, 75% are randomly masked, and only the remaining 25% of the patch embeddings are fed into the encoder. The encoder consists of 20 Transformer blocks, each comprising an 8-headed multi-head self-attention layer and a MLP layer. A learnable class token is prepended to the sequence of visible patch embeddings, and the encoder outputs a latent representation based solely on the unmasked patches.

The decoder is designed with a shallower structure than the encoder and consists of a total of 12 transformer blocks with an embedding dimension of 128. The encoder output is linearly projected to match the decoder's input dimension. A learnable mask token is inserted at the masked positions, and positional encodings are added before feeding the full sequence into the Transformer decoder. The decoder is trained to reconstruct the full patch sequence, with predictions made on a per-patch basis. The reconstruction loss is computed only over the masked patches and is defined as the mean squared error between the decoder outputs and the original patch inputs. This architecture enables the model to learn semantically meaningful representations by inferring the full input from partial observations.

During pretraining, we use the AdamW optimizer with a learning rate of 0.001. The model is trained for a total of 300 epochs. A summary of the architecture and key hyperparameters is provided in Table I.

TABLE I
MODEL AND TRAINING HYPERPARAMETERS

Parameter	Value
Input Size	(30, 100)
Patch Size	(5, 5)
Input Channel	1
Masking Ratio	0.75
Epoch	300
Batch Size	512
Accumulation Step	1
Encoder Dim	256
Encoder Depth	20
Encoder Head	8
Decoder Dim	128
Decoder Depth	12
Decoder Head	8
Optimizer	AdamW
Learning Rate	1e-3

D. Evaluation Schema

After the self-supervised pretraining phase, the encoder was frozen, and a 2-layer MLP classifier was trained using the latent representation extracted from the encoder's class token. This linear probing procedure was used to evaluate the classification performance of the representations learned by the encoder. During this stage, only the classifier weights were updated, while the encoder remained fixed. The classifier was optimized using cross-entropy loss.

Model performance was evaluated using three commonly adopted metrics: Accuracy (*ACC*), Macro F1 Score (*MF1*), and Cohen's Kappa (*Kappa*). Each metric was computed as follows:

$$ACC = \frac{\sum_{c=1}^{C} TP_c}{N}$$

where TP_c is the number of true positives for class c, and N is the total number of test epochs.

$$MF1 = \frac{\sum_{c=1}^{C} F1_c}{C}$$

where $F1_c$ is the F1 score for class c, and C is the total number of sleep stages.

$$Kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed agreement, and p_e is the expected agreement by chance.

To ensure the generalizability and reliability of the results, 31 subjects were designated as a fixed test set, and 5-fold group subject cross-validation [26] was conducted on the remaining 122 subjects. All experiments were performed using the same random seed, model architecture, training parameters, and data split strategy to ensure consistency and reproducibility.

III. RESULTS

A. Comparison of Time-Frequency Representation Methods

TABLE II
PERFORMANCE COMPARISON OF MAE
USING DIFFERENT TIME-FREQUENCY REPRESENTATIONS

Transform Type	Accuracy	Macro F1 Score	Cohen's Kappa
Superlet	**75.87% ± 1.39**	**63.23% ± 1.74**	**0.64 ± 0.02**
STFT	74.81% ± 1.27	61.74% ± 1.76	0.62 ± 0.02
CWT	75.45% ± 1.42	62.84% ± 1.16	0.63 ± 0.02

* STFT: Short-Time Fourier Transform analysis; CWT: Continuous Wavelet Transform

Table II summarizes the linear evaluation performance of MAE models trained using different TFR methods: Superlet, STFT, and CWT. The Superlet-based MAE model consistently achieved the highest scores across all three-evaluation metrics—*ACC* of 75.87% ± 1.39, *MF1* of 63.23% ± 1.74, and *Kappa* of 0.64 ± 0.02.

In particular, the Superlet model outperformed the STFT model by approximately 1.06%, and the CWT model by around 0.42% in terms of *ACC*. It also yielded the highest average values for *MF1* and *Kappa*, along with relatively smaller standard deviations, suggesting more consistent and robust classification performance.

B. Comparison with Other Self-supervised Methods

TABLE III
COMPARISON WITH OTHER SSL METHODS
FOR LINEAR EVALUATION USING SINGLE-EPOCH EEG

Model	Accuracy	Macro F1 Score	Cohen's Kappa
Ours	**75.87% ± 1.39**	**63.23% ± 1.74**	**0.64 ± 0.02**
MAEEG [18]	72.29% ± 4.56	62.58% ± 6.02	0.62 ± 0.08
TS-TCC [27]	69.27% ± 8.15	54.09% ± 18.89	0.55 ± 0.15
BENDR [15]	70.73% ± 8.57	62.04% ± 8.03	0.60 ± 0.11

Table III compares the performance of the proposed Superlet-based MAE model with representative SSL methods for single-epoch EEG. The proposed model outperformed all baselines in all three evaluation metrics. Specifically, our model achieved an *ACC* improvement of approximately 3.58% over MAEEG [18], 6.6% over TS-TCC [27], and 5.14% over BENDR [15]. In terms of *MF1* and *Kappa*, the proposed model also recorded the highest average scores among all compared methods, along with relatively low standard deviations, indicating more consistent and reliable performance.

These results demonstrate that Superlet scalograms effectively capture the diverse time-frequency characteristics of EEG signals, facilitating robust and generalizable representation learning in a self-supervised framework.

C. Reconstruction Visualization

Figure 1 illustrates the reconstruction capability of the MAE model when 75% of the input patches are masked.

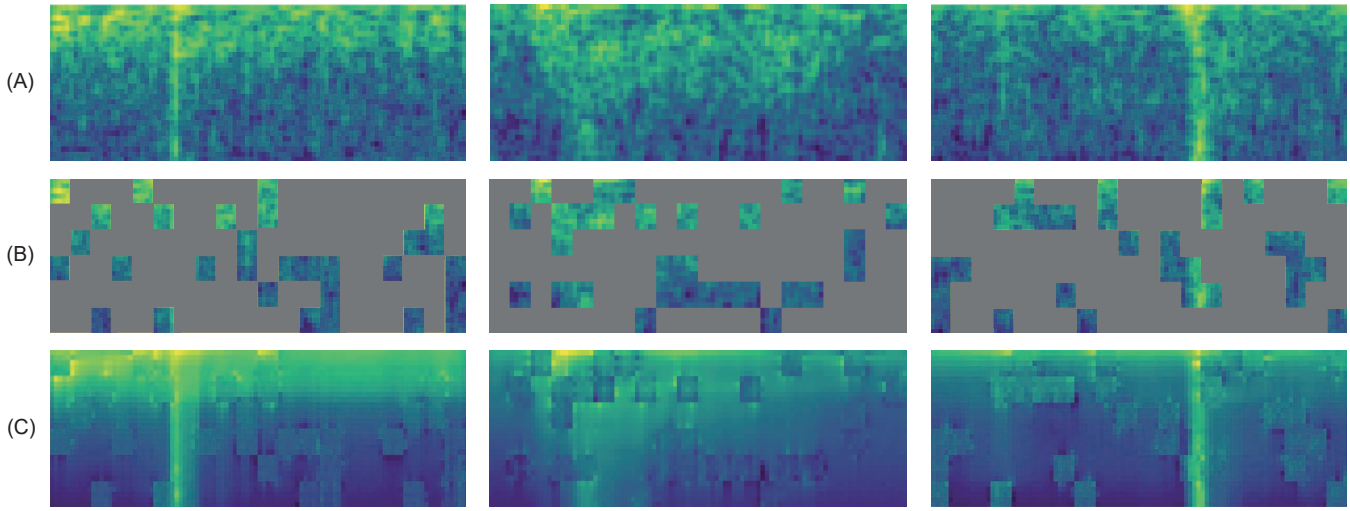

Fig. 1. Visualization of MAE reconstruction performance (mask ratio = 0.75). (A) original scalogram, (B) masked scalogram, (C) reconstructed scalogram.

Figure 1(A) shows the original Superlet scalogram used as input, while Figure 1(B) shows the scalogram with randomly applied masking. Figure 1(C) shows the reconstructed output generated by the model.

Despite the high masking ratio, the reconstructed scalogram retains the key time-frequency features observed in the original input. The overall structure and patterns remain largely intact, indicating that the model successfully inferred and recovered the missing information.

These results highlight the effectiveness of MAE in capturing the temporal and spectral patterns of EEG signals, even when substantial portions of the input are occluded.

D. Hypnogram

The sleep stage classification results for subject #SC4632E0 are visualized in Figure 2. Figure 2(A) shows the ground-truth hypnogram based on manual annotations by sleep experts, illustrating the temporal transitions and durations of each sleep stage.

Figure 2 (B) shows the hypnogram predicted by the proposed model. It closely mirrors the structure and transition patterns observed in the expert-labeled hypnogram, indicating that the model effectively captures the overall sleep architecture. Figure 2 (C) shows the softmax probabilities output by the model, stacked by sleep stage over time. In most segments, the model shows a clear preference for a single class, reflecting high confidence in its predictions. However, during transitional periods between N1 and REM stages, multiple classes show similarly elevated probabilities, suggesting increased uncertainty. This observation is consistent with previous studies, which have reported that these stages are inherently ambiguous and often yield low inter-rater agreement even among human experts [5], [28].

These results demonstrate that the proposed model is capable of effectively learning and representing the temporal dynamics of sleep architecture using only single-channel EEG data.

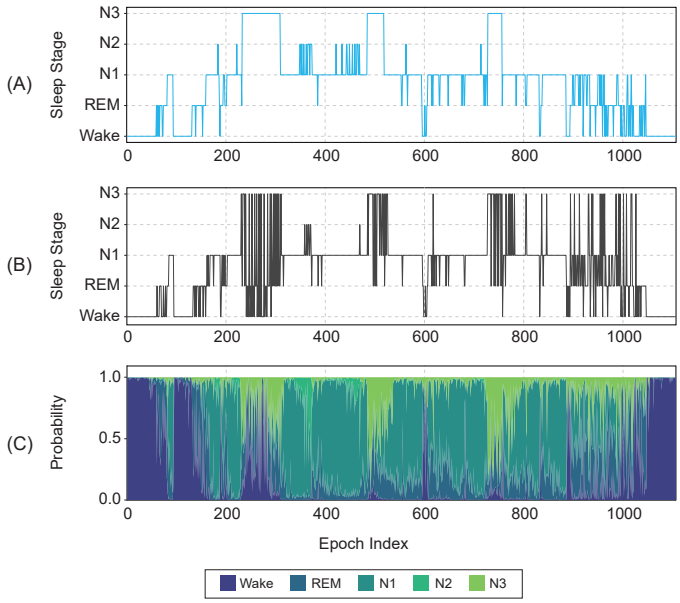

Fig. 2. Visualization of sleep stage classification for subject #SC4632E0 in Sleep-EDFX. (A) ground-truth hypnogram based on manual annotations by sleep experts, (B) predicted hypnogram generated by the proposed MAE classifier, (C) softmax probability distribution across sleep stage over time.

IV. Discussion

In this study, we proposed the first SSL framework that combines Superlet TFRs with a MAE architecture to learn effective representations from sleep EEG signals. Experimental results demonstrated that the Superlet-based MAE consistently outperformed models utilizing conventional time-frequency methods such as STFT and CWT across all evaluation metrics. These findings suggest that the high-resolution spectral features provided by Superlet deliver richer information during

training, enabling the encoder to learn more expressive representations. Notably, these performance gains were achieved using only single-channel EEG inputs, underscoring the practical impact of the Superlet representation on model effectiveness.

Compared with other SSL approaches such as BENDR [15], TS-TCC [27], and MAEEG [18], the proposed method achieved superior results. Specifically, it outperformed BENDR by approximately 5.14%, TS-TCC by 6.6%, and MAEEG by 3.58% in terms of *ACC*. It also achieved the highest scores for both *MF1* and *Kappa* metrics, demonstrating the effectiveness of combining Superlet scalograms with the MAE framework for the learning of EEG representation.

Traditional contrastive learning approaches are heavily highly on data augmentation strategies [27]. In EEG analysis, common augmentations include noise addition, jittering, scaling, and masking can significantly affect performance depending on their configuration [16]. This sensitivity often limits the stability and reproducibility of contrastive learning methods. In contrast, the MAE architecture mitigates these limitations while still enabling robust and semantically rich representation learning.

As shown in Figure 1, the model successfully preserved key time-frequency patterns even when 75% of the input patches were masked. Moreover, Figure 2 indicates that the highest classification performance was obtained at a 75% masking ratio, emphasizing the importance of an appropriate masking strategy in MAE-based EEG representation learning.

To assess real-world generalizability, future studies should incorporate EEG data collected under diverse clinical settings and sensor configurations. Since this study was limited to the Sleep-EDFX dataset, its cross-dataset adaptability remains unverified. Additionally, the model's robustness against noisy EEG signals, such as those acquired in clinical or mobile environments, has not yet been evaluated. These aspects require further investigation to ensure real-world applicability.

Although Superlet inherently provides rich and high-resolution time-frequency information, we intentionally reduced the input resolution in this study to improve training stability and computational efficiency. This downscaling may have caused the loss of fine-grained features. In future work, we plan to leverage the full spectral resolution of Superlet representations to enable more powerful and comprehensive representation learning.

Furthermore, the current framework is limited to static, single-epoch analysis and cannot capture the temporal transitions between sleep stages. Incorporating temporal context and modeling stage transitions using sequential or recurrent structures will be an important future direction. Despite these limitations, this study represents the first attempt to apply high-resolution Superlet scalograms within a MAE-based SSL framework for sleep EEG signals. It offers a new perspective in EEG representation learning.

Future directions include expanding the method to multi-channel EEG and other physiological signals (e.g., electrooculography, electromyography), exploring cross-dataset transfer learning, and developing multimodal MAE architectures with improved generalizability and clinical utility.

ACKNOWLEDGMENT

This work was supported by a National Research Foundation of Korea (NRF) Grant funded by the Korean government (Ministry of Science and ICT, MSIT) (No. 2022R1A2C1013205)

REFERENCES

[1] J. M. Siegel, "Clues to the functions of mammalian sleep," *Nature*, vol. 437, no. 7063, pp. 1264–1271, 2005.

[2] M. W. Mahowald and C. H. Schenck, "Insights from studying human sleep disorders," *Nature*, vol. 437, no. 7063, pp. 1279–1285, 2005.

[3] S. D. Davis, E. Eber, A. C. Koumbourlis *et al.*, *Diagnostic tests in pediatric pulmonology*. Springer, 2015.

[4] M. M. Grigg-Damberger, "The aasm scoring manual: a critical appraisal," *Current opinion in pulmonary medicine*, vol. 15, no. 6, pp. 540–549, 2009.

[5] R. S. Rosenberg and S. Van Hout, "The american academy of sleep medicine inter-scorer reliability program: sleep stage scoring," *Journal of clinical sleep medicine*, vol. 9, no. 1, pp. 81–87, 2013.

[6] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1285–1296, 2018.

[7] A. M. Eldaraa, H. Baali, A. Bouzerdoum, S. B. Belhaouari, T. Alam, and A. S. A. Rahman, "Classification of sleep arousal using compact cnn," in *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*. IEEE, 2020, pp. 247–253.

[8] A. Supratak, H. Dong, C. Wu, and Y. Guo, "Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.

[9] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "An attention-based deep learning approach for sleep stage classification with single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.

[10] C.-H. Lee, H.-J. Kim, Y.-T. Kim, H. Kim, J.-B. Kim, and D.-J. Kim, "Sleepexpertnet: high-performance and class-balanced deep learning approach inspired from the expert neurologists for sleep stage classification," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 6, pp. 8067–8083, 2023.

[11] C.-H. Lee, H. Kim, H.-j. Han, M.-K. Jung, B. C. Yoon, and D.-J. Kim, "Neuronet: A novel hybrid self-supervised learning framework for sleep stage classification using single-channel eeg," *arXiv preprint arXiv:2404.17585*, 2024.

[12] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[13] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[14] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[15] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data," *Frontiers in Human Neuroscience*, vol. 15, p. 653659, 2021.

[16] C. Yang, D. Xiao, M. B. Westover, and J. Sun, "Self-supervised eeg representation learning for automatic sleep staging," *arXiv preprint arXiv:2110.15278*, 2021.

[17] V. Kumar, L. Reddy, S. Kumar Sharma, K. Dadi, C. Yarra, R. S. Bapi, and S. Rajendran, "muleeg: a multi-view representation learning on eeg signals," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 398–407.

[18] H.-Y. S. Chien, H. Goh, C. M. Sandino, and J. Y. Cheng, "Maeeg: Masked auto-encoder for eeg representation learning," *arXiv preprint arXiv:2211.02625*, 2022.

[19] V. V. Moca, H. Bârzan, A. Nagy-Dăbâcan, and R. C. Mureșan, "Time-frequency super-resolution with superlets," *Nature communications*, vol. 12, no. 1, p. 337, 2021.

[20] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[21] L. Aguiar-Conraria and M. J. Soares, "The continuous wavelet transform: Moving beyond uni-and bivariate analysis," *Journal of economic surveys*, vol. 28, no. 2, pp. 344–375, 2014.

[22] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[23] T.-S. Han, J.-W. Heo, H. Kim, C.-H. Lee, H. Huh, E.-K. Choi, and D.-J. Kim, "Diffusion-based electrocardiography noise quantification via anomaly detection," 2025. [Online]. Available: https://arxiv.org/abs/2506.11815

[24] M. X. Cohen, "A better way to define and describe morlet wavelets for time-frequency analysis," *NeuroImage*, vol. 199, pp. 81–86, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811919304409

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[26] T.-T. Wong and P.-Y. Yeh, "Reliable accuracy estimates from k-fold cross validation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1586–1594, 2020.

[27] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan, "Time-series representation learning via temporal and contextual contrasting," 2021. [Online]. Available: https://arxiv.org/abs/2106.14112

[28] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 400–410, 2019.