VEGA: Vision Enhanced Generative Augmentation for Cross-Subject Electrocardiogram Generalization

Tianyi Shi[1,†][0009−0000−7380−6898], Siyang Zheng[1,†], Zhu Meng[1], Zhe Cui[1], Jin Huang[2], Changrui Ren[2], and Zhicheng Zhao[1,3,*]

[1] Beijing University of Posts and Telecommunications, Beijing, China
{sty0622,zhengsiyang,bamboo,cuizhe}@bupt.edu.cn
[2] Beijing Academy of Blockchain and Edge Computing, Beijing, China
{JinHuang2024,ChangruiRen}@outlook.com
[3] Beijing Key Laboratory of Network System and Network Culture, Beijing, China
zhaozc@bupt.edu.cn

Abstract. Electrocardiogram (ECG) signals are critical physiological indicators for assessing human health status and diagnosing cardiovascular diseases. However, substantial inter-subject feature variability poses a major challenge to cross-subject generalization in ECG analysis. Existing methods primarily rely on domain adaptation and subject-invariant feature learning, which are often task-specific and may introduce potential data leakage. To address these issues, we propose Vision Enhanced Generative Augmentation (VEGA), a novel paradigm that enables models to quickly adapt to different subjects through data generation without introducing any data leakage. Specifically, VEGA leverages rich semantic priors from a generative model pre-trained on natural images to guide the synthesis of realistic and diverse ECG signals. To adapt to the visual generator, one-dimensional (1D) ECG signals are first converted into two-dimensional (2D) visual representation. Then, to align the numerical distributions between the visualized ECG signals and natural images, VEGA adopts a lightweight domain adaptation strategy by fine-tuning only the normalization layers of the generator with a small amount of external ECG data. Through augmented ECG data generation, VEGA effectively bridges the feature distribution gap between training and testing subjects. We evaluate VEGA on multiple ECG tasks, including fatigue detection, ventricular premature beat (VPB) detection, and generative heartbeat prediction. Across all tasks and datasets, VEGA consistently improves cross-subject generalization performance, demonstrating its effectiveness as a unified generative augmentation paradigm for ECG analysis. Our code is available at https://github.com/Yuyun1011/VEGA-release.

Keywords: Electrocardiogram · Cross-subject generalization · Generative augmentation · Masked autoencoder.

† Equal contribution.
* Corresponding author.

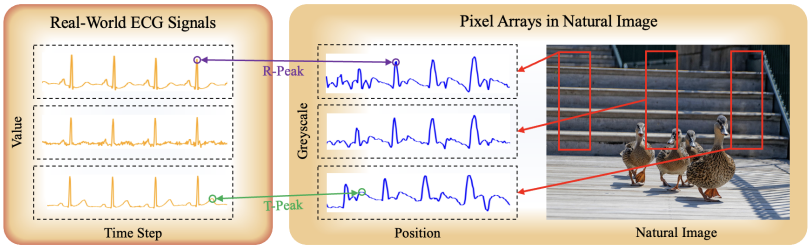

Fig. 1. Morphological similarity between real ECG signals from BUT-QDB dataset [20] and pixel arrays (the average of the column-wise pixel arrays within the red box) derived from natural images in the COCO dataset [16].

1 Introduction

ECG signals are a fundamental physiological modality widely used in human condition monitoring and clinical diagnosis [4]. Nowadays, deep neural network–based approaches have achieved remarkable performance across a variety of ECG analysis tasks [25]. However, cross-subject variability in the morphology of the ECG poses a significant challenge to model generalization [8], and during training, deep models tend to capture subject-specific characteristics that correlate with individual identity rather than task-relevant discriminative features [6].

From a perspective of data distribution, the generation of synthetic samples that approximate the feature distribution of unseen subjects offers a principled way to mitigate the gaps in cross-subject generalization [12]. However, existing generative models such as diffusion models [15] and normalizing flows [21] typically require large-scale ECG datasets to learn high-fidelity signal generators, which is often impractical in subject-limited clinical settings. To overcome this limitation, we propose Vision Enhanced Generative Augmentation (VEGA), a novel framework that leverages structural priors learned from natural images to guide the generation of ECG signals, enable effective data augmentation, and improve cross-subject robustness under limited ECG data. As illustrated in Fig. 1, natural images contain rich spatial structures and repetitive patterns whose spatial morphology and statistical distribution share similarities with the temporally periodic waveforms of ECG signals, especially in their regularity behavior and salient local transitions such as R-peaks and T-peaks. Therefore, by pre-training on large-scale natural image datasets [16], it is reasonable that a visual model could capture structural regularities shared between modalities and consequently demonstrate the ability to model and generate ECG signals.

In VEGA, a visual masked autoencoder (MAE) [13] is adopted as the generator. To facilitate effective processing by visual MAE, as shown in Fig. 2 (d), the 1D ECG signals are first transformed into a 2D representation and reshaped to occupy a certain proportion of the input images. The remaining proportions are masked, prompting the MAE to reconstruct the missing region in an auto-

regressive manner. The generated output is then serialized back into the 1D signal and used for data augmentation, while the original task labels are preserved. To further enhance the quality of generation, we introduce a frequency adaptive alignment strategy. Specifically, the multilayer perceptron (MLP) and attention layers, which contain rich visual priors within the Transformer architecture [24, 11, 26], are frozen, while parameter-efficient fine-tuning in the time and frequency domains is conducted exclusively on the Layer Normalization (LN) layers. Evaluated on fatigue detection, VPB detection and generative heartbeat prediction tasks, VEGA enables more efficient generation of high-quality ECG signals and improves cross-subject generalization for different methods effectively, establishing a unified generative augmentation framework for robust ECG analysis.

2 Related Work

Recent methods have investigated the transformation of ECG signals into 2D visual forms to leverage advances in visual architectures. Adib et al. [1] and Bedin et al. [3] introduced diffusion model for ECG generation, which demanded substantial ECG data and considerable computational resources. Yoon et al. adopted a visual MAE framework for 2D generative reconstruction of ECG signals [27]. Nevertheless, they did not exploit the structural priors learned from natural images. Chen et al. introduced an image pre-trained MAE framework for generic sequence forecasting [5], revealing the structural correspondence between visual and sequential data. Nevertheless, their adaptation during fine-tuning remained limited, lacking dedicated mechanisms to align pre-trained visual representations with sequence-specific morphology. Unlike them, we propose an ECG signal generation approach that leverages visual priors and employs a frequency adapter to align the ECG and natural image modalities.

3 Methodology

As illustrated in Fig. 2 (a), the morphological variability between subjects in ECG signals induces pronounced distribution shifts in the learned latent representation space, thus impairing cross-subject generalization. To address this issue, we propose VEGA, as shown in Fig. 2 (b). By leveraging visual priors to generate diverse augmented ECG samples from the training set, VEGA explicitly guides the model to focus on task-relevant discriminative structures rather than implicitly learning subject-specific identity cues.

Pre-Training with Images. Prior to ECG adaptation, as shown in Fig. 2 (c), the MAE is pre-trained on ImageNet [7] to acquire generic visual representations. We directly adopt publicly available pre-trained MAE weights, thus avoiding training from scratch. MAE is pre-trained in a self-supervised masked reconstruction framework, where images are patchified, heavily masked, encoded using only visible patches, and reconstructed by a lightweight decoder, thereby learning transferable structural priors that support subsequent ECG generation.

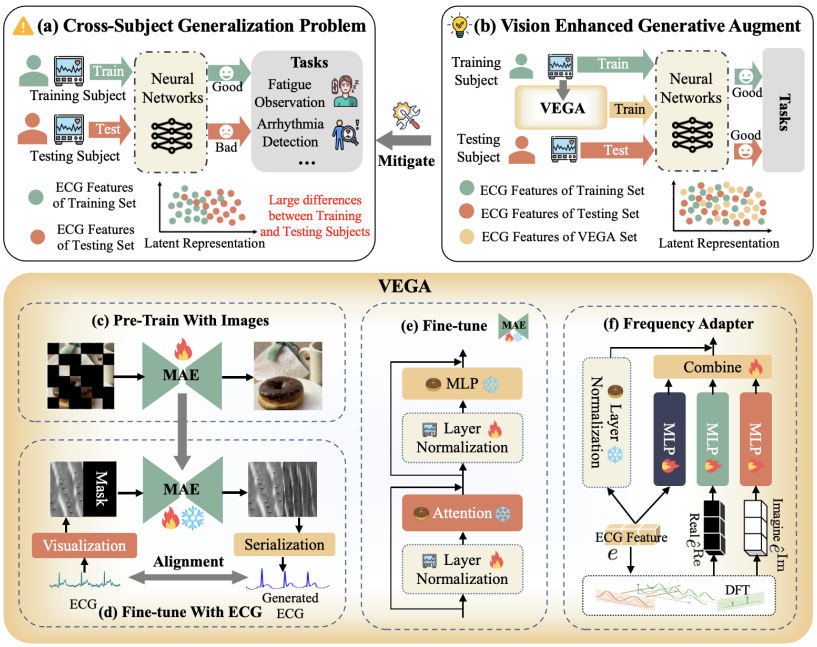

Fig. 2. Architecture of VEGA. To address cross-subject generalization (a), VEGA performs visual prior guided generative augmentation on ECG signals from training subjects, reducing the representation gap between training and test data and thereby improving model robustness (b). Specifically, MAE is first pre-trained on large-scale natural image datasets (c) and subsequently fine-tuned on external ECG data for modality adaptation (d)(e)(f).

Visualization and Serialization. As illustrated in Fig. 2 (d), the ECG fine-tuning process is conducted entirely within the 2D representation space. Given an ECG signal $s \in \mathbb{R}^T$, it is divided into $\lfloor T/K \rfloor$ consecutive segments of length K. The segments are stacked column-wise to form $Z_{\text{raw}} \in \mathbb{R}^{K \times \lfloor T/K \rfloor}$, enabling simultaneous modeling of intra-cycle morphology and inter-cycle dynamics, while preserving local continuity analogous to spatial locality in images. To align with the input statistics of the pre-trained MAE, we apply instance normalization, which yields $Z_{\text{norm}} = (Z_{\text{raw}} - \mu(Z_{\text{raw}}))/\sigma(Z_{\text{raw}})$, where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation. Z_{norm} is then rendered as a three-channel grayscale image $Z_{\text{img}} \in \mathbb{R}^{K \times \lfloor T/K \rfloor \times 3}$ by duplicating values across channels. To match the patch configuration of the pre-trained MAE, which operates on $M \times M$ patches of size $R \times R$, Z_{img} is resized by bilinear interpolation to $(MR,\ qR)$, where the number of visible patch columns is defined as $q = \lfloor \eta \cdot M \cdot T/(T+V) \rfloor$, where V denotes the length of the masked ECG segment to be generated and $\eta \in [0, 1]$ controls the visible ratio (setting $\eta = 0.4$ performs best in our experiments), and the remaining patches are masked for reconstruction. This transformation

bridges 1D ECG signals and 2D visual representations, enabling effective transfer of pre-trained visual priors for ECG generation. After generation, the image is rescaled to the original ECG layout via bilinear interpolation, the three channels are averaged, inverse normalization restores the amplitude scale, and the result is reshaped into a 1D sequence to obtain the generated ECG segment.

Fine-tuning. As shown in Fig. 2 (e), to adapt the pre-trained MAE to ECG signals while preserving its visual priors, a parameter-efficient fine-tuning strategy is adopted. Specifically, all self-attention and MLP blocks in the Vision Transformer [9] backbone are frozen, and only the affine parameters of the LN layers are updated. Self-attention and MLP layers encode transferable structural representations learned from natural images. Since fully updating them on limited ECG data may overwrite these priors, we freeze these layers. In contrast, LN performs feature-wise rescaling and shifting, thereby enabling adaptation of feature distributions to the ECG modality without altering the underlying structural transformations. This lightweight adaptation strategy facilitates stable cross-modal transfer and effective ECG generation with minimal parameter updates.

Frequency Adapter. To further enhance modality adaptation, as shown in Fig. 2 (f), we introduce a frequency adapter strategy instead of directly updating the LN parameters. Let $e \in \mathbb{R}^d$ denote the input feature to a frozen LN layer. The original LN transformation is kept unchanged, producing $e_{\mathrm{LN}} = \mathrm{LN}(e)$. Parallel to this frozen branch, a three-branch adapter is constructed to explicitly incorporate spectral information. First, the original feature is retained as one branch. In addition, the discrete Fourier transform (DFT) of e is computed along the feature dimension $\hat{e}_k = \sum_{n=0}^{d-1} e_n \exp\left(-j\frac{2\pi kn}{d}\right), \quad k = 0, 1, \ldots, d-1$, where $j = \sqrt{-1}$. The transformed representation $\hat{e} \in \mathbb{C}^d$ is decomposed into its real and imaginary components $\hat{e}^{\mathrm{Re}} = \mathrm{Re}(\hat{e}), \quad \hat{e}^{\mathrm{Im}} = \mathrm{Im}(\hat{e})$. Three parallel inputs are thus obtained, namely e, $\hat{e}^{\mathrm{Re}}$, and $\hat{e}^{\mathrm{Im}}$. Each branch is processed by an identical MLP $\phi(\cdot)$ with shared architecture but independent parameters, yielding $z_1 = \phi(e), \quad z_2 = \phi(\hat{e}^{\mathrm{Re}}), \quad z_3 = \phi(\hat{e}^{\mathrm{Im}})$. The three outputs are fused through a linear aggregation function $z_{\mathrm{fuse}} = W[z_1 \,\|\, z_2 \,\|\, z_3] + b$, where $\|$ denotes concatenation. Finally, z_{fuse} is added residually to e_{LN} to obtain the output $e_{\mathrm{out}} = e_{\mathrm{LN}} + z_{\mathrm{fuse}}$. This design preserves the pre-trained normalization behavior while enabling frequency-aware feature adaptation through explicit modeling of real and imaginary spectral components.

Alignment. During ECG fine-tuning, a direct reconstruction alignment objective is adopted to ensure that the generated signal remains faithful to the original morphology. Specifically, given the original ECG segment $s \in \mathbb{R}^T$ and its generated counterpart $\tilde{s} \in \mathbb{R}^T$, we minimize the mean squared error (MSE) loss $\mathcal{L}_{\mathrm{align}} = \frac{1}{T}\sum_{t=1}^{T}(\tilde{s}_t - s_t)^2$. By enriching the training distribution with subject-diverse yet task-consistent synthetic samples, VEGA implicitly reduces the model's reliance on subject-specific low-level patterns and encourages the learning of invariant representations that are predictive of the task but independent of individual identity.

Table 1. Experimental results of cross-subject fatigue detection task on Drozy dataset [17] with five-fold validation. Mean represents the mean value of cross-validation (larger is better ↑), and Std. represents the standard deviation (smaller is better ↓). Bold numbers indicate the best values for each metric.

Method	Type	VEGA	Indicator	AUC	ACC	PRE	REC	F1
FanRBT [10]	Spesific	✗	Mean ↑	0.778	0.628	0.532	0.628	0.562
			Std. ↓	0.148	0.198	0.287	0.198	0.254
		✓	Mean ↑	**0.820**	**0.755**	**0.735**	**0.920**	**0.808**
			Std. ↓	**0.130**	**0.104**	**0.151**	**0.067**	**0.073**
MuMuT [19]	Spesific	✗	Mean ↑	0.700	0.724	0.687	0.724	0.691
			Std. ↓	0.238	0.176	0.236	0.176	0.205
		✓	Mean ↑	**0.784**	**0.755**	**0.718**	**0.949**	**0.812**
			Std. ↓	**0.168**	**0.133**	**0.147**	**0.065**	**0.097**
ResNet1D [14]	General	✗	Mean ↑	0.689	0.725	0.782	0.725	0.694
			Std. ↓	0.150	**0.106**	**0.092**	**0.106**	**0.148**
		✓	Mean ↑	**0.778**	**0.780**	**0.808**	**0.780**	**0.757**
			Std. ↓	**0.118**	0.119	0.103	0.119	0.151

4 Experiments

The ECG signals are resampled to 360Hz. Sample lengths are 3600 for fatigue detection and VPB detection tasks, and 270 (one heartbeat) for generative heartbeat prediction. For samples with multiple heartbeats, we set $K = 220$ to approximate one heartbeat; if only one heartbeat is present, $K = T = 270$. All experiments are conducted on a single NVIDIA A100-80G GPU.

Fatigue Detection. The fatigue detection experiments are conducted on the Drozy dataset [17]. Fourteen subjects are randomly partitioned into five groups for a five-fold cross-validation. To prevent data leakage, the BUP-QDB dataset [20] is used exclusively to fine-tune VEGA. In each fold, VEGA performs generative augmentation on every training sample, and the synthesized data are combined with the original training set to train the downstream models. As shown in Table 1, FanRBT and MuMuT are two specialized networks for fatigue detection, while ResNet1D is a classic general-purpose network. Obviously, after VEGA augmentation, various metrics, including area under curve (AUC), accuracy (ACC), precision (PRE), recall (REC), and F1-Score (F1) have improved significantly. Notably, for specialized networks with inductive biases aligned to ECG characteristics, the augmented data enhances both accuracy and robustness. In contrast, for ResNet1D, the expanded data distribution increases optimization sensitivity, resulting in higher variance across folds despite improved average performance. To explain the mechanism of VEGA, we visualize the latent representations of fatigue samples before the classification layer of the FanRBT method. As shown in Fig. 3, in the original training, the distribution of samples from the same class differs significantly between the training set and the test set, while VEGA effectively bridges this distribution gap.

VPB Detection. VPB, which are early heartbeats originating from the ventricles, can disrupt normal cardiac function and, when frequent, may lead

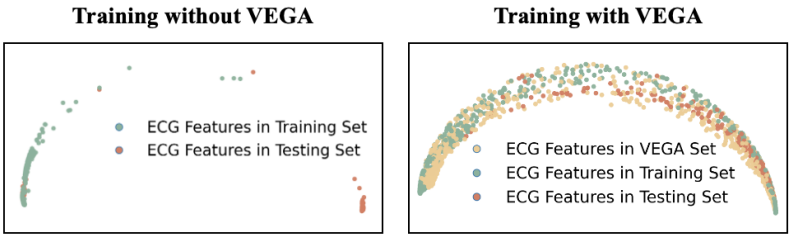

Fig. 3. Visualization of latent representations of ECG signals. The left and right subplots show the feature distributions training without and with VEGA, respectively.

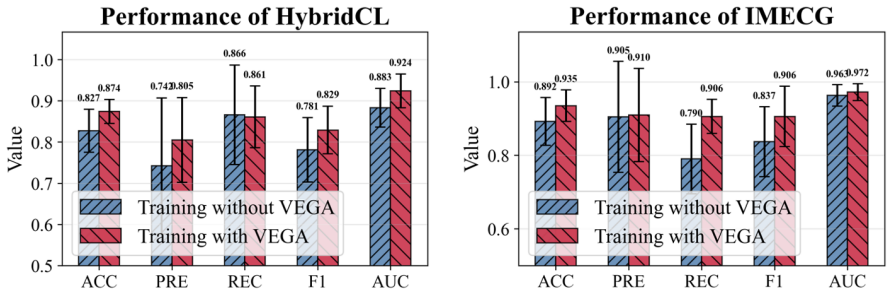

Fig. 4. VPB detection performance of HybridCL [2] and IMECG [23] training with and without VEGA.

to complications such as cardiomyopathy, heart failure, or life-threatening arrhythmias [22]. The VPB detection experiments are conducted on the MIT-BIH dataset [18]. Forty seven subjects are randomly partitioned into four groups for a four-fold cross-validation. To prevent data leakage, the BUP-QDB dataset [20] is used exclusively to fine-tune VEGA. In each fold, VEGA augments each training sample and adds the synthesized data to the original set for downstream model training. As shown in Fig. 4, HybridCL and IMECG are two specialized networks for VPB detection. The height of each bar represents the mean value of the metrics (AUC, ACC, PRE, REC, and F1) in the cross-validation folds, while the error bars denote the standard deviation. It is clear that after training with VEGA, for all methods, the majority of metrics achieve higher bars and shorter error bars, indicating both improved performance and stability.

Generative Heartbeat Prediction. To compare the generation quality and training efficiency of VEGA with mainstream generative models, this experiment adopts the generative heartbeat prediction task. The MIT-BIH dataset is divided by subject into 80% for the training set and 20% for the testing set. All methods use a single heartbeat of 270 points as a condition to generate the next heartbeat of 270 points. During the prediction stage, the generated signals from VEGA are aligned with the predicted heartbeats, rather than with the original

Table 2. Experimental results of generative heartbeat prediction task on MIT-BIH dataset [18] with four-fold validation. Training ratio indicates the utilization rate of the training set. Training parameters refer to the categories of parameters that need to be updated, where full means all parameters are updated, none means no parameters need to be updated, that is, training free, LN means only the parameters of the LN layers are updated and LN+FA represents the frequency adapter update method proposed in this paper. MSE and mean absolute error (MeAE) are used as the evaluation metrics. Bold numbers indicate the best values for each metric.

Method	Mechanism	Training Ratio	Training Parameters	MSE	MeAE
NF [21]	Normalizing Flows	100%	Full	0.1442	0.2184
DDPM [15]	Diffusion	100%	Full	0.1172	0.1893
VEGA	MAE	0%	None	0.1277	0.1980
		10%	Full	0.1151	0.2021
		10%	LN	0.1139	0.1853
		10%	LN+FA	**0.1092**	**0.1828**

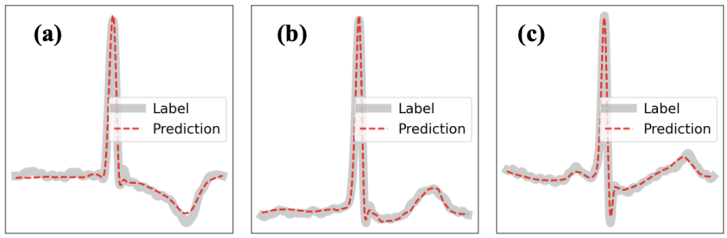

Fig. 5. Visualization of heartbeat prediction with different morphologies, showing (a) T wave inversion, (b) absence of P wave, and (c) normal signal.

heartbeats during augmentation. As shown in Table 2, VEGA with few-shot training, can surpass the other generative methods such as DDPM and NF with full-data training. If all parameters of visual MAE are fully updated, the prior information stored in the MLP and attention layers during visual pre-training would be disrupted, thereby degrading the generative prediction performance. Furthermore, it is not difficult to observe that adjusting the parameters of the LN layers via the frequency adapter is more effective than directly modifying the LN parameters. We visualized the predicted heartbeats, as shown in Fig. 5. Whether it is inversion of the T-wave (a), disappearance of the P-wave (b), or normal waveform (c), VEGA achieves accurate prediction results.

5 Conclusions

In this paper, a novel framework of vision enhanced generative augmentation for the ECG data, termed VEGA, is presented. By leveraging a visual MAE pre-trained on natural image datasets and subsequently fine-tuned with only a

small set of ECG samples, VEGA enables high-quality ECG generation. During fine-tuning, a frequency adapter is introduced to update the Transformer backbone in a parameter-efficient manner by modulating only the LN layers, yielding improved adaptation performance. The synthesized signals are combined with the original training data to train downstream models, effectively reducing cross-subject latent distribution discrepancies and enhancing generalization across various ECG analysis tasks.

Acknowledgments. This work was supported by Chinese National Natural Science Foundation (62401069), Beiing Advanced Innovation Center for Future Blockchain and Privacy Computing (GJJ-24-021) and Key Research and Development Program of Xizang (XZ202601ZY0042).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adib, E., Fernandez, A.S., Afghah, F., Prevost, J.J.: Synthetic ecg signal generation using probabilistic diffusion models. IEEe Access **11**, 75818–75828 (2023)
2. Alamatsaz, N., Tabatabaei, L., Yazdchi, M., Payan, H., Alamatsaz, N., Nasimi, F.: A lightweight hybrid cnn-lstm explainable model for ecg-based arrhythmia detection. Biomedical Signal Processing and Control **90**, 105884 (2024)
3. Bedin, L., Cardoso, G., Duchateau, J., Dubois, R., Moulines, E.: Leveraging an ecg beat diffusion model for morphological reconstruction from indirect signals. Advances in Neural Information Processing Systems **37**, 84409–84446 (2024)
4. Chen, G., Shi, T., Xie, B., Zhao, Z., Meng, Z., Huang, Y., Dong, J.: Swindae: Electrocardiogram quality assessment using 1d swin transformer and denoising autoencoder. IEEE journal of biomedical and health informatics **27**(12), 5779–5790 (2023)
5. Chen, M., Shen, L., Li, Z., Wang, X.J., Sun, J., Liu, C.: Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. arXiv preprint arXiv:2408.17253 (2024)
6. Chen, P.J., Chien, W.S., Lee, C.C.: Disentangle heart rate signals for improved stress detection. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
8. Ding, C., Yao, T., Wu, C., Ni, J.: Advances in deep learning for personalized ecg diagnostics: A systematic review addressing inter-patient variability and generalization constraints. Biosensors and Bioelectronics **271**, 117073 (2025)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Fan, J., Dong, L., Sun, G., Zhou, Z.: A deep learning approach for mental fatigue state assessment. Sensors **25**(2), 555 (2025)

11. Ghiasi, A., Kazemi, H., Borgnia, E., Reich, S., Shu, M., Goldblum, M., Wilson, A.G., Goldstein, T.: What do vision transformers learn? a visual exploration. arXiv preprint arXiv:2212.06727 (2022)
12. Golany, T., Radinsky, K.: Pgans: Personalized generative adversarial networks for ecg synthesis to improve patient-specific deep ecg classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 557–564 (2019)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
17. Massoz, Q., Langohr, T., François, C., Verly, J.G.: The ulg multimodality drowsiness database (called drozy) and examples of use. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–7. IEEE (2016)
18. Moody, G.B., Mark, R.G.: The impact of the mit-bih arrhythmia database. IEEE engineering in medicine and biology magazine **20**(3), 45–50 (2001)
19. Mu, S., Liao, S., Tao, K., Shen, Y.: Intelligent fatigue detection based on hierarchical multi-scale ecg representations and hrv measures. Biomedical Signal Processing and Control **92**, 106127 (2024)
20. Nemcova, A., Smisek, R., Opravilová, K., Vitek, M., Smital, L., Maršánová, L.: Brno university of technology ecg quality database (but qdb). PhysioNet **101**, e215–e220 (2020)
21. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015)
22. Shi, T., Zheng, S., Meng, Z., Cui, Z., Huang, J., Ren, C., Zhang, B., Zhao, Z.: Tell me: Tackle electrocardiogram with large language model effectively. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
23. Tao, R., Wang, L., Xiong, Y., Zeng, Y.R.: Im-ecg: an interpretable framework for arrhythmia detection using multi-lead ecg. Expert Systems with Applications **237**, 121497 (2024)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
25. Velandia, H., Pardo, A., Vera, M.I., Vera, M.: Systematic review of artificial intelligence and electrocardiography for cardiovascular disease diagnosis. Bioengineering **12**(11), 1248 (2025)
26. Yeh, C., Chen, Y., Wu, A., Chen, C., Viégas, F., Wattenberg, M.: Attentionviz: A global view of transformer attention. IEEE Transactions on Visualization and Computer Graphics **30**(1), 262–272 (2023)
27. Yoon, T., Kang, D.: Efficient pretraining of ecg scalogram images using masked autoencoders for cardiovascular disease diagnosis. Scientific Reports **15**(1), 24444 (2025)