MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Electric Motor Cogging Torque Prediction with Vision Transformer Models

Sun, Siyuan; Wang, Ye; Koike-Akino, Toshiaki; Yamamoto, Tatsuya; Sakamoto, Yusuke; Wang, Bingnan

TR2025-059 May 20, 2025

Abstract

Motor performances such as cogging torque and torque ripple are difficult to predict accurately with surrogate models. In this work, we propose Vision Transformer (ViT) based models to tackle the problem. We adopt a ViT model pre-trained on image classification tasks, and fine-tune it with a dataset prepared for interior permanent magnet motor designs. Each motor design is represented by a 2d image and fed into the ViT model for making predictions on cogging torque. To further improve the data efficiency of the model, we customize it by utilizing the motor design parameter information to initialize the class token of the ViT model. We show that the proposed method significantly outperforms established deep convolutional neural network (CNN) based models, and achieves high accuracy on cogging torque prediction on the test dataset.

IEEE International Electric Machines and Drives Conference (IEMDC) 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Electric Motor Cogging Torque Prediction with Vision Transformer Models

Siyuan Sun
Department of Computer Science,
Iowa State University,
Ames, IA 50011, USA

Ye Wang
Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA 02139 USA

Toshiaki Koike-Akino
Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA 02139 USA

Tatsuya Yamamoto
Advanced Technology R&D Center
Mitsubishi Electric Corporation
Amagasaki, Hyogo 661-8661 Japan

Yusuke Sakamoto
Advanced Technology R&D Center
Mitsubishi Electric Corporation
Amagasaki, Hyogo 661-8661 Japan

Bingnan Wang
Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA 02139 USA

Abstract: Motor performances such as cogging torque and torque ripple are difficult to predict accurately with surrogate models. In this work, we propose Vision Transformer (ViT) based models to tackle the problem. We adopt a ViT model pre-trained on image classification tasks, and fine-tune it with a dataset prepared for interior permanent magnet motor designs. Each motor design is represented by a 2d image and fed into the ViT model for making predictions on cogging torque. To further improve the data efficiency of the model, we customize it by utilizing the motor design parameter information to initialize the class token of the ViT model. We show that the proposed method significantly outperforms established deep convolutional neural network (CNN) based models, and achieves high accuracy on cogging torque prediction on the test dataset.

***Index Terms*—Electric Motors, Surrogate Model, Machine Learning, Vision Transformer.**

I. Introduction

Electric machines are widely used in households and various industries, and their technologies and design principles are well established. However, the requirements for motor design and customization, especially for new applications such as transportation electrification and factory automation, always pose new challenges to motor designers. Parameter sweeping or iterative optimization methods are often utilized to evaluate a large number of design candidates before identifying the optimal design for a specific task. The accurate analysis of a motor design typically relies on numerical simulations based on finite-element analysis (FEA), which are time-consuming, especially when various operating points are evaluated for one design. Surrogate model based optimization has been investigated to speed up the process [1]. Due to the highly nonlinear nature, the accuracy of conventional surrogate models is not sufficient for the prediction of certain motor performances such as torque profile and efficiency map. In recent years, machine learning and deep learning methods have been developed and applied to many applications including motor design [2], [3], due to their powerful capability to emulate highly nonlinear functions. For example, several studies [4]–[7] have proposed neural network-based approaches to address various critical tasks, such as generating innovative motor designs, predicting physical responses for given designs, and detecting potential faults during operation. In particular, due to the success of convolutional neural networks (CNNs) for image recognition, one popular approach is to represent a motor design with a 2D image, which is fed into a CNN-based model, to predict the motor performance [8]. However, the highly sensitive cogging torque and torque ripple prediction for permanent magnet motors, remain a challenge even for deep CNN models.

The Vision Transformer (ViT) is first been proposed in [9] as a variant of the transformer [10]. It turns images into a token sequence, which is suitable for the transformer to handle. The ViT models have been primarily used to solve general image-related tasks, achieving superior performance compared with other existing deep learning models. For instance, they have been widely applied in object detection [11]–[13],object segmentation [14]–[16] and image generation [17]–[19]. In the motor design field, Shimizu et al. [20] introduced the ViT as part of the motor design process, demonstrating its potential in this domain. However, the full capabilities of Vision Transformers in motor design remain largely unexplored. In this paper, we propose a novel Vision Transformer-based surrogate model to estimate the physical responses of motor designs more accurately and efficiently, advancing the use of modern deep learning techniques in motor design optimization. We show that ViT models can achieve significantly improved accuracy compared with deep CNN models, demonstrated by cogging torque prediction of interior permanent magnet synchronous motors (IPMSMs). In addition, we develop an effective way of combining the motor design information in the form a 2D image and a list of parameters to facilitate the training process of ViT models. Specifically, we utilize

This work was done when S. Sun was with Mitsubishi Electric Research Laboratories as research intern. Corresponding Author: Bingnan Wang (bwang@merl.com).

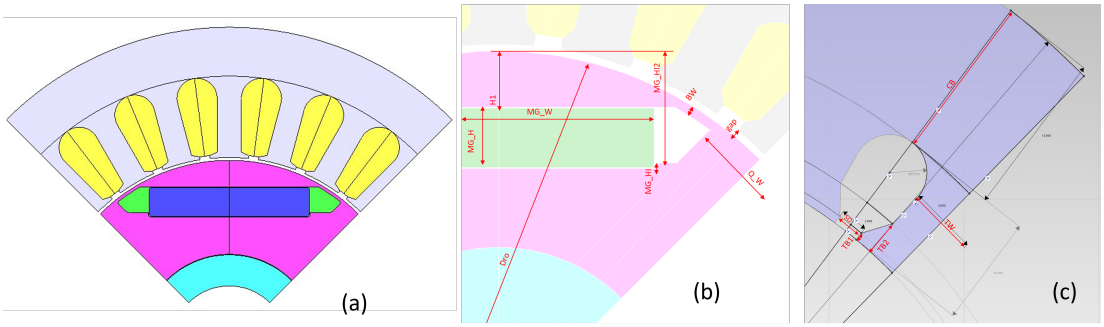

Fig. 1. (a) A section of the magnetic design for an example IPM motor; (b) tunable design parameters on rotor, and (c) tunable design parameters on the stator.

the motor design parameters to initialize the class token of the ViT model, and show that this approach improves the data efficiency of the model, and achieves improved prediction accuracy.

II. Problem Defintion

In this work, we develop surrogate models to evaluate the cogging torque of 4-pole 24-slot IPMSM. The topology of the motor is shown in Fig. 1. The values of 13 design parameters, as indicated in Fig. 1(b) and (c) are tunable. To create the dataset for the study, a total of 19,373 design candidates are generated, each have a distinct combination of the 13 parameters. FEM simulations are then performed for each of the design candidate to obtain the torque waveform under no-load condition. The obtained torque waveform is further decomposed into the following Fourier series including the dominating harmonic terms:

$$\mathrm{T}(\theta) = A_{12}\cos(12\theta) + B_{12}\sin(12\theta) + A_{24}\cos(24\theta) + B_{24}\sin(24\theta) \quad (1)$$

This way, instead of recording the waveform by the torque at each rotor angle, we can represent it with only four Fourier coefficients:$A_{12}, B_{12}, A_{24}, B_{24}$. In addition, the induced voltage is also calculated. For each entry in the compiled dataset, the input includes the value of 13 design parameters, and corresponding 2D cross section image as shown in Fig. 1(a), and the output includes the EMF, and four Fourier coefficients for cogging torque.

To predict the cogging torque, the Fourier coefficients are first predicted by a surrogate model, which are used to recover the torque waveform according to Eq.(1). Finally the cogging torque is determined by the difference between maximum and minimum value of the torque waveform: $T_c = \max(T(\theta)) - \min(T(\theta))$. Based on previous studies [21], the performance of cogging torque prediction with this approach is generally better than directly predicting cogging torque value using the same model. We aim to develop models that make predictions for the cogging torque, with the objective of reducing the root-mean-square error (RMSE) of the prediction.

III. Vision Transformer for Cogging Torque Prediction

One popular method of deep learning-based surrogate models for motor design is to represent the magnetic design of a motor design candidate as a standard RGB image, and feed it into a deep CNN model for performance prediction. A CNN model extracts features by applying local spatial filters to adjacent pixels, and by applying this operation over many layers, the resulting features can cover larger areas. To capture features representing the motor design across the whole image, the CNN-based network needs to be quite deep, with many layers, and the extracted features can become very abstract. While the approach generally performs better than other existing surrogate models, the accuracy for cogging torque and torque ripple prediction is still insufficient.

A. Vision Transformer for Motor Design

Recently, Vision Transformer [10] architectures demonstrated strong capabilities in image-related tasks. ViT adopts the multi-head self-attention mechanism to grasp the correlation between different input parts of an image. In this section, we will introduce why the Vision Transformer is more proficient for Motor Design.

1) Multi-head Self-Attention Layer: In a transformer-based model, each input vector represents a token. The model processes a sequence of tokens as input, and its self-attention layer enables each token to attend to every other token, including itself. This mechanism allows the model to effectively capture dependencies and contextual relationships within the sequence.

Generally, each token will be transformed to query token Q, key token K, and value token V. These transformations are achieved through three learned matrices W_Q, W_K and W_V

$$Q = XW_Q, K = XW_K, V = XW_V. \quad (2)$$

Here, the query token Q represents the current token that searches for relevant information. The key K represents all tokens in the sequence used for comparison. The value V contains the actual information carried out by each token.

The self-attention mechanism then compares each query Q with all keys K to compute a relevance score, which determines how much attention the token should pay to others.

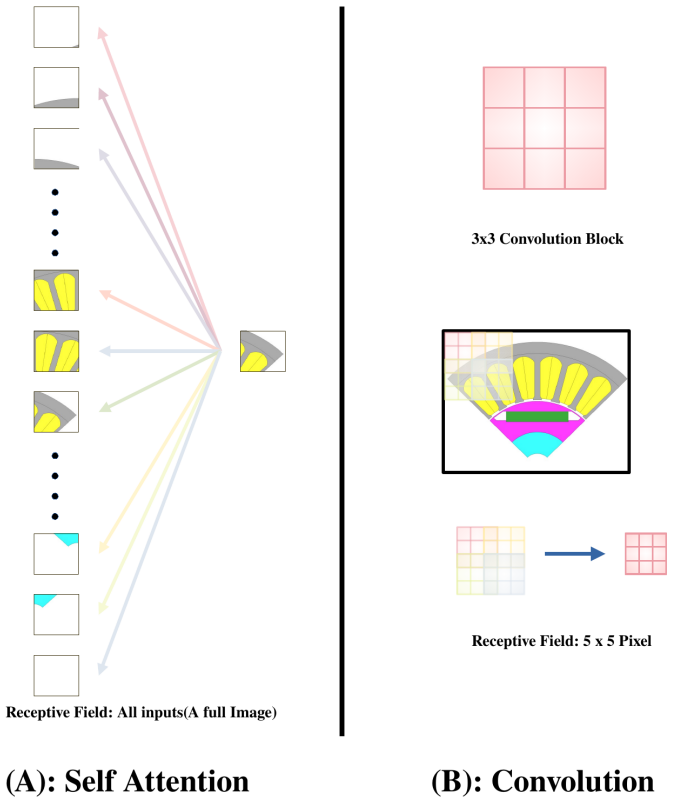

Fig. 2. Illustration of receptive field for self attention and convolution. Left: Self attention. The receptive field is the full input image. Right: Convolution. The receptive field is 5×5 pixels for a 3×3 convolution.

Finally, the output is a transformed representation of the input, where each token is encoded based on its relevance score and the corresponding value V. This process enables the model to capture meaningful contextual relationships across the sequence, contributing to the model's ability to understand complex patterns. Furthermore, we can align several self-attention layers in parallel to create a multi-head self-attention layer. Each self-attention layer can learn different features in a multi-head self-attention layer, leading to a more comprehensive understanding of the input data.

2) Vision Transformer: The Multi-head Self-Attention Layer is originally designed for sequence data, such as text or time series. To adapt this mechanism for image data, some modifications are necessary. This led to the proposal of the Vision Transformer (ViT).

In ViT, each input image I is divided into a sequence of fixed-size 2D patches. This sequence of 2D patches is then passed through a convolution layer to generate a sequence of tokens, along with a special token known as the CLS token. This results in a token sequence X derived from the image, where the CLS token is a learnable embedding that serves as a context aggregator, gathering information from all other tokens in the sequence.

This token sequence X is then treated as input and passed through a series of self-attention layers to obtain an encoded vector $X^{'}$. The information contained in the CLS token of $X^{'}$ is then used for further prediction tasks.

3) Why ViT is more suitable for Motor Design?: Motor design images are unique in that they typically share the same format, with variations primarily in the size and shape of different parts. Therefore, the global features such as overall structure and spatial relationships between parts play a more significant role compared with local features at pixel levels.

Compared with the convolution layer in CNN-based methods, the self-attention layer in ViT can grasp global features more thoroughly. This difference is evident when comparing the receptive fields of these two approaches. In deep learning, the receptive field indicates the connection between the output (feature) and the input. Figure 2 compares the receptive fields of ViT and CNN-based networks. For the CNN-based network, the receptive field size is related to its kernel size of convolution operation. For instance, a 3×3 convolution kernel can cover a 5×5 pixel area. To cover a larger area, a very deep network is required. On the other hand, as shown on the left, one patch interacts with all other patches in a transformer. Thus, the receptive field for the transformer-based network encompasses the entire input space from the first to the last layer. This extensive receptive field makes the

transformer better at learning global and structural features of inputs compared to traditional CNN-based networks [22].

With this understanding, in this paper, we apply ViT-based models to predict the physical response of a motor design. The input is an RGB image representing a motor design, which is passed through the ViT-based model to obtain an encoded sequence. The information contained in the "CLS token" is then extracted. Finally, a linear layer transforms the "CLS token" to produce the final prediction, which includes the Fourier coefficients for cogging torque prediction, as well as the induced voltage, labeled as EMF.

B. Parameter as prior knowledge

One concern for training ViT-based models is the limited amount of training data available, which can largely hinder the performance of network if training from scratch. In this paper, we implement two strategies to address this problem. First, we start from models pre-trained on ImageNet for image classification tasks, and fine-tune the models on our IPM motor dataset. However, most pre-trained networks are trained on ImageNet for image classification tasks, which differ greatly from motor design tasks. Therefore, we need an efficient way to fine-tune pre-trained ViT models on motor design data.

To address this issue, we effectively combine the design information in the form of a 2D image and a list of parameters into the training process. In particular, we propose using the parameter list as sample-specific information for CLS token initialization.

As mentioned in section II, a motor design sample can be described either by an image of the magentic design, or a list of design parameters. One nature thought to improve the performance of a surrogate model is to combine the information of the design parameters and the images. However, how to effectively combine the two types of pint is not straightforward, as the dimensions of these two inputs are largely different. A image is usually 3-dimension ($224\times224\times3$ for all images in the dataset) and list of parameters is a vector (13×1 in our case). Simply concatenate the two inputs will render the parameter list ineffective due to its much lower dimension. Simply upscaling the parameter list to the same dimension of the image is not a good solution either, as this will largely distort the information included in the parameter. Considering the special structure of Vision Transformer, we propose to utilize the parameter list as sample-specifc information and use it to initialize the "CLS token", which can help the Vision Transformer to better converge on motor design data. The CLS-token is a special token that gathers information from all other patches. The information embedded in the CLS token will be directly utilized to make the final prediction. Generally, there are two methods to initialize the CLS token: initializing it as a zero token or initializing it with pre-trained parameters. However, these methods do not contain sample-specific information, which is crucial for fast convergence. To address this problem, we consider the parameter list as sample-specific information and encode it as the initialization of CLS token.

Specifically, we employ a linear transformer to convert the parameter sequence into the same size as a token in the Vision Transformer. Then, we assign the encoded parameter sequence as the initialization for the CLS token and pass the entire token list to the multi-head attention layer (MSA layer). In the MSA layer, the information from the parameter list interacts with all the tokens in the image token list. The MLP layer is co-trained with the Vision Transformer, which help generate better initialization for the CLS token. This architecture shown in Fig. 3.

This setting provides several benefits for Vision Transformer on motor design task. First, the upscaling does not distort the information of parameter sequence. The size of the CLS token is very small comparing with the entire token lists, which means that the parameter list does not need to be upscaled to a large dimension. This preserves the information contained in the parameter list. Moreover, the parameter list can interact with all parts of the image, ensuring that the model fully integrates the information from the parameter list into the image representation. This enhances the model's understanding of the input data especially when dealing with various design patterns. Additionally, the parameter-based CLS token is fully customized based on input designs, providing the model with tailored information for each sample. This customization enables the model to predict challenging samples better, enhancing its overall performance and robustness.

IV. Numerical Tests & Results

Model	Parameter Number
VGG16	138M
VGG19	144M
ResNet50	23.9M
ResNet101	42.8M
ResNet152	58.5M
ViT-B/16	86M
ViT-B/32	88M
ViT-L/16	304M
ViT-L/32	305M

TABLE I. Parameter Number for each network

A. Models and Hyperparameters

To compare the performance of the ViT models, we choose 5 CNN-based models as baseline: VGG16, VGG19 [23], ResNet50, ResNet101 and ResNet152 [24]. For ViT models, we evaluate five of the variants: ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32 [9] and ViT-B/32 with parameter initialized CLS token. Table I shows the hyperparameter number for each baseline. The parameter numbers for all baselines except ViT-L/16 and ViT-L/32 are much larger and are on the same order of magnitude(around 100M parameters). ViT-L/16 and ViT-L/32 have much larger parameter numbers(around 300M parameters). For both the CNN-based models and Vision Transformer models, we train for a total of 300 epochs with a

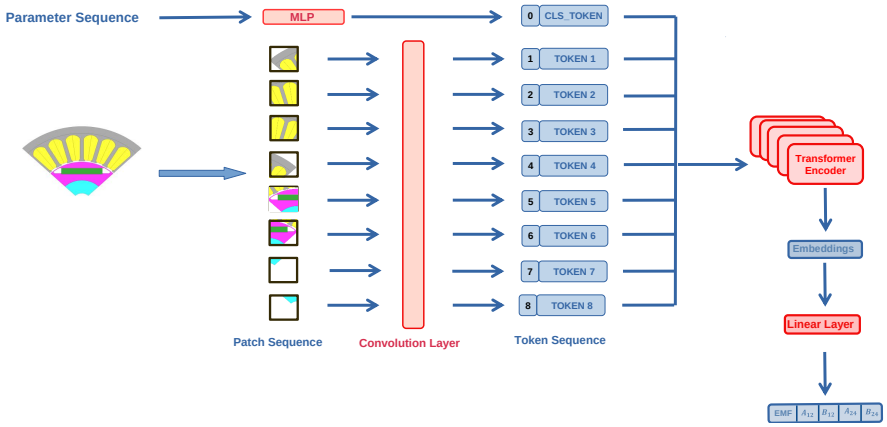

Fig. 3. Illustration of Proposed ViT model with parameter as prior Knowledge

batch size of 128. For the CNN-based methods, the learning rate is set to 0.0001, while for the Vision Transformer methods, we use an initial learning rate of 0.009 with a cosine annealing scheduler to gradually adjust the learning rate during training.

B. Results

In this section, we conduct numerical experiments to validate the proposed methods and compare the ViT-based models against state-of-the-art CNN models. The IPM dataset is split into training (70%), validation (10%), and test (20%) datasets. For comprehensive evaluation, all nine models introduced in Section IV-A are compared with our proposed Vision Transformer model equipped with a parameter-initialized CLS token. For fair comparison, all models are pre-trained on ImageNet and fine-tuned with the IPM motor training dataset.

The prediction results evaluated by RMSE on test data are shown in Table II. The experiment supports two major claims in our paper. First, the Vision Transformer outperforms CNN-based methods. For EMF prediction, all models perform well, with RMSE around 0.1. For cogging torque prediction on the test dataset, all CNN-based models perform similarly, with VGG16 (RMSE 0.905) and ResNet50 (RMSE 0.979) models having slightly better performance than the others. In contrast, all ViT based models perform much better, with RMSE below 0.3. The best performing model is ViT-L/16, which achieves RMSE of 0.215.

And Our proposed method significantly improves Vanilla Vision Transformer (ViT) performance. Using the ViT-B/32 model as the backbone, we evaluated our approach on the Flat and Vshape datasets. The results indicate that our method consistently outperforms Vanilla Vision Transformer models. Specifically, on the Flat dataset, the ViT-B/32 model with an initialized CLS token achieves an RMSE of 0.083 on EMF and an RMSE of 0.182 on torque ripple, which outperform all Vanilla Vision Transformer models. On the Vshape dataset, the ViT-B/32 model with an initialized CLS token achieves an RMSE of 0.102 on EMF and an RMSE of 0.133 on torque ripple, outperforming all Vanilla Vision Transformer models. These consistent improvements demonstrate that our proposed methods significantly enhance the efficiency of the ViT model in fine-tuning for model design datasets.

C. Visualization

To further demonstrate the effectiveness of the proposed models, we plot model prediction vs. true values for all test data in Fig. 4. In addition to RMSE, we also evaluate the model prediction performance with R-squared (R^2) value, which indicates how well a model predicts a variable. The R^2 value ranges from 0 to 1, with values closer to 1 signifying more accurate predictions. The results are reported in Figure 4

For EMF prediction, all models perform well on test data, while the ViT models perform even better with almost perfect predictions as indicated by the low RMSE values and R^2 approaching 1. Cogging torque prediction performance is much worse for deep CNN models VGG16 and ResNet50, with many outlier points away from the diagonal. In contrast, the ViT-L/16 model with a pre-trained CLS token performs significantly better, with R^2 value reaching 0.997. With a parameter-initiated CLS token, a smaller ViT-B/32 model further improves the performance, with reduced RMSE and R^2 of 0.998.

D. Ablation study

In this section, we want to further check the efficiency of parameter-initialized CLS token with the other two CLS token initialization methods. One commonly used approach is to take the pre-trained CLS token, which contains information from the dataset used in the training process, and fine-tune them on the motor dataset. Another approach is to initialize the CLS token to zero, which means the token does not contain any prior information, and tune it on the motor dataset. Finially, in our proposed approach, we utilize the motor design parameters, and initialize the CLS token with these design information through an MLP network. The comparison of the three methods is done using the smaller ViT-B/32 model, and the results are shown in Table III.

As we can see, with parameter-initialized CLS token, the model outperforms the other two token initialization methods.

Model	A_{12}	B_{12}	A_{24}	B_{24}	EMF	Torque Ripple
ResNet152	0.004 ± 0.002	0.498 ± 0.452	0.002 ± 0.001	0.230 ± 0.204	0.130 ± 0.087	0.995 ± 0.897
ResNet101	0.005 ± 0.004	0.519 ± 0.468	0.008 ± 0.003	0.230 ± 0.200	0.138 ± 0.095	1.010 ± 0.904
ResNet50	0.020 ± 0.005	0.492 ± 0.440	0.006 ± 0.003	0.223 ± 0.192	0.141 ± 0.095	0.979 ± 0.873
VGG19	0.002 ± 0.001	0.454 ± 0.403	0.002 ± 0.001	0.222 ± 0.194	0.171 ± 0.114	0.938 ± 0.829
VGG16	3.7e-3 ± 1.8e-3	0.453 ± 0.404	9e-4 ± 7e-4	0.213 ± 0.185	0.175 ± 0.119	0.905 ± 0.806
ViT-B-16	1e-3 ± 9e-4	0.122 ± 0.085	6e-4 ± 5e-4	0.065 ± 0.044	0.105 ± 0.068	0.258 ± 0.179
ViT-B-32	1e-3 ± 9e-4	0.128 ± 0.095	6e-4 ± 5e-4	0.072 ± 0.049	0.128 ± 0.077	0.272 ± 0.201
ViT-L-16	1e-3 ± 9e-4	0.106 ± 0.077	6e-4 ± 5e-4	0.059 ± 0.039	0.090 ± 0.058	0.215 ± 0.151
ViT-L-32	1e-3 ± 9e-4	0.148 ± 0.112	6e-4 ± 5e-4	0.075 ± 0.052	0.142 ± 0.090	0.309 ± 0.231

TABLE II. Performance Comparison for predicting EMF and Torque Ripple. 5 CNN-based methods: VGG16,VGG19, ResNet50, ResNet101, ResNet152. 4 ViT-based methods:ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32 with pretrained "CLS-token". We report RMSE as evaluation metric.

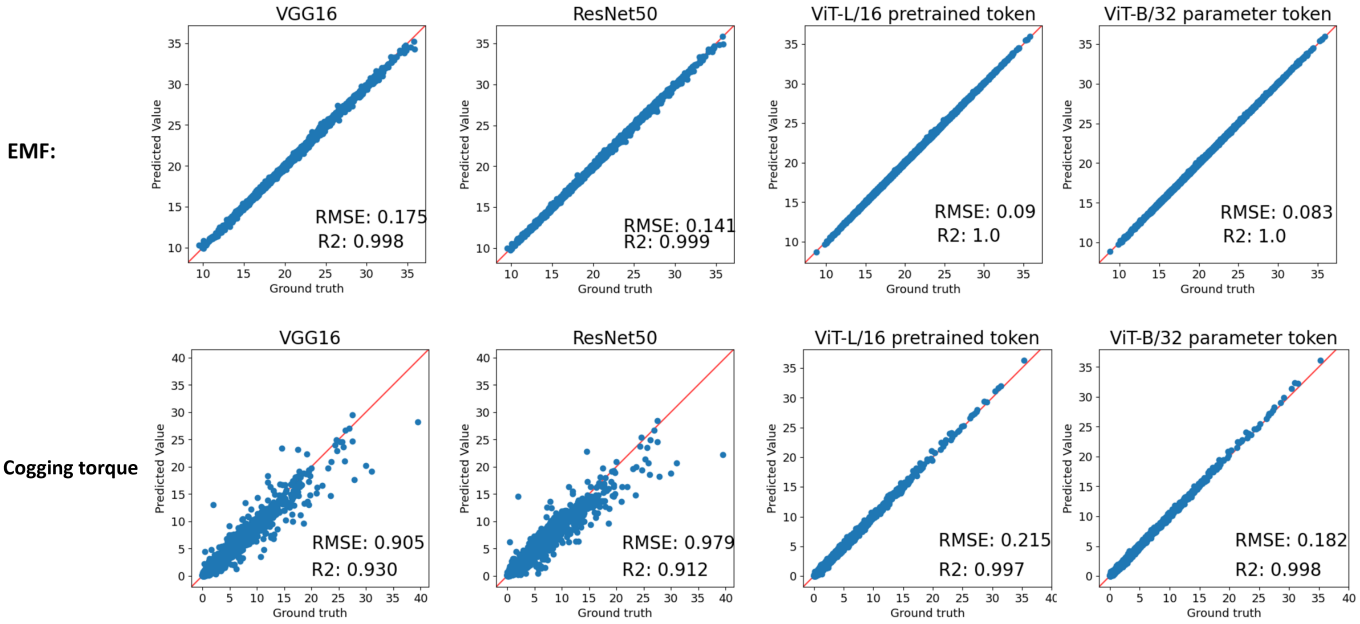

Fig. 4. Visualization for four models on the Flat dataset. Four models are VGG16, ResNet50, ViT-L/16 with pretrained token and ViT-B/32 with parameter token. For each figure, we also report corresponding RMSE and R-squared value.

CLS Token	A_{12}	B_{12}	A_{24}	B_{24}	EMF	Torque Ripple
Pre-trained	0.001 ± 0.0009	0.128 ± 0.095	6e-4 ± 5e-4	0.072 ± 0.049	0.128 ± 0.077	0.272 ± 0.201
Zero initialized	0.001 ± 0.0009	0.123 ± 0.089	6e-4 ± 5e-4	0.069 ± 0.046	0.127 ± 0.077	0.260 ± 0.187
Parameter-initialized	0.001 ± 0.0009	0.085 ± 0.064	6e-4 ± 5e-4	0.056 ± 0.038	0.083 ± 0.051	0.182 ± 0.134

TABLE III. Performance comparison for predicting EMF and torque ripple with different initialization of "CLS-token". We compared three initializations: pretrained-initialized token, zero-initialized token, and parameter-initialized token. The backbone network is ViT-B/32.

The model with zero-initialized token also performs slightly better than the model with pretrained token initialization.

These results support our claim in the methods section. Pretrained datasets, such as ImageNet, differ significantly from the motor design dataset. Consequently, the pretrained-initialized CLS token can hinder the fine-tuning process due to the discrepancy in dataset characteristics. The zero-initialized token, which lacks prior information, neither aids nor obstructs the fine-tuning process. In contrast, our proposed parameter-initialized token integrates sample-specific information, significantly enhancing the efficiency fine-tuning process.

V. Conclusion

In this work, we developed Vision Transformer-based deep learning models serving as surrogate for accurate cogging torque prediction. We adopted pre-trained ViT models and fine-tune them on the IPM motor dataset, and showed that ViT models, with larger receptive field, can capture important features more effectively than deep CNN based-models and achieve much high prediction accuracy. We further utilized sample-specific motor design parameters to initialize the CLS token in the ViT model, which further improves the prediction performance, even with a smaller model.

REFERENCES

[1] R. C. P. Silva, T. Rahman, M. H. Mohammadi, and D. A. Lowther, "Multiple operating points based optimization: Application to fractional slot concentrated winding electric motors," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 2, pp. 1719–1727, 2017.

[2] S. Doi, H. Sasaki, and H. Igarashi, "Multi-objective topology optimization of rotating machines using deep learning," *IEEE transactions on magnetics*, vol. 55, no. 6, pp. 1–5, 2019.

[3] A. Khan, V. Ghorbanian, and D. Lowther, "Deep learning for magnetic field estimation," *IEEE Transactions on Magnetics*, vol. 55, no. 6, pp. 1–4, 2019.

[4] W. Kirchgässner, O. Wallscheid, and J. Böcker, "Estimating electric motor temperatures with deep residual machine learning," *IEEE Transactions on Power Electronics*, vol. 36, no. 7, pp. 7480–7488, 2020.

[5] Y.-m. You, "Multi-objective optimal design of permanent magnet synchronous motor for electric vehicle based on deep learning," *Applied Sciences*, vol. 10, no. 2, p. 482, 2020.

[6] Y. Shimizu, S. Morimoto, M. Sanada, and Y. Inoue, "Automatic design system with generative adversarial network and convolutional neural network for optimization design of interior permanent magnet synchronous motor," *IEEE Transactions on Energy Conversion*, vol. 38, no. 1, pp. 724–734, 2022.

[7] A. Choudhary, D. Goyal, and S. S. Letha, "Infrared thermography-based fault diagnosis of induction motor bearings using machine learning," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1727–1734, 2020.

[8] T. Aoyagi, Y. Otomo, H. Igarashi, H. Sasaki, Y. Hidaka, and H. Arita, "Prediction of current-dependent motor torque characteristics using deep learning for topology optimization," in *2021 23rd International Conference on the Computation of Electromagnetic Fields (COMPUMAG)*, 2022.

[9] 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.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 6000–6010, Curran Associates Inc., 2017.

[11] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 2, pp. 60–65, Ieee, 2005.

[12] T. Chen, S. Saxena, L. Li, D. J. Fleet, and G. Hinton, "Pix2seq: A language modeling framework for object detection," *arXiv preprint arXiv:2109.10852*, 2021.

[13] Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, and W. Liu, "You only look at one sequence: Rethinking transformer in vision through object detection," 2021.

[14] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," 2019.

[15] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," 2021.

[16] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," 2021.

[17] X. Wang, C. Yeshwanth, and M. Nießner, "Sceneformer: Indoor scene generation with transformers," 2021.

[18] Y. Jiang, S. Chang, and Z. Wang, "Transgan: Two pure transformers can make one strong gan, and that can scale up," 2021.

[19] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *International conference on machine learning*, pp. 1691–1703, PMLR, 2020.

[20] Y. Shimizu and K. Akatsu, "Deep learning-based automatic design system for ipmsm with variable magnet properties," in *2024 International Conference on Electrical Machines (ICEM)*, pp. 1–6, IEEE, 2024.

[21] Y. Sakamoto, Y. Xu, B. Wang, T. Yamamoto, and Y. Nishimura, "Electric motor surrogate model combining subdomain method and neural network," in *2023 24th International Conference on the Computation of Electromagnetic Fields (COMPUMAG)*, pp. 1–4, 2023.

[22] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, (Red Hook, NY, USA), Curran Associates Inc., 2024.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.