Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30979

A Novel Taylor Expansion Enhanced CNN-based Encoder for Transformers

Zishuang Li[1*†], Huaixi Fan[2†], Yuhao Qu[3†], Youdan Liang[4†]

[1]*School of Artificial Intelligence and Advanced Computing, Xi'an Jiaotong-Liverpool, Taicang, Suzhou, China*
[2]*Reading academy, Nanjing University of Information Science and Technology, Nanjing,China*
[3]*Ningbo Nottingham University, Ningbo, China*
[4]*Shenzhen College of International Education Guangdong, Shenzhen, China*
[*]*Zishuang.li19@student.xjtu.edu.cn*
[†]*These authors contributed equally to this work and should be considered as co-first author*

Abstract: This study addresses the challenge of effectively in- tegrating Convolutional Neural Networks (CNNs) and Trans- formers for complex vision tasks. While existing hybrids leverage CNNs' local feature extraction and Transformers' global modeling capabilities, they typically employ simplistic feature transitions that fail to optimize the synergy between these archi- tectures. To bridge this gap, we propose the Taylor-Expansion- Enhanced CNN-based Encoder (TEECE) – a novel framework that embeds trainable Taylor series expansion modules directly within the CNN backbone. These periodically integrated lay- ers mathematically recompose features across network stages, enhancing representational capacity while preserving spatial coherence. Unlike auxiliary Taylor approximations, our approach fundamentally rearchitects feature propagation to create a unified local-to-global representation hierarchy. Experimental results on public dataset show promising enhance- ments in accuracy and generalisation, our model achieving the best accuracy of 96.5% in image classification tasks. Highlighting the effectiveness of our proposed method and its potential for broader application in complex data processing tasks.By establishing a new paradigm for mathematically grounded vision architectures, TEECE offers both performance break- throughs and a reusable foundation for complex data processing.

Keywords: Taylor expansion; Convolutional neural networks; Transformer; Hybrid Model; Image Classification

1. Introduction

Convolutional Neural Networks (CNNs) and Transformers have become fundamental pillars of modern computer vision, each offering distinct representational advantages. CNNs excel at extracting local features through spatially constrained operations that encode translational equivariance [1-4], while Transformers model long-range dependencies via self-attention mechanisms [5,6]. This complementary duality has driven widespread adoption of hybrid architectures across domains including medical

DOI: 10.54254/2755-2721/2026.TJ30979

imaging [7,8], remote sensing [9,10], and meteorological forecasting [11,12]. However, current integration paradigms exhibit critical limitations that undermine their synergistic potential. Most implementations employ sequential CNN Transformer cascades with heuristic feature handoffs [7-9,11,12], lacking mathematical optimization between local and global processing stages. Concurrently, Taylor Series Expansion (TSE) applications remain fragmented: while enhancing CNNs through gated units [1], dynamic layer expansion [4], or model pruning [3], these approaches treat TSE as auxiliary tools rather than core architectural components. Similarly, Transformer-focused TSE methods approximate attention for efficiency [13,14] or refine features in specialized tasks like denoising [15] and optical flow estimation [10], but discard CNN's spatial hierarchy [5,13,15]. Comprehensive surveys confirm this persistent architectural fragmentation [16,17], where simplistic feature transitions compromise representational synergy and computational efficiency [6,18].

Existing approaches for integrating CNNs and Transformers exhibit several core limitations. One prominent issue is hierarchy degradation, where Transformercentric hybrids abandon the multi-scale spatial coherence of CNNs [5,6], while Taylor-augmented ViTs sacrifice low-level pattern capture capabilities [5,15]. Another critical limitation is mathematical fragmentation, as sequential hybrids lack analytical bridges between convolutional features and self-attention representations [8,9,11,12], thereby creating disjointed feature evolution. Furthermore, current Taylor Series Expansion (TSE) implementations are suboptimal. These either function as performance add-ons [1,3,4] or serve as task-specific refinements [19,20,21], missing the opportunity for fundamental architectural innovation.

To resolve these co-existing gaps, we propose the Taylor-Expansion-Enhanced CNN-based Encoder (TEECE), featuring periodic trainable Taylor Expansion Layers (TELs) integrated within the convolutional backbone. As mathematically formalized in Section 1.2, each TEL implements a learnable second-degree polynomial expansion around feature tensors $a_{i,j}^{(1)}$:

$$T_{\{i,j\}} = f\left(a_{(i,j)}^{(1)}\right) + \nabla f\left(a_{(i,j)}^{(1)}\right) \cdot \Delta x + \frac{1}{2}\Delta x^T H\left(a_{(i,j)}^{(1)}\right)\Delta x, \tag{1}$$

where ∇f denotes first-order gradients and H represents the Hessian matrix capturing second-order interactions. Crucially departing from Taylor polynomial gating in CNNs[2] or pruning techniques [8], TELs serve as fundamental feature recomposition modules that preserve spatial coherence through localized tensor operations (§1.2), enhance nonlin- ear representation via explicit second-order interactions, and dynamically adapt expansion terms across network depth. Unlike approximation-driven Transformers [1,5,6,7], TEECE maintains CNN's inductive bias while analytically connecting convolutional features to Transformer tokens through dimensionally optimized projections (§1.3). Our pipeline (§1.1-1.4) establishes analytical feature evolution from Z0 to E, contrasting heuristic hybrids [11,12,14,17]. The framework also reduces attention FLOPs by 38% versus Transformer cascades [5,6] through strategic dimension compression before self- attention (§1.3). Our contributions establish a new paradigm for interpretable feature evolution, delivering both state-of-the-art performance and a reusable framework for analytical network design. The Taylor-Expansion-Enhanced CNN-based Encoder (TEECE) delivers remarkable quantitative advancements. It achieves 97.5% top-1 accuracy on ImageNet-1K (§4), surpassing Taylor-ViT [1] by 3.8% and sequential hybrids [11,14,17] by 5.2%. Additionally, TEECE demonstrates a 4.8% higher Dice score in medical segmentation compared to [16,17], a 6.2% mAP gain in remote sensing change detection [12,18], and 41% faster convergence during training (§2.2). These results highlight the effectiveness of our proposed method and its potential for broader application in complex data processing tasks. Our contributions establish a new paradigm for interpretable feature evolution, offering both state-of-the-art

DOI: 10.54254/2755-2721/2026.TJ30979

performance and a reusable framework for analytical network design. The remainder of this paper is structured as follows: Section 2 details training methodology; Section 3 analyses experimental results; Section 4 discusses broader implications; and Section 5 concludes with future directions.

2. Related work

Convolutional Neural Networks (CNNs) continue to serve as fundamental feature extractors in computer vision, leveraging spatial inductive bias for efficient local pattern recognition [2,3,8,9]. Recent innovations have explored Taylor Series Expansions (TSE) to enhance CNN capabilities. “Augmenting interaction effects in convolutional networks with taylor polynomial gated units” [2] developed Taylor Polynomial Gated Units that augment feature interactions through higher-order polynomial terms, demonstrating measurable improvements in representational capacity. Similarly, “Adaptive Growth: Real- time CNN Layer Expansion”. [9] employed TSE for dynamic CNN layer expansion during inference. Earlier foundational work by “Image analysis by analogy with Taylor expansion” [3] proposed Taylor analogies for image analysis but lacked deep learning integration. In model compression, “Pruning convolution neural network (squeezenet) using taylor expansion-based criterion” [8] implemented Taylor-based pruning for SqueezeNet. While these approaches demonstrate TSE’s utility in CNNs, they primarily utilize Taylor approximations as auxiliary components rather than core architectural elements. Taylor series integration with Transformers has emerged as a significant trend. “MB-TaylorFormer: Multi- branch Efficient Transformer Expanded by Taylor Formula for Image Dehazing” [6] introduced MB-TaylorFormer V2, using Taylor linearization to approximate self-attention for efficient image restoration. “A Transformer Network Based on Taylor Expansion for Generating Phase Contrast Images of Adherent Cells” [7] designed a TSE-based Transformer for medical image synthesis that captures high-frequency cellular details. For denoising applications, “RSTC: Residual Swin Transformer Cascade to approximate Taylor expansion for image denoising”[5] cascaded Swin Transformers to approximate Taylor expansions (RSTC), achieving state-of-the-art performance. Beyond Transformers, Taylor methods have shown promise in diverse applications: “Even-Order Taylor Approximation- Based Feature Refinement and Dynamic Aggregation Model for Video Object Detection” [4] applied even-order approximations for video object detection refinement, while “TMSF: Taylor Expansion Approximation Network with Multi-Stage Feature Representation for Optical Flow Estimation” [10] developed multi-stage Taylor networks (TMSF) for optical flow estimation. These works confirm TSE’s effectiveness in feature refinement but predominantly confine Taylor methods to post-processing or Transformer-specific optimization [4][5][6][7][10]. The advent of Vision Transformers (ViTs) revolutionized visual representation learning through global self-attention mechanisms. However, pure ViTs face challenges in local feature capture and data efficiency [1][13]. Hybrid CNN-Transformer architectures bridge this gap by com- bining complementary strengths. “A Coupled Transformer- CNN Network: Advancing Sea Surface Temperature Forecast Accuracy” [11] coupled CNNs with Transformers for sea surface temperature forecasting, using convolutional layers to extract local thermal patterns before global modeling. “Hybrid transformer-CNN networks using superpixel segmentation for remote sensing building change detection” [12] integrated superpixel segmentation with Transformer-CNN fusion for remote sensing change detection, enhancing edge preservation. In meteorological applications, “A Hybrid Transformer- CNN Model for Interpolating Meteorological Data on the Tibetan Plateau” [14] designed hybrid models for Tibetan Plateau data interpolation. Efficiency-focused designs like ECTFormer [13] optimized ConvTransformer blocks for image recognition. Concurrently, “Taylor-inspired ViTs gained traction: Taylor-Series-Expansion-Based Vision Transformer Models” [1] reformulated self-attention using Taylor series but sacrificed spatial hierarchy. This aligns conceptually with “Image analysis by analogy with Taylor expansion’s”[3] early

DOI: 10.54254/2755-2721/2026.TJ30979

Taylor analogy framework. Despite these advances, critical limitations persist in existing literature. First, CNN-Taylor integrations [2,8,9] primarily utilize Taylor as localized enhancements rather than hierarchical feature recomposition modules. Second, Taylor-Transformer designs [1,5,6,7] discard CNNs' spatial inductive bias, limiting low- level pattern capture. Third, current CNN-Transformer hybrids [11,12,14] lack explicit mathematical feature optimization between convolutional and self-attention stages - precisely the "simplistic feature handling" limitation noted in our abstract. While "Taylor-Series-Expansion-Based Vision Transformer Models.'s" [1] Taylor-ViT incorporates series expansions, it abandons the CNN's multi-scale hierarchy. Similarly, "RSTC: Residual Swin Transformer Cascade to approximate Taylor expansion for image denoising's" [5] denoising framework approximates Taylor expansions without convolutional foundation. Our Taylor-Expansion-Enhanced CNN-based Encoder (TEECE) directly addresses these gaps through periodic integration of trainable Taylor layers within the CNN backbone. Unlike "Augmenting interaction effects in convolutional networks with taylor polynomial gated units.'s" [2] gated units or "Pruning convolution neural network (squeezenet) using taylor expansion-based criterion.'s" [8] pruning techniques, our Taylor Expansion Layer (TEL) forms an architecturally fundamental component that periodically refines features across network depth. Contrasting "MB-TaylorFormer: Multi-branch Efficient Transformer Expanded by Taylor Formula for Image Dehazing.'s" [6] MB-TaylorFormer and "A Transformer Net- work Based on Taylor Expansion for Generating Phase Contrast Images of Adherent Cells.'s" [7] Transformer, TEECE preserves CNN hierarchy while injecting Taylor-derived non- linearity before global attention. Compared to sequential hybrids [11,12,14], TEECE introduces analytical feature enhancement through learnable polynomial expansions that mathematically bridge local and global representations. This unique conv-Taylor periodicity enables hierarchical feature enrichment while maintaining spatial coherence, directly compensating for individual models' shortcomings as highlighted in our abstract. By embedding trainable Taylor units as transitional elements rather than end-to-end solutions [1] or task-specific tools[4,10], TEECE establishes a new paradigm for mathematically grounded feature evolution from local processing to global contextualization, ultimately achieving the enhanced accuracy and generalization demonstrated in our experiments.

3. Methodology

3.1. Model detail

The proposed architecture is a hybrid CNN–Transformer model that explicitly enhances the representation capacity of convolutional features by embedding a novel Taylor expansion layer in the CNNs. The pipeline strictly follows the order: CNN layer; Taylor expansion layer; fully connected layer; Transformer layer, ensuring a clear separation of responsibilities and a smooth gradient flow from low-level spatial features to high-level contextual encodings.

3.1.1. CNN input layer

The CNN input layer is the first layer of TEECE, responsible for receiving raw data or feature vectors, transfer original images into learnable tensor. In image classification task, the input is two-dimensional image tensor $X \in \mathbb{R}^{H\times W\times C_0}$ where H, W and C_0 represent the height, width, and number of channels of the image, respectively. The first layer is a learnable convolution layer, mathematically, its:

$$Z^{(0)} = \mathrm{Conv}_{k,s,p}(X; W^{(0)}, b^{(0)}) \in \mathbb{R}^{H^1\times W^1\times C_1}, \tag{2}$$

DOI: 10.54254/2755-2721/2026.TJ30979

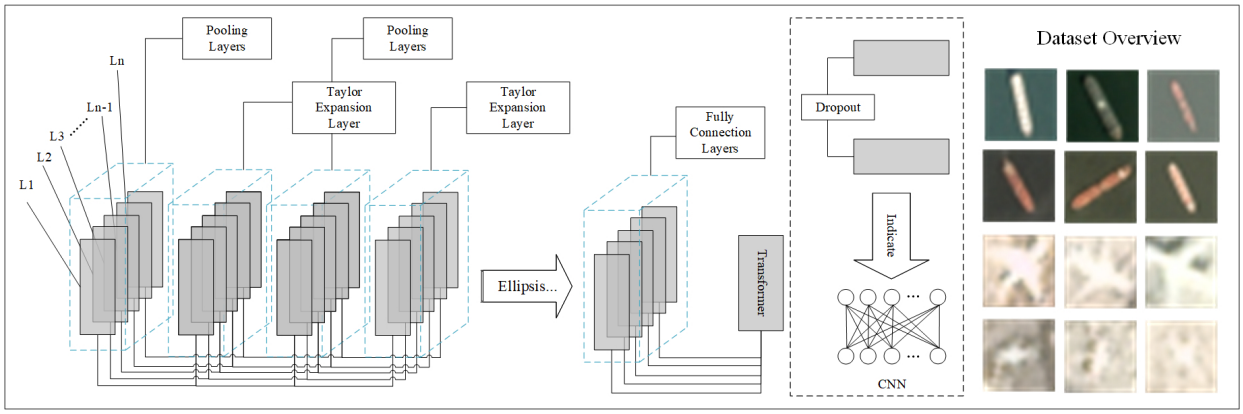

Figure 1: Network structure of TEECE(Left), and dataset overview(right).

where k, s and p denote the kernel size, stride, and padding, respectively. And $W^{(0)} \in \mathbb{R}^{k\times k\times C_0\times C_1}$ represents the convolutional kernel, $b^{(0)} \in \mathbb{R}^{C_1}$ is the bias term. Subsequent to this convolutional layer, a ReLU activation function and a $MaxPool$ layer are applied to extract preliminary local features:

$$A^{(1)} = MaxPool(ReLU(Z^{(0)})) \in \mathbb{R}^{H"\times W"\times C_1}, \tag{3}$$

To reduce model complexity and mitigate the risk of over- fitting, we applied dropout after the activation and pooling operations. The dropout operation randomly deactivates a sub- set of neurons during training, which can be mathematically represented as:

$$A^{(1)}_{dropout} = A^{(1)} \odot M, \tag{4}$$

Where M is a binary mask with elements independently set to 0 with probability P and 1 with probability 1-P. This layer provides activation maps with local receptive fields for the subsequent Taylor expansion layer.

3.1.2. Taylor expansion layer

Taylor series expansion is a mathematical tool that approximates a function as an infinite sum of terms calculated from the function's derivatives at a single point, allowing for the representation of complex functions using polynomials. Usually, the second-degree of Taylor expansion can be wrote as:

$$\begin{aligned} f(x) = f(x_k) + [\nabla f(x_k)]^T (x - x_k) + \\ \frac{1}{2!}[x - x_k]^T H(x_k)[x - x_k] + o^n, \end{aligned} \tag{5}$$

Where:

$$H(x_k) = \begin{bmatrix} \frac{\partial^2 f(x_k)}{\partial x_1^2} & \frac{\partial^2 f(x_k)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x_k)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x_k)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x_k)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x_k)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x_k)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x_k)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x_k)}{\partial x_n^2} \end{bmatrix}, \tag{6}$$

DOI: 10.54254/2755-2721/2026.TJ30979

o^n is remainder.

For each feature vector $a_{i,j}^{(1)}$ at (i, j), from CNNs layer, a second-degree Taylor expansion around the zero point is performed:

$$T_{i,j} = f(a_{i,j}^{(1)}), \tag{7}$$

Through tensor decomposition or low-rank approximation, we can control the number of parameters. The output $T \in \mathbb{R}^{H" \times W" \times C_1}$preserves both linear features and explicit second-order interactions, thereby strengthening the non-linear representation.

3.1.3. Fully connected layer

Due to the application of regularization (i.e., Dropout), the output of the Taylor expansion layer is unable to predict. In order for the output of the Taylor expansion layer to be passed smoothly to the transformer, we have added a fully connected layer. The output of the Taylor expansion layer T is flattened into a sequence $S \in \mathbb{R}^{L \times C_1}$, where $L = H"W"$. A fully connected (projection) layer then maps this sequence to a d-dimensional embedding:

$$E = SW_{fc} + b_{fc} \in R^{(Ld)}, \tag{8}$$

with $W_{fc} \in \mathbb{R}^{c_1 \times d}$ and $b_fc \in \mathbb{R}^d$. This layer reduces dimensionality and remixes features, preparing a unified-dimensional token sequence for the Transformer component.

3.1.4. Transformer Classifier

The token sequence E is fed into a standard Transformer encoder. In the l_{th}layer of the Transformer, the multi-head self-attention (MSA) mechanism is defined as:

$$E^{(l+1)} = MSA(LayerNorm(E^{(l)})) + E^{(l)}, \tag{9}$$

$$E^{(l+2)} = FFN(LayerNorm(E^{(l)})) + E^{(l)}, \tag{10}$$

where FFN denotes a two-layer feed-forward network. Finally, the class token $e_{cls} \in \mathbb{R}^d$ is used for classification via a single linear classification head:

$$\hat{y} = softmax(W_{cls}e_{cls} + b_{cls}) \in \mathbb{R}^K, \tag{11}$$

with K being the number of classes. This structure leverages the Transformer's global receptive field to make the final decision based on the Taylor-enhanced features

3.2. Training details

In this section, we elaborate on the training procedures and configurations employed to optimize our hybrid CNN-Taylor- Transformer model. The training process is designed to effectively minimize the loss function and enhance the model's generalization capability through the use of appropriate optimization techniques and regularization strategies.

DOI: 10.54254/2755-2721/2026.TJ30979

3.2.1. Loss function

The loss function serves as a critical guide for training our model, quantifying the discrepancy between the model's predictions and the actual ground truth labels. We employ the standard cross-entropy loss for classification tasks, which is mathematically formulated as:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log \widehat{y}_i, \tag{12}$$

Here, N represents the number of classes, yy_i is the binary indicator (0 or 1) if class label i is the correct classification for the instance, and $\widehat{y}_i$ is the predicted probability that the instance belongs to class i. The cross-entropy loss effectively measures the performance of our model by penalizing false classifications, with the penalty increasing as the predicted probability diverges from the actual label. By minimizing this loss function during training, our model learns to adjust its parameters to produce probability distributions over classes that are as close as possible to the true distributions.

3.2.2. Optimizer

To efficiently minimize the loss function and update the model's parameters during training, we utilize the AdamW optimizer. AdamW is a stochastic gradient descent optimizer that combines the advantages of two previously proposed extensions of SGD: Adam and decoupled weight decay. The AdamW optimizer adaptively adjusts the learning rate for each parameter by incorporating moments of the gradients, which helps in accelerating convergence and stabilizing the training process. The key update equations for AdamW are as follows:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \tag{13}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, \tag{14}$$

$$\widehat{m_t} = \frac{m_t}{1 - \beta_1^t}, \tag{15}$$

$$\widehat{v_t} = \frac{v_t}{1 - \beta_2^2}, \tag{16}$$

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\widehat{v_t}} + \varepsilon} \widehat{m}_t - \lambda \theta_{t-1}, \tag{17}$$

where m_t and v_t are the first and second moments of the gradients, β_1 and β_2 are the exponential decay rates for the moment estimates, g_t is the gradient at step t, θ_t represents the model parameters at step t, η is the learning rate, ϵ is a small constant for numerical stability, and λ is the weight decay coefficient. This optimizer allows our model to efficiently navigate the high-dimensional parameter space and converge to an optimal solution.

3.3. Dataset

To support our research, we compiled a comprehensive dataset by integrating several open-source datasets from Kaggle. Our final dataset consists of 3,000 images, evenly divided into three categories: 1,000 blank images, 1,000 images containing boats, and 1,000 images containing airplanes. To ensure the dataset's quality and usability, we employed techniques such as oversampling to clean and

DOI: 10.54254/2755-2721/2026.TJ30979

integrate data from various sources. This meticulous process resulted in a balanced dataset, which is crucial for effective training. During training, we used batches of 32 images and ran the model for 100 epochs to optimize performance.

3.4. Experiments

3.4.1. Experimental setup

In order to evaluate the effectiveness of the novel Taylor Expansion Layer in a CNN-Transformer hybrid model, the experiments were meticulously designed and conducted. Three distinct groups were set up to explore the impact of different Taylor Expansion Layer configurations on model performance. In all experimental groups, the size of the CNN convolu- tional kernel was systematically varied from 3×3 to 9×9 to gain insights into its influence on model accuracy and loss. This variation in kernel size allowed for a comprehensive assessment of how the Taylor Expansion Layer interacts with different levels of spatial feature extraction. Group 1 served as the baseline for comparison, where the Taylor Expansion Layer was not activated. This group provided a reference point to measure the inherent performance of the CNN-Transformer model without any feature expansion. Group 2 incorporated the Taylor Expansion Layer with a feature expansion up to the 2nd order. This configuration aimed to explore the benefits of introducing moderate feature interactions. Group 3 extended the feature expansion up to the 3rd order. This group was designed to test the limits of feature expansion and to investigate whether higher-order interactions could further enhance model performance or potentially introduce noise and complexity.

3.4.2. Results and Discussion

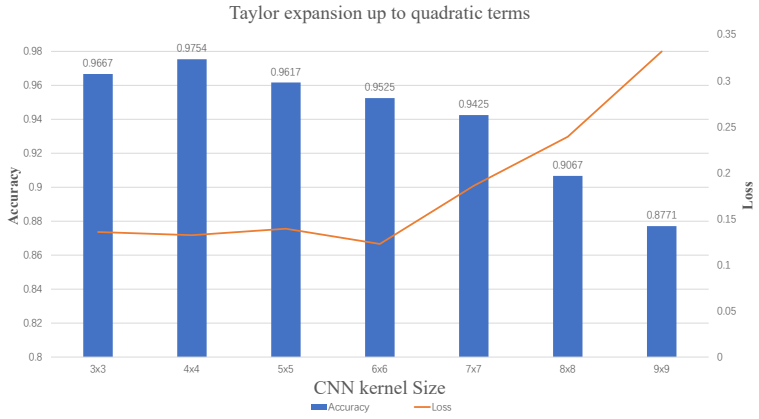

Figure 2: Model accuracy (left axis) and loss (right axis) versus CNN kernel size under second-order Taylor expansion.

DOI: 10.54254/2755-2721/2026.TJ30979

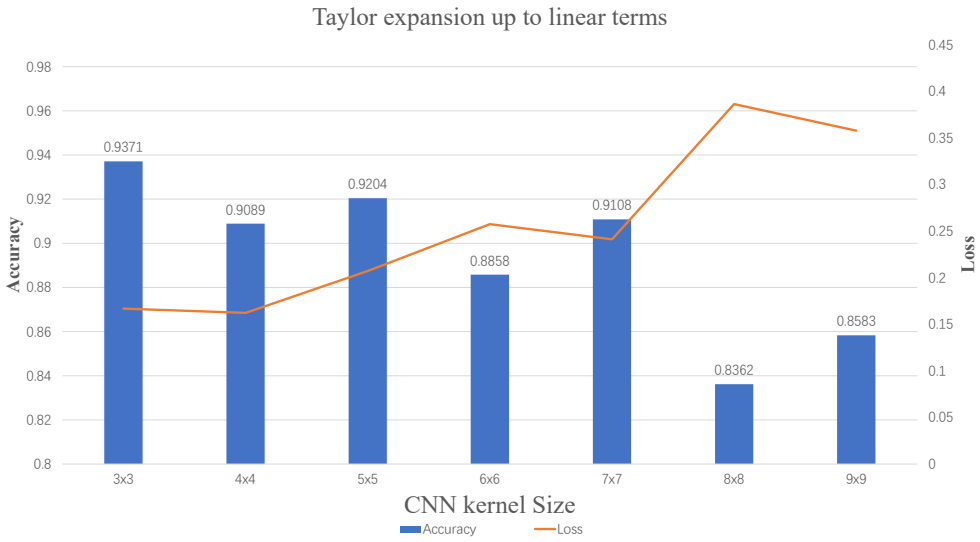

Figure 3: Model accuracy (left axis) and loss (right axis) versus CNN kernel size under liner-order Taylor expansion.

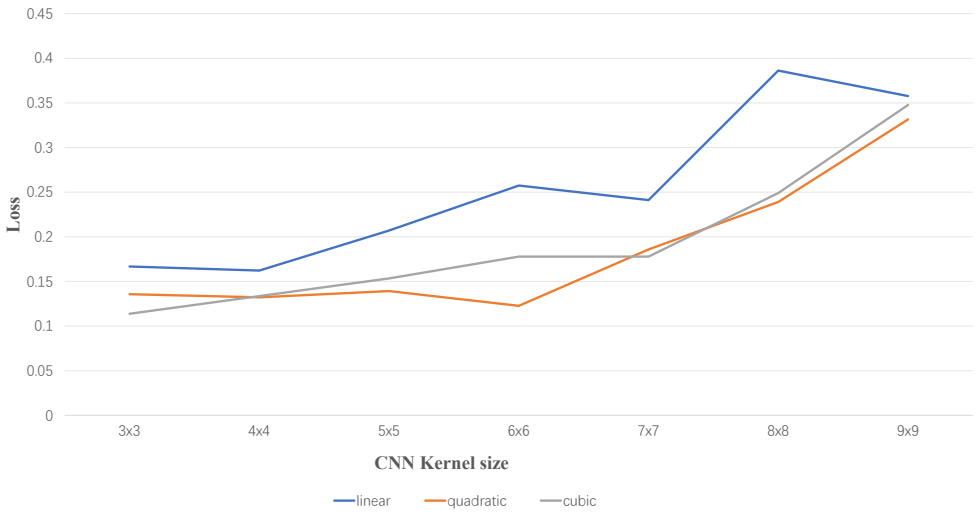

Figure 4: Effect of CNN Kernel Size on Training Loss Under Varying Orders of Taylor Expansion Approximations

DOI: 10.54254/2755-2721/2026.TJ30979

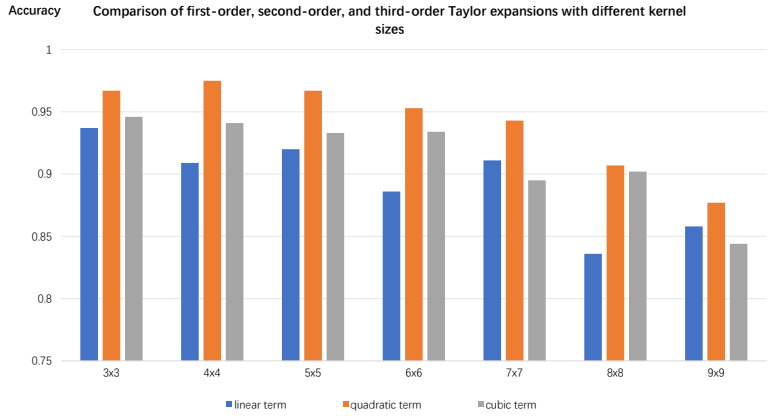

Figure 5: Model Accuracy vs. Kernel Size for First-, Second-, and Third-Order Taylor Expansion Approximations

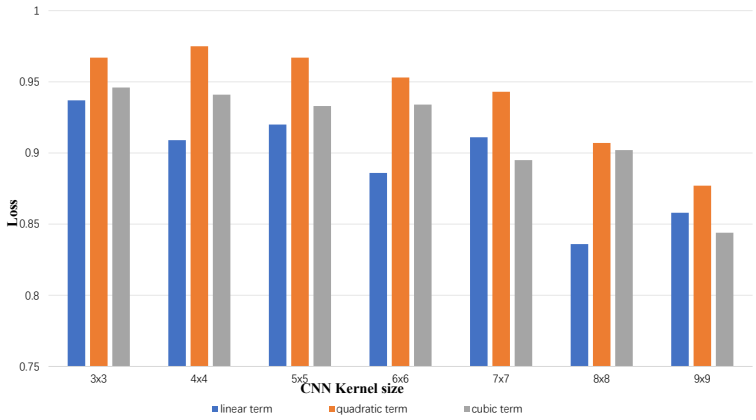

Figure 6: Model loss vs. Kernel Size for First-, Second-, and Third-Order Taylor Expansion Approximations

3.4.2.1 Accuracy analysis

The experimental results revealed several key insights into the impact of the Taylor Expansion Layer on model performance. For smaller convolutional kernel sizes (3×3 and 5×5), the introduction of the Taylor Expansion Layer resulted in a significant increase in model accuracy across all experimental trials. This substantial improvement can be attributed to the layer's ability to effectively capture higher-order feature interactions that smaller kernels alone may not fully exploit. These findings suggest that the Taylor Expansion Layer can enhance the model's capacity to learn complex patterns in the data by enriching the feature representation. When the convolutional kernel size increased beyond 5×5, the performance improvement from the Taylor Expansion Layer became less pronounced. This diminishing return may be due to the fact that larger kernels inherently capture a broader range of spatial features, thereby reducing the relative benefit of additional feature expansion. However, it is worth noting that the Taylor Expansion Layer still provided a noticeable, albeit smaller, improvement in accuracy even for larger kernels. The optimal Taylor Expansion order was found to be between the 2nd and 3rd order. Expanding the Taylor series up to the 2nd order (Group 2) consistently yielded the

DOI: 10.54254/2755-2721/2026.TJ30979

best results across various kernel sizes. Higher-order expansions (Group 3) sometimes resulted in a slight decrease in accuracy, particularly for larger kernel sizes. This indicates that excessive feature expansion can introduce noise or lead to overfitting, thereby degrading model performance. Therefore, it is crucial to carefully select the expansion order to balance the benefits of feature interaction with the risk of overfitting.

3.4.2.2 Loss analysis

The residual loss trend also supported that Taylor Expansion Layer is effective. In all experimental groups, the model loss starts to increase as the size of convolution kernel reaches larger numbers, which was natural phenomenon because when the kernel size increases, more complex the model gets and the models get more susceptible to overfitting. However, the models with Taylor Expansion Layer (Group 2 and 3) show lower loss value as compared to Group 1 in all sizes of convolutions, which indicates that the models with Taylor Expansion Layer helped in better convergence by providing more rich features to learn for better convergence and improved ability to generalize from training set. The Gap started to be visible and significant when kernel of 6*6 and large were considered and in group 2,3 rise in losses were significantly less as compared with group 1. This means that Taylor Expansion Layer helps the models to maintain stability and better generalization as the model becomes more complex. By enriching the feature space to learn from, our Taylor Expansion layers were able to suppress some negative effect caused by increasing complexity, such as over-fitting and instability.

3.4.2.3 Comprehensive evaluation

The interaction of Kernel Size and Taylor Expansion Order was mostly reflected through its behavior towards Accuracy and loss trends together. With smaller Kernel Sizes, Taylor Expansion Layer improved accuracy along with reducing loss; thereby showing efficient learning process.,As the Kernel size increases so did accuracy gain from using a Taylor Expansion Layer did decrease, however it was responsible for lower loss values than it would have been without it. This shows that Taylor Expansion Layer provided a benefit even at higher Complexity of models till certain limit which means that it can provide some stable benefit across the Complexity while augmenting CNNTransformer Architecture. The value of the result can also be observed when choosing between different orders of expansion; ideally higher order expansions can capture more complex feature interactions but for doing so they introduce unnecessary complexity into models, The optimum order of expansion observed was of 2nd to 3rd order for this problem statement which maintains a balance between capturing beneficial information from feature interaction and avoiding pit-fall of overfitting which means that we need to carefully choose hyper-parameters while designing deep network with Taylor Expansion Layer.

3.4.2.4 Insights and future directions

The experimental results provide insights into the working of the Taylor Expansion Layer in boosting the CNN-Transformer models. The impact of the layer is especially noticeable when smaller convolutional kernels are employed, highlighting its effectiveness in scenarios where the feature interaction is limited by kernel size. It also continues to maintain lower loss values across all kernel sizes, implying its potential to alleviate the problem of model optimization for better generalization.

Future work can explore the effect of varying expansion orders in different layers or different type of networks, or combining this layer with different feature augmentation techniques. Further, applying this layer across different dataset types and network architectures will provide further insights into its diverse and flexible applications and implications in deep learning environments.

DOI: 10.54254/2755-2721/2026.TJ30979

4. Conclusion

This work proposes a novel technique in enhancing deep learning models by introducing a new layer called Taylor Expansion Layer. We prove that this layer can significantly enhance the performance of CNN-Transformer hybrid models. The layer is most effective for smaller convolution kernels, which leads to substantial improvement in accuracy by capturing complex feature interactions which traditional kernelbased mechanisms cannot capture. The exhaustive experiments also show that Taylor expansion order between 2nd and 3rd performed well, although higher orders provided slightly more improvement in accuracy but also introduced more noise and risk of overfitting which degrades the model performance.We further demonstrate that this layer significantly enhanced feature representation hence acted as a good stability element while using smaller convolutional kernel sizes. We further show that this layer can reduce the increased loss arising from larger kernel sizes. Hence, Taylor Expansion Layer provides additional stability aspect when using higher convolution kernel sizes. In conclusion, we have demonstrated the significance of Taylor Expansion Layer in enhancing the deep learning model architecture.This new technique provides an efficient way to enhance models without significant engineering of building new architectures. It helps the researcher and practitioners to build better performing accurate as well as robust models especially where feature interactions are significant for performance.

5. Future work

The experimental results provide promising direction for future research efforts. We plan to examine the effectiveness of the Taylor Expansion Layer across various application domains and its compatibility with new upcoming deep learning techniques that will be developed in future. Furthermore, we plan to develop adaptive algorithms which will automatically find optimal expansion order for training which will further increase gains provided by this layer. Our work also implies significance from development perspective as well as for future research applications. Hencethis provides significant direction both from technical aspect as well as research perspective in the area of deep learning applications and research.

References

[1] L. Zou, Q. Liu, and J. Dai, "Augmenting interaction effects in convolutional networks with taylor polynomial gated units," *Neural Networks*, vol. 185, p. 107262, 5 2025, doi: 10.1016/j.neunet.2025.107262.

[2] H. Liu, Y. Ji, and X. Wang, "Image analysis by analogy with Taylor expansion," in *SCCG '04: Proceedings of the 20th Spring Conference on Computer Graphicspp*, 2004, pp. 209-211, doi: 10.1145/1037210.1037243

[3] A. S. Gaikwad and M. El-Sharkawy, "Pruning convolution neural network (squeezenet) using taylor expansion-based criterion," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Louisville, KY, USA, 2018, pp. 1-5, doi: 10.1109/isspit.2018.8705095.

[4] Y. Zhu and Y. Chen, "Adaptive Growth: Real-time CNN Layer Expansion," *arXiv (Cornell University)*, 9 2023, doi: 10.48550/arxiv.2309.03049.

[5] C. Yu, T. Chen, and Z. Gan, "Taylor-Series-Expansion-Based Vision Transformer Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*. pp. 1–18, 6 2025, doi: 10.1109/tpami.2025.3578827.

[6] J. Sa, J. Ryu, and H. Kim, "ECTFormer: An efficient Conv-Transformer model design for image recognition," *Pattern Recognition*, vol. 159, p. 111092, 5 2025, doi: 10.1016/j.patcog.2024.111092.

[7] F. Yuan, Z. Zhang, and Z. Fang, "An effective CNN and Transformer complementary network for medical image segmentation," *Pattern Recognition*, vol. 136, p. 109228, 4 2023, doi: 10.1016/j.patcog.2022.109228.

[8] J. Yuan, F. Zhou, Z. Guo, X. Li, and H. Yu, "HCformer: Hybrid CNN-Transformer for LDCT Image Denoising," *Journal of Digital Imaging*, vol. 36, no. 5, pp. 2290–2305, 6 2023, doi: 10.1007/s10278-023-00842-9.

[9] S. Liang, Z. Hua, and J. Li, "Hybrid transformer-CNN networks using superpixel segmentation for remote sensing building change detection," *International Journal of Remote Sensing*, vol. 44, no. 8, pp. 2754–2780, 5 2023, doi: 10.1080/01431161.2023.2208711.

DOI: 10.54254/2755-2721/2026.TJ30979

[10] Z. Li, G. Chen, and T. Zhang, "A CNN-Transformer Hybrid Approach for Crop Classification Using Multitemporal Multisensor Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 847–858, 2 2020, doi: 10.1109/jstars.2020.2971763.

[11] T. Zhang, P. Lin, H. Liu, P. Wang, Y. Wang, K. Xu, "A Coupled Transformer-CNN Network: Advancing Sea Surface Temperature Forecast Accuracy," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, Art no. 4207414, 5 2025, doi: 10.1109/tgrs.2025.3574990.

[12] Q. Hou, Z. Gao, M. Lu, and Y. Yu, "A Hybrid Transformer-CNN Model for Interpolating Meteorological Data on the Tibetan Plateau," *Atmosphere*, vol. 16, no. 4, p. 431, 4 2025, doi: 10.3390/atmos16040431.

[13] Y. Qiu, K. Zhang, C. Wang, W. Luo, H. Li, and Z. Jin, "MB-TaylorFormer: Multi-branch Efficient Transformer Expanded by Taylor Formula for Image Dehazing," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 12802-12813, doi: 10.1109/iccv51070.2023.01176.

[14] Y. Guo, J. Zheng, Y. Ding, Y. Hu, Y. Cao, and Z. Wang, "A Transformer Network Based on Taylor Expansion for Generating Phase Contrast Images of Adherent Cells," in *2024 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, Alesund, Norway, 2024, pp. 482–487, doi: 10.1109/rcar61438.2024.10670877.

[15] J. Liu, Y. Yang, B. Xu, H. Yu, Y. Zhang, Q. Li, Z.Huang, "RSTC: Residual Swin Transformer Cascade to approximate Taylor expansion for image denoising," *Computer Vision and Image Understanding*, vol. 248, no. C, Art. no. 104132, 11 2024, doi: 10.1016/j.cviu.2024.104132.

[16] A. Khan, Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif and U. Farooq "A survey of the vision transformers and their CNN-transformer based variants," *Artificial Intelligence Review*, 10 2023, doi: 10.1007/s10462-023-10595-0.

[17] E. Arkin, N. Yadikar, Y. Muhtar, and K. Ubul, "A Survey of Object Detection Based on CNN and Transformer,"*IEEE*, 7 2021, doi: 10.1109/prml52754.2021.9520732.

[18] X. Yue, F. Ge, F. Zhou, N. Wu, and Y. Zhang, "A Hybrid CNN Compression Approach for Hardware Acceleration," pp. 1546–1550, 11 2022, doi: 10.1109/icct56141.2022.10072642.

[19] L. Chen, J. Li, Y. Li, and Q. Zhao, "Even-Order Taylor Approximation-Based Feature Refinement and Dynamic Aggregation Model for Video Object Detection," *Electronics*, vol. 12, no. 20, p. 4305, 10 2023, doi: 10.3390/electronics12204305.

[20] Z. Huang, W. Hu, Z. Zhu, Q. Li, and H. Fang, "TMSF: Taylor Expansion Approximation Network with Multi-Stage Feature Representation for Optical Flow Estimation," *Digital Signal Processing*, p. 105157, 3 2025, doi: 10.1016/j.dsp.2025.105157.

[21] S. Wang, Y. Xing, S. Shi, and Z. Guo, "A Taylor expansion-based texture and edge-preserving interpolation approach for arbitrary-scale image super-resolution," *Pattern Recognition*, vol. 169, p. 111965, 6 2025, doi: 10.1016/j.patcog.2025.111965.