Going Deep: Deep Visual Prompting with LoTeP

Zijie Zhao[1*], Yuji Wang[1*], Yanru Wu[1], Haohua Wang[1], Enming Zhang[1], Wai Kin Victor Chan[1†], and Yang Li[2†]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, China
{zhaozj24, wu-yr21, yuji-wan24, wanghh24}@mails.tsinghua.edu.cn
chanw@sz.tsinghua.edu.cn
[2] School of AI, Chinese University of Hong Kong (Shenzhen), China
yangl@cuhk.edu.cn

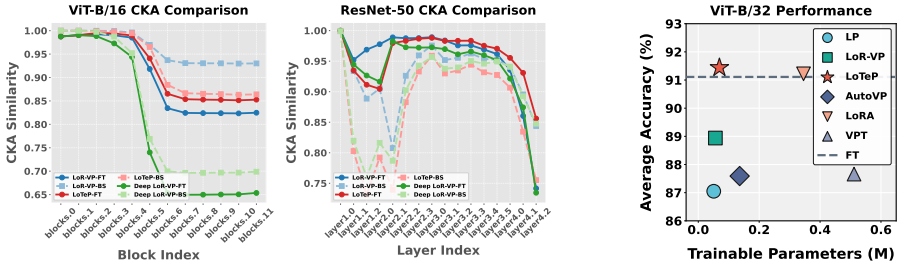
Abstract. Visual Prompting (VP) has emerged as a parameter-efficient paradigm for adapting pre-trained models to downstream tasks. While existing VP methods predominantly operate in the input space, extending VP to deep activations confronts a fundamental Capacity-Integrity Paradox: increasing prompt capacity for better task adaptation inevitably triggers a parameter explosion and severe semantic disruption, where deep prompts override pre-trained representations. To strike a delicate tradeoff within this Capacity-Integrity Paradox, we conducted a series of preliminary experiments. Interestingly, these experiments reveal a strong correlation among prompt channels: applying shared prompts across grouped channels not only drastically reduces parameters but also improves performance. Motivated by this structural correlation and the need to systematically master the tradeoff, we propose Low-Rank Tensor Visual Prompting (LoTeP). By modeling deep prompts as low-rank tensors, LoTeP elegantly achieves extreme parameter compression. Concurrently, it enables precise control over the prompt capacity via layer-wise rank decaying, effectively preserving the semantic integrity of deep activations. Extensive experiments demonstrate the superior generality, effectiveness, and efficiency of our approach. Overall, LoTeP consistently outperforms the state-of-the-art LoR-VP by an average margin of over 2.5% across all evaluated scenarios, while adding less than 0.1% of the backbone parameters over the LoR-VP baseline.

Keywords: Visual Prompting · Tensor Decomposition · Visual Model Adaptation

1 Introduction

In recent years, fine-tuning pre-trained large models for downstream tasks has demonstrated formidable capabilities in Natural Language Processing (NLP) [5,25] and Computer Vision (CV) [12,36]. To circumvent the prohibitive computational costs of full fine-tuning, prompt learning was introduced as a parameter-efficient alternative in NLP [24, 26]. Inspired by this success, Visual Prompting

* Equal contribution.
† Corresponding authors: Wai Kin Victor Chan and Yang Li.



(a) (b)

Fig. 1: Motivation and superiority of LoTeP. **(a)** CKA similarity of intermediate features across blocks for ResNet-50 and ViT-B/16. LoR-VP suffers from signal attenuation: its deep features remain excessively close to the pre-trained backbone rather than the full fine-tuned (FT) oracle, indicating insufficient task adaptation. Deep LoR-VP causes severe semantic disruption, indicated by a sharp drop in CKA similarity. By leveraging rank-decaying, our LoTeP successfully prevents this disruption, effectively preserving semantic integrity and aligning closely with FT. **(b)** Performance versus parameter efficiency on 10 downstream datasets using the ViT-B/32 backbone. LoTeP strikes an optimal balance, achieving state-of-the-art average accuracy that even surpasses FT, while introducing fewer than 0.1M trainable parameters.

(VP) [1,7,18,38] has emerged as a highly effective paradigm in CV by introducing learnable parameters strictly in the input pixel space.

Despite its broad compatibility, this input-only paradigm faces a fundamental bottleneck: the severe attenuation of prompt signals as they propagate through deep layers. Our Centered Kernel Alignment (CKA) analysis reveals that input-space VP struggles to align deep-layer representations with fully fine-tuned (FT) models but rather remains aligned with pretrained base (BS) models, rendering it ineffective at extracting task-relevant semantics. This inherently limits deep feature adaptation and leads us to question whether prompts can be applied directly to deep activations.

However, directly applying existing VP methods to network activations is impractical. The first major obstacle is parameter inefficiency. The explosive growth in channel dimensions of deeper-layer activations would cause a corresponding surge in VP parameters, which violates the principle of parameter efficiency and complicates optimization. The second challenge is the severe risk of semantic disruption. Unconstrained deep prompts can aggressively alter intermediate representations and overwrite invaluable pre-trained knowledge. As evidenced by our CKA analysis in Fig. 1(a) , applying LoR-VP to deep activations (Deep LoR-VP) causes a significant drop in feature similarity to both the pre-trained backbone and the FT model. Collectively, these phenomena reveal a fundamental **Capacity-Integrity Paradox** in current VP research: equipping deep layers with sufficient prompting capacity for task-specific adaptation

inevitably clashes with the imperative to preserve the semantic integrity of pre-trained representations, while simultaneously risking parameter explosion.

To address these challenges, we first conducted a preliminary study with a specially designed channel-grouping prompting strategy. The channel-wise redundancy observed in this preliminary study directly inspired us to propose a more generalized and continuous solution: **Lo**w-Rank **Te**nsor Visual **P**rompting (LoTeP). This method models the VP as a low-rank tensor, which fundamentally resolves the parameter explosion. Concurrently, the tensor rank serves as a natural proxy to govern the prompt's expressiveness. By introducing a rank-decaying schedule, we effectively restrict the excessive degrees of freedom in deeper layers, thereby preventing the disruption of pre-trained semantics. Furthermore, under certain structural assumptions, we provide a theoretical analysis that formally corroborates our empirical design, revealing the intrinsic low-rank property of deep prompts. Without a significant increase in parameters, our LoTeP method robustly outperforms existing state-of-the-art (SOTA) VP approaches. Across 38 comprehensive experimental settings, it achieves an overall average improvement of over 2.5% requiring an additional parameter budget of less than 0.1% relative to the backbone when compared to the state-of-the-art LoR-VP.

Our contributions can be summarized as follows:

- We systematically investigate and articulate the fundamental challenges of applying VPs to deep-layer activations, explicitly identifying and validating the dual obstacles of parameter inefficiency and semantic disruption.
- We propose LoTeP, a deep prompting framework, and provide a formal theoretical analysis that mathematically uncovers the intrinsic low-rank property of deep VPs. By modeling prompts as low-rank tensors and employing a rank-decaying strategy, our method effectively regularizes the prompt capacity to mitigate semantic disruption while maintaining parameter efficiency.
- We conduct comprehensive experiments on diverse model architectures and downstream datasets. Results show that our method achieves a SOTA performance improvement of over 2.5% with an average of less than 0.1% more parameters than LoR-VP and negligible training and inference overhead, demonstrating its excellent generalization across both models and tasks.

2 Related Works

Visual Prompting. Prompt learning is a new paradigm that has emerged in recent years in the field of Natural Language Processing (NLP) for adapting pre-trained large models to downstream tasks [24,26]. The success of prompt learning in the NLP domain has sparked researchers' interest in whether it could be introduced to the field of computer vision (CV). Based on this idea, Visual Prompting (VP) [1] method was quickly proposed. Various methods have already been applied to VP design. For instance, CLIP-VP [1] adopts the method of adding a prompt frame to the image edges. In contrast, AutoVP [38] and ILM-VP [7] use a method of scaling the image and then padding it with prompts back to the standard size. LoR-VP [18], on the other hand, employs the technique of

adding a prompt of the same size as the image, approximating this prompt with a low-rank matrix. Besides research on VP paradigms, other works have explored different applications: BlackVIP [33] applies VP to black-box model optimization, while C-AVP [6] generates a class-wise VP to improve model robustness. All these existing methods restrict prompts to the input pixel space, and our work explores how to extend visual prompting into deep network activations through a low-rank tensor formulation.

Low Rank Tensor. Low-rank tensor decomposition is widely adopted in deep learning. By approximating original tensors, it extracts core representations and filters noise while reducing parameter overhead [19]. In past research, low-rank tensors have been utilized in the area of model compression, including Convolutional Neural Network (CNN) [11, 35], Recurrent Neural Network (RNN) [39, 42, 43], and Transformer [32], as well as in the fields of information denoising [40], adversarial defense [44], model acceleration [30] and fine-tuning [2, 41].

3 Preliminary Study

3.1 Exploring Channel Grouping in Deep VP

Existing VP methods primarily append learnable parameters to the input pixel space. If we were to naively replicate this input-level VP paradigm directly onto deep feature maps (*i.e.*, Naive Deep VP), it would inevitably fall victim to the aforementioned Capacity-Integrity Paradox. This naturally begs the question:

Is this paradox entirely insurmountable?

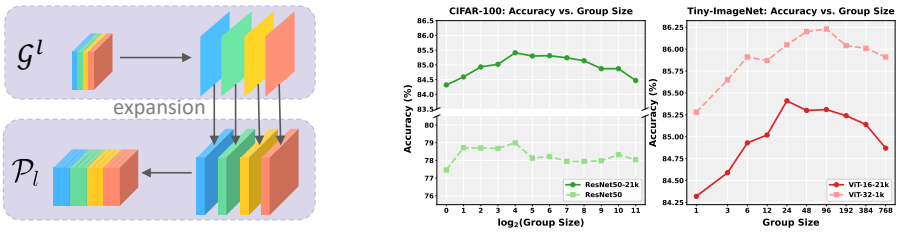
Recent studies on feature representations reveal a high degree of similarity and redundancy among feature channels [3, 13, 23]. This observation implies that when designing deep visual prompts, explicitly modeling the correlation and redundancy between channels could not only trim down the parameter overhead but also potentially boost the VP performance.

Driven by this insight, we conduct a preliminary study introducing a channel-grouped deep VP, denoted as Grouped VP. Let $\mathcal{X}_l \in \mathbb{R}^{C\times H\times W}$ be the l-th intermediate feature map of a pre-trained vision model, where C is the number of channels, and $H \times W$ represents the spatial resolution. A standard, or naive, deep visual prompt $\mathcal{P}_l \in \mathbb{R}^{C\times H\times W}$ introduces independent parameters for every single channel. Specifically, the visual prompt $\mathcal{P}_l$ is added element-wise to the input feature, and the resulting representation is then processed by the layer block f_l:

$$\mathcal{X}_{l+1} = f_l(\mathcal{X}_l + \mathcal{P}_l) \qquad l = 0, 1, \ldots, L-1 \tag{1}$$

$$y = \mathrm{Head}(\mathcal{X}_L) \tag{2}$$

The colors • and • indicate learnable and frozen parameters, respectively. This naive formulation introduces $\mathcal{O}(C \times H \times W)$ parameters, becoming computationally prohibitive and prone to overfitting as C grows larger in deeper layers.



(a) Illustration of Grouped VP (b) Effect of Group Size

Fig. 2: Overview of Grouped VP. (a) Architecture of Grouped VP, which divides prompt parameters $\mathcal{P}_l$ into g groups to reduce the $\mathcal{O}(C \times H \times W)$ complexity. **(b)** Effect of group size on accuracy. Intermediate sizes achieve the best trade-off between parameter efficiency and representation capacity.

To mitigate this, Grouped VP evenly divides the C feature channels into g sequential groups, where each group contains C/g channels. We then learn a compact prompt representation $\mathcal{G}^l \in \mathbb{R}^{g \times H \times W}$. Within each group, the prompt parameters are strictly shared across the channel dimension. The parameter expansion process can be formulated as:

$$\mathcal{P}_l = \Big[\underbrace{\mathcal{G}_1^l, \ldots, \mathcal{G}_1^l}_{C_1} \Big\| \ \ldots \ \Big\| \underbrace{\mathcal{G}_g^l, \ldots, \mathcal{G}_g^l}_{C_g}\Big] \tag{3}$$

The operator $\|$ denotes concatenation along the channel dimension.

By doing so, Grouped VP inherently enforces strong prompt correlations along the channel dimension and aggressively reduces the prompt parameter size to $\mathcal{O}(g \times H \times W)$. This coupled efficiency and structural regularization against disruption is a crucial stepping stone towards optimizing overall performance.

3.2 Empirical Observations on Grouped VP

To investigate the efficacy of the proposed grouping mechanism, we instantiate our Grouped VP using the SOTA prompting method, LoR-VP, yielding Grouped LoR-VP. To ensure the generalizability of our findings, we conduct extensive preliminary experiments across different network architectures, including both CNNs and Vision Transformers (ViTs). We evaluate the performance on two image classification benchmarks: CIFAR-100 and Tiny-ImageNet.

In our experiments, we systematically vary the number of groups g and compare the classification performance against the naive application of LoR-VP directly onto the deep feature maps (which corresponds to $g = C$). The results are summarized in Fig. 2 for clear comparison.

From the empirical results, we observe an inverted U-shaped curve that perfectly manifests the aforementioned Capacity-Integrity Paradox. Optimal accuracy consistently emerges at an intermediate g rather than at the extremes.

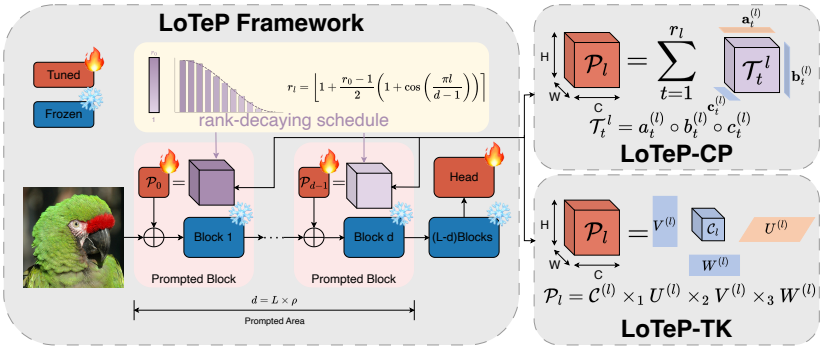


Fig. 3: Illustration of LoTeP. **Left:** The framework of LoTeP, applying VP on deep activations with a rank-decaying schedule. **Right (top):** LoTeP-CP, a low-CP-rank tensor represented as a sum of several rank-one tensors. **Right (bottom):** LoTeP-TK, a low-Tucker-rank tensor reconstructed from a low-rank core tensor and factor matrices. Orange parts indicate zero-initialized parameters.

On one hand, an overly small g severely restricts the prompt's representational capacity, hindering adequate task adaptation. Conversely, maximizing g (equivalent to the naive deep LoR-VP) triggers severe semantic disruption and fails to preserve pre-trained integrity. The peak at the intermediate g essentially represents an empirical sweet spot, demonstrating that a viable tradeoff between capacity and integrity is indeed achievable.

4 Methodology

4.1 Low-rank Tensor Visual Prompting (LoTeP)

Inspired by the empirical observations from our preliminary study, we recognize that introducing structural correlation serves as an effective regularization to achieve a viable tradeoff. However, relying on rigid, discrete channel grouping restricts the prompt's representational capacity too inflexibly. To overcome this limitation, we propose **Low-rank Tensor Visual Prompting (LoTeP)**, a novel design that explicitly formulates the deep visual prompt as a low-rank tensor. The overall framework is presented in Fig. 3. This tensor-based modeling allows us to seamlessly integrate low-rank properties across both channel and spatial dimensions into a cohesive mathematical framework. Based on two definitions of rank, we propose two variants of LoTeP: LoTeP-TK and LoTeP-CP.

LoTeP-TK. The prompt tensor $\mathcal{P}_l$ is modeled as a $[r_l, r_l, r_l]$ Tucker rank tensor [10], that is:

$$\mathcal{P}_l = \mathcal{C}^{(l)} \times_1 U^{(l)} \times_2 V^{(l)} \times_3 W^{(l)} \tag{4}$$

where $\times_i$ is i-mode product and $\mathcal{C}^{(l)} \in \mathbb{R}^{r_l \times r_l \times r_l}$, $U^{(l)} \in \mathbb{R}^{r_l \times c}$, $V^{(l)} \in \mathbb{R}^{r_l \times h}$, $W^{(l)} \in \mathbb{R}^{r_l \times w}$, i.e.,

$$(\mathcal{P}_l)_{i,j,k} = \sum_{t_1=1}^{r_l} \sum_{t_2=1}^{r_l} \sum_{t_3=1}^{r_l} \mathcal{C}^{(l)}_{t_1,t_2,t_3} U^{(l)}_{t_1,i} V^{(l)}_{t_2,j} W^{(l)}_{t_3,k} \tag{5}$$

LoTeP-CP. Alternatively, $\mathcal{P}_l$ is modeled as a r_l CANDECOMP/PARAFAC (CP) rank tensor [19], that is:

$$\mathcal{P}_l = \sum_{t=1}^{r_l} \mathbf{a}^{(l)}_t \circ \mathbf{b}^{(l)}_t \circ \mathbf{c}^{(l)}_t \tag{6}$$

$$(\mathcal{P}_l)_{i,j,k} = \sum_{t=1}^{r_l} a^{(l)}_{t,i}\, b^{(l)}_{t,j}\, c^{(l)}_{t,k} \tag{7}$$

in which $\circ$ is outer product and $\forall t, \mathbf{a}^{(l)}_t \in \mathbb{R}^c, \mathbf{b}^{(l)}_t \in \mathbb{R}^h, \mathbf{c}^{(l)}_t \in \mathbb{R}^w$.

For both of these VP variants, we initialize the parameters corresponding to the channel dimension as zero vectors or matrices, while the parameters for the spatial dimensions are randomly initialized. This allows the prompt tensor, $\mathcal{P}_l$, to be initialized as a zero tensor, ensuring that the expressive power of the pre-trained model is not disrupted at the start of training.

Advantages of the Tensor Formulation. This unified design overcomes the limitations of Grouped VP by enabling unique channel-wise expressivity and breaking inter-group isolation via a shared low-rank subspace. Concurrently, it inherits LoR-VP's ability to exploit spatial redundancy. By intrinsically capturing these multi-dimensional correlations within a single tensor, LoTeP unlocks superior representational capacity.

Furthermore, the tensor formulation critically addresses the parameter explosion problem in deep layers. Directly applying spatial-only methods (*e.g.*, Deep LoR-VP) requires independent low-rank matrices for every channel, leading to a prohibitive complexity of $\mathcal{O}(C \cdot r_l(H+W))$. In contrast, by jointly factorizing the channel and spatial dimensions, LoTeP decouples the prompt's complexity from the raw channel count C. For instance, LoTeP-CP drastically reduces the overall complexity to $\mathcal{O}(r_l(C+H+W))$, enabling truly efficient deep prompting when $r_l \ll C$.

4.2 Rank Decaying for Semantic Integrity

Although low-rank tensors provide effective layer-wise regularization, this alone cannot fully guarantee semantic integrity. As features propagate deeper, they encode increasingly invaluable pre-trained knowledge. Thus, explicit depth-wise intervention is necessary to prevent deep prompts from causing catastrophic semantic disruption.

Initially, we introduce a ratio $\rho \in (0, 1]$ to strictly truncate the prompting depth at layer $d = \lfloor L \cdot \rho \rceil$. However, this hard cut-off is overly rigid. To elegantly master the Capacity-Integrity Paradox—balancing early task capacity with strict

deep semantic preservation—we dynamically navigate this tradeoff. Since tensor rank dictates the capacity, we propose a cosine rank-decaying schedule to gracefully phase out the prompts.

Formally, the layer-wise rank r_l transitions smoothly from the initial rank r_0 at the first layer ($l = 1$) down to a zero-capacity prompt (complete phase-out) at depth d:

$$r_l = \left\lfloor 1 + \frac{r_0 - 1}{2}\left(1 + \cos\left(\frac{\pi l}{d-1}\right)\right)\right\rceil \quad l = 0, 1, 2, ..., (d-1) \tag{8}$$

This cosine schedule operationalizes our tradeoff strategy: a slow initial decay provides shallow layers with sufficient capacity for task adaptation, while a seamless phase-out gracefully concedes capacity in deeper layers to protect pre-trained semantics.

4.3 Theoretical Analysis of the Low-Rank Property

To better understand the behavior of deep visual prompts, we briefly analyze their intrinsic low-rank property from the perspective of gradient flow. In deep Convolutional Neural Networks (CNNs), the backpropagated gradients are inherently constrained by the network's linear operations, bottleneck structures, and pre-output average pooling layers. Since visual prompts are updated entirely via linear combinations of these gradients during training, the converged prompt matrices naturally tend to reside within a low-rank subspace.

To formalize this intuition, we establish a theoretical upper bound for the channel-wise rank of prompts in CNNs.[3]

Assumption 1. *Let $\mathcal{M}$ be a CNN model with depth L, and let $\mathcal{K}_i$ be the 4D weight tensor of the first convolution layer in the i-th residual block. We assume the weight tensors are constrained such that their CP rank is bounded by a constant r, i.e., Rank$(\mathcal{K}_i) \leq r$, $\forall i \leq L$.*

This assumption is not merely a theoretical construct, but reflects the realistic over-parameterization of modern CNNs, serving as the foundational premise for some research in tensor-based model compression [22, 35].

Based on Assumption 1 and the chain rule of backpropagation, we can bound the rank of the learned prompts as follows:

Proposition 1. *Under Assumption 1, the rank of the spatially flattened prompt matrix P_i at the i-th layer, learned on model $\mathcal{M}$, is bounded as:*

$$Rank(P_i) \leq (1 + L - i)r + 1 \qquad P_i \in \mathbb{R}^{c_i \times h_i w_i} \tag{9}$$

This theoretical bound reveals two important insights that guide our method design. First, Eq. (9) indicates that the rank upper bound strictly decreases as

[3] Due to page limits, analyses for other models are provided in the appendix.

the layer index i increases. This mathematical property aligns with our motivation for the **rank-decaying schedule** proposed in Sec. 4.2, suggesting that deep prompts inherently possess a lower capacity bound than shallow ones. Second, while LoR-VP has empirically shown that visual prompts exhibit spatial redundancy, our analysis provides a complementary perspective: deep prompts are also low-rank along the channel dimension. Unifying these spatial and channel-wise low-rank properties naturally motivates explicitly modeling deep visual prompts as a low-rank tensor (LoTeP).

5 Experiments

Our experiments follow the standard evaluation protocol for visual prompting, measuring the transfer performance of pre-trained models on downstream tasks. To comprehensively validate the proposed LoTeP, we conduct extensive empirical evaluations focusing on the following aspects: (1) Demonstrating the model and task generality of LoTeP across a diverse set of architectures and ten distinct downstream datasets in Sec. 5.1; (2) Investigating the out-of-distribution robustness of our method in Sec. 5.2; (3) Showcasing the extreme efficiency of LoTeP in terms of parameter usage, training time, *etc.* (Sec. 5.3); and (4) Performing detailed ablation studies to validate our core components and analyze the impact of key hyperparameters on the final performance(Sec. 5.4).

Backbones. We conduct experiments across nine representative pre-trained models spanning various architectural paradigms.

(1) CNN Models: We include **ResNet-18**, **ResNet-50** [14] pre-trained on ImageNet-1K, and **ResNet-50-P** pre-trained on ImageNet-21K-P [36], along with a modernized CNN, **ConvNeXt-B** [28], pre-trained on ImageNet-21K.

(2) Transformer Models: We select **ViT-B/16**, **ViT-B/32** [12] pre-trained on ImageNet-21K and fine-tuned on ImageNet-1K, and **ViT-B/16-P** pre-trained on ImageNet-21K-P [36]. Additionally, the hierarchical **Swin-B** [27] pre-trained on ImageNet-21K is included.

(3) Hybrid Models: We evaluate the hybrid architecture **CoAtNet-2** [9] pre-trained on ImageNet-21K.

Baselines. We compare our proposed method with seven commonly used approaches, categorized into three fine-tuning protocols:

(a) Traditional fine-tuning methods:
- **Full Fine-tuning (FT)**: updates *all* backbone and head parameters.
- **Linear Probing (LP)**: only trains a linear classification head.

(b) Parameter-efficient fine-tuning:
- **LoRA** [16]: injects trainable low-rank matrices for adaptation.
- **VPT-Deep** [17]: introduces learnable prompts into the input space of every Transformer layer.

(c) Visual Prompting (VP):
- **ILM-VP** [7]: introduces Iterative Label Mapping.
- **AutoVP** [38]: an end-to-end approach for VP configuration selection.
- **LoR-VP** [18]: a low-rank approximation-based VP method.

Table 1: Accuracy and parameter counts across nine models on CIFAR-100 and Tiny-ImageNet. Reference baselines (FT, LP, LoRA) are marked in gray with the best results among them in **Bold**. The best-performing VP method is highlighted in **Bold**. † indicates backbones where FT surpasses our methods on both datasets. CoNX and CANet denote ConvNeXt and CoAtNet, respectively.

			ResNet			ViT			Others		
	Method	Param	18†	50†	50-P	B/16	B/32	B/16-P	Swin-B	CoNX-B	CANet-2
CIFAR-100	FT	575.39M	**82.51**	**83.46**	**85.96**	**92.15**	90.11	91.16	**91.84**	**92.04**	**88.52**
	LP	1.00M	63.29	74.19	83.34	87.12	86.21	88.90	87.37	88.01	80.02
	LoRA	11.05M	77.10	80.34	85.11	91.38	**90.50**	**92.15**	90.11	91.11	87.42
	ILM-VP[CVPR23]	1.32M	24.87	38.92	31.51	50.50	40.10	41.49	65.78	70.25	68.75
	AutoVP[ICLR24]	9.97M	63.67	73.88	83.76	88.84	85.96	88.58	86.83	86.78	88.06
	LoR-VP[ICLR25]	1.05M	69.93	74.86	84.27	89.01	88.65	89.69	90.42	90.87	88.05
	LoTeP-CP	1.13M	74.51	80.39	85.68	91.59	**90.57**	91.69	91.29	**92.21**	**90.37**
	LoTeP-TK	1.13M	**75.31**	**80.82**	**86.00**	**91.82**	90.39	**91.88**	**91.98**	92.01	90.30
Tiny-Imagenet	FT	576.39M	**72.34**	**81.56**	**80.02**	**91.24**	86.59	88.21	**88.52**	**88.26**	**87.41**
	LP	2.00M	65.17	76.52	77.40	89.25	83.95	82.75	86.54	87.81	69.27
	LoRA	12.05M	71.90	80.45	79.58	91.19	**87.40**	**88.50**	86.93	88.13	85.62
	ILM-VP[CVPR23]	1.32M	14.13	37.74	27.81	75.92	32.58	26.17	56.28	78.55	77.27
	AutoVP[ICLR24]	18.39M	59.78	77.04	75.67	89.53	82.43	82.75	84.81	84.44	83.65
	LoR-VP[ICLR25]	2.05M	68.28	77.55	78.42	90.53	85.85	84.40	88.28	88.50	85.07
	LoTeP-CP	2.13M	70.95	80.41	**79.67**	91.90	87.61	87.43	88.47	**89.70**	87.03
	LoTeP-TK	2.14M	**71.29**	**80.63**	79.56	**91.93**	**87.71**	**87.84**	**88.84**	89.63	**87.21**

Notably, since VPT is architecture-specific, it is exclusively evaluated on the ViT-B/32 backbone in our task generality experiments.

Datasets. We follow the setup in LoR-VP and use the same set of downstream datasets: Tiny-ImageNet [21], CIFAR-10/100 [20], OxfordPets [34], EuroSAT [15], Food101 [4], DTD [8], OxfordFlowers [31], SVHN [29], and GTSRB [37].

Experimental Setup. For our methods, we report two LoTeP variants in experiments: LoTeP-CP, LoTeP-TK. All input ranks (r_0) for these LoTeP variants are set to 6. All VP methods were reproduced using the same configurations as in their respective original papers, and the rank of LoRA is set to 8. Furthermore, no data augmentation is applied other than image resizing. Images for ILM-VP and AutoVP are resized according to their original experimental settings, while a standard resolution of 224×224 is used for all other methods.

5.1 Main Results

Model Generality The results are presented in Tab. 1. Our analysis reveals that:

1. **Comparison with other VP methods:** Within the VP framework, our LoTeP-CP and LoTeP-TK methods achieve new SOTA performance across a wide array of models, including CNNs, Transformers, and hybrid architectures of various scales. Surpassing the previous SOTA results by approximately 2.3%, these findings compellingly demonstrate the strong generalizability of our approach across diverse model architectures.

Table 2: Accuracy and parameter counts of different methods on ViT-B/32 and ResNet18 for the 10 downstream datasets. The best-performing VP method is highlighted in **Bold**, and the second-best performing VP method is underlined. The traditional fine-tuning baselines are marked in gray, with the best results among them in **Bold**. Tiny-IN denotes Tiny-ImageNet

		Tiny-IN	EuroSAT	Pets	Food	DTD	Flowers	CIFAR10	CIFAR100	SVHN	GTSRB	Average
ViT-B/32	FT	86.59	97.78	90.90	83.57	**75.43**	98.36	**98.58**	90.11	**95.01**	**94.76**	91.11
	LP	83.95	95.67	91.90	82.18	69.83	97.98	96.51	86.21	83.09	83.21	87.05
	LoRA	**87.40**	**97.87**	90.49	**85.01**	**75.43**	97.63	98.30	**90.50**	94.87	94.70	**91.22**
	VPT	83.54	95.90	**92.27**	82.29	72.11	**98.47**	96.02	86.22	81.48	88.29	87.66
	ILM-VP	32.58	88.12	78.91	48.24	42.65	64.27	85.27	40.10	80.81	67.88	62.88
	AutoVP	82.43	96.25	92.12	82.86	70.81	98.42	95.45	85.96	85.24	86.39	87.59
	LoR-VP	85.85	96.25	**92.18**	83.51	72.49	98.58	97.52	88.65	86.31	88.07	88.94
	LoTeP-CP	87.61	97.22	91.96	**85.03**	76.06	99.04	98.15	**90.57**	94.06	93.34	91.30
	LoTeP-TK	**87.71**	**97.48**	91.88	84.67	**76.17**	**99.12**	**98.18**	90.39	**94.26**	**94.52**	**91.44**
ResNet18	FT	**72.34**	**98.82**	**87.83**	**75.42**	65.49	72.78	**96.50**	**82.51**	**96.58**	**97.95**	**84.62**
	LP	65.17	93.82	87.26	50.65	60.08	78.11	85.93	63.29	65.04	77.38	72.67
	LoRA	71.90	97.37	84.14	68.53	**66.41**	**79.00**	92.44	77.10	94.63	96.33	82.79
	ILM-VP	14.13	85.23	65.48	14.79	35.30	27.91	65.51	24.87	75.15	52.04	46.04
	AutoVP	59.78	93.01	82.65	54.15	54.82	73.79	87.81	63.67	83.74	81.52	73.49
	LoR-VP	68.28	93.93	**89.67**	60.92	65.90	80.52	88.64	69.93	85.66	78.09	78.15
	LoTeP-CP	70.95	**95.81**	89.32	62.94	**67.07**	79.49	91.83	74.51	90.72	93.90	81.65
	LoTeP-TK	**71.29**	95.78	89.34	**62.97**	66.70	**80.61**	**92.01**	**75.31**	**92.88**	**94.97**	**82.19**

2. **Comparison between the two LoTeP variants:** There is no definitive better-performing method between LoTeP-CP and LoTeP-TK; instead, they exhibit a trade-off between expressive power and optimization stability. On one hand, LoTeP-TK is more expressive, a property derived from the fact that a tensor of CP-rank r has a Tucker rank no greater than [r, r, r]. On the other hand, this increased capacity potentially weakens parameter independence, leading to optimization difficulties when vectors in the factor matrices have opposing gradient directions. This theoretical dichotomy is mirrored in our empirical findings: while LoTeP-TK excels on smaller models, both methods perform on par with each other on larger-scale models.
3. **Comparison with other fine-tuning methods:** LoTeP method presents a compelling alternative to both FT and LoRA. It is exceptionally parameter-efficient: excluding the task-specific classification head, LoTeP requires only $\sim$1% of the trainable parameters used by LoRA, and substantially fewer than FT. Nevertheless, LoTeP consistently matches or exceeds LoRA's performance across various backbones. Furthermore, it surpasses FT on four of the six large models tested, while remaining competitive on the other two, effectively bridging the historical performance gap between visual prompting and full-weight updating. It is noted, however, that on small-scale models where intrinsic capacity is the primary constraint, FT and LoRA still hold an absolute advantage.

Task Generality As shown in Tab. 2, our LoTeP outperforms all other VP methods on nine out of ten datasets, surpassing the runner-up (LoR-VP) by an average of 2.50% on ViT-B/32 and 4.04% on ResNet18. The only exception is OxfordPets, where our expressive model, along with FT and LoRA, likely

Table 3: Out-of-Distribution Generalization Performance. Evaluation of the out-of-distribution generalization performance using the ImageNet-21K pre-trained ConvNeXt-B, with visual prompting applied on ImageNet-1K, and tested across four out-of-distribution datasets.

Method	Source	Target			
	ImageNet-1K	ImageNet-R	ImageNet-Sketch	ImageNet-A	ImageNet-V2
LP	84.61	58.56	45.84	38.03	74.32
AutoVP [ICLR24]	78.75	45.71	34.16	22.99	66.94
LoR-VP [ICLR25]	84.66	58.88	<u>46.22</u>	39.36	74.65
LoTeP-CP	<u>84.75</u>	**59.77**	**46.76**	<u>43.57</u>	**74.78**
LoTeP-TK	**84.83**	<u>59.61</u>	46.15	**44.41**	<u>74.73</u>

suffers from overfitting due to the fine-grained dataset's high similarity to the pre-training data.

Compared to traditional baselines, LoTeP-TK achieves a remarkable 91.44% average accuracy on ViT-B/32, successfully surpassing both full FT (91.11%) and LoRA (91.22%). Conversely, FT and LoRA maintain a pronounced advantage on ResNet18. This disparity stems from the inherent mechanism of Visual Prompting: since VP only optimizes the input space, its performance upper bound is fundamentally bottlenecked by the frozen backbone's representational capacity. ResNet18's limited capacity restricts VP's effectiveness, whereas methods like FT and LoRA bypass this limit by directly updating internal weights.

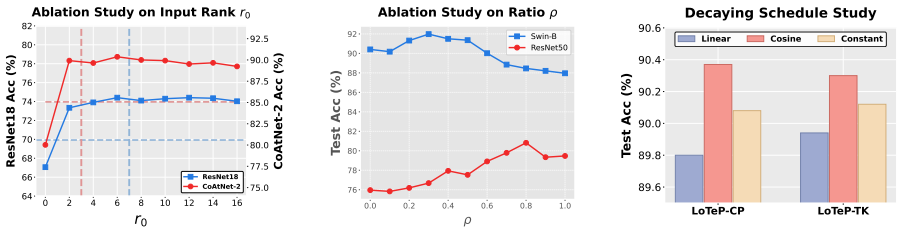
5.2 Robustness Analysis

To evaluate the robustness of our method against out-of-distribution (OOD) data, we employ a ConvNeXt-B backbone pre-trained on ImageNet-21K. We adapt the model to the ImageNet-1K dataset using various VP methods and evaluate the transferred models on four standard OOD benchmarks: ImageNet-R, ImageNet-Sketch, ImageNet-A, and ImageNet-V2.

As shown in Tab. 3, while all competitive methods achieve comparable accuracy on the clean ImageNet-1K validation set (ranging from 84.61% to 84.83%), the unified LoTeP framework demonstrates significantly superior OOD generalization. Setting aside a negligible 0.07% deficit by one variant on ImageNet-Sketch, LoTeP consistently outperforms the previous SOTA, LoR-VP, across all OOD scenarios. Most notably, on the notoriously challenging ImageNet-A dataset—which comprises natural adversarial examples—LoTeP yields a remarkable accuracy of up to 44.41%, surpassing LoR-VP by a substantial margin of +5.05%. These results suggest that deep tensor modulation can help reduce overfitting to source-specific visual artifacts and improve robustness in the evaluated OOD settings.

Table 4: Training and Inference Efficiency. Comparison of training and inference efficiency across different methods. Evaluated using ViT-B/16 on Tiny-ImageNet. Note that due to out-of-memory issues at a batch size of 256, the reported data for VPT is evaluated at a batch size of 128 (on a larger GPU at a batch size of 256, its GPU usage is 24.53GB).

Network Dataset	Method	Epochs	Time	# Tunable Params	GPU Usage	Latency	Accuracy
ViT-B/16 Tiny-ImageNet	LoRA[ICLR22]	20	2.89h	449K	22.08GB	**5.26ms**	91.19
	VPT[ECCV22]	20	3.48h	538K	12.60GB	6.33ms	91.10
	AutoVP[ICLR24]	100	8.05h	2,318K	14.55GB	8.80ms	89.53
	LoR-VP[ICLR25]	20	1.65h	**159K**	**14.34GB**	5.74ms	90.53
	LoTeP-CP	20	**1.57h**	168K	**14.34GB**	5.81ms	91.90
	LoTeP-TK	20	1.59h	169K	**14.34GB**	5.77ms	**91.93**



(a) (b) (c)

Fig. 4: Ablation and sensitivity studies. (a) Evaluation of initial rank r_0 on CIFAR-100 using ResNet-18 and CoAtNet-2. The horizontal dashed line represents the accuracy of the LoR-VP baseline, and the vertical dashed line indicates the rank at which the parameter count exceeds that of LoR-VP. **(b)** Test accuracy on CIFAR-100 is evaluated across different values of ρ using ResNet-50 and Swin-B backbones. **(c)** Performance of constant, linear, and cosine decaying functions is evaluated for LoTeP variants on the CoAtNet-2 backbone.

5.3 Efficiency Analysis

As shown in Tab. 4, our proposed LoTeP achieves a favorable trade-off between efficiency and accuracy. Compared to baselines like LoRA and VPT, LoTeP significantly reduces training time and requires drastically fewer tunable parameters (∼168K). During inference, both LoTeP variants maintain highly competitive latency and low GPU usage. Despite this remarkable computational efficiency, LoTeP delivers state-of-the-art performance, with LoTeP-TK achieving the highest accuracy of 91.93% on Tiny-ImageNet, successfully outperforming heavier methods like LoRA and VPT.

5.4 Ablation Study

Influence of r_0 in LoTeP. To further investigate the effect of input rank on LoTeP, we conduct experiments on CIFAR-100 using LoTeP-CP with two backbones: ResNet-18 and CoAtNet-2. The results are presented in Fig. 4(a).

From the analysis of the results, we can draw the following conclusions: (1) An increase in the number of parameters leads to a more significant performance improvement for small models, whereas this advantage is very slight in large models. (2) For a lower rank, our method achieves superior results to LoR-VP with fewer parameters, which highlights its architectural efficiency.

Impact of Prompting Depth Ratio ρ. As a core hyperparameter in LoTeP, the depth ratio ρ governs not only the total parameter count but also the effective intervention depth of the prompts. To clearly illustrate its impact, we evaluate LoTeP-TK on the CIFAR-100 dataset using both ResNet-50 and Swin-B backbones. The results are presented in Fig. 4(b). Our findings reveal two distinct trends depending on the model's inherent capacity: (1) For relatively smaller architectures like ResNet-50, a larger ρ consistently yields better performance, as the network directly benefits from the increased representational capacity provided by deeper prompts. (2) Conversely, for larger models like Swin-B, an excessively high ρ degrades performance. We attribute this to the formidable feature extraction capabilities of large-scale pre-trained models. Their deep-layer representations encode invaluable semantics. Intervening too deeply into these layers risks overriding this precious pre-trained knowledge, ultimately leading to severe semantic disruption rather than effective task adaptation.

Effectiveness of the Rank-Decaying Schedule. To validate our layer-wise rank allocation, we compare constant (fixed rank), linear, and our cosine decaying schedules using CoAtNet-2 on CIFAR-100 (Fig. 4(c)). The results highlight the importance of balancing capacity and regularization. Comparing the cosine and constant schedules confirms that rank decaying is essential: while a constant rank offers high capacity, it lacks the strong regularization needed in deep layers to prevent the disruption of pre-trained semantics. Furthermore, comparing the cosine and linear schedules reveals that reducing the rank too abruptly (linear) deprives early layers of the capacity required for task adaptation. The cosine schedule optimally maintains high capacity in shallow layers while seamlessly increasing regularization depth-wise. Besides, the success of this decaying schedule provides solid empirical evidence for our theoretical analysis in Proposition 1, where the intrinsic rank capacity of prompts is shown to naturally decrease with network depth.

6 Conclusion

Visual prompting has emerged as a powerful parameter-efficient technique for adapting pre-trained models to specific downstream tasks. Existing methods, however, predominantly restrict prompts to the input pixel space, as extending them to deep network activations triggers a prohibitive **Capacity-Integrity Paradox**: increasing capacity for task adaptation inherently causes parameter explosion and severe semantic disruption to pre-trained knowledge. Addressing this fundamental paradox, our study introduces Low-Rank Tensor Visual

Prompting (LoTeP), which formulates deep prompts as low-rank tensors and employs a progressive rank-decaying schedule. This design elegantly masters the tradeoff, facilitating extreme parameter compression while gracefully preserving the semantic integrity of deep representations. Extensive experiments across nine network architectures and ten datasets consistently demonstrate the state-of-the-art effectiveness, efficiency, and robustness of our method.

7 Acknowledgment

This work is supported by the Natural Science Foundation of China (Grant 62371270)

References

1. Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274 (2022)
2. Bershatsky, D., Cherniuk, D., Daulbaev, T., Mikhalev, A., Oseledets, I.: Lotr: Low tensor rank weight adaptation. arXiv preprint arXiv:2402.01376 (2024)
3. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. In: International Conference on Learning Representations (ICLR) (2022)
4. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Chen, A., Lorenz, P., Yao, Y., Chen, P.Y., Liu, S.: Visual prompting for adversarial robustness. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
7. Chen, A., Yao, Y., Chen, P.Y., Zhang, Y., Liu, S.: Understanding and improving visual prompting: A label-mapping perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19133–19143 (2023)
8. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
9. Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. Advances in neural information processing systems **34**, 3965–3977 (2021)
10. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM journal on Matrix Analysis and Applications **21**(4), 1253–1278 (2000)
11. Denton, E.L., Zaremba, W., Bruna, J., LeCun, Y., Fergus, R.: Exploiting linear structure within convolutional networks for efficient evaluation. Advances in neural information processing systems **27** (2014)
12. 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. In: International Conference on Learning Representations (2021)

13. Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C.: Ghostnet: More features from cheap operations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1580–1589 (2020)
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. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
17. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
18. Jin, C., Li, Y., Zhao, M., Zhao, S., Wang, Z., He, X., Han, L., Che, T., Metaxas, D.N.: Lor-vp: Low-rank visual prompting for efficient vision model adaptation. In: The Thirteenth International Conference on Learning Representations (2025)
19. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM review **51**(3), 455–500 (2009)
20. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
21. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS231n Course Project (2015)
22. Lebedev, V., Ganin, Y., Rakhuba, M., Oseledets, I., Lempitsky, V.: Speeding-up convolutional neural networks using fine-tuned cp-decomposition. In: International Conference on Learning Representations (ICLR) (2015)
23. Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning filters for efficient convnets. In: International Conference on Learning Representations (ICLR) (2017)
24. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4582–4597 (2021)
25. Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., Raffel, C.A.: Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. Advances in Neural Information Processing Systems **35**, 1950–1965 (2022)
26. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM computing surveys **55**(9), 1–35 (2023)
27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
28. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
29. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: Neural Information Processing Systems (NIPS) Workshop on Deep Learning and Unsupervised Feature Learning. Granada (2011)

30. Nguyen, L.T., Quélennec, A., Tartaglione, E., Tardieu, S., Nguyen, V.T.: Activation map compression through tensor decomposition for deep learning. Advances in Neural Information Processing Systems **37**, 130384–130407 (2024)
31. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008)
32. Noach, M.B., Goldberg, Y.: Compressing pre-trained language models by matrix decomposition. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. pp. 884–889 (2020)
33. Oh, C., Hwang, H., Lee, H.y., Lim, Y., Jung, G., Jung, J., Choi, H., Song, K.: Blackvip: Black-box visual prompting for robust transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24224–24235 (2023)
34. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
35. Phan, A.H., Sobolev, K., Sozykin, K., Ermilov, D., Gusak, J., Tichavskỳ, P., Glukhov, V., Oseledets, I., Cichocki, A.: Stable low-rank tensor decomposition for compression of convolutional neural network. In: European Conference on Computer Vision. pp. 522–539. Springer (2020)
36. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972 (2021)
37. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural networks **32**, 323–332 (2012)
38. Tsao, H.a., Hsiung, L., Chen, P.Y., Liu, S., Ho, T.y.: Autovp: An automated visual prompting framework and benchmark. In: International Conference on Learning Representations (2024)
39. Winata, G.I., Madotto, A., Shin, J., Barezi, E.J., Fung, P.: On the effectiveness of low-rank matrix factorization for lstm model compression. arXiv preprint arXiv:1908.09982 (2019)
40. Xue, J., Zhao, Y., Liao, W., Chan, J.C.W.: Nonlocal low-rank regularized tensor decomposition for hyperspectral image denoising. IEEE Transactions on Geoscience and Remote Sensing **57**(7), 5174–5189 (2019)
41. Yang, Y., Zhou, J., Wong, N., Zhang, Z.: Loretta: Low-rank economic tensor-train adaptation for ultra-low-parameter fine-tuning of large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 3161–3176 (2024)
42. Yang, Y., Krompass, D., Tresp, V.: Tensor-train recurrent neural networks for video classification. In: International conference on machine learning. pp. 3891–3900. PMLR (2017)
43. Ye, J., Wang, L., Li, G., Chen, D., Zhe, S., Chu, X., Xu, Z.: Learning compact recurrent neural networks with block-term tensor decomposition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9378–9387 (2018)
44. Zollicoffer, G., Vu, M.N., Nebgen, B., Castorena, J., Alexandrov, B., Bhattarai, M.: Lorid: Low-rank iterative diffusion for adversarial purification. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 23081–23089 (2025)