GLPI: A Global Layered Prompt Integration approach for Explicit Visual Prompt

Yufei Gao[1,3]
yfgao@zzu.edu.cn
Bin Fu[1]
binfu@zzu.gs.edu.cn
Lei Shi[1,3]
shilei@zzu.edu.cn
Chengming Liu[1,3]
cmliu@zzu.edu.cn
Yucheng Shi[†2]
ieycshi@zzu.edu.cn

[1] School of Cyber Science and Engineering
Zhengzhou University
Zhengzhou, China

[2] School of Computer and Artificial Intelligence
Zhengzhou University
Zhengzhou, China

[3] SongShan Laboratory
Zhengzhou, China

Abstract

In the era of large models, prompt learning of pre-trained visual models has shown significant flexibility in various downstream tasks. Explicit Visual Prompt (EVP) serves as an outstanding unified framework applicable to foreground segmentation, achieving superior performance by fine-tuning with features from frozen patch embeddings and high-frequency components. However, in the process of training large models with EVP, the approach of freezing parameters and centrally updating the prompt embeddings may pose difficulties for long-distance backpropagation. These challenges can affect the generalization performance of the model, potentially limiting its ability to fully adapt and represent across different tasks and data. Besides, compared with other tuning methods, EVP requires more steps to perform competitively. To address these issues, this paper proposes Global Layered Prompt Integration(GLPI), which filters and combines the prompt information of adjacent encoder layers with adaptive threshold values to obtain an integration prompt closer to the downstream tasks. The optimal prompts with global information are constructed to enable the model to process images from a wider range of perspectives. Extensive experiments conducted on foreground segmentation tasks demonstrate that GLPI outperforms EVP and other advanced approaches.

1 Introduction

Prompt is originally proposed in Natural Language Processing(NLP) [30]. The study referenced by [30] demonstrates the remarkable generalization ability of GPT-3 in downstream transfer learning tasks. Prompt learning can be categorised as discrete prompt [35] and continuous prompt [16] from the point of view whether the parameters of the prompts are

†Corresponding author

optimized. The model parameters corresponding to the word in discrete prompt are fixed after optimization during pre-trained, and are not further optimized in the downstream application. The model parameters corresponding to the words in continuous prompt can be optimized for specific tasks and data in downstream applications. This optimization process is called prompt tuning. In situations where corresponding downstream data can be used to assist prompt tuning, continuous prompt can demonstrate the advantage of being tailored to specific tasks and data. So continuous prompt is widely used by subsequent visual prompt learning methods.

More recently, prompt [23, 24] has been adapted for visual tasks. VPT [24] incorporates a set of learnable embedding vectors into each Transformer layer. The generalizability and feasibility of visual prompt are also investigated through extensive experiments on multiple recognition tasks across multiple domains and backbone architectures. In subsequent work [17, 19, 39] which are shown in table 1, the researchers still focus on the initialization process of prompt. Inspired by continuous prompt learning methods such as Prefix-Tuning [18] in the field of NLP, [8, 24, 34] by concatenating additional sequences of optimizable vectors as prompts on either the original input sequences or each layer of labeled embeddings of the Transformer structure, and updating only the parameters of the prompt embeddings during the training phase. This treatment leads to a few problems: First, prompt embedding in Computer Vision(CV) is significantly different from token embedding in NLP. Prompt embedding is not pre-trained and requires more optimization steps than token embedding [21]. Second, in large pre-trained models, EVP freezes the parameters during training and only updates the prompt embeddings, which can result in the model's generalization performance being compromised due to the difficulties of long-distance backpropagation [27]. Various reasons suggest that it is necessary to design a better training process for prompt embedding depending on the task.

Model	Method	Integration Prompts
VPT [24]	Concatenated Optimizable Vector Sequences	No
OpenPrompt [17]	Adding Pixel-Level Optimizable Perturbations	No
NOAH [39]	Network Structure Search	No
VPTM [19]	Task Reconstruction	No
Ours	Integrated Prompt Learning	Yes

Table 1: Comparison of several recent prompt learning methods. Integration Prompts indicates whether integration prompts are used in the process.

Long Short Term Memory (LSTM) [28] is a variant of RNNs that can effectively solve long-distance dependencies problem due to its specific structure and gating mechanism. LSTM introduces three gates: input gate, output gate, and forget gate. These gates use the sigmoid activation function to regulate the flow of information, thereby selectively remembering or ignoring information in the input sequence. Through these gating mechanisms, LSTM can handle long sequences more efficiently and propagate gradients more stably, thus alleviating the difficulties of long-distance backpropagation.

Inspired by LSTM, this paper proposes Global Layered Prompt Integration which consists of two parts: memory gate and forget gate. Based on EVP [34], memory gates and forget gates are used to control prompt embeddings information flow and update between different encoder layers. Before the prompt embeddings are input to the encoder layer, forget gate is used to filter the currently generated prompt embeddings, and adaptively generate the weights according to different training phases, which are used to decide how much con-

tent needs to be forgotten in the current prompt embeddings. We also use the memory gate to filter the prompt embeddings of the previous encoder layers and adaptively generate weights to determine how much content should be reserved in the previous prompt embeddings. Finally, the filtered prompts from neighboring layers are combined and fed into the current encoder layer.

Foreground segmentation represents a pivotal research objective in computer vision, with broad applications across diverse domains. Recent deep learning algorithms have been extremely successful in performing this task. The disadvantage of deep neural networks for segmentation is that they require large labeled datasets to train. This prerequisite is one of the main reasons that led researchers to adopt data augmentation methods in order to minimize manual labeling efforts while maintaining highly accurate results. As shown in Figure 1, previous study [7] has shown that random extraction of augmented statistics from uniform distributions and perturbation of original intermediate features can explicitly learn domain invariant representations, thereby enhancing the generalization performance of the model.

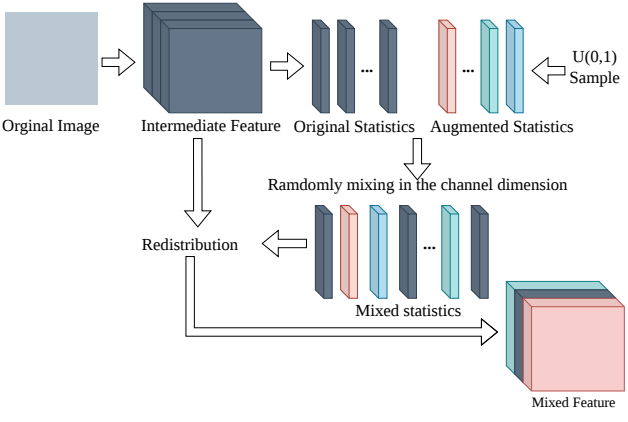


Figure 1: An overview of the feature perturbation method(TriD) proposed in [7].

The contributions of this paper are as follows:

1. We propose a training approach for prompt embeddings in EVP, namely Global Layered Prompt Integration(GLPI). It can make additional adjustments to the prompt embeddings in EVP during training. By efficiently integrating prompt information from various encoder layers, it optimizes and enhances the model's generalization performance.

2. To find a balance between computational overhead and segmentation performance, different variants of GLPI are proposed. The objective is achieved by adjusting the number of GLPIs and the manner in which they are connected. The effectiveness of the approach is proved by ablation experiments.

3. GLPI outperforms EVP in all six datasets for the three different tasks, and the visualisation results are placed in Section 4.2. Not only that, it achieves better results in ISTD [10], CAMO [32], COD10K [4] compared to other task-specific advanced approaches.

2 Related Work

Pre-Trained Model: In 2009 Pan and Yang introduced transfer learning [25], which enables the learning of new problem-solving skills with a limited number of samples, rather than requiring the training of models from scratch with a vast quantity of data. Humans are capable of employing previously acquired knowledge to address uncharted challenges. In this way, transfer learning provides a viable solution for data scarcity. It is subsequently adopted with great alacrity in computer vision. Since 2012, a series of CNN models and Transformer models have been pre-trained in ImageNet with very good results for fine-tuning for different downstream tasks. Pre-trained models (PTMs) are used for almost all CV tasks.

Explicit Visual Prompt: Liu [34] proposes a unified framework for many foreground segmentation tasks that does not require any task-specific design. The authors draw upon the pre-training and prompt tuning protocols that are commonly employed in the field of NLP to propose a creative visual prompt model, namely the Explicit Visual Prompt (EVP). In contrast to previous visual prompts, which are typically implicit embeddings at the dataset level, EVP focuses on processing the explicit visual content of each image features and uses these features as prompts from frozen patch embeddings and high-frequency components. This is achieved through the execution of adjustable parameters. The pre-trained model is frozen and a few additional parameters are used to learn task-specific knowledge in this method. Despite the limited number of tunable parameters introduced, EVP outperforms full fine-tuning and other parametric efficient fine-tuning methods.

3 Method

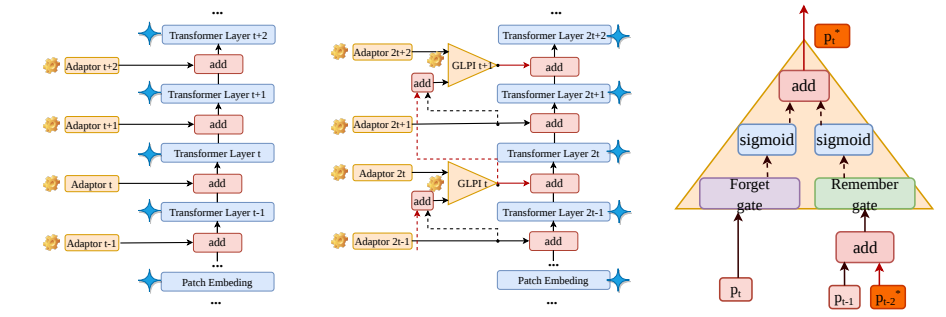
In this section, Global Layered Prompt Integration (GLPI) is presented. This approach allows the model to take advantage of prompts from a broader perspective by integrating different layers of prompt embeddings. The GLPI is divided into four groups according to the model processing workflow, with parameters shared within the groups.

3.1 GLPI between the Encoders

In the Transformer model, residual connection is used to alleviate the problem of reduced training accuracy [33]. Building upon this concept, GLPI implements a more sophisticated connectivity mechanism surrounding each pair of neighbouring encoder layers. In particular, GLPI is designed to operate on the prompt portion of each vector.

3.2 Inside of GLPI

Figure 2(c) provides a comprehensive illustration of the internal structure of GLPI. The GLPI consists of two gated units: a memory gate and a forget gate. Each gated unit contains a single-layer feedforward neural network. It is shared between all encoder layers within the same stage according to the stage of feature extraction. In a manner analogous to ResNet, the memory gate utilizes previous prompt embeddings prior to the input to the encoder, thereby emphasizing and leveraging the original information. Instead, the forget gate operates on the prompt embedding of this layer, thereby reducing its impact and preventing extreme results. Then the GLPI combines the output of the two units to generate the final outcome. Thus, between several neighboring encoders, the GLPI considers all the original and the



(a) (b) (c)

Figure 2: (a) EVP original processing flow, mainly through adaptors to complete the update of the prompts; (b) The GLPI proposed in this paper provides additional updates to the prompts during the training process by linking the adaptors of different layers; (c) Detailed internal structure of GLPI. In this paper, yellow screws are used to represent layers that need to be adjusted during training, and blue stars are used to represent layers that are frozen during training.

integration prompt embeddings, encouraging the model to forget some of the current state while allowing it to maintain past information.

As shown in Figure 2, we use p for prompt embedding. p_t represents the original prompt input for the t encoder, p_t^* represents the integration prompt input for the t encoder, p_{t-1} represents the original prompt input for the t-1 encoder, and p_{t-2}^* represents the integration prompt input for the t-2 encoder. Then, p_t, p_{t-1} and p_{t-2}^* are processed to form a complete output for prompt adjustment.

Prompts are updated as follows:

$$p_t^* = GLPI(p_t, p_{t-1}, p_{t-2}^*) \tag{1}$$

The following equation demonstrates how GLPI handles prompt embeddings:

$$p_t^* = \sigma_F(W_F(p_t)) + \sigma_R(W_R(p_{t-1} + p_{t-2}^*)) \tag{2}$$

The weights of the forget unit and remember unit are represented by W_F and W_R respectively. The activation functions σ_F, σ_R are also included in this equation.

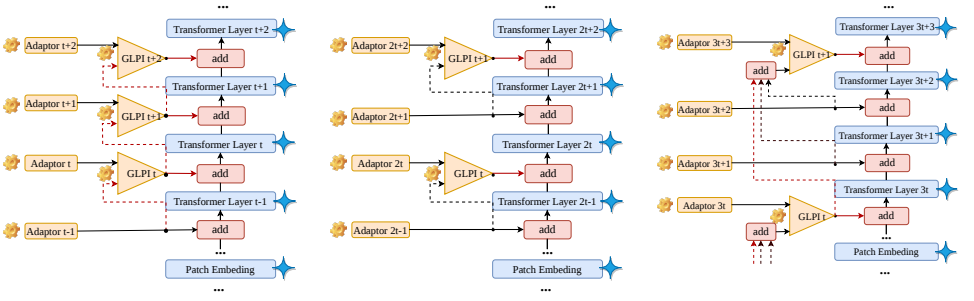
3.3 Variants of GLPI

In this section, we provide several possible GLPI structures, which are shown and explained in Figure 3.

Full Connection: GLPI is performed for each of the two adjacent encoder layers.

Half Connection: A simplified version of full connection, where the number of GLPI is halved and only non-repeating adjacent encoder layers are integrated.

One Third Integrated Connection: The number of GLPI is one third of full connection, and the prompt information of the adjacent three encoder layers and the integration prompt information of the previous layer can be obtained at once.



(a) (b) (c)

Figure 3: (a) The structure of Full Connection; (b) The structure of Half Connection; (c) The structure of One Third Integrated Connection.

Model	Updata Method
Ours_{fc}	$p_t^* = \sigma_F(W_F(p_t)) + \sigma_R(W_R(p_{t-1}^*))$
Ours_{hc}	$p_{2t}^* = \sigma_F(W_F(p_{2t})) + \sigma_R(W_R(p_{2t-1}))$
Ours_{otic}	$p_{3t+3}^* = \sigma_F(W_F(p_{3t+3})) + \sigma_R(W_R(p_{3t+2} + p_{3t+1} + p_{3t}^*))$

Table 2: Different settings for GLPI.

4 Experiment

4.1 Datasets

The approach is evaluated on various datasets for three foreground segmentation tasks: camouflaged object detection, shadow detection, forgery detection. A summary of the fundamental characteristics of all datasets is presented in Table3.

Camouflaged Object Detection. COD10K [4] as the largest camouflaged object detection dataset, contains 3040 training samples and 2026 testing samples. CAMO [32] provides a variety of images with both naturally and artificially camouflaged objects. The third dataset CHAMELEON [26] consists of 76 testing images from the Internet. In accordance with the methodology proposed by [4, 6], the combined dataset is subjected to training, with the three aforementioned datasets subsequently employed for testing purposes. We use commonly used metrics: S-measure (S_m), average E-measure (E_ϕ), weighted F-measure (F_β^ω), and MAE to evaluate performance.

Shadow Detection. SBU [31] as the largest shadow dataset with annotations, contains 4089 training samples and 638 testing samples respectively. The ISTD [10] contains triple samples for shadow detection and removal. Only shadow images and shadow masks are used to train our approach. Following the methodology proposed by [14, 15, 41], both datasets are subjected to training and testing at a resolution of 400×400. In this study, the Balanced Error Rate (BER) is employed as the evaluation metric.

Forgery Detection. CASIA [8] is a comprehensive forgery detection dataset. It contains 5123 training and 921 testing samples, which have been manipulated by means of stitching and copying. We followed the protocol of previous work [11, 22, 36], training and evaluating with the size of 256×256. We use F_1 scores and pixel-level Area Under the Receiver Operating Characteristic Curve (AUC) for evaluation.

Task	Dataset Name	# Train	# Test
Forgery Detection	CAISA [8]	5,123	921
Shadow Detection	ISTD [10]	1,330	540
	SBU [31]	4,089	638
Camouflaged Object Detection	COD10K [4]	3,040	2,026
	CAMO [32]	1,000	250
	CHAMELEON [26]	-	76

Table 3: A summary of the datasets considered in this work is presented above. The number of images in the training set (*#Train*) and the testing set (*#Test*) for each dataset are shown.

4.2 Implementation Details

All experiments are conducted on NVIDIA T4 GPU with 12GB memory. The AdamW [13] optimizer and Cosine decay are used for all experiments. The minibatch size is equal to 4. The initial learning rates are set to 2e-4 and 5e-4 for camouflaged combination datasets and others datasets respectively. The model is trained for 40 epochs on the camouflaged combination datasets [4, 26, 32] and 50 epochs on the other datasets. During the data augmentation phase, the augmented statistics are randomly extracted from the uniform distribution, the original intermediate features are disturbed, and the feature mixing of channel dimension is carried out. Binary Cross Entropy(BCE) loss is used for forgery detection, balanced BCE loss is used for shadow detection, and BCE loss and IOU loss are used for camouflaged object detection. All experiments are performed using SegFormer-B4 [5] pre-trained on the ImageNet-1k [9] dataset.

4.3 Main Results

Comparison with task-specific approaches: A comparison between our approach and other task-specific approaches is presented in Tables 4, 5 and 6. In all tables, the best results are bolded in black, and those that are not the best results but are better than EVP are bolded in blue. Due to our more rational prompt optimization strategy, GLPI achieves better results than the EVP in all six datasets for the three different tasks, and the best results are achieved in three of the datasets. We also show some visual comparisons of four datasets with the EVP separately in Figure 4. It indicates that our approach is capable of generating more accurate masks than EVP.

Method	SBU [31]	ISTD [10]
	BER↓	BER↓
DSC[7]	5.59	3.42
DSD[10]	3.45	2.17
MIMT[31]	3.15	1.72
FDRNet[15]	**3.04**	1.55
EVP[34]	4.31	1.35
Ours	**3.86**	**1.16**

Table 4: Comparison with advanced approaches on shadow detection.

Method	CAISA [8]	
	F1↑	AUC↑
SPAN[33]	0.382	0.838
PSCCNet[23]	0.554	0.875
TransForensics[11]	0.627	0.837
ObjectFormer[19]	0.579	**0.882**
EVP[34]	0.636	0.862
Ours	**0.642**	**0.877**

Table 5: Comparison with advanced approaches on forgery detection.

Comparison with the efficient tuning approaches: The efficacy of our approach is evaluated by comparing it to other efficient tuning approaches. This is also a widely used method for judging downstream task adaptation. As shown in Table 7, we set up VPT [24] and AdaptFormer [29] in the same way as [34]. It can be seen that GLPI is superior to the previous approaches in all aspects, which also proves the effectiveness of GLPI.

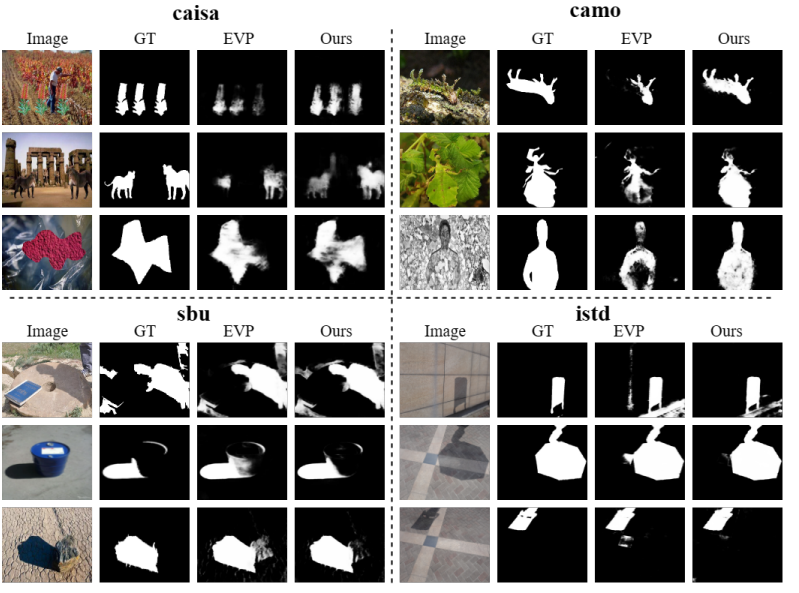


Figure 4: Visualization of EVP and Ours. We show the segmentation results of: EVP and Ours on CAISA [8]dataset for forgery detection (Top-left), on CAMO [32] dataset for camouflaged object detection (Top-right), on SBU [31] and ISTD [10] dataset for shadow detection (Bottom).

Method	CHAMELEON [26]				CMAO [6]				COD10K [9]			
	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^\omega\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^\omega\uparrow$	MAE↓	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^\omega\uparrow$	MAE↓
RankNet[63]	0.846	0.913	0.767	0.045	0.712	0.791	0.583	0.104	0.767	0.861	0.611	0.045
JCOD[9]	0.870	0.924	-	0.039	0.792	0.839	-	0.082	0.800	0.872	-	0.041
PFNet[6]	0.882	**0.942**	0.810	0.033	0.782	0.852	0.695	0.085	0.800	0.868	0.660	0.040
FBNet[21]	**0.888**	0.939	**0.828**	0.032	0.783	0.839	0.702	0.081	0.809	0.889	0.684	0.035
EVP[34]	0.871	0.917	0.795	0.036	0.846	0.895	0.777	0.059	0.843	0.907	0.742	0.029
Ours	**0.879**	**0.921**	**0.811**	**0.032**	**0.851**	**0.906**	**0.792**	**0.055**	**0.844**	**0.909**	**0.750**	**0.027**

Table 6: Comparison with advanced approaches on camouflaged object detection.

Method	Trainable Param.(M)	ISTD [10] BER↓	CAISA [8] F1↑	CAISA [8] AUC↑	CAMO [32] $S_\alpha\uparrow$	CAMO [32] $E_\phi\uparrow$	CAMO [32] $F_\beta^\omega\uparrow$	CAMO [32] MAE↓
Full-tuning	64.00	2.42	0.465	0.754	0.837	0.887	0.778	0.060
Only Decoder	3.15	4.36	0.396	0.722	0.783	0.827	0.671	0.088
VPT-Deep [24]	3.27	1.73	0.588	0.847	0.833	0.884	0.751	0.068
AdaptFormer [25]	3.21	1.85	0.602	0.855	0.830	0.877	0.750	0.068
EVP[34]	3.70	1.35	0.636	0.862	0.846	0.895	0.777	0.059
Ours	4.47	**1.16**	**0.642**	**0.877**	**0.851**	**0.906**	**0.792**	**0.055**

Table 7: Comparison with advanced efficient tuning approaches.

4.4 Ablation Study

We try to answer the question: which structure contributes mostly to prompting tuning.We adjusted the number of GLPIs and how they are connected in the four stages segformer used for multi-scale feature extraction to determine which structure could strike a balance between

computational effort and performance.

As shown in Table 8, $Ours_{fc}$ has better segmentation effect than Ours on some datasets, but it requires more GLPI and more computation to achieve this effect; $Ours_{hc}$ has similar complexity as Ours, but it doesn't make use of the integration prompts to deal with the data from a more global perspective, and therefore its performance is inferior to that of Ours. Compared with Ours, although the complexity of $Ours_{otic}$ is reduced, its effect in all datasets is not as good as Ours. After comparing various structures, the method proposed in this paper achieves a balance between segmentation effect and training complexity.

Method	Trainable Param.(M)	GLPI Numbers	FLOPs Param.(G)	Integration Prompts	ISTD [11] BER↓	CAISA [8] F1↑	CAISA [8] AUC↑	CMAO [3] $S_\alpha\uparrow$	CMAO [3] $E_\phi\uparrow$	CMAO [3] $F_\beta^\omega\uparrow$	CMAO [3] MAE↓
EVP[32]	3.70	0	21.69	No	1.35	0.636	0.862	0.846	0.895	0.777	0.059
$Ours_{fc}$	4.47	41	23.42	Yes	1.20	**0.647**	**0.880**	**0.852**	0.906	0.786	0.056
$Ours_{hc}$	4.47	19	22.57	No	1.25	0.636	0.866	0.850	0.901	0.783	0.057
$Ours_{otic}$	4.47	13	22.29	Yes	1.21	0.634	0.870	0.849	0.901	0.784	0.057
Ours	4.47	19	22.57	Yes	**1.16**	0.642	0.877	0.851	**0.906**	**0.792**	**0.055**

Table 8: Comparison of segmentation results by different numbers and structure of GLPI. GLPI Numbers are based on the mit-b4-evp model. Integration Prompts indicates whether integration prompts are used in the process. When calculating FLOPs, the input image size is 256.

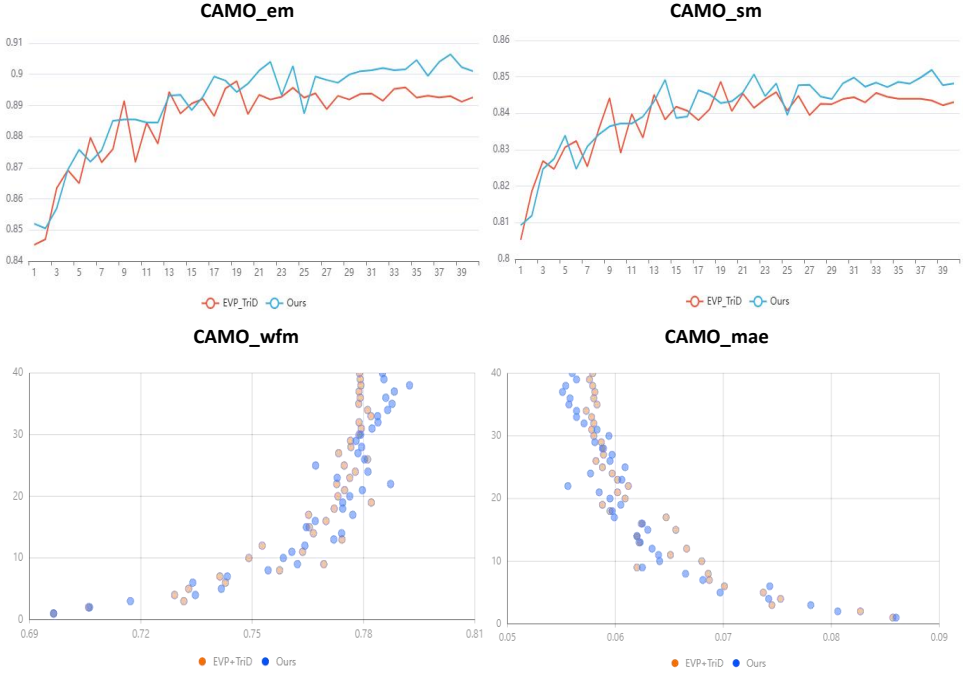


Figure 5: This table illustrates the changes in various metrics for EVP_{TriD} and *Ours* on the CAMO [32]. The upper two graphs have the training epoch as the horizontal axis and the index value as the vertical axis. The lower two graphs have the index value as the horizontal axis and the epoch as the vertical axis.

As shown in Figure 5, this experiment compares the trend of index changes between EVP and Ours on the CAMO dataset under exactly the same experimental environment. It can be seen from the graph that Ours has a more stable metric change during the first 15 training epochs compared to EVP. Moreover, Ours' metrics are consistently higher than EVP in the last 10 epochs, indicating that our method can improve the model's generalization performance, achieving higher results in different metrics.

In Table 9, the data related to EVP_{TriD} and EVP_{glpi} are presented in detail. Corresponding ablation experiments are conducted on three datasets, ISTD [10], CAISA [8] and CAMO [32]. It can be inferred from the table that the data augmentation method achieves excellent results in most datasets, but it is not applicable to all datasets. In contrast, our GLPI has yielded positive results across all datasets in this experiment, which fully demonstrates the broad application prospects of GLPI.

Method	Trainable Param.(M)	ISTD [10]	CAISA [8]		CAMO [32]			
		BER↓	F1↑	AUC↑	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE↓
EVP[34]	3.70	1.35	0.636	0.862	0.846	0.895	0.777	0.059
EVP_{TriD}	3.70	1.31	0.633	0.864	0.848	0.897	0.782	0.058
EVP_{glpi}	4.47	1.23	0.639	0.873	0.851	0.900	0.787	0.057
Ours	4.47	**1.16**	**0.642**	**0.877**	**0.851**	**0.906**	**0.792**	**0.055**

Table 9: Regarding the ablation experiments involving different influencing factors in this paper. In the ablation experiments discussed in this table, EVP_{TriD} represents the scenario where only data augmentation with TriD is applied to EVP, while EVP_{glpi} indicates the case where only GLPI is utilized in conjunction with EVP.

5 Conclusion

In conclusion, our study introduces the Global Layered Prompt Integration (GLPI) as a new way to enhance continuous prompt adjustment. By means of selective filtering and combining of prompt embeddings, GLPI is able to achieve more efficient prompt updates, which leads to an improvement in model performance. By conducting experiments on six datasets of foreground segmentation, we demonstrate that GLPI can improve the results with EVP. This provides compelling evidence for the capacity of GLPI to significantly enhance model performance across a range of object detection tasks. Finally, GLPI can be used not only for EVP, but also in combination with existing visual prompt methods such as VPT to quickly adjust prompts. We believe that our approach has the potential to facilitate the development of more effective and efficient models. We anticipate that further exploration and refinement of this approach will be beneficial in future studies.

Acknowledgements

This work was supported by the Nature Science Foundation of China (62006210), the Key Scientific and Technology Project in Henan Province of China (221100210100, 221100211200-02), the Project of Joint Graduate-student Education Base in Henan Province (YJS2023JD04), the Key Project of Collaborative Innovation in Nanyang (22XTCX12001), the Research Foundation for Advanced Talents of Zhengzhou University (32340306), Pre-research Project

of Songshan Laboratory (YYJC022022001), and Supported Project by Songshan Laboratory (232102210154).

References

[1] Yunqiu Lv Bowen Liu Tong Zhang Aixuan Li, Jing Zhang and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 10071–10081, 2021.

[2] Ziyang Chen, Yongsheng Pan, Yiwen Ye, Hengfei Cui, and Yong Xia. Treasure in distribution: A domain randomization based multi-source domain generalization for 2d medical image segmentation. *Medical Image Computing and Computer Assisted Intervention*, page 1735–1780, 2023.

[3] Han Cheng, Wang Qifan, Cui Yiming, Cao Zhiwen, Wang Wenguan, Qi Siyuan, and Liu. Dongfang. E2vpt: An effective and efficient approach for visual prompt tuning. In *International Conference on Computer Vision (ICCV)*, pages 17445–17456, 2023.

[4] Guolei Sun Ming-Ming Cheng Jianbing Shen Deng-Ping Fan, Ge-Peng Ji and Ling Shao. Camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 2777–2787, 2020.

[5] Zhiding Yu Anima Anandkumar Jose M Alvarez Enze Xie, Wenhai Wang and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *In Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[6] Ziqi Wei Xin Yang Xiaopeng Wei Haiyang Mei, Ge-Peng Ji and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 8772–8781, 2021.

[7] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):955–968, 2020.

[8] W. Wang J. Dong and T. Tan. Casia image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pages 422–426, 2013.

[9] Richard Socher Li-Jia Li Kai Li Jia Deng, Wei Dong and Li Fei-Fei. Imagenet:a large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[10] Xiang Li Jifeng Wang and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 1788–1797, 2018.

[11] Shicai Yang Di Xie Jing Hao, Zhixin Zhang and Shiliang Pu. Transforensics: image forgery localization with dense self-attention. In *IEEE/CVF International Conference on Computer Vision*, page 15055–15064, 2021.

[12] Jingjing Chen Xintong Han Abhinav Shrivastava Ser-Nam Lim Junke Wang, Zuxuan Wu and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 2364–2373, 2022.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Xiaowei Hu Chi-Wing Fu Xuemiao Xu-Jing Qin Lei Zhu, Zijun Deng and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *European Conference on Computer Vision (ECCV)*, pages 122–137, 2018.

[15] Zhanghan Ke Lei Zhu, Ke Xu and Rynson WH Lau. Miti-gating intensity bias in shadow detection via feature decomposition and reweighting. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4682–4691, 2021.

[16] Constant N. Lester B, Al-Rfou R. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.Online*, pages 3045–3059, 2021.

[17] Zhou T Li H, Feng C M. Prompt-driven efficient open-set semi-supervised learning. *arXiv preprint arXiv:2209.14205*, 2022.

[18] Liang P. Li X L. 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*, pages 4582–4597, 2021.

[19] Zhang X. Liao N, Shi B. Rethinking visual prompt learning as masked visual token modeling. *arXiv preprint arXiv:2303.04998*, 2023.

[20] XU Ke MA Lizhuang LIN Jiaying, TAN Xin and WH Rynson. Frequency-aware camouflaged object detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1551–6857, 2023.

[21] Chi-Liang Liu, Hao Wang, Nuwa Xi, Sendong Zhao, and Bing Qin. Global prompt cell: A portable control module for effective prompt tuning. In *Natural Language Processing and Chinese Computing*, pages 657–668, 2023.

[22] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Psccnet: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022.

[23] Max Vladymyrov Mark Sandler, Andrey Zhmoginov and Andrew Jackson. Fine-tuning image transformers using learnable memory. In *Proceedings ofthe Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12145–12154, 2022.

[24] Bor-Chun Chen Claire Cardie Serge Belongie-Bharath Hariharan Menglin Jia, Luming Tang and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

[25] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22(10): 1345–1359, 2009.

[26] J Błaszczyk Tomasz Depta Adam Kornacki Przemysław Skurowski, Hassan Abdulameer and P Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6), 2018.

[27] J Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61: 85–117, 2015.

[28] Jürgen Schmidhuber Sepp Hochreiter. Long short-term memory. *Neural Comput*, 9(8): 1735–1780, 1997.

[29] Zhan Tong Jiangliu Wang Yibing Song Jue Wang Shoufa Chen, Chongjian Ge and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *Neural Information Processing Systems (NeurIPS)*, page 16664–16678, 2022.

[30] Nick Ryder Melanie Sub-biah Jared D Kaplan Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry Amanda Askell et al. Tom Brown, Benjamin Mann. Language models are few-shot learners. In *In Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[31] Chen-Ping Yu Minh Hoai and Dimitris Samaras Tom´as F Yago Vicente, Le Hou. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *Computer Vision – ECCV 2016*, pages 816–832, 2016.

[32] Zhongliang Nie Minh-Triet Tran Trung-Nghia Le, Tam V Nguyen and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. In *Computer Vision and Image Understanding*, page 45–56, 2019.

[33] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[34] C. M. Pun W. Liu, X. Shen and X. Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19434–19445, 2023.

[35] Kandpal N et al. Wallace E, Feng S. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[36] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2354–2363, 2022.

[37] Zhenye Jiang Syomantak Chaudhuri Zhenheng Yang Xuefeng Hu, Zhihan Zhang and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *Computer Vision – ECCV 2020*, pages 312–328, 2020.

[38] Yuchao Dai Aixuan Li Bowen Liu Nick Barnes Yunqiu Lv, Jing Zhang and Deng-Ping Fan. Simultaneously localize,segment and rank the camouflaged objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 11591–11601, 2021.

[39] Liu Z. Zhang Y, Zhou K. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022.

[40] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W.H. Lau. Distraction-aware shadow detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5171, 2019.

[41] Liang Wan Song Wang Wei Feng Zhihao Chen, Lei Zhu and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5610–5619, 2020.