Prompt-Conditioned Vision-Language Models for Detecting Unseen Backdoor Images

Kyle Stein[1], Andrew Arash Mahyari[1,2], Guillermo Francia, III[3], Eman El-Sheikh[3]
[1] Department of Intelligent Systems and Robotics, University of West Florida, Pensacola, FL, USA
[2] Florida Institute For Human and Machine Cognition (IHMC), Pensacola, FL, USA
[3] Center for Cybersecurity, University of West Florida, Pensacola, FL, USA
ks209@students.uwf.edu, amahyari@ihmc.org, gfranciaiii@uwf.edu, eelsheikh@uwf.edu

***Abstract*—Backdoor attacks pose a critical threat to computer vision by embedding hidden triggers into inputs, causing models to misclassify them into target labels. While extensive research has focused on mitigating these attacks in object recognition models through model cleansing, much less attention has been given to detecting backdoored samples directly. Given the vast datasets used in training, manual inspection for backdoor triggers is impractical, and even state-of-the-art defense mechanisms fail to fully neutralize their impact. To address this gap, we introduce a novel method to detect unseen backdoored image types during both training and inference. Leveraging the transformative success of conditional prompt tuning in Vision Language Models (VLMs), our approach trains learnable text prompt prefixes to differentiate clean images from those with hidden backdoor triggers. Furthermore, we shift the learned prefix based on the image features for each sample through a lightweight, image-conditioned network. Experiments demonstrate the exceptional efficacy of this method, achieving an impressive average accuracy of 84% across two renowned datasets for detecting unseen backdoor triggers, establishing a new standard in backdoor defense.**

***Index Terms*—Adversarial Attacks, Backdoor Attacks, Vision-Language Model, Parameter-Efficient Tuning, Prompt Tuning.**

I. Introduction

Deep neural networks have revolutionized the field of computer vision, achieving human-level performance on tasks such as image classification [1] and object detection [2]. This remarkable progress has led to widespread deployment of DNNs in safety-critical applications ranging from autonomous vehicles and medical diagnosis to biometric authentication [3], [4]. However, as these models become increasingly integrated into real-world systems, their vulnerability to adversarial manipulation poses significant security risks. With backdoor attacks [5], an adversary stealthily embeds a hidden trigger into inputs so that, at inference time, any trigger-embedded example is misclassified into an attacker-chosen target class. For instance, a backdoored traffic sign recognition system might classify stop signs as speed limit signs when a specific sticker is present, while correctly recognizing normal signs with no trigger present (See Fig. 1). This dual behavior makes backdoor attacks exceptionally difficult to detect through standard validation procedures.

Current state-of-the-art defenses predominantly focus on model-level mitigation [6], [7], [8], [9], [10], attempting to detect or remove backdoors after training. These post-hoc approaches suffer from several critical limitations. First, they require the model to have already been trained on poisoned data, allowing the backdoor to be embedded before any defense mechanism is applied. Second, many defenses incur substantial computational costs for model cleansing or retraining. Third, they often assume prior knowledge about trigger characteristics which are to be classified and removed, limiting their applicability in practice. Fourth, recent adaptive attacks have demonstrated that many model-level defenses can be bypassed through careful trigger design or training strategies. Finally, these methods typically degrade model performance on clean samples while attempting to mitigate backdoor effects, presenting an undesirable trade-off between security and utility.

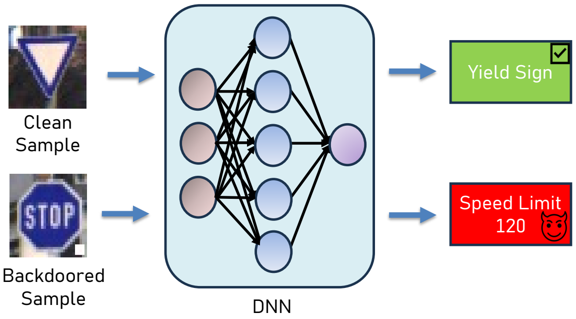

Fig. 1: Overview of backdoor attacks. A backdoored model correctly classifies clean inputs but misclassifies any input containing the hidden trigger to a target class specified by the adversary.

In contrast to model-level defenses, we introduce a *pre-training data cleansing* paradigm that identifies and filters poisoned samples *before* downstream training, preventing backdoor injection at its source. This proactive approach avoids the computational overhead of post-hoc model purification and integrates seamlessly into existing training pipelines. Furthermore, a critical challenge in backdoor detection is generalization to novel, unseen attack types. Adversaries continuously develop new trigger designs and injection strategies, and detectors trained only on known attacks may fail catastrophically against new variants. To address this, we leverage Vision-Language Models (VLMs) [11] and conditional prompt tuning [12] to generalize to unseen backdoor triggers in a zero-shot manner. We show that CLIP's rich multimodal representations, learned from hundreds of millions of image-text pairs, can capture subtle differences between clean and backdoored images, even for trigger types never seen during training. Our main contributions can be summarized as:

- We introduce a novel pre-training backdoor detection framework that identifies unseen backdoor attack types before model training, preventing backdoor injection at its source.
- We design an image-conditioned meta-network that enables instance-specific prompt adjustments, capturing subtle per-sample trigger characteristics.
- Through extensive experiments on CIFAR-10 and GTSRB with six diverse backdoor attacks, we demonstrate that our method achieves state-of-the-art performance in detecting unseen backdoors.

II. Related Work

A. Backdoor Attacks and Defenses

Backdoor attacks [5] poison training data with inputs stamped by a trigger so the model behaves normally on clean samples but misclassifies any input containing the trigger; the idea extends beyond vision to NLP [13]. Early work like BadNets [5] used simple patterns, while trojaning [14] reverse-engineers neuron activations to craft potent triggers. Subsequent variants increase stealth, for example steganographic, pixel-level embeddings [15], or naturalness, such as warping-based triggers in WaNet [16].

Defenses span the adversarial machine learning (AML) lifecycle but largely focus on model-level mitigation, leaving data vetting underexplored [17]. Training-time filters attempt to remove poisons [18], [19], [20], [21]: VisionGuard compares softmax responses under transformations [19], Deep k-NN prunes anomalies via feature-space voting [20], and Holmes ensembles external detectors on labels and top-k logits [21]. Traditional approaches assume a compromised model and clean it post-training [22], [23], [24], which is reactive and computationally costly.

B. Vision-Language Models and Prompt Tuning

Vision-Language Models (VLMs) such as CLIP [11], have revolutionized multimodal learning by learning unified representations from paired image-text data. CLIP, trained on 400 million image-text pairs from the internet, demonstrates remarkable zero-shot transfer capabilities and has become a foundation model for numerous downstream tasks. The joint training of image and text encoders creates a shared embedding space where semantically similar concepts cluster together, regardless of modality. Prompt tuning has emerged as an efficient adaptation strategy for large pre-trained models [25]. Rather than fine-tuning entire networks, prompt tuning learns a small set of continuous vectors prepended to input sequences while keeping the model frozen. CoOp [26] first applied prompt tuning to CLIP for few-shot image recognition, demonstrating significant improvements over hand-crafted prompts. CoCoOp [12] extended this by conditioning prompts on individual image features through a lightweight meta-network, enabling instance-specific adaptations.

III. Preliminaries

A. Problem Setup

We consider a data-centric detection setting where the goal is to flag backdoored images *before* any downstream model is trained. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be the underlying dataset, where $x_i \in \mathcal{X}$ is an image and $y_i \in \{1, \ldots, K\}$ is its semantic class label. An adversary selects a subset $\mathcal{D}_s \subset \mathcal{D}$ with poisoning rate $\alpha = |\mathcal{D}_s|/|\mathcal{D}|$ and applies an image transform $G_X : \mathcal{X} \to \mathcal{X}$ and optional label map G_Y to obtain poisoned pairs $(x', y') = (G_X(x), G_Y(y))$. This yields a pool containing both clean images and images with hidden backdoor triggers. For detection, we define $\mathcal{D}_{\text{det}} = \{(x_i, z_i)\}_{i=1}^{N}$, where $z_i \in \{\text{clean}, \text{backdoored}\}$ is a backdoor label indicating whether x_i contains a trigger, regardless of its semantic class y_i. Our objective is to learn a binary decision function $c(x) \in \{\text{clean}, \text{backdoored}\}$ that generalizes to unseen backdoor attack families. We focus on detecting backdoored inputs $G_X(x)$ rather than defending or retraining a downstream classifier.

B. VLMS

Models such as CLIP [11], trained on hundreds of millions of image–text pairs, map both modalities into compact, information-rich embeddings within a shared space. This joint training enables strong zero-shot transfer where short natural-language prompts can act as lightweight classifiers when compared to image features through cosine similarity. The same property makes VLMs a natural fit for prompt tuning [25], where both the image encoder and the text encoder are frozen and a small set of continuous prompt tokens is learned to steer the text side toward a downstream objective. This yields efficient adaptation with minimal additional parameters and good robustness under distribution shift.

Specifically, a VLM comprises a token embedding layer $E(\cdot)$, a text encoder $f_t(\cdot)$, and an image encoder $f_I(\cdot)$ that project inputs into a common feature space, typically followed by normalization and a temperature-scaled similarity for scoring. The prompt design controls the adaptation capacity without touching the backbone, letting us exploit the model's broad multimodal prior while keeping compute low. In our setting, we construct text prompts for the binary labels $\{\text{“clean”}, \text{“backdoored”}\}$ and tune only a small text-side adapter to separate clean from backdoored images, while reusing the strong visual representations learned by CLIP.

IV. Proposed Method

Our goal is to identify unseen backdoored types injected by adversaries to images by harnessing conditional prompt tuning of a frozen VLM, specifically CLIP [11]. CLIP's backbone consists of an image encoder $f_I : \mathcal{X} \to \mathbb{R}^D$ and text encoder $f_t : \mathcal{T} \to \mathbb{R}^d$. On top, we learn a small set of *continuous prompt tokens* and a lightweight *image-conditioned meta-network* that adapts those tokens on a per-sample basis. The learned prompts are prepended to class names {“clean”, “backdoored”} and encoded by the frozen f_t. For an input image x, we embed it with f_I and compare its normalized feature V to normalized, *instance-conditioned* text features $\{T_{x,\text{clean}}, T_{x,\text{backdoored}}\}$ using a scaled cosine similarity. Figure 2 depicts the overall architecture of our detector.

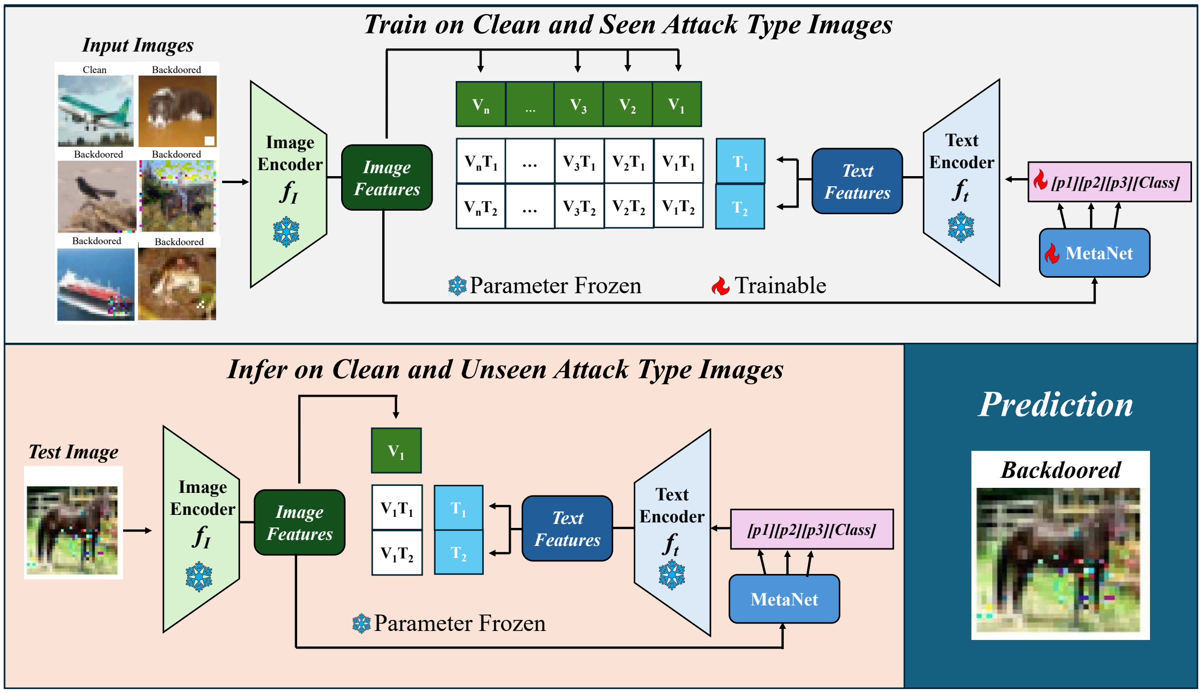

Fig. 2: The overall architecture of the proposed method on classifiying unseen backdoor attack images.

A. *Instance-Conditioned Prompt Tuning*

We initialize three continuous prompt vectors $p_1, p_2, p_3 \in \mathbb{R}^d$ with the word embeddings of the phrase “a photo of”, where $d = 512$ matches CLIP's token embedding dimension; we denote the number of prompt tokens by $m = 3$. For each backdoor label $c \in \{\text{clean}, \text{backdoored}\}$ we form a text prompt by concatenating these learnable tokens with the token embedding of the class name, i.e., $\text{prompt}_c = [p_1, p_2, p_3, E(c)]$, where $E(\cdot)$ is CLIP's token embedding layer. Passing this sequence through the frozen text encoder and normalizing yields a class embedding $T_c = f_t(\text{prompt}_c)/\|f_t(\text{prompt}_c)\|_2 \in \mathbb{R}^d$.

Each training image x_j is encoded by the frozen image encoder and then normalized as $V_j = f_I(x_j)/\|f_I(x_j)\|_2 \in \mathbb{R}^D$. To capture subtle, per-image trigger patterns, we introduce a lightweight meta-network $\text{MetaNet} : \mathbb{R}^D \rightarrow \mathbb{R}^d$ that leverages the image feature V_j and outputs a shift vector $\delta_j = \text{MetaNet}(V_j)$. We apply the same shift to each of the m prompt slots, i.e., $p'_i = p_i + \delta_j$ for $i = 1, \ldots, m$, which encourages coherent edits in text space while keeping the parameter count small. The instance-conditioned prompt for label c and image x_j is then $[p'_1, p'_2, p'_3, E(c)]$, which we feed to the frozen text encoder and normalize to obtain an instance-conditioned text feature:

$$T_{j,c} = \frac{f_t\big([p'_1, p'_2, p'_3, E(c)]\big)}{\big\|f_t\big([p'_1, p'_2, p'_3, E(c)]\big)\big\|_2}. \tag{1}$$

We score each backdoor label c with a temperature (scale) parameter $\tau > 0$ using cosine similarity between the image and text embeddings, i.e., $s_{j,c} = \tau\,(V_j \cdot T_{j,c})$. Let $z_j \in \{\text{clean}, \text{backdoored}\}$ denote the backdoor label of x_j in the detection dataset. The detector is trained with a standard cross-entropy objective over the two backdoor labels:

$$L = -\frac{1}{N}\sum_{j=1}^{N} \log \frac{\exp(s_{j,z_j})}{\sum_{c\in\{\text{clean},\text{backdoored}\}} \exp(s_{j,c})}. \tag{2}$$

B. *Inference*

During inference, the soft prompts $\{p_i\}$ and the MetaNet parameters are frozen. For each test image x_j, we compute the normalized image feature V_j, obtain the shift $\delta_j = \text{MetaNet}(V_j)$, and form the instance-conditioned prompts and text embeddings $T_{j,c}$ for $c \in \{\text{clean}, \text{backdoored}\}$ exactly as in training. We then compute scores $s_{j,c} = \tau\,(V_j \cdot T_{j,c})$ and predict the backdoor label via $\hat{z}_j = \arg\max_{c\in\{\text{clean},\text{backdoored}\}} s_{j,c}$. This procedure aligns clean and backdoored images with their corresponding image-conditioned text embeddings and enables robust detection of unseen backdoor triggers, without ever predicting the underlying semantic class y.

V. Experimental Results

A. *Implementation Details*

Attack Models. We comprehensively evaluate our detector on six established backdoor schemes representing diverse threat models. BadNets-SQ [5] embeds a checkerboard pattern in the bottom-right corner, creating a visible, localized trigger. BadNets-PX [5] randomly modifies individual pixels scattered throughout the image, producing a subtle, distributed trigger pattern. Trojan-SQ [14] employs a square trigger with label-specific patterns, where different source classes use distinct triggers. Trojan-WM [14] applies watermark-style triggers blended with the original image content using alpha compositing. ℓ_2-inv [15] generates imperceptible ℓ_2-bounded perturbations optimized for specific source-target class pairs.

TABLE I: Unseen Attack Type Classification Accuracy (%).

Dataset	Method	Trojan-WM	Trojan-SQ	ℓ_2-inv	ℓ_0-inv	Badnets-SQ	Badnets-PX	Average
CIFAR-10	Simple-CNN	64.86	75.10	51.56	49.93	49.94	50.10	56.92
	Deep-CNN	76.37	58.69	55.77	50.18	50.14	50.04	56.87
	ResNet-18	84.52	75.40	53.42	51.17	50.00	54.29	61.47
	Proposed	**98.41**	**98.62**	**88.40**	**98.05**	**83.24**	**63.03**	**88.79**
GTSRB	Simple-CNN	82.86	60.80	53.58	50.84	50.00	50.03	58.02
	Deep-CNN	83.93	61.61	53.17	51.45	50.01	50.04	58.36
	ResNet-18	74.40	87.52	52.75	69.29	50.01	50.02	64.00
	Proposed	**98.49**	**97.41**	**74.20**	**79.98**	**69.13**	**59.71**	**79.15**

ℓ_0-inv [15] creates sparse ℓ_0-constrained perturbations affecting only a small fraction of pixels while remaining visually imperceptible. These attacks span a wide spectrum of visibility and spatial distribution

Datasets. We conduct experiments on two widely-used computer vision benchmarks. CIFAR-10 [27] contains 50,000 training and 10,000 test images across 10 object classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck). GTSRB [28]: The German Traffic Sign Recognition Benchmark comprises 39,209 training and 12,630 test images spanning 43 traffic sign categories. These datasets provide complementary challenges: CIFAR-10 offers controlled conditions with uniform resolution, while GTSRB presents real-world complexity with varying image sizes, lighting conditions, and background clutter.

Baseline Models. We compare our VLM-based detector against three CNN architectures trained from scratch on the same binary detection task: Simple-CNN [29] consists of three convolutional layers and a fully connected, Deep-CNN [28] consists of six convolutional layers with a dropout layer and fully connected layer for classification, and ResNet-18 [30].

Training Configuration. We freeze CLIP's pre-trained ViT-B/16 encoders (both image and text), learning only the prompt embeddings and MetaNet parameters. Our learnable components comprise three prompt tokens (initialized with embeddings of "a photo of") and a lightweight MetaNet with two fully-connected layers (768→256→512 dimensions) with ReLU activations. We optimize using Adam with learning rate 2×10^{-3} for prompts and 1×10^{-4} for MetaNet, applying weight decay $\lambda = 1 \times 10^{-4}$. Training proceeds for 10 epochs with batch size 128 and cosine similarity temperature (scale) $\tau = 100$.

B. Main Results

Table I reports the detection accuracy for each experiment's unseen backdoor type on CIFAR-10 and GTSRB. Our prompt-conditioned VLM consistently outperforms CNN baselines and a ResNet-18 detector trained from scratch. On CIFAR-10, our proposed prompt-tuned VLM detector achieves an average accuracy of 88.79%, outperforming the ResNet-18 baseline by 27.3 points. On GTSRB, average accuracy reaches 79.15%, a 15-point improvement over ResNet-18.

Regarding per-attack analysis, for both datasets, visible and spatially coherent triggers, such as Trojan-WM and Trojan-SQ, are detected with very high accuracy (CIFAR-10: 98.41/98.62; GTSRB: 98.49/97.41). Imperceptible or global perturbations are more challenging but still show strong gains over baselines: on CIFAR-10, ℓ_2-inv reaches 88.40% and ℓ_0-inv 98.05%; on GTSRB, ℓ_2-inv and ℓ_0-inv achieve 74.20% and 79.98%, respectively. However, local but low-magnitude pixel triggers (BadNets-PX) are the most difficult unseen backdoor attacks to detect (63.03% on CIFAR-10; 59.71% on GTSRB). This is likely because their sparse, distributed alterations resemble sensor noise or natural texture and thus require more aggressive instance adaptation or ensembling. In contrast, structured corner triggers (BadNets-SQ) are substantially easier (83.24% and 69.13%), reflecting the model's ability to align text embeddings toward localized triggers when guided by the image-conditioned shift.

These results demonstrate that conditioning continuous prompts on image embeddings enables robust generalization to novel backdoor patterns. By training on five attack types and holding out the sixth for inference, our method consistently surpasses standard CNNs and ResNet-18 across both datasets.

VI. Discussion and Limitations

From a deployment perspective, the detector's compute and parameter overhead are minimal since only three learnable prompt tokens and a small MLP on top of frozen CLIP encoders are trainable. Furthermore, we leverage the smaller CLIP variant (ViT-B/16) for this study, making our proposed method feasible as a pre-training filter.

The empirical gains indicate that instance-conditioned prompt tuning leverages CLIP's broad prior effectively while staying data-efficient and stable. It works especially well on clearly structured triggers, likely because the image-conditioned shift can pull the text prototypes toward those localized artifacts. The more difficult cases are sparse, low-magnitude perturbations (BadNets-PX) that look like sensor noise or normal texture, particularly under domain shifts (e.g., GTSRB). Closing this gap will require methods that better capture fine, high-frequency details without becoming brittle to natural variation.

Several limitations temper the present findings. We test a leave-one-attack-out setup over six families; a stronger, adaptive attacker could design triggers to shrink our margin or imitate the "clean" direction. Results carry from CIFAR-10 to GTSRB, but larger domain shifts may need recalibrating τ or light template ensembling. The shared per-image shift δ keeps the model small but can miss pixel-distributed triggers. Very low poisoning rates and label noise can bias training toward "clean" images.

VII. Conclusion

In this paper we propose a novel, preventative detector that flags previously unseen, backdoored images before they can contaminate training data or slip through at inference. Experimental results demonstrate that our proposed image-conditioned VLM can detect a variety of unseen backdoor threats by leveraging the knowledge learned from previously seen backdoor triggers. Future work will focus on improving detection rates on sparse pixel-level triggers, exploring additional parameter-efficient fine-tuning mechanisms, and validating more robust and general datasets with additional triggers.

VIII. Acknowledgment

This work is partially supported by the UWF Argo Cyber Emerging Scholars (ACES) program funded by the National Science Foundation (NSF) CyberCorps® Scholarship for Service (SFS) award under grant number 1946442. Any opinions, findings, and conclusions or recommendations expressed in this document are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4804–4814.

[2] K. Stein, A. A. Mahyari, G. Francia III, and E. El-Sheikh, "Transductive one-shot learning meet subspace decomposition," *arXiv preprint arXiv:2504.00348*, 2025.

[3] K. Sundararajan and D. L. Woodard, "Deep learning for biometrics: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, pp. 1–34, 2018.

[4] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A survey of deep learning applications to autonomous vehicle control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 712–733, 2020.

[5] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[6] L. Song, X. Yu, H.-T. Peng, and K. Narasimhan, "Universal adversarial attacks with natural triggers for text classification," in *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2021.

[7] Y. K. Singla, S. Parekh, S. Singh, C. Chen, B. Krishnamurthy, and R. R. Shah, "Minimal: mining models for universal adversarial triggers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 330–11 339.

[8] Y. Shi, M. Du, X. Wu, Z. Guan, J. Sun, and N. Liu, "Black-box backdoor defense via zero-shot image purification," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[9] D. Yuan, S. Wei, M. Zhang, L. Liu, and B. Wu, "Activation gradient based poisoned sample detection against backdoor attacks," *arXiv preprint arXiv:2312.06230*, 2023.

[10] S. Wei, M. Zhang, H. Zha, and B. Wu, "Shared adversarial unlearning: Backdoor mitigation by unlearning shared adversarial examples," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[12] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.

[13] X. Chen, A. Salem, D. Chen, M. Backes, S. Ma, Q. Shen, Z. Wu, and Y. Zhang, "Badnl: Backdoor attacks against nlp models with semantic-preserving improvements," in *Proceedings of the 37th Annual Computer Security Applications Conference*, 2021, pp. 554–569.

[14] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018.

[15] S. Li, M. Xue, B. Z. H. Zhao, H. Zhu, and X. Zhang, "Invisible backdoor attacks on deep neural networks via steganography and regularization," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2088–2105, 2020.

[16] A. Nguyen and A. Tran, "Wanet–imperceptible warping-based backdoor attack," *arXiv preprint arXiv:2102.10369*, 2021.

[17] B. Wu, S. Wei, M. Zhu, M. Zheng, Z. Zhu, M. Zhang, H. Chen, D. Yuan, L. Liu, and Q. Liu, "Defenses in adversarial machine learning: A survey," *arXiv preprint arXiv:2312.08890*, 2023.

[18] D. Tang, X. Wang, H. Tang, and K. Zhang, "Demon in the variant: Statistical analysis of {DNNs} for robust backdoor contamination detection," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 1541–1558.

[19] Y. Kantaros, T. Carpenter, K. Sridhar, Y. Yang, I. Lee, and J. Weimer, "Real-time detectors for digital and physical adversarial inputs to perception systems," in *Proceedings of the ACM/IEEE 12th International Conference on Cyber-Physical Systems*, 2021, pp. 67–76.

[20] N. Peri, N. Gupta, W. R. Huang, L. Fowl, C. Zhu, S. Feizi, T. Goldstein, and J. P. Dickerson, "Deep k-nn defense against clean-label data poisoning attacks," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 55–70.

[21] J. Wen, "Holmes: to detect adversarial examples with multiple detectors," *arXiv preprint arXiv:2405.19956*, 2024.

[22] M. Zhu, S. Wei, L. Shen, Y. Fan, and B. Wu, "Enhancing fine-tuning based backdoor defense with sharpness-aware minimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4466–4477.

[23] M. Guo, Y. Yang, R. Xu, Z. Liu, and D. Lin, "When nas meets robustness: In search of robust architectures against adversarial attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 631–640.

[24] Z. Yue, B. Lin, Y. Zhang, and C. Liang, "Effective, efficient and robust neural architecture search," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.

[25] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[26] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[27] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[28] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural networks*, vol. 32, pp. 323–332, 2012.

[29] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625–1634.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.