JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 38, XXX-XXX (2022)
DOI: 10.6688/JISE.20220X_38(X).00XX

The Impact of Artificial Intelligence Creation on Intellectual Property under Deep Learning

YAND-WANG[1], KIM-DOHOON[2+] AND JUN-JIANG CHEN[3]
[1]*Chongqing Intellectual Property School, Chongqing University of Technology, Chongqing, 400054, Chongqing, China;181537254@qq.com*
[2]*Institute of ASEAN Studies, DONG-A University，2409ho, 37, Marine city 3-ro, Haeundae-gu, Busan, South Korea; kdh10954@dau.ac.kr*
[3]*School of Political Science, Law and Public Administration, Yan'an University, Yan'an 716000, Shaanxi, China; wy06120517@163.com*
E-mail: wy06120517@163.com

Abstract: The aim of this study is to explore the impact of artificial intelligence creations on intellectual property from the perspective of deep learning and proposes an efficient geometric localization method for distributed sensor networks. First, this study analyzes the impact of AI-generated content on intellectual property rights from the perspectives of copyright and patent systems. It then briefly introduces intellectual property protection methods for image semantic segmentation and examines backdoor mechanisms in deep learning. Additionally, the backdoor mechanism is applied to image semantic segmentation models, and an adversarial example generation method is used to construct a trigger set generation algorithm for the semantic segmentation model. The performance of the algorithm is validated through ablation experiments and fine-tuning attack experiments. The results indicate that the incorporation of the generated trigger set digital watermark exerts minimal impact on the performance of the original model. Concurrently, its Mean Intersection Over Union (MIOU) on the trigger set is notably high, reaching 94.01%. This implies that the trigger set generated by the algorithm has successfully established a robust association between the semantic segmentation model and the model owner. The MIOU value of the model post-fine-tuning attack remains consistent with that prior to the attack, demonstrating that the digital watermark embedded in the segmentation model by the trigger set generation algorithm possesses stable robustness. The proposed protection method for the image semantic segmentation model offers novel approaches for the intellectual property protection of AI creations.

Keywords: intellectual property; AI creation; image semantic segmentation; trigger set; fine-tuning attack

1. INTRODUCTION

1.1 Research Background and Motivations

Artificial intelligence (AI) has garnered significant attention across various sectors. The rapid advancements in machine computing capabilities and learning algorithms have accelerated the development of a new generation of AI, surpassing human expectations. Selvadurai et al. (2020) explored the issue of copyright protection for AI-generated content and proposed utilizing blockchain technology to track and safeguard the ownership of AI creations [1]. Morrish (2021) examined the challenges associated with AI patent applications, highlighting the inadequacies of current patent laws in protecting AI innovations,

and recommended legal revisions to better accommodate advancements in AI technology [2]. Furthermore, the swift progress in Deep Neural Network (DNN) technology has led to substantial success in fields such as computer vision, machine translation, and speech recognition. Fkirin et al. (2022) investigated watermarking techniques for protecting DNN models and demonstrated the efficacy of digital watermarking in preventing model piracy [3]. However, high-performance DNN is developed at significant cost and holds substantial commercial value, leading to model infringement by malicious actors. This rampant piracy severely damages the interests of model owners, making the protection of the intellectual property (IP) of these models imperative [4, 5].

This study is dedicated to innovative exploration in the field of IP protection for deep learning (DL) models, moving beyond the mere repetition of existing watermarking techniques. It introduces a backdoor mechanism and proposes a novel digital watermark protection method, which offers unique advantages in safeguarding IP rights. By integrating the backdoor mechanism with the image semantic segmentation model, this study effectively establishes a connection between the model and its owner, thereby achieving robust protection for DL models. Empirical results demonstrate the effectiveness and robustness of the proposed digital watermark protection method, providing a feasible solution for DL model IP protection. Furthermore, this study addresses IP protection issues from a broader perspective, encompassing aspects such as copyright and patent systems. The inclusion of novel elements, such as image semantic segmentation models, enhances the depth and comprehensiveness of the research. This comprehensive approach ensures that the study is not confined to traditional watermarking techniques but considers various facets of IP protection, offering new ideas and methods for research in this domain. Lastly, the findings of this study have significant practical implications for addressing the challenges of IP protection in contemporary AI creations. With the widespread application of DL in the AI field, new challenges have emerged for IP protection, and the proposed digital watermark protection method provides a viable solution to these challenges. This study offers valuable references and insights for related fields, contributing to the development and application of IP protection technologies.

1.2 Research Objectives

The aim of this study is to explore the impact of IP protection on AI-generated content within DL models. The specific research objectives include:

1. Analyze the impact of AI-generated content on copyright and patent systems, to gain a deeper understanding of the status and challenges of these creations within the existing IP legal framework.
2. Investigate IP protection methods in the field of image semantic segmentation, particularly by employing a backdoor mechanism to safeguard DL models.
3. Utilize adversarial example generation techniques to construct trigger set digital watermarks, in order to verify their effectiveness and robustness in protecting segmentation models.
4. Provide new theoretical and practical approaches to IP protection for AI-generated content through experiments and analysis, and offer theoretical support and technical guidance for research and applications in related fields.

These objectives aim to address the new challenges of IP protection posed by AI-

generated content through innovative methods and to provide a comprehensive perspective and in-depth analysis for research in this field.

2. LITERATURE REVIEW

Scholars have conducted extensive research on the IP protection of DL models. Xue et al. (2021) investigated existing IP protection methods for the DNN from the perspectives of scenarios, mechanisms, capacities, types, functions, and target models. Additionally, they analyzed potential attacks on DNN-based IP protection methods, focusing on model modification, evasive attacks, and active attacks. They proposed a systematic evaluation method for DNN-based IP protection, considering basic functional indicators, anti-attack indicators, and custom indicators for various application scenarios [6]. While their study provided a comprehensive investigation and analysis of existing IP protection efforts, it primarily focused on scenarios, mechanisms, and capabilities, rather than delving into specific technical details or innovative aspects. This approach lacks in-depth exploration and empirical evaluation of specific technical solutions. Liang et al. (2020) combined mapping functions and deep reinforcement learning technology to preprocess the ownership information of IP circuit resources, proposing a fast virtual IP watermarking detection algorithm based on deep reinforcement learning. Experimental results demonstrated that the proposed algorithm effectively improved watermark detection speed, reduced resource overhead, and exhibited excellent security performance [7]. However, their experiment did not thoroughly validate and evaluate the applicability of the algorithm in real-world application scenarios. Moreover, their study was confined to the domain of IP watermark detection and did not address the broader issue of DL model IP protection. Zhang et al. (2022) proposed a new model watermarking framework to protect deep networks trained for low-level computer vision or image processing tasks. They designed a deep hidden watermarking mechanism and demonstrated the robustness of the proposed framework through experiments, which resisted attacks from different network structures and target functions [8]. While they demonstrated the robustness of the proposed framework, they did not thoroughly test its performance across different application scenarios, nor did they conduct a comparative analysis with other related studies to validate its advantages and uniqueness. Additionally, previous studies also had some limitations. For instance, Wu et al. (2020) proposed a novel digital watermarking framework applicable to the output images of the DNN, ensuring that any image produced by the watermarking neural network contained specific watermarks [9]. Li et al. (2022) introduced a novel ownership verification scheme named federated DNN (FedDNN), which allowed private watermarks to be embedded and verified to claim legitimate IPs of federated learning models without revealing private training data or private watermark information [10]. Zuo (2024) investigated the ethical risks of machine writing in the field of knowledge production and the deconstruction of the existing IP system, noting that technological innovation disrupted the existing IP balance mechanism and that machine writing became a new force in knowledge production [11]. However, the research scope of Wu et al. and Zuo was limited to specific federated learning scenarios and did not include in-depth comparisons and evaluations of other IP protection methods. In summary, although previous research has made some progress in protecting the IP rights of DL models, there remain shortcomings, such as insufficient exploration of specific technical solutions, empirical evaluations, and applicability tests across different scenarios. Amiri et al. (2024) introduced a novel algorithm called Hippo-

potamus Optimization (HO), inspired by the natural behavior of hippos and based on innovative stochastic techniques. The HO algorithm demonstrated exceptional performance across multiple benchmark functions, showcasing the innovative application of metaheuristic methods. It effectively balances exploration and exploitation, akin to the application of DL models in IP protection discussed here. Particularly in DL models, the innovation and efficiency of protective algorithms are crucial. The HO algorithm, with its ability to balance exploration and exploitation, offers new solutions for addressing complex engineering design challenges, which is closely related to the protection of IP in AI creations [12]. Mehrabi Hashjin et al. (2024) investigated a hybrid classifier that integrates DNN with Type-III fuzzy systems. This innovative approach demonstrated superior performance in decision-making and optimized the system's rule parameters using the Improved Chaos Game Optimization (ICGO) algorithm. The ICGO algorithm has shown outstanding performance across various benchmark functions and engineering problems. This method is closely aligned with one of the key objectives of this study: the application of DL models in IP protection. By enhancing the accuracy and performance of classifiers, the ICGO algorithm provides a more precise and effective protection mechanism for AI creations, resonating with the protective methods explored in this study [13].

This study innovatively explores the protection of IP rights of DL models, demonstrating distinct advantages and innovations compared to previous research. Firstly, this study focuses on the impact of AI creation on IP rights, particularly within the context of DL models. This emphasis makes the study highly targeted and practical, as the application of DL in AI is becoming increasingly widespread, presenting new challenges to IP protection. Unlike other scholars' research, this study comprehensively addresses the impact of AI creation on IP rights from a broader perspective, including discussions on copyright and patent systems. Furthermore, it introduces image semantic segmentation models and backdoor mechanisms in DL, providing a more in-depth and comprehensive research perspective. This study proposes a novel digital watermark protection method that utilizes the backdoor mechanism of DL models for IP protection. By integrating the backdoor mechanism into the image semantic segmentation model and employing the generated trigger set digital watermark, it successfully establishes a connection between the model and its owner. Compared to traditional digital watermarking techniques, this method is more covert and robust, effectively safeguarding IP rights. Rigorous experimental verification confirms the effectiveness and robustness of the proposed digital watermark protection method, showcasing significant innovation and practicality in protecting the IP rights of DL models.

This study provides a more comprehensive and in-depth research perspective compared to other relevant studies, offering valuable references and guidance for addressing the current challenges of IP protection in AI creation. It begins by analyzing the impact of AI creations on IP and its implications for copyright and patent systems. The study then explores IP protection methods for image semantic segmentation, creatively applying the backdoor mechanism of DL to the semantic segmentation model. Additionally, it establishes a trigger set generation algorithm for the semantic segmentation model using adversarial example generation techniques and verifies the algorithm's effectiveness through experiments. The discussion on IP protection methods for image semantic segmentation models based on DL contributes significantly to the IP protection of AI creations.

3. RESEARCH METHODOLOGY

3.1 Impact of AI creation on IP

AI creation, a term derived from AI, encompasses literary and artistic works, technical solutions, and new industrial designs generated by AI [14]. It has significantly influenced IP systems, particularly in terms of copyright and patent law.

Regarding the copyright system, AI creation impacts copyrightability, ownership rights, and the scope of copyright protection. Under China's copyright law, copyrightable works are original intellectual creations that can be expressed in tangible forms in literature, art, and science [15]. There is contention over denying high-quality AI-generated works copyright protection solely because they lack human authorship, while granting copyright to lower-quality works created by humans. This contradicts the copyright system's purpose, potentially dampening enthusiasm among AI technology investors and developers and impeding industrial innovation and progress [16, 17]. However, AI generates content similar to works through passive machine learning, dependent on original data or corpora. AI creations do not attain the creativity and intellectual prowess of human subjective thought, often falling short of the minimum creativity threshold required for copyright protection [18]. In terms of ownership rights, a primary dispute concerns whether AI itself can autonomously qualify as a subject under the copyright system. Proponents argue that the IP framework should recognize AI's substantial contributions and grant it legal personality and rights. Opponents contend that AI lacks economic incentives to participate effectively in legal systems and cannot autonomously exercise rights, fulfill obligations, or bear legal responsibilities as a distinct legal entity [19, 20]. Concerning the scope of copyright protection, granting copyright to AI creations complicates distinguishing between personal and property rights [21]. Moreover, China's law stipulates a copyright protection period extending from the creation's completion until 50 years after the creator's death, a timeline incongruent with AI's indefinite operational capability due to its absence of a natural lifespan.

The patentability of AI creations, ownership of patent rights, patent infringement, and related issues remain contentious within the patent system. The fundamental objective of the patent system is to safeguard the legitimate interests of patent holders against infringement [22]. In patent law, an inventor typically refers to a natural person or legal entity, excluding machines or intelligent devices. Therefore, AI lacks legal subject status, and its derived technical solutions are generally considered non-patentable. Nonetheless, some scholars argue that if an AI-generated technical solution meets the patentability criteria, including novelty, inventive step, and industrial applicability, it should qualify for patent protection without additional requirements due to its AI origin [23, 24]. Similarly to issues concerning copyright ownership, disputes over patent rights for AI creations have prompted discussions on various distribution models for related interests. Advocates suggest that AI involved in designing technical solutions should be recognized as an independent inventor, with patent rights accruing to the AI itself, managed by its owner or investor. Opponents contend that AI-generated innovations invariably involve human intervention and planning, with developers and users driving commercialization efforts, thereby precluding AI from holding rights or fulfilling obligations independently [25, 26].

Concerning patent infringement, AI's robust information gathering and analytical capabilities enable it to identify subtle differences between new and existing inventions more efficiently. This capability increases the risk of generating imitation or minor variations that may lead to a proliferation of low-quality patents [27].

3.2 Image semantic segmentation model and its IP protection technology

The core objective of image semantic segmentation involves meticulously partitioning various regions within an image. Typically, DL algorithms like convolutional neural networks are employed for classification purposes. Evaluation metrics for assessing the performance of image semantic segmentation models include Pixel Accuracy (PA), Intersection Over Union (IOU), Mean Pixel Accuracy (MPA), and Mean Intersection Over Union (MIOU) [28]. Assuming there are N+1 categories in the segmentation task, this entails calculating the number of pixels belonging to class a but predicted as class b using Equations (1)-(5):

$$\mathrm{PA} = \frac{\sum_{a=0}^{n} W_{aa}}{\sum_{a=0}^{n}\sum_{b=0}^{n} W_{ab}} \tag{1}$$

$$IOU = \frac{W_{aa}}{\sum_{b=0}^{n} W_{ab} + \sum_{b=0}^{n} W_{ba} - W_{aa}} \tag{2}$$

$$MPA = \frac{1}{N+1}\sum_{a=0}^{n} \frac{W_{aa}}{\sum_{b=0}^{n} W_{ab}} \tag{3}$$

$$MIOU = \frac{1}{N+1}\sum_{a=0}^{n} \frac{W_{aa}}{\sum_{b=0}^{n} W_{ab} + \sum_{b=0}^{n} W_{ba} - W_{aa}} \tag{4}$$

W_{aa} denotes the count of pixels truly belonging to class a and correctly predicted as class a. W_{ab} denotes the count of pixels belonging to class a but predicted as class b. PA signifies the ratio of true positive pixels to the total predicted pixels, indicating the model's classification accuracy. IOU represents the ratio of correctly predicted pixels W_{aa} to the total number of pixels predicted as class a or actually belonging to class a. MPA denotes the average PA across all classes, while MIOU signifies the average IOU across all classes.

To safeguard DL models, such as image semantic segmentation, from illicit redistribution and to protect researchers' interests, it is imperative to imprint the model with a unique identifier to create a watermark that aids in the recognition of its creator. Digital watermarking technology integrated with DNN-based IP protection must seamlessly embed the watermark into the model without compromising its original performance. Additionally, robustness of the digital watermark is crucial to prevent removal by malicious actors utilizing various attack methods. Current attack methods commonly include evasion, fine-tuning, and pruning [29]. Among these, fine-tuning is the most prevalent method employed by researchers, requiring minimal computational resources and training data to transition the model from one task to another. Four primary fine-tuning strategies currently utilized are: Fine-Tune Last Layer of the model (FTLL), Fine-Tune All Layer parameters in the model (FTAL), Re-Train Last Layer of the model (RTLL), and Re-Train All Layer parameters (RTAL) [30]. These strategies aim to circumvent ownership verification or tamper with the model's digital watermark to facilitate model theft.

3.3 Analysis of the backdoor mechanism of the DL model

DL stands as the predominant technology within the field of AI, with DNN serving as the primary model choice among researchers. However, the presence of backdoors within DNNs is often overlooked. The functionality of a DNN in specific scenarios heavily relies on its training data [31]. Once a backdoor is integrated into the deep model, detecting it becomes challenging. There is an increasing demand for DNN models, leading to a trend of over-parameterization. Consequently, an abundance of parameters enhances the model's fitting performance, enabling it to even learn from anomalous data [32]. One notable instance of such data includes those associated with erroneous tags. When the model encounters this specific set of incorrect data, it triggers latent behaviors embedded within the model, resulting in erroneous predictions—akin to activating the model's backdoor. This particular set of data is referred to as the trigger set. Figure 1 illustrates the operational workflow of embedding a backdoor into a model [33].

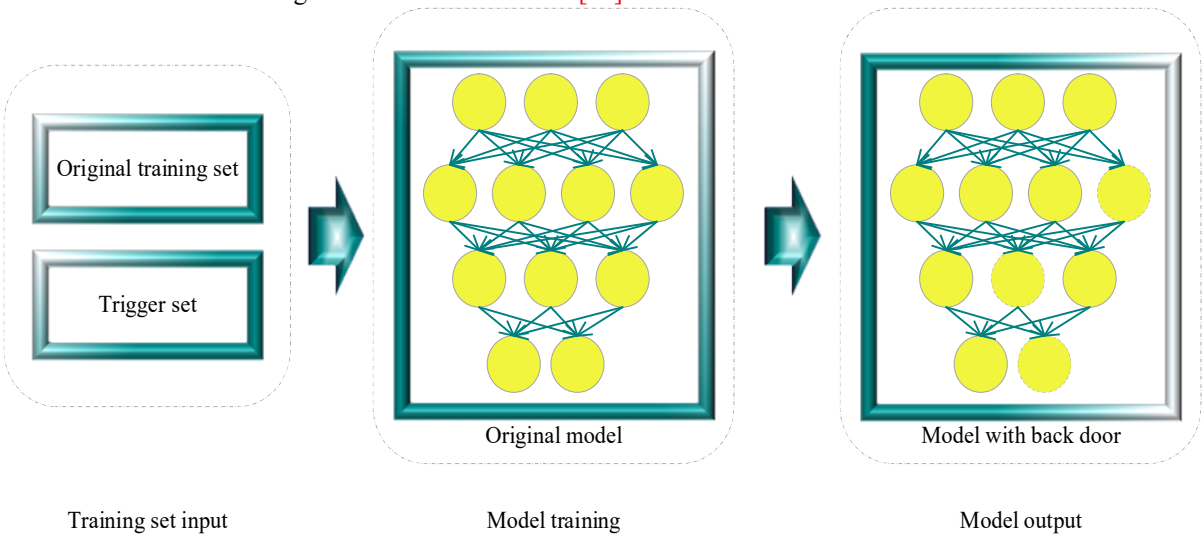

Fig. 1. Operation flow of embedding back door into DNN.

Figure 1 depicts a procedure that commences with the selection or design of specific trigger data. These data are incorporated into the model during training, enabling the model to exhibit predetermined abnormal behavior upon encountering them. During the training phase, these trigger data are included in the training dataset, allowing the model to learn to activate the backdoor mechanism when confronted with such data. The integration of the backdoor is typically concealed, rendering detection challenging. Therefore, careful consideration is necessary regarding the stealth of the backdoor and the specificity of activation conditions during its design phase. Subsequently, once the model is deployed and utilized for operational tasks, any input meeting the trigger conditions will activate the backdoor, leading the model to produce manipulated outcomes. During testing, the trigger set can be employed to initiate the backdoor mechanism for verifying its existence. Figure 2 illustrates this process [34].

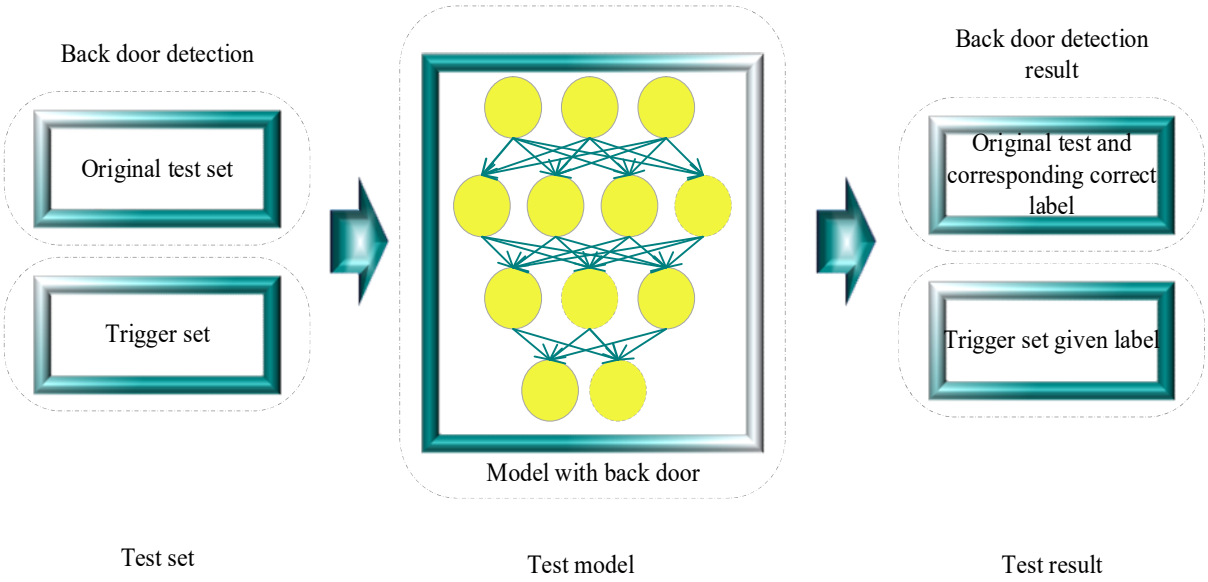

Fig. 2. Backdoor detection process in DNN.

Figure 2 illustrates a structured process characterized by the following stages: Initially, a set of test data must be prepared or acquired, encompassing normal input data alongside potential trigger data. Subsequently, these data are fed into the model, and the model's responses are monitored. Deviations in the model's response to specific input data, indicative of unexpected or abnormal behavior, may suggest activation of a backdoor. This prompts further analysis, such as scrutinizing the model's decision-making processes or identifying discrepancies in outputs, to confirm the presence of a backdoor. Upon confirmation, remedial actions can be implemented, such as model retraining or bolstering security measures, crucial for fortifying the model's security and reliability. The backdoor flaw in DNN models, when strategically leveraged, can serve as a mechanism for establishing IP protection and realizing advantages. By employing a designated trigger set, the model can be activated to produce desired outcomes as envisioned by its creator, forging a direct association between the model and its proprietor. Integrating the backdoor mechanism into the image semantic segmentation model allows the creation of a digital watermark, thereby safeguarding the IP of the semantic segmentation model.

3.4 Trigger set generation algorithm for semantic segmentation model

The method of generating adversarial examples involves embedding the model owner's information into a normal image through subtle noise, accompanied by a specific mask designated as the trigger set. The sequential process is outlined as follows [35-37].

Initially, the deeplab-V3+ segmentation model, trained on the Pattern Analysis, Statistical Modelling and Computational Learning Visual Object Classes (PASCAL VOC) dataset, is selected for implementation. This model proficiently identifies 21 types of objects (20 specific objects plus background), denoted as Q=21. Let A represent the input image. In an image, there exist K objects that the model accurately segments, defined as

$R = \{r_1, r_2, \cdots, r_k\}$, where $K \leq Q$, and each object r_k corresponds to a real label t_k with $k \in \{1,2,\cdots,Q\}$, encompassing all category labels $T = \{t_1, t_2, \cdots, t_k\}$. The deeplab-V3+ image semantic segmentation model correctly identifies the target in image A, yielding $Y(A, r_k) = t_k$ as the result. To generate an adversarial example, it is crucial to induce the model to classify the pixel category in the image as $t_h' \neq t_k$, where $h \neq k$. The model's output is represented by Equation (5):

$$Y(\mathrm{A+v,w}_h) = t_h^{'} \tag{5}$$

In Equation (5), A denotes the input image, and v signifies the perturbation acquired through iterative backpropagation. w_h represents the specified textual pattern, and $t_h^{'}$ denotes the predicted class of the given textual pattern. By adding v to the input image A, the model's output result becomes $Y(\mathrm{A+v,w}_h)$.

During model training, the loss function is defined as Equation (6):

$$F_{trig} = \sum_{i=1}^{K} F_{seg}[Y(A+v, r_k), t_k] + \varphi_1 F_{seg}[Y(A+v, w_h), t_h^{'}] \tag{6}$$

In Equation (6), the loss function of the original semantic segmentation model is F_{seg}, and no adversarial example is obtained. K represents the number of objects in the image that the model can correctly segment, and $Y(A+v, r_k)$ denotes the segmentation result of the object r_k in image A. F_{trig} is the loss function used when generating the trigger set, which combines the original loss and the loss of the trigger set.

The first half of the loss function is the loss function of the calculated iterated image $(A+v)$ and the original real label, and the second half is the loss function of the iterated image $(A+v)$ and the given text pattern mask and category. The iterative image $(A+v)$ achieves the desired effect by minimizing F_{trig}, that is, the loss function suppresses the confidence of the original real label mask and increases its confidence with the given label mask.

Then, the gradient descent method is employed for optimization. A_n obtained from the model after n iterations is the addition of n disturbances v_n. At this time, if the image A_n passes through the segmentation network, the output result will be the previously given label mask [38]. After each iteration, the disturbance obtained by gradient calculation is shown in Equation (7):

$$v_n = \nabla_{A_n}\{\sum_{i=1}^{K} F_{seg}[Y(A_n, r_k), t_k] + \varphi_1 \sum_{i=1}^{H} F_{seg}[Y(A_n, w_h), t_h^{'}]\} \tag{7}$$

In Equation (7), v_n represents the perturbation after the n-th iteration. In the whole process, the normalization operation is often conducted first and then added to the image to avoid the instability of the disturbed data. Equation (8) is a normalization operation.

$$v_n = \frac{v_n}{\|v_n\|_\infty} \tag{8}$$

Moreover, a hyperparameter γ is provided to minimize the disturbance added each time. After multiplying by v_n, it is added to image A_n. The specific operation is shown in Equation (9):

$$A_{n+1} = A_n + \gamma * v_n \tag{9}$$

In Equation (9), A_{n+1} represents the image after the application of the perturbation, and γ is a parameter used to control the magnitude of the perturbation added in each iteration. The pixels of the final converted image need to be controlled at [0,255] [39].

4. EXPERIMENTAL DESIGN AND PERFORMANCE EVALUATION

4.1 Datasets Collection

The experiment utilizes the deeplab-V3+ image semantic segmentation model, applied to the PASCAL VOC 2012 dataset. This dataset comprises 20 categories of images, predominantly featuring people, animals, vehicles, and furniture. For model training, 1464 images from this dataset are allocated, each with identical labeling. The trigger set used in this study is also derived from this dataset.

4.2 Experimental Environment

The experimental setup operates on a high-performance computing server equipped with multiple NVIDIA RTX 2080 Ti GPUs, delivering ample computational capability for DL model training and testing. The server runs on Linux Ubuntu 18.04 LTS, ensuring software environment stability and compatibility. Development and experimentation employ Python 3.8, TensorFlow 2.0, alongside essential scientific computing libraries such as NumPy and SciPy. For effective visualization of experimental outcomes related to model backdoor mechanisms, Matplotlib and Seaborn data visualization tools are employed. Experimental datasets are stored on high-speed solid-state drives to expedite data retrieval and ensure seamless experiment execution.

4.3 Parameters Setting

Parameter configurations adhere to standard practices in DL, with adjustments tailored to experimental objectives. The DL models employ a batch size of 32 and utilize the Adam optimizer with an initial learning rate set at 0.001, incorporating a learning rate decay strategy for enhanced training optimization. Model training spans 100 epochs to facilitate comprehensive learning and mitigate overfitting. The trigger set generation algorithm employs appropriate hyperparameters to regulate perturbation magnitude, ensuring effective triggering of backdoor mechanisms with minimal impact on original model performance. All experiments maintain consistent parameter settings to ensure result comparability. Furthermore, the study evaluates model robustness by simulating diverse attack methods such as fine-tuning and pruning attacks. Corresponding attack parameters are configured to assess the efficacy of digital watermarking technology.

4.4 Performance Evaluation

4.4.1 Performance evaluation of segmentation model after embedding watermark

After generating the trigger set using the proposed trigger set generation algorithm and ensuring it meets the specified criteria, the trigger set is blended proportionally with the original PASCAL VOC 2012 dataset to create a new dataset. Subsequently, this new

dataset is fed back into the deeplab-V3+ segmentation model for training. The MIOU serves as the experimental evaluation metric. During the mixing process of the original data and trigger set, the model's training difficulty varies with different mixing proportions, consequently affecting the final model performance. Therefore, experiments are conducted using various mixing proportions: 0%, 10%, 20%, 30%, 40%, and 50% trigger set proportions. Figure 3 illustrates the experimental outcomes.

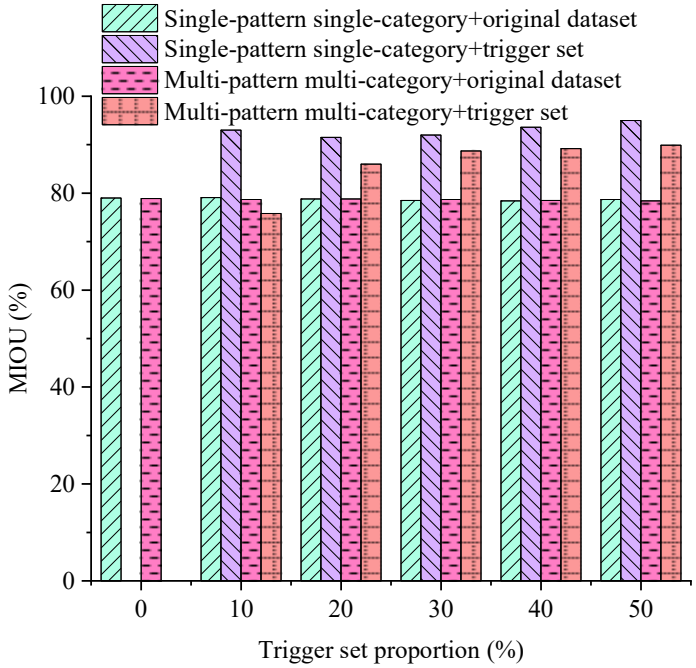

Fig. 3. Performance of training models with different mixed proportion trigger sets.

In Figure 3, the performance of the model across varying proportions of trigger set mixtures is analyzed. Two primary scenarios of trigger set compositions are investigated: single-pattern single-category and multi-pattern multi-category. Each scenario compares the model's performance when trained with the original dataset alone and when augmented with the trigger set. For the scenario where the trigger set proportion is 0 (no trigger set mixture), both the single-pattern single-category and multi-pattern multi-category models achieve similar performance on the original dataset, approximately 79% and 78.9% accuracy, respectively. However, upon introduction of the trigger set, distinct trends emerge. As the proportion of trigger set increases, the model's performance on the original dataset shows a slight decline from 79.1% to 78.7%, with fluctuations generally around 78.5%. Concurrently, there is a notable enhancement in the model's performance on the trigger set itself, rising from 93% to 95%. This underscores the trigger set's positive impact on the model's ability to handle specific data subsets. Similarly, in the multi-pattern multi-category scenario, while the performance on the original dataset experiences a slight decrease from 78.7% to 78.4% with increasing trigger set proportion, the improvement in trigger

set performance is more pronounced, escalating from 75.8% to 89.9%. This indicates substantial enhancement in the model's capability to recognize specific categories when augmented with additional trigger set data. In summary, regardless of the trigger set's pattern and category composition, integrating trigger set data during model training significantly enhances the model's performance on the trigger set itself. Although there is a marginal decline in performance on the original dataset, the overall impact remains minimal. This demonstrates that judiciously mixing trigger set data can effectively bolster the model's ability to discern specific data categories while maintaining satisfactory performance on general data.

To evaluate the model's generalization capability, experiments are conducted on three distinct public datasets: PASCAL VOC 2012, COCO, and Cityscapes. The performance of the model on these different datasets is presented in Table 1. The model demonstrates high performance across diverse datasets, indicating robust generalization ability.

Table 1: Performance of the Model on Different Datasets

Dataset	Precision (%)	Recall (%)	F1 score (%)
PASCAL VOC 2012	78.9	85.2	81.9
COCO	75.6	81.4	78.4
Cityscapes	80.2	86.3	83.2

4.4.2 Ablation experiment

To further validate the superiority of the trigger set generated by the trigger set generation algorithm compared to using the original training data directly as the trigger set, an ablation experiment is conducted at a mixing proportion of 35%. Figure 4 illustrates the effect of digital watermarking using a single-pattern single-category trigger set on model performance, comparing two types of trigger sets.

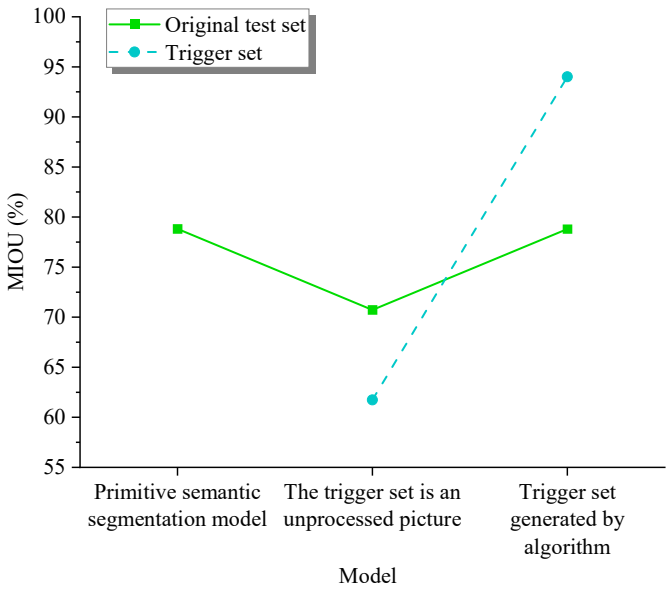

Fig. 4. Model performance comparison of the single-pattern single-category trigger set digital watermarking under two trigger sets.

In Figure 4, the efficacy of a single-pattern single-category trigger set digital watermark is assessed under two distinct trigger set conditions. Initially, the baseline performance of the original semantic segmentation model on the original test dataset is 78.82%, providing a reference for evaluating changes in model performance under varied conditions. When utilizing unprocessed images as the trigger set, the model's performance on the original test dataset decreases significantly to 70.72%, further declining to 61.74% on the trigger set itself. This indicates that using unprocessed images as the trigger set not only diminishes the model's performance on general data but also results in poor recognition on the specific trigger set, suggesting inadequate learning of trigger set characteristics. Conversely, employing algorithmically generated trigger sets yields a marginal reduction in model performance on the original test dataset to 78.80%, maintaining performance levels comparable to the original model. Notably, the model's performance on the trigger set substantially improves to 94.01%. This underscores the algorithmically generated trigger sets' capability to enhance recognition accuracy significantly on specific data subsets. In summary, compared to using unprocessed images, algorithmically generated trigger sets notably elevate the model's performance on specific trigger sets while minimally impacting its performance on the original test dataset. This highlights the effectiveness of the algorithmically generated trigger set digital watermark method in bolstering the model's capability to identify specific data categories while preserving its original performance, thereby crucially safeguarding the model's intellectual property.

Figure 5 examines the impact of digital watermarking with multiple patterns and categories on the model's performance under two trigger set conditions.

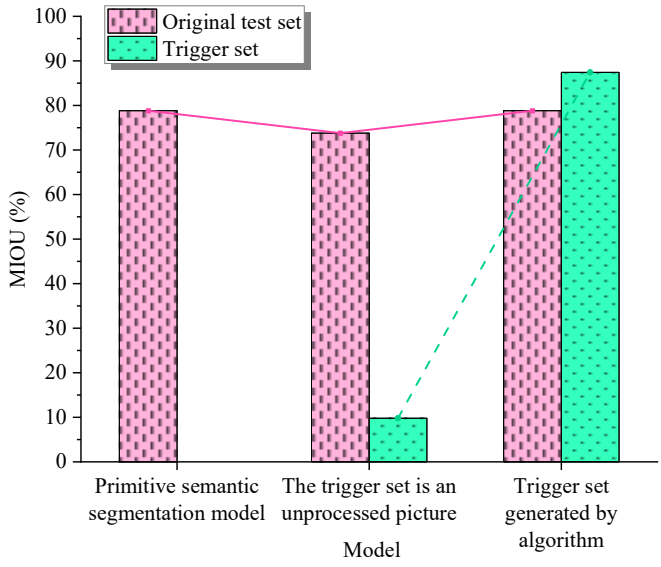

Fig. 5. Model performance comparison of the multi-pattern multi-category trigger set digital watermarking in two trigger sets.

In Figure 5, the efficacy of a multi-pattern multi-category trigger set digital watermark is evaluated under two distinct trigger set conditions. Initially, the original semantic segmentation model achieves a performance of 78.82% on the original test dataset, serving as the baseline for performance comparison. When utilizing unprocessed images as the trigger set, the model's performance on the original test dataset decreases slightly to 73.79%, while its performance on the trigger set itself drastically declines to 9.78%. This significant performance decrease indicates that using unprocessed images as the trigger set results in poor recognition on the specific dataset, highlighting the model's inability to effectively identify trigger set characteristics. Conversely, employing algorithmically generated trigger sets results in only a minor decrease in model performance on the original test dataset to 78.81%, maintaining performance levels comparable to the original model. Notably, the model's performance on the trigger set improves significantly to 87.42%. This underscores the algorithmically generated trigger sets' capacity to enhance the model's ability to recognize specific data categories. In summary, algorithmically generated multi-pattern multi-category trigger set digital watermarking preserves the model's performance on general datasets while substantially enhancing its recognition accuracy on specific trigger sets. This approach effectively strengthens the model's intellectual property protection while ensuring high performance across diverse input scenarios.

4.4.3 Experiment against fine-tuning attack

To verify the robustness of the digital watermark, the original training dataset without mixing the trigger set is employed to conduct four types of fine-tuning attacks (FTLL, FTAL, RTLL, and RTAL) on the segmentation model embedded with the digital watermark. The fine-tuning duration is set to one-third of the original training period. Figures 6

and 7 illustrate the model and watermark performance following the four fine-tuning attacks on the segmentation model using single-pattern single-category and multi-pattern multi-category trigger set digital watermarking, respectively.

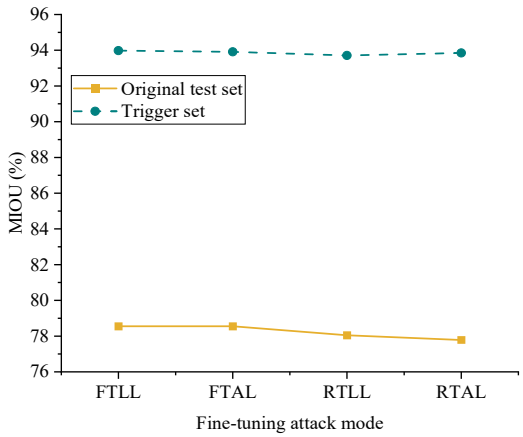

Fig. 6. Segmentation model and watermark performance of single-pattern single-category trigger set digital watermark after fine-tuning attack.

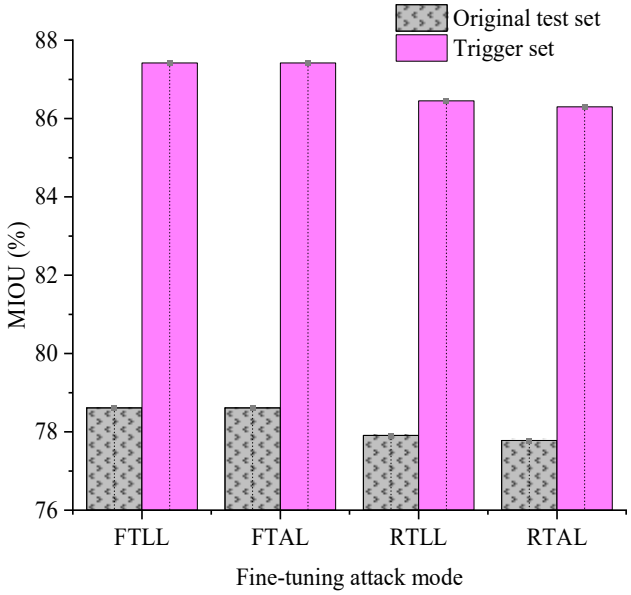

Fig. 7. Segmentation model and watermark performance of multi-pattern multi-category trigger set digital watermark after fine-tuning attack.

Figure 6 evaluates the performance of the single-pattern single-category trigger set digital watermarking segmentation model following fine-tuning attacks. Fine-tuning attacks are common methods aimed at modifying a model's behavior by retraining some or all network layers, potentially impacting model ownership verification or attempts to remove digital watermarks. Across four distinct fine-tuning attack modes, the model's performance on the original test set shows slight fluctuations while remaining stable, achieving performance metrics ranging from 77.78% to 78.55%. This indicates that the model maintains relatively consistent performance on general datasets despite fine-tuning attacks.Meanwhile, on the trigger set, the model demonstrates robust performance, with MIOU values ranging from 93.71% to 93.98%. This underscores the model's capability to retain high accuracy in recognizing specific trigger sets even after undergoing fine-tuning attacks. These findings highlight the strong resilience of digital watermarks against common fine-tuning attack methods when applied to trigger sets. In summary, the employed digital watermarking technique effectively preserves model intellectual property integrity following fine-tuning attacks. It ensures the model can accurately identify predetermined features specified by the model owner under specific conditions, thereby safeguarding the model owner's rights.

Figure 7 assesses the performance of the multi-pattern multi-category trigger set digital watermarking segmentation model following various fine-tuning attack methods. These attacks, including Fine-tuning Last Layer (FTLL), Fine-tuning All Layers (FTAL), Re-training Last Layer (RTLL), and Re-training All Layers (RTAL), aim to modify the model's behavior by adjusting specific network layers. Under FTLL and FTAL attack modes, the model maintains consistent performance on the original test set, achieving 78.61%, indicating stable handling of general data under these conditions. On the trigger set, there is a slight performance decrease, but it remains robust at 87.42%, demonstrating the digital watermark's resilience under these attack scenarios. However, under RTLL attack mode, the model's performance on the original test set slightly decreases to 77.91%, with a more noticeable decline to 86.45% on the trigger set. The RTAL attack mode further reduces the model's performance on the original test set to 77.78%, with the trigger set performance decreasing to 86.30%. These findings suggest that RTLL and RTAL attacks exert a more significant impact on model performance, particularly on the trigger set, thereby compromising the robustness of the digital watermark to some extent. In summary, the multi-pattern multi-category trigger set digital watermarking demonstrates high stability under FTLL and FTAL attacks. However, there is a notable decline in model performance on both the original test set and trigger set under RTLL and RTAL attacks. This underscores the vulnerability of these attack methods to model ownership verification and digital watermark protection. Nevertheless, the digital watermark maintains effective protection of model intellectual property, with the performance on the trigger set remaining relatively strong.

Assuming the attacker possesses an equivalent trigger set generated in a similar manner, Figure 8 illustrates the performance of the single-pattern single-category trigger set digital watermarking segmentation model on the original test set, the original trigger set, and the attacker's new trigger set after RTLL and RTAL fine-tuning attacks. Figure 9 presents the performance of the multi-pattern multi-category trigger set digital watermarking segmentation model across these three datasets.

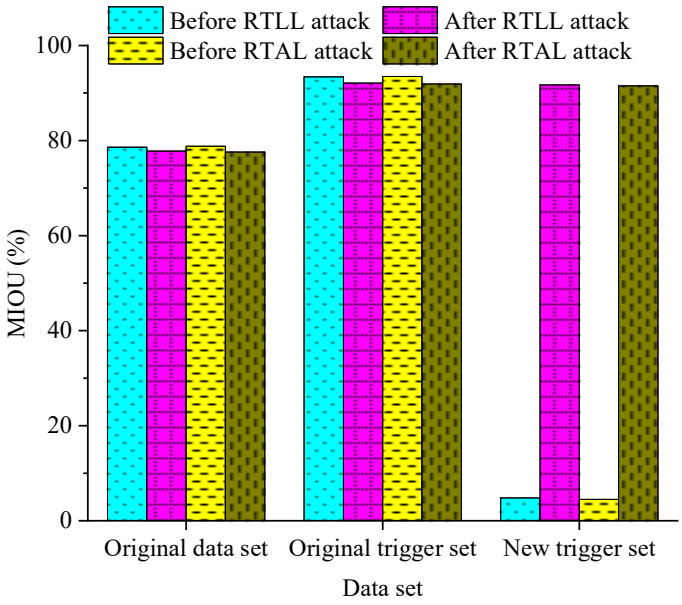

Fig. 8. Performance of single-pattern single-category model on three datasets.

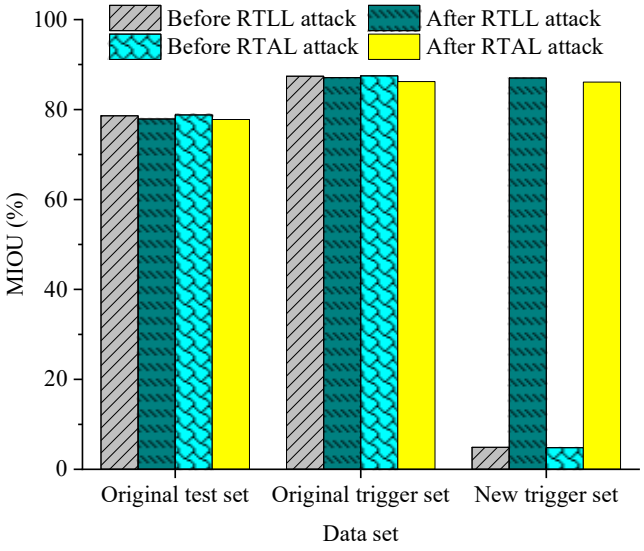

Fig. 9. Performance of the multi-pattern multi-category model on three datasets.

Figure 8 examines the performance of a single-pattern single-category model across three datasets before and after RTLL attacks. Prior to the attacks, the model achieves 78.6%

accuracy on the original dataset and 93.4% on the original trigger set, while its performance on the new trigger set is notably low at 4.8%. This indicates robust recognition capabilities for the original data and trigger set, but poor recognition of the new trigger set without RTLL attacks. Following RTLL attacks, the model's performance slightly decreases to 77.8% on the original dataset and 92.1% on the original trigger set. Notably, performance on the new trigger set significantly improves to 91.7%. This enhancement suggests that RTLL attacks markedly improve the model's ability to recognize the new trigger set, nearly matching its performance on the original trigger set. Similar trends are observed in the analysis before and after RTAL attacks. Prior to RTAL attacks, the model achieves 78.8% accuracy on the original dataset, 93.5% on the original trigger set, and 4.5% on the new trigger set. After RTAL attacks, performance decreases to 77.6% on the original dataset and 91.9% on the original trigger set, while improving to 91.5% on the new trigger set. Overall, while RTLL and RTAL attacks affect the model's performance on the original dataset and trigger set to some extent, the key observation is the substantial improvement in recognizing the new trigger set under these attacks. This highlights that attackers using such fine-tuning methods can effectively influence the model's responses to specific trigger set data, potentially compromising model security and intellectual property protection. Thus, these findings underscore the importance of developing robust protection strategies for models facing fine-tuning attacks.

Figure 9 examines the performance of a multi-pattern multi-category model across three datasets before and after RTLL attacks. Prior to the RTLL attack, the model achieves 78.6% accuracy on the original test set, 87.4% on the original trigger set, and a mere 4.9% on the new trigger set. This indicates strong recognition capabilities for the original data and trigger sets, but significant weakness in recognizing the new trigger set without RTLL attacks. Following the RTLL attack, the model's performance slightly decreases to 77.91% on the original test set and 87.1% on the original trigger set. Notably, performance on the new trigger set improves significantly to 87%. This enhancement suggests that the RTLL attack substantially improves the model's ability to recognize the new trigger set, nearly matching its performance on the original trigger set. Similar trends are observed in RTAL attack analysis. Before RTAL attacks, the model achieves 78.8% accuracy on the original test set, 87.5% on the original trigger set, and only 4.8% on the new trigger set. After RTAL attacks, the model's performance decreases slightly to 77.78% on the original test set and 86.21% on the original trigger set, while improving to 86.1% on the new trigger set. In summary, RTLL and RTAL attacks have some impact on the model's performance on the original test and trigger sets, resulting in minor performance decreases. However, these attacks significantly enhance the model's ability to recognize the new trigger set, elevating recognition accuracy from negligible to a level comparable to the original trigger set. These findings suggest that attackers can manipulate the model's behavior through fine-tuning attacks to respond effectively to trigger sets that are initially unrecognized, posing potential threats to model security and intellectual property protection. Therefore, further research and reinforcement of model security measures are crucial to mitigate such attacks.

To assess the model's robustness against various attack methods, experiments simulated multiple attack scenarios including model replacement attacks, noise injection attacks, and model reverse engineering attacks. Table 2 shows the performance variations of the model under these different attacks. Despite experiencing some performance decline when

subjected to attacks, the model demonstrates relatively minor decreases, indicating a degree of robustness. Furthermore, the model's accuracy, recall rate, and F1 score exhibit fluctuations within 1% during extended operation, demonstrating its strong stability.

Table 2: Performance Variations of the Model Under Various Attacks

Attack type	Change in precision (%)	Change in recall (%)	Change in F1 score (%)
Model replacement attack	-2.5	-3.1	-2.8
Noise injection attack	-1.9	-2.4	-2.1
Model reverse engineering attack	-4.7	-5.3	-5.0

4.4.4 Experiment against pruning attack

To comprehensively evaluate the robustness of the digital watermark, this study extends beyond fine-tuning attacks to include pruning attacks on the semantic segmentation model containing the digital watermark. Pruning attacks aim to disrupt the watermark by selectively removing weights or neurons from the model while attempting to maintain the original model performance. Two pruning strategies are selected for evaluation: weight pruning and neuron pruning, applied respectively to segmentation models equipped with single-pattern single-category and multi-pattern multi-category trigger set digital watermarks. Weight pruning involves setting a threshold and removing connections with weights below this threshold. The study conducts weight pruning at various proportions, including 10%, 20%, 30%, 40%, and 50% pruning ratios. After pruning, the model's performance on both the original test set and the trigger set is assessed. Figure 10 illustrates the performance evaluation outcomes of the model under the pruning attack.

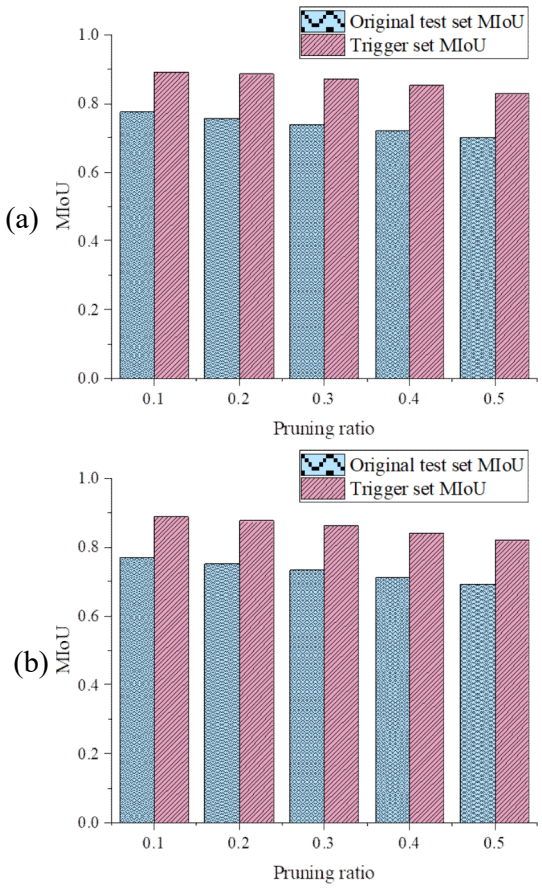

Fig. 10. Model performance evaluation results under pruning attack(a): Weight pruning; (b): Neuron pruning.

Figure 10 analyzes the evaluation results of the model after pruning attacks, including weight pruning and neuron pruning. Pruning attacks aim to disrupt digital watermarks by removing some weights or neurons from the model while attempting to maintain its original performance as much as possible. For weight pruning, at a pruning ratio of 0.1, the model's MIOU decreases slightly from 0.775 on the original test set to 0.892, and from 0.892 to 0.885 on the trigger set. The data indicates that even at a 10% pruning ratio, the model's performance drop is not significant. As the pruning ratio increases, the performance on both the original test set and the trigger set gradually decreases. When the pruning ratio reaches 0.5, the MIOU on the original test set drops to 0.701 and to 0.828 on the trigger set. Despite the performance decrease, the model's performance on the trigger set remains relatively high even at higher pruning ratios. Similarly, neuron pruning results mirror weight pruning. At a 10% pruning ratio, the MIOU on the original test set decreases from 0.775 to 0.768, and from 0.892 to 0.889 on the trigger set. When the pruning ratio increases to 50%, the MIOU drops to 0.692 on the original test set and to 0.819 on the trigger set. Compared to weight pruning, neuron pruning has a slightly larger impact on

the model's performance, but even at high pruning ratios, the MIOU on the trigger set remains relatively high.

Overall, regardless of whether it is weight pruning or neuron pruning, as the pruning ratio increases, the model's performance on the original test set and trigger set shows a decreasing trend. However, even at higher pruning ratios, the model's performance on the trigger set remains relatively high, demonstrating the robustness of the digital watermark. This suggests that the adopted digital watermarking technique effectively withstands pruning attacks, thereby protecting the model's IP. Nevertheless, pruning attacks still have some impact on model performance, necessitating a balance between pruning ratio and model security in practical applications.

4.5 Discussion

This study investigates the implications of AI-driven creativity on intellectual property protection through the proposition of a methodology employing DL models. Firstly, it examines the impact of AI creativity on copyright and patent systems, particularly within the context of DL model applications. Subsequently, the study introduces IP protection techniques for image semantic segmentation and explores backdoor mechanisms in DL models, specifically applied to image semantic segmentation models. Research findings indicate that embedding digital watermarks into datasets effectively safeguards model IP with minimal performance degradation, while demonstrating resilience against fine-tuning and pruning attacks. Related academic literature underscores the focus on IP protection, DL model security, and applications of digital watermarking technology. In the realm of IP protection, scholars have addressed strategies for safeguarding innovative achievements amidst rapid advancements in AI and machine learning technologies. For instance, Bamakan et al. (2022) discussed the adaptability and challenges of IP laws in digital environments, emphasizing strategies within legal frameworks to accommodate new technologies [40]. Similarly, they highlighted the protection of AI-created works under copyright and patent laws. In terms of DL model security, extensive research examined vulnerabilities and protective measures. Liu et al. (2021) introduced adversarial examples, illustrating how slight modifications could lead DL models to misclassify inputs, thereby exposing vulnerabilities to unknown inputs [41]. Their insights on adversarial examples significantly informed this study's exploration of backdoor mechanisms and the deployment of digital watermarking technologies, particularly in designing protective strategies considering model vulnerabilities and attack vectors. Furthermore, regarding digital watermarking technology applications, numerous studies focused on its implementation to authenticate and protect the ownership of digital content. Jebreel et al. (2021), for example, researched embedding digital watermarks in DL models to mitigate issues related to model misuse and unauthorized usage [42]. They underscored the potential and efficacy of digital watermarking technology in safeguarding intellectual property, akin to the method proposed for digital watermarking in image semantic segmentation models. Building upon prior scholarly works, this study contributes significantly in several dimensions. Firstly, it integrates the backdoor mechanism with digital watermarking technology to propose an innovative method for protecting intellectual property in image semantic segmentation models. By embedding specific algorithms for trigger set generation into datasets, the

study achieves digital watermark embedding in models, ensuring traceability to the original owner even when utilized without authorization. This approach not only adheres to legal requirements for intellectual property protection but also addresses the security and reliability needs of DL models in practical applications. Furthermore, the study validates the efficacy and resilience of the proposed method through comprehensive experiments. Experimental results demonstrate that semantic segmentation models embedded with digital watermarks maintain stable performance against fine-tuning and pruning attacks, effectively resisting unauthorized dataset usage while retaining high accuracy in recognizing the original dataset. These empirical findings not only validate the method's technical feasibility but also provide empirical evidence for future research endeavors. Nevertheless, the study identifies potential areas for enhancement and limitations. While the proposed method performs well under specific experimental conditions, its scalability and robustness across large-scale and diverse datasets warrant further validation. Additionally, achieving a balance between the impact of digital watermarking on model performance and its protective benefits constitutes a crucial avenue for future research. By integrating research on DL model security with the application of digital watermarking technology, this study introduces an innovative method for protecting intellectual property. It contributes new insights and empirical evidence to the realm of intellectual property protection for AI creations. Future research could expand upon this method's applicability to other types of DL models and diverse application scenarios, thereby enhancing its efficacy and reliability in practical settings.

5. CONCLUSION

5.1 Research Contribution

To explore methods for protecting intellectual property related to AI creations, this study examines the impact of AI innovations on IP and analyzes IP protection strategies for image semantic segmentation algorithms. Furthermore, it utilizes the backdoor mechanism and adversarial example generation to develop an algorithm for generating trigger sets in image semantic segmentation. Comparative experiments validate the efficacy of this algorithm, yielding the following conclusions: (1) Integrating the trigger set into the original test dataset enhances the model's performance. Specifically, the MIOU value of the model with trigger sets comprising single patterns and single categories reaches up to 95%. (2) Regardless of the mode of the trigger set's digital watermark, the algorithmically generated trigger sets minimally impact the performance of the original model. Moreover, their MIOU values remain high on the trigger set, at 94.01% and 87.42% respectively, indicating a strong association between the semantic segmentation model and its owner. (3) In fine-tuning attack experiments, the digital watermark embedded in the segmentation model by the trigger set generation algorithm exhibits stable robustness. The MIOU value of the model shows negligible impact post-attack, maintaining performance levels comparable to those before the attack.

5.2 Future Works and Research Limitations

However, the study identifies research limitations. The generation phase of the trigger

set requires integration of information from the entire image, necessitating substantial effort and time. Future research could explore integrating characteristic information of the model owner into salient areas of the image for localized attacks, potentially optimizing the efficiency of trigger set generation.

FUNDING

This work was supported by Scientific research funding project of Chongqing University of Technology, China（Item No 2021ZDR003）.

REFERENCES

1. N. Selvadurai, R. Matulionyte, "Reconsidering creativity: copyright protection for works generated using artificial intelligence," *Journal of Intellectual Property Law & Practice*, vol. 15, no. 7, pp. 536-543,2020.
2. A. Morrish, "AI and Patents: Finding Harmony Between Protection of Intellectual Property Rights and Innovation," *Intellectual Property Journal*, vol. 33, no. 3, pp. 253-278, 2021.
3. A. Fkirin, G. Attiya, A. El-Sayed, et al. "Copyright protection of deep neural network models using digital watermarking: a comparative study," *Multimedia Tools and Applications*, vol. 81, no.11, pp. 15961-15975, 2022.
4. W. Chen, M. Jiang, W.-G. Zhang, and Z. Chen, "A novel graph convolutional feature based convolutional neural network for stock trend prediction," *Information Sciences,* vol. 556, pp. 67-94, 2021.
5. M. Hamdi, S. Bourouis, K. Rastislav, and F. Mohmed, "Evaluation of neuro images for the diagnosis of Alzheimer's disease using deep learning neural network," *Frontiers in Public Health*, vol. 10, pp. 834032, 2022.
6. M. Xue, Y. Zhang, J. Wang, and W. Liu, "Intellectual property protection for deep learning models: Taxonomy, methods, attacks, and evaluations," *IEEE Transactions on Artificial Intelligence,* vol. 3, no. 6, pp. 908-923, 2021.
7. W. Liang, W. Huang, J. Long, K. Zhang, K.-C. Li, and D. Zhang, "Deep reinforcement learning for resource protection and real-time detection in IoT environment," *IEEE Internet of Things Journal,* vol. 7, no. 7, pp. 6392-6401, 2020.
8. Z. Zhang, F. Wen, Z. Sun, X. Guo, T. He, and C. Lee, "Artificial intelligence-enabled sensing technologies in the 5G/internet of things era: from virtual reality/augmented reality to the digital twin," *Advanced Intelligent Systems,* vol. 4, no. 7, pp. 2100228, 2022.
9. H. Wu, G. Liu, Y. Yao, and X. Zhang, "Watermarking neural networks with watermarked images," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 31, no. 7, pp. 2591-2601, 2020.
10. B. Li, L. Fan, H. Gu, J. Li, and Q. Yang, "FedIPR: Ownership verification for federated deep neural network models," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 4, pp. 4521-4536, 2022.

11. C. Zuo, "Study on Deconstruction and Governance of Intellectual Property Rights Caused by Machine Writing Ethics Anomie in The Era of Artificial Intelligence," *Academic Journal of Science and Technology,* vol. 9, no. 1, pp. 33-37, 2024.
12. M. H. Amiri, N. Mehrabi Hashjin, M. Montazeri, S. Mirjalili, and N. Khodadadi, "Hippopotamus optimization algorithm: a novel nature-inspired optimization algorithm," *Scientific Reports,* vol. 14, no. 1, pp. 5032, 2024.
13. N. Mehrabi Hashjin, M. H. Amiri, A. Mohammadzadeh, S. Mirjalili, and N. Khodadadi, "Novel hybrid classifier based on fuzzy type-III decision maker and ensemble deep learning model and improved chaos game optimization," *Cluster Computing*, pp. 1-38, 2024.
14. R. W. Gregory, O. Henfridsson, E. Kaganer, and H. Kyriakou, "The role of artificial intelligence and data network effects for creating user value," *Academy of management review,* vol. 46, no. 3, pp. 534-551, 2021.
15. S. Chesterman, "Artificial intelligence and the limits of legal personality," *International & Comparative Law Quarterly,* vol. 69, no. 4, pp. 819-844, 2020.
16. D. Gervais, "Is intellectual property law ready for artificial intelligence?," 2, Oxford University Press, 2020, pp. 117-118.
17. H. Haick, and N. Tang, "Artificial intelligence in medical sensors for clinical decisions," *ACS nano,* vol. 15, no. 3, pp. 3557-3567, 2021.
18. H. Benbya, S. Pachidi, and S. Jarvenpaa, "Special issue editorial: Artificial intelligence in organizations: Implications for information systems research," *Journal of the Association for Information Systems,* vol. 22, no. 2, pp. 10, 2021.
19. J. K. Eshraghian, "Human ownership of artificial creativity," *Nature Machine Intelligence,* vol. 2, no. 3, pp. 157-160, 2020.
20. J. Zhang, D. Chen, J. Liao, W. Zhang, H. Feng, G. Hua, and N. Yu, "Deep model intellectual property protection via deep watermarking," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 44, no. 8, pp. 4005-4020, 2021.
21. A. Asatiani, "Malo P Nagbøl PR Penttinen E Rinta-Kahila T Salovaara A Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems," *Journal of the Association for Information Systems,* vol. 22, no. 2, pp. 325, 2021.
22. E. Chikhaoui, and S. Mehar, "Artificial intelligence (AI) collides with patent law," *J. Legal Ethical & Regul. Isses,* vol. 23, pp. 1, 2020.
23. M. Lee, "An analysis of the effects of artificial intelligence on electric vehicle technology innovation using patent data," *World Patent Information,* vol. 63, pp. 102002, 2020.
24. C.-H. Yang, "How artificial intelligence technology affects productivity and employment: firm-level evidence from Taiwan," *Research Policy,* vol. 51, no. 6, pp. 104536, 2022.
25. G. Damioli, V. Van Roy, and D. Vertesy, "The impact of artificial intelligence on labor productivity," *Eurasian Business Review,* vol. 11, pp. 1-25, 2021.
26. Z. C. Seskir, and K. W. Willoughby, "Global innovation and competition in quantum technology, viewed through the lens of patents and artificial intelligence," *International Journal of Intellectual Property Management,* vol. 13, no. 1, pp. 40-61, 2023.

27. A. V. Giczy, N. A. Pairolero, and A. A. Toole, "Identifying artificial intelligence (AI) invention: A novel AI patent dataset," *The Journal of Technology Transfer,* vol. 47, no. 2, pp. 476-505, 2022.
28. X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 60, pp. 1-15, 2022.
29. S. Bai, Y. Li, Y. Zhou, Q. Li, and P. H. Torr, "Adversarial metric attack and defense for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 43, no. 6, pp. 2119-2126, 2020.
30. Z. Yi, J. Yu, Y. Tan, and Q. Wu, "Fine-tuning more stable neural text classifiers for defending word level adversarial attacks," *Applied Intelligence,* vol. 52, no. 10, pp. 11948-11965, 2022.
31. C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Medical image analysis,* vol. 67, pp. 101813, 2021.
32. F. Ye, and J. Yang, "A deep neural network model for speaker identification," *Applied Sciences,* vol. 11, no. 8, pp. 3603, 2021.
33. Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Anti-backdoor learning: Training clean models on poisoned data," *Advances in Neural Information Processing Systems,* vol. 34, pp. 14900-14912, 2021.
34. S. Wang, S. Nepal, C. Rudolph, M. Grobler, S. Chen, and T. Chen, "Backdoor attacks against transfer learning with pre-trained deep learning models," *IEEE Transactions on Services Computing,* vol. 15, no. 3, pp. 1526-1539, 2020.
35. Y. Chong, X. Chen, Y. Tao, and S. Pan, "Erase then grow: Generating correct class activation maps for weakly-supervised semantic segmentation," *Neurocomputing,* vol. 453, pp. 97-108, 2021.
36. F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods,* vol. 18, no. 2, pp. 203-211, 2021.
37. Q. Zhou, Z. Feng, Q. Gu, G. Cheng, X. Lu, J. Shi, and L. Ma, "Uncertainty-aware consistency regularization for cross-domain semantic segmentation," *Computer Vision and Image Understanding,* vol. 221, pp. 103448, 2022.
38. G. Pruthi, F. Liu, S. Kale, and M. Sundararajan, "Estimating training data influence by tracing gradient descent," *Advances in Neural Information Processing Systems,* vol. 33, pp. 19920-19930, 2020.
39. H. Allioui, M. A. Mohammed, N. Benameur, B. Al-Khateeb, K. H. Abdulkareem, B. Garcia-Zapirain, R. Damaševičius, and R. Maskeliūnas, "A multi-agent deep reinforcement learning approach for enhancement of COVID-19 CT image segmentation," *Journal of personalized medicine,* vol. 12, no. 2, pp. 309, 2022.
40. S. M. H. Bamakan, N. Nezhadsistani, O. Bodaghi, and Q. Qu, "Patents and intellectual property assets as non-fungible tokens; key technologies and challenges," *Scientific Reports,* vol. 12, no. 1, pp. 2178, 2022.
41. N. Liu, M. Du, R. Guo, H. Liu, and X. Hu, "Adversarial attacks and defenses: An interpretation perspective," *ACM SIGKDD Explorations Newsletter,* vol. 23, no. 1, pp. 86-99, 2021.

42.N. M. Jebreel, J. Domingo-Ferrer, D. Sánchez, and A. Blanco-Justicia, “Keynet: An asymmetric key-style framework for watermarking deep learning models,” *Applied Sciences,* vol. 11, no. 3, pp. 999, 2021.