DUNK: DUAL-SPACE NON-SEMANTIC KNOWLEDGE CONSOLIDATION FOR CONTINUAL IMAGE FORGERY LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image Forgery Localization (IFL) aims to locate manipulated regions at the pixel level. However, with the rapid evolution of generative artificial intelligence and image editing tools, newly emerging manipulation patterns continuously cause severe distribution shifts, making existing static IFL methods suffer from catastrophic forgetting when adapted to streaming forgery data. Therefore, in this paper, we focus on a challenging practical task called Continual Image Forgery Localization (CIFL), which requires continuously adapting to new manipulation patterns while preserving fine-grained localization ability for historical ones. To this end, we propose a Dual-Space non-semantic Knowledge Consolidation (DUNK) framework to maintain and consolidate historical authentic-forgery discrepancy knowledge in both spatial and frequency spaces. Specifically, a Dual-space authentic-forgery Prototype Memory (DPM) module is designed to compactly preserve structured authentic and forged prototypes without storing massive raw samples. Then, a Dual-space Adaptive Prototype Replay (DAPR) module adaptively coordinates spatial and frequency prototype replay according to their task-specific reliability, thereby preserving complementary historical forgery cues. Finally, a Multi-Perspective Knowledge Distillation (MPKD) module constrains the consistency between old and new models from mask, boundary, and feature perspectives to alleviate the degradation of fine-grained localization ability. Extensive experiments on multiple forgery localization benchmarks demonstrate the superiority of our method against existing IFL and continual learning methods in both localization accuracy and anti-forgetting ability. **Our code will be released soon.**

1 INTRODUCTION

Image Forgery Localization (IFL) aims to localize manipulated regions in images at the pixel level. To this end, existing IFL methods (Kwon et al., 2022; Liu et al., 2022; Guillaro et al., 2023; Zhu et al., 2025c; Kong et al., 2025) have made remarkable progress based on massive pre-collected manipulated images for forgery traces modeling. However, with the rapid evolution of image editing (Brooks et al., 2023) and Artificial Intelligence Generative Content(AIGC) (Pan et al., 2023; Sun et al., 2024; Rombach et al., 2022), the constant adaptation to new forgery methods typically limit their general performance as the continuously derived unprecedented forgery patterns due to the well-known catastrophic forgetting challenge (De Lange et al., 2021; Masana et al., 2022). Therefore, in this paper, we focus on such a practical but challenging task called Continual Image Forgery Localization (CIFL), as shown in Figure 1(a), which aims to continuously adapt to new forged methods (*e.g.*, forged by PhotoShop (Dong et al., 2013), GAN (Novozamsky et al., 2020), and MLLM (Huang et al., 2025)) while retaining the localize ability for historical forged content.

To tackle the above issue, Continual Semantic Segmentation(CSS) (Cermelli et al., 2020; Zhu et al., 2025b; Douillard et al., 2021; Zhu et al., 2025a) becomes a feasible solution. It alleviates the catastrophic forgetting of old semantic segmentation knowledge caused by continuous adaptation to semantic drift derived from new images by balancing pixel-level semantic information. Unfortunately, the derived preservation of semantic knowledge is often useless or even harmful on the CIFL task, which intensifies the confusing between old and new forged content localization and brings catastrophic forgetting for knowledge of old forgery traces. Specifically, as shown in Figure 1(b), existing methods

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

inevitably preserve semantic knowledge (*e.g.* Swan features) entangled with forged traces. However, such knowledge typically mislead the capturing of new forged content (*e.g.*, the forged tiger) by confusing erroneous forged traces, leading a more significant catastrophic forgetting. Therefore, the core challenge of CIFL task involves in preserving the discrimination between authentic and manipulated regions, which is also called non-semantic knowledge due to its independence from specific categories. However, when it comes to the confused distribution of forgery traces in both spatial and frequency domains, it is extremely challenging to force the model to accumulate non-semantic discriminative knowledge between forged and real content which is not associated with any specific categories. On the one hand, the manipulation instinctively disrupt the local consistency in the spatial domains with texture inconsistency, boundary artifacts and unnatural transitions across regions (Xu et al., 2023; Dong et al., 2022; Wu et al., 2019). For the other, some imperceptible methods (*e.g.*, copy-move, i.e., splicing, copy-move) typically leave subtle frequency-domain artifacts with spectral perturbations and compression traces in the frequency domain (Kwon et al., 2022; Tan et al., 2024). Therefore, dynamically capturing and consolidating various types of forgery traces as well as continuously balancing multi-domain forged discriminative knowledge accumulated in previous IFL tasks remains an unsolved challenge.

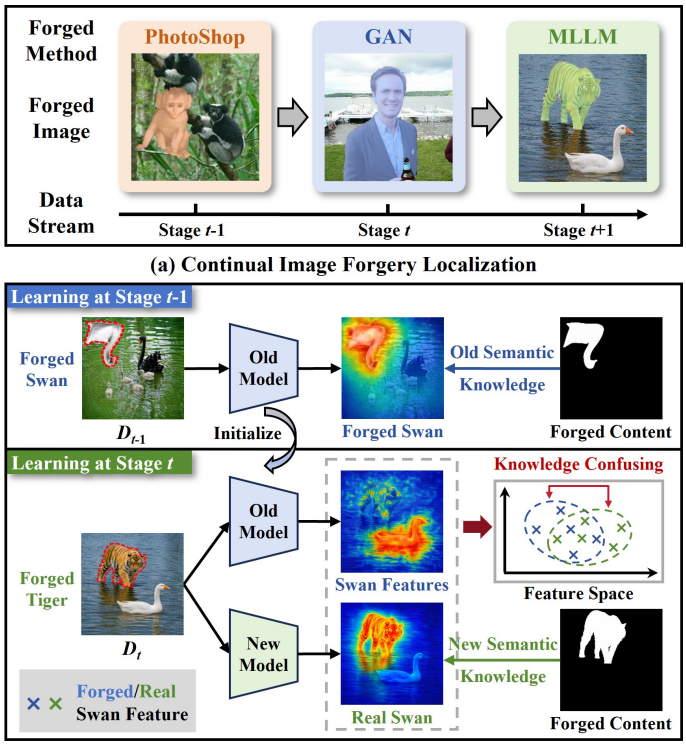


Figure 1: The definition and challenge of Continual Image Forgery Localization(CIFL) task. (a) CIFL focus on the newly forged methods are sequentially emerged. (b) When sequentially learning on the CIFL task, the entangled old semantic knowledge typically confusing the new forged content localization, which severely hinders the LReID performance. The experimental results are obtained from the latest IFL method (Kong et al., 2025).

To tackle the above issue, we propose a **Du**al-space **N**on-semantic **K**nowledge consolidation(DUNK) framework for CIFL task. In order to preserve the non-semantic discriminative knowledge of forgery traces in both spatial and frequency domains, we first construct a Dual-space authentic–forgery Prototype Memory (DPM) module, which extracts the forgery and real prototype representations of similar authentic and forgery patterns from historical tasks in both the spatial and frequency space, thus avoiding reliance on old data while robustly aggregating forgery traces. Besides, considering the inconsistencies in differential spatial structure and frequency statistics arising from various forgery methods, we introduce a Dual-space Adaptive Prototype Replay (DAPR) module, which enables the model to adaptively coordinate the replay strength of spatial-space and frequency-space prototypes according to the reliability of both spaces in the current task, thereby preserving the complementary discriminative information of historical manipulated patterns. Finally, to balance the multi-level distributed forgery traces, we further introduce a Multi-Perspective Knowledge Distillation (MPKD) module, which collaboratively constrains the consistency between the old and new models from region-level forgery response, boundary-level forgery delineation, and feature-level forgery abnormality, (namely mask, boundary, and feature) so as to alleviate the degradation of forgery localization ability during continual learning.

In summary, the main contributions of this paper are as follows: (1) A **Du**al-Space **N**on-Semantic **K**nowledge consolidation (DUNK) framework is proposed to solve the Continual Image Forgery Localization (CIFL) task, which requires continuously adapting to newly emerging forgery tasks while preserving historical non-semantic knowledge. (2) A Dual-space authentic-forgery Prototype Memory (DPM) module is proposed to compactly preserve structured authentic and forgery prototypes in both the spatial and frequency domains, without relying on massive raw sample replay. (3) A Dual-space Adaptive Prototype Replay (DAPR) module is proposed to adaptively coordinate the replay strength of spatial-space and frequency-space prototypes according to the reliability of the two spaces in the current task, thereby more effectively preserving complementary historical forgery cues under varying task characteristics. (4) A Multi-Perspective Knowledge Distillation (MPKD) module is proposed to collaboratively constrain the consistency between the old and new models from the perspectives of mask prediction, boundary delineation, and feature representation, boosting the capacity of balancing old and new non-semantic discriminative knowledge of forgery traces.

2 Related Work

2.1 Image Forgery Localization

Image Forgery Localization (IFL) aims to locate manipulated regions at the pixel level. Existing methods mainly exploit spatial or frequency-aware forensic cues. Spatial-dominant approaches (Hu et al., 2020; Sun et al., 2023; Li et al., 2024; Zhu et al., 2025c) capture texture, boundary, structural, and multi-scale inconsistencies in manipulated regions. Recent methods further improve feature modeling through adaptive architectures: AdaIFL (Li et al., 2024) dynamically activates expert groups for different forgery patterns, while Mesorch (Zhu et al., 2025c) combines CNNs and Transformers to integrate microscopic forensic traces with macroscopic context. Frequency-aware methods (Kwon et al., 2022; Han et al., 2025; Li et al., 2025; Nam et al., 2025; Wong et al., 2025) exploit spectral and compression artifacts complementary to spatial cues. In particular, M2SFormer (Nam et al., 2025) jointly models multi-frequency and multi-scale information, while ADCD-Net (Wong et al., 2025) adaptively leverages RGB and DCT features for document forgery localization. Despite these advances, most IFL methods follow a static training-then-inference paradigm. As forgery techniques (Rombach et al., 2022; Pan et al., 2023; Ramesh et al., 2022) continuously evolve, adapting to new manipulation patterns may cause catastrophic forgetting of previously learned ones, motivating Continual Image Forgery Localization.

2.2 Continual Semantic Segmentation

Continual Semantic Segmentation (CSS) (Cermelli et al., 2020; Zhu et al., 2025a; Douillard et al., 2021; Cha et al., 2021; Zhang et al., 2022; Yin et al., 2025; Lin et al., 2025) shares a similar dense prediction nature with Continual Image Forgery Localization (CIFL), as both require pixel-level prediction. A key challenge in CSS is background shift (Cermelli et al., 2020), where previously learned classes may be absorbed into the background label during incremental training, aggravating catastrophic forgetting. Distillation-based methods such as MiB (Cermelli et al., 2020) and PLOP (Douillard et al., 2021) preserve historical predictions through background modeling and multi-scale distillation. Replay-based methods (Jiang et al., 2025; Maracani et al., 2021; Zhu et al., 2023) revisit historical samples, while recent approaches (Chen et al., 2025; Zhu et al., 2025a;b; Chen et al., 2023) adopt compact prototype replay to reduce storage cost and representation drift. Despite their effectiveness, these methods primarily preserve semantic knowledge, whereas CIFL requires maintaining non-semantic authentic–forgery discrepancies across both spatial and frequency spaces.

Different from these methods, we propose a Dual-Space Non-semantic Knowledge consolidation(DUNK) framework that preserves authentic and forgery prototypes in both spatial and frequency spaces for non-semantic knowledge consolidation. Moreover, an adaptive prototype replay strategy and a multi-perspective knowledge distillation module is introduced to further balance the differential content within spatial and frequency domains according to their task-specific reliability, eliminating the catastrophic forgetting of non-semantic knowledge.

3 METHOD

3.1 PRELIMINARY

This paper focuses on the Continual Image Forgery Localization (CIFL) task, which aims to continuously learn fine-grained forgery localization from a sequence of datasets. Formally, we consider a task sequence $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T\}$, where each dataset $\mathcal{D}_t$ is produced by different image manipulation methods. When sequentially learning on the current (t-th) dataset $\mathcal{D}_t$, the previous $t-1$-th datasets are unavailable. After each training stage t, the model is evaluated on all seen tasks $\{D_1, \ldots, D_t\}$.

3.2 OVERVIEW OF DUNK

The pipeline of our proposed DUNK framework is shown in figure 2, which consists of a Dual-space authentic-forgery Prototype Memory (DPM) module, a Dual-space Adaptive Prototype Replay (DAPR) module, and a Multi-Perspective Knowledge Distillation (MPKD) module.

3.3 BASELINE METHOD

The baseline method of our DUNK is derived from the traditional IFL method PIM (Kong et al., 2025). Given a batch $(x_i, y_i)_{i=1}^{B}$ sampled from the current task $\mathcal{D}_t$, where x_i is the input image and y_i is the corresponding manipulation mask. Based on y_i, we further extract its manipulated boundary y_i^b for continual training. At stage t, the backbone $\phi^t(\cdot)$ extracts the spatial representation $z_i^t = \phi^t(x_i)$. Then, a prediction head $\varphi^t(\cdot)$ is employed to generate the manipulation map M_i^t with the boundary map B_i^t, as well as the reconstructed image R_i^t, respectively. Following PIM, the baseline is optimized with the following multi-task objective:

$$\mathcal{L}_{base} = \mathcal{L}_M + \lambda_B \mathcal{L}_B + \lambda_C \mathcal{L}_C + \lambda_R \mathcal{L}_R, \tag{1}$$

where $\mathcal{L}_M$ and $\mathcal{L}_B$ supervise manipulation region and boundary prediction, $\mathcal{L}_C$ encourages discriminative forgery feature aggregation, and $\mathcal{L}_R$ facilitate the global statistic inconsistency through reconstruction. Despite som progress on the IFL task, it typically suffers from catastrophic forgetting problem of non-semantic knowledge when continuously learning on new data with adaptation to newly emerging manipulation patterns. However, optimizing L_{base} on the current task alone overwrites historical authentic–forgery discrepancies. DUNK therefore consolidates such knowledge through prototype memory, adaptive replay, and fine-grained localization distillation.

3.4 DUAL-SPACE AUTHENTIC-FORGERY PROTOTYPE MEMORY

Directly storing historical images to preserve discrepancy knowledge is undesirable due to the increasing storage cost and potential privacy concerns. Therefore, we propose a **Dual-space Authentic-Forgery Prototype Memory (DPM)** module to compactly preserve the authentic-forgery discrepancies learned from each task. Considering forgery traces exhibit complementary characteristics in spatial and frequency domains, DPM summarizes the authentic and forged distributions into compact prototype statistics for subsequent knowledge consolidation.

Spatial Prototype Preservation (SPP). Given the spatial representation $z_i^t = \phi^t(x_i)$, we first construct spatial prototypes to preserve the authentic-forgery discrepancies in the learned feature space. Using the manipulation mask y_i, we separate the spatial representation z_i^t into forged and authentic features, where $z_i^t(p)$ denotes the feature at spatial location p. Since different manipulation regions may exhibit diverse forgery patterns, the forged features are further grouped by K-Means, while the authentic features are summarized as a reference cluster. For each spatial cluster $A_k^{S,t}$, its prototype center $\mu_k^{S,t}$ and diagonal variance $\sigma_k^{S,t}$ are computed as

$$\mu_k^{S,t} = \frac{1}{|A_k^{S,t}|} \sum_{(i,p)\in A_k^{S,t}} z_i^t(p), \qquad \sigma_k^{S,t} = \frac{1}{|A_k^{S,t}|} \sum_{(i,p)\in A_k^{S,t}} \left(z_i^t(p) - \mu_k^{S,t}\right)^2. \tag{2}$$

In this way, the spatial statistics compactly characterize different authentic-forgery feature distributions learned from the current task.

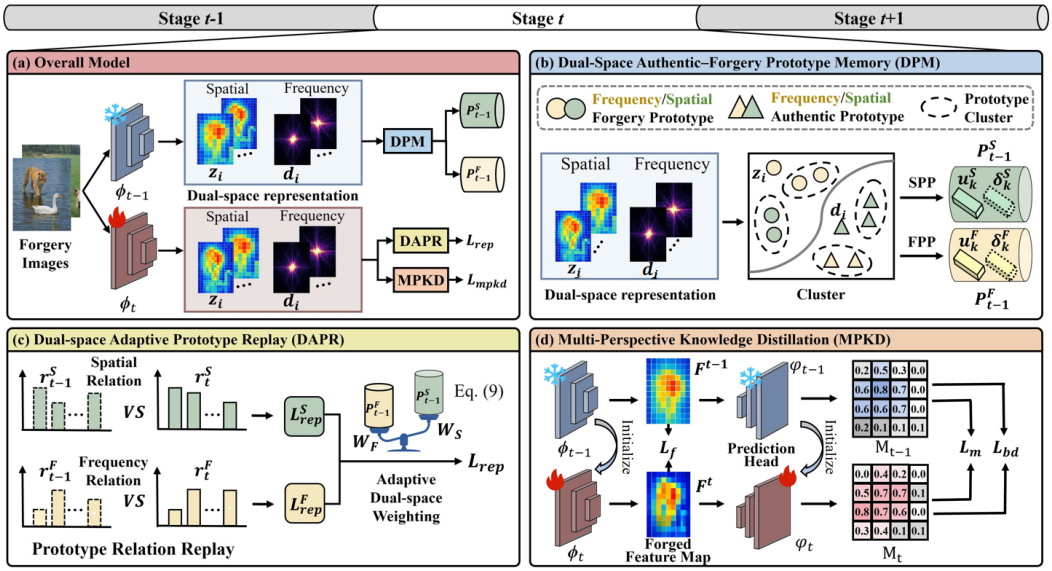


Figure 2: Overview of proposed Dual-Space Non-Semantic Knowledge consolidation (DUNK). The Dual-Space Authentic–Forgery Prototype Memory (DPM) module stores authentic/forgery prototypes in both spatial and frequency spaces after training on each stage. Then, the Dual-space Adaptive Prototype Replay (DAPR) module adaptively replays historical prototype relations. Finally, the Multi-Perspective Knowledge Distillation (MPKD) module exploits the mask, boundary, and global feature to distill multi-grained forgery discrimination to balance non-semantic knowledge.

Frequency Prototype Preservation (FPP). While spatial prototypes characterize discrepancies in the spatial space, some subtle manipulation traces are more evident in the frequency domain. Therefore, we further construct complementary frequency prototypes from the same current-task samples. Given a region mask y_i, its Fourier-spectrum descriptor is defined as

$$d_i^t = \log\left(1 + \left|\mathcal{F}\big(g(x_i) \odot y_i\big)\right|\right), \tag{3}$$

where $g(\cdot)$ denotes grayscale conversion and $\mathcal{F}(\cdot)$ denotes the 2D Fourier transform. Similar to the spatial representation, forged frequency descriptors are clustered to capture diverse spectral patterns, while authentic descriptors are summarized as a reference cluster. For each frequency cluster $A_k^{F,t}$, we obtain

$$\mu_k^{F,t} = \frac{1}{|A_k^{F,t}|} \sum_{i \in A_k^{F,t}} d_i^t, \qquad \sigma_k^{F,t} = \frac{1}{|A_k^{F,t}|} \sum_{i \in A_k^{F,t}} (d_i^t - \mu_k^{F,t})^2. \tag{4}$$

Accordingly, the spatial and frequency distributions of task $\mathcal{D}_t$ are compactly represented as

$$P_t^q = \left\{(\mu_k^{q,t}, \sigma_k^{q,t})\right\}_{k=1}^{K_q}, \qquad q \in \{S, F\}, \tag{5}$$

where S and F denote the spatial and frequency spaces, respectively, and K_q denotes the number of prototypes retained in the corresponding space. After training on the t-th stage, the resulting dual-space prototypes are added to the memory bank and serve as historical references for subsequent knowledge consolidation.

3.5 Dual-Space Adaptive Prototype Replay

Although DPM compactly stores historical authentic-forgery discrepancies, these prototypes need to be effectively exploited during new-task learning. Therefore, we propose a **Dual-space Adaptive Prototype Replay (DAPR)** module to preserve historical authentic-forgery discrepancy knowledge by maintaining the prototype-relation structures in both spatial and frequency spaces.

At stage t, we freeze the old model and measure the relations between forged representations and the historical prototypes preserves by DPM. For space $q \in \{S, F\}$, we uniformly denote the forged

representation as $v_n^{q,t} \in \{z_i^t(p), d_i^t\}$, and $(\mu_k^q, \sigma_k^q) \in \bigcup_{s<t} P_s^q$ denotes the k-th historical prototype. Since different manipulation patterns induce varying spatial structures and frequency statistics, forgery representations exhibit unequal variation across dimensions. As Euclidean distance weights all dimensions equally, it may be dominated by unstable ones. Therefore, we exploit the prototype variance and measure the distance $d_{n,k}^{q,t}$ using the diagonal Mahalanobis distance:

$$d_{n,K}^{q,t} = \frac{1}{C_q} \sum_{c=1}^{C_q} \frac{(v_{n,c}^{q,t} - \mu_{k,c}^q)^2}{\sigma_{k,c}^q}. \tag{6}$$

Based on the top-K nearest historical prototypes selected by the previous model, the relation distributions of the previous and current models are defined as

$$r_n^{q,t-1} = \rho(-d_{n,K}^{q,t-1}/\tau), \qquad r_n^{q,t} = \rho(-d_{n,K}^{q,t}/\tau), \tag{7}$$

where $\rho(\cdot)$ denotes softmax and τ is the temperature. The replay loss in each space is then formulated as

$$\mathcal{L}_{\text{rep}}^q = \frac{1}{N_q} \sum_{n=1}^{N_q} \text{KL}\left(r_n^{q,t-1} \| r_n^{q,t}\right), \qquad q \in \{S, F\}. \tag{8}$$

Since the reliability of spatial and frequency cues varies across manipulation types, we adaptively combine their replay losses as

$$\mathcal{L}_{\text{rep}} = w_S \mathcal{L}_{\text{rep}}^S + w_F \mathcal{L}_{\text{rep}}^F, \qquad w_q = \frac{R_q}{R_S + R_F}, \tag{9}$$

where

$$R_q = \frac{N_q}{B} \exp(-H_q), \qquad q \in \{S, F\}. \tag{10}$$

Here, H_q denotes the average normalized entropy of the historical relation distributions, N_q denotes the number of valid replay representations, and B is the batch size. Thus, DAPR assigns larger replay weights to spaces providing more reliable historical prototype relations.

3.6 Multi-Perspective Knowledge Distillation

Although DAPR preserves historical prototype relations, prototype-level constraints alone are insufficient to maintain fine-grained localization behavior. Therefore, we further introduce a **Multi-Perspective Knowledge Distillation (MPKD)** module to preserve historical localization knowledge from mask, boundary, and feature perspectives. Given the same input x_i, the current model and the frozen old model produce (M_i^t, B_i^t, z_i^t) and $(M_i^{t-1}, B_i^{t-1}, z_i^{t-1})$, respectively. The mask and boundary distillation losses are defined as

$$\mathcal{L}_m = \left\|\rho(M_i^t)_{\text{fg}} - \rho(M_i^{t-1})_{\text{fg}}\right\|_2^2, \qquad \mathcal{L}_{bd} = \left\|\rho(B_i^t)_{\text{fg}} - \rho(B_i^{t-1})_{\text{fg}}\right\|_2^2. \tag{11}$$

where $(\cdot)_{\text{fg}}$ denotes the foreground-class channel. For feature distillation, we constrain only informative regions Ω_i, including manipulated pixels, boundary pixels, and high-confidence forged regions predicted by the old model:

$$\mathcal{L}_f = \frac{1}{|\Omega_i|} \sum_{p \in \Omega_i} \left\|z_i^t(p) - z_i^{t-1}(p)\right\|_2^2. \tag{12}$$

The overall MPKD loss is

$$\mathcal{L}_{\text{mpkd}} = \mathcal{L}_m + \mathcal{L}_{bd} + \lambda_f \mathcal{L}_f. \tag{13}$$

By jointly preserving region responses, boundary structures, and feature representations, MPKD alleviates the degradation of fine-grained localization ability during continual adaptation.

3.7 Overall Objective

The overall training objective can be calculated as follows:

$$\mathcal{L} = \mathcal{L}_{base} + \lambda_{kd}\mathcal{L}_{mpkd} + \lambda_{re}\mathcal{L}_{rep}. \tag{14}$$

Therefore, DUNK continuously adapts to newly emerging manipulation patterns while mitigating the forgetting of historical non-semantic discriminative knowledge.

Table 1: Performance comparison of our proposed DUNK method against existing Image Forgery Localization methods and Continual Learning methods.

Method	Pub.	CASIAv2		IMD2020		AutoSplice		MMTD-Set		SID-Set		AVG↑	
		IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
CAT-Net Kwon et al. (2022)	*IJCV'22*	1.6	3.8	4.7	9.1	12.2	24.7	24.8	28.3	72.1	81.6	23.1	29.5
PSCC-Net Liu et al. (2022)	*TCSVT'22*	2.8	6.1	11.6	21.7	15.1	22.8	33.7	39.3	79.2	85.7	28.5	35.2
TruFor Guillaro et al. (2023)	*CVPR'23*	1.7	5.1	10.6	15.9	21.1	27.8	34.8	39.2	82.3	91.6	30.1	35.9
Mesorch Zhu et al. (2025c)	*AAAI'25*	6.5	10.8	15.4	22.3	5.9	10.6	46.3	63.2	80.1	89.6	30.9	39.3
PIM Kong et al. (2025)	*TPAMI'25*	4.4	9.1	15.7	26.9	33.2	42.8	42.6	51.3	66.1	79.7	32.5	42.0
Joint-Train	-	64.8	74.7	67.2	76.9	86.5	92.6	79.8	91.8	79.2	90.6	75.5	85.3
PIM+LwF Li & Hoiem (2017)	*TPAMI'17*	15.8	31.9	22.7	36.1	43.1	54.8	46.6	56.3	76.2	88.6	40.9	53.5
+PLOP Douillard et al. (2021)	*CVPR'21*	16.9	31.8	17.6	31.5	52.3	68.4	56.7	70.2	52.8	68.1	39.3	54.0
+DKP Xu et al. (2024)	*CVPR'24*	19.5	32.7	20.4	34.8	48.6	63.7	57.9	71.6	60.2	75.4	41.3	55.6
+Adapter Zhu et al. (2025b)	*AAAI'25*	18.4	34.2	18.9	33.4	56.4	72.1	60.4	75.3	54.5	70.5	<u>41.7</u>	<u>57.1</u>
+CoGaMiD Zhu et al. (2025a)	*NeurIPS'25*	21.7	34.1	21.9	42.1	30.8	43.9	49.2	61.7	68.7	80.6	38.5	52.5
DUNK(Ours)	This paper	38.7	53.1	35.6	46.9	79.2	87.8	76.4	86.2	75.1	87.6	**61.0**	**72.3**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets and metrics. To verify the effectiveness of our method, we construct a sequential benchmark using five representative image forgery datasets, including CASIAv2 (Dong et al., 2013), IMD2020 (Novozamsky et al., 2020), AutoSplice (Jia et al., 2023), MMTD-Set (Xu et al., 2025), and SID-Set (Huang et al., 2025)following the order *CASIAv2* $\rightarrow$ *IMD2020* $\rightarrow$ *AutoSplice* $\rightarrow$ *MMTD-Set* $\rightarrow$ *SID-Set*. These datasets cover diverse manipulation patterns, ranging from traditional copy-move and splicing operations to recent AI-editing forgery images and each dataset is regarded as a new task. We evaluate the performance of different methods using Intersection over Union (IoU), F1-score (F1), and Average Forgetting (AF). The configurations of the datasets and the details of the evaluation metrics are presented in Appendix B.

Implementation Details. We train the proposed DUNK on 8 NVIDIA GeForce RTX 4090 GPUs. During training, the Adam optimizer is adopted to optimize the whole model, with the learning rate and weight decay set to 1×10^{-4} and 1×10^{-4}, respectively. Both the training and testing batch sizes are set to 8. We train the model for 60 epochs in the first dataset and 40 epochs for the remains. In the base objective, λ_B, λ_C, and λ_R are set to 1.0, 0.001, and 0.1, following PIM (Kong et al., 2025) .The distillation and replay loss weights are set to $\lambda_{kd} = 1.0$,$\lambda_f = 0.2$, and $\lambda_{re} = 2.0$; for DAPR, top-$K = 3$ and the relation temperature τ is 0.1.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Compared Methods. Various state-of-the-art IFL approaches including CAT-Net (Kwon et al., 2022), PSCC-Net (Liu et al., 2022), TruFor (Guillaro et al., 2023), Mesorch (Zhu et al., 2025c) and PIM (Kong et al., 2025) are employed for evaluation on CIFL task, where each method is training on each dataset sequentially. In addition, we further compared with several continual learning(CL) methods, including LwF (Li & Hoiem, 2017), PLOP (Douillard et al., 2021), DKP (Xu et al., 2024), Adapter (Zhu et al., 2025b) and CoGaMiD (Zhu et al., 2025a) for knowledge anti-forgetting. Details of the baseline adaptation are provided in Appendix C. In addition, the Joint-Train result, the upper bound of CIFL models where all datasets are given at once for training, is also reported. The best results are highlighted in **bold** and the second-best results are <u>underlined</u>.

Compared with IFL methods: As shown in Tab. 1 and Tab. 2, our proposed DUNK framework achieves the SOTA performance compared with traditional IFL methods, reaching **61.0%/72.3%** on the average IoU/F1 accuracy and **18.5%/17.3%** on the average AF(IoU/F1) score, respectively. For example, compared with the baseline IFL method PIM, DUNK brings significant gains of **28.5%/30.3%** on the average IoU/F1 accuracy, and substantially reduces the AF(IoU/F1) score by **30.9%/33.7%**, demonstrating stronger localization accuracy and anti-forgetting ability in the CIFL scenario. This is because DUNK preserves historical authentic-forgery discrepancies via prototype

memory, adaptive replay, and multi-perspective distillation, mitigating the overwrite of previously learned manipulation patterns.

Compared with CL methods: As shown in Tab. 1 and Tab. 2, our proposed DUNK framework also achieves SOTA overall performance compared with CL methods, reaching **61.0%/72.3%** on the average IoU/F1 accuracy and **18.5%/17.3%** on the average AF(IoU/F1) score,respectively. For example, compared with the Adapter (Zhu et al., 2025b) in localization accuracy, DUNK improves the average IoU/F1 accuracy by **19.3%/15.2%**; meanwhile, compared with the strongest CL baseline CoGaMiD (Zhu et al., 2025a) in anti-forgetting performance, DUNK reduces the AF(IoU/F1) score by **6.1%/4.0%**. This is because our DUNK explicitly preserves non-semantic authentic-forgery discrepancy knowledge in both spatial and frequency spaces through dual-space prototype memory, adaptive prototype-relation replay, and multi-perspective localization distillation. To further validate the order robustness of DUNK, additional results under *Order-2* and *Order-3* are provided in the AppendixG.

Table 2: Comparison of our DUNK method with existing methods on AF (IoU/F1) performance on different datasets.

Method	CASIAv2	IMD2020	AutoSplice	MMTD-Set	AF↓
CAT-Net Kwon et al. (2022)	80.5/86.3	81.7/83.4	76.2/69.8	63.6/64.1	75.5/75.9
PSCC-Net Liu et al. (2022)	60.1/67.5	64.8/63.2	67.4/67.7	54.3/52.6	61.7/62.8
TruFor Guillaro et al. (2023)	54.6/66.2	64.1/65.8	48.7/55.3	43.5/43.9	52.7/57.8
Mesorch Zhu et al. (2025c)	48.3/60.7	57.4/61.2	82.6/83.1	45.8/32.5	58.5/59.4
PIM Kong et al. (2025)	51.9/62.4	48.2/51.7	54.5/49.1	42.8/40.6	49.4/51.0
PIM+LwF Li & Hoiem (2017)	45.8/51.6	36.2/39.7	19.3/27.5	31.9/25.4	33.3/36.1
+PLOP Douillard et al. (2021)	41.7/46.9	31.8/35.2	18.6/22.4	29.7/23.8	30.5/32.1
+DKP Xu et al. (2024)	39.6/44.1	29.4/33.6	16.2/18.5	27.9/21.6	28.3/29.5
+Adapter Zhu et al. (2025b)	57.2/50.8	37.5/38.1	41.9/37.4	43.7/35.2	45.1/40.4
+CoGaMiD Zhu et al. (2025a)	35.4/28.6	24.9/27.3	12.5/9.7	25.1/19.4	24.6/21.3
DUNK(Ours)	34.7/28.1	18.5/24.9	8.6/8.2	12.3/7.8	**18.5/17.3**

4.3 Ablation Studies

In this section, ablation studies are conducted to verify the effectiveness of each component.

Effectiveness of SPP: As shown in Tab 3, the utilization of the SPP brings **24.3%/27.2%** performance improvements in average IoU/F1 accuracy compared with the Base method. This implies that explicitly preserving spatial authentic-forgery prototypes can effectively maintain historical manipulation patterns, where the prototype-based spatial knowledge provides reliable guidance for retaining texture inconsistency, boundary artifacts, and local structural discrepancies, thereby facilitating the anti-forgetting ability of the model during continual adaptation. **Effectiveness of FPP:** As shown in Tab 3, the introduction of the FPP brings **20.3%/22.6%** performance improvements in average IoU/F1 accuracy compared with the Base method.Especially when together using SPP and FPP modules, our DUNK further achieves **58.2%/70.6%**. It indicates that our DUNK can effectively preserve complementary forgery knowledge in both spatial and frequency spaces, where the frequency prototype preservation captures subtle statistical irregularities that are difficult to model only in the spatial space, verifying its effectiveness. **Effectiveness of MPKD module:** As Tab 3 shows, after introducing MPKD our DUNK achieves **2.8%** and **1.7%** performance gains, respectively. The above results illustrate that by collaboratively constraining the consistency between old and new models from mask prediction, boundary delineation, and high-level feature representation, MPKD can alleviate the degradation of fine-grained localization ability during continual learning.

Table 3: Ablation Studies of each module.

Base	DPM		DAPR	MPKD	IoU	F1
	SPP	FPP				
✓	-	-	-	-	32.5	42.0
✓	✓	-	✓	-	56.8	69.2
✓	-	✓	✓	-	52.8	64.6
✓	✓	✓	✓	-	58.2	70.6
✓	✓	✓	✓	✓	**61.0**	**72.3**

Hyper-parameter Analysis. Our DUNK involves several important hyper-parameters: λ_{re} , λ_{kd} in Eq.(14), λ_f in Eq.(13) and K in Eq.(6). As shown in Fig. 4, DUNK achieves the best overall performance when $\lambda_{re} = 2.0$, $\lambda_{kd} = 1.0$, $K = 3$, and $\lambda_f = 0.2$, indicating that moderate replay and distillation constraints are crucial for balancing new-task adaptation and historical knowledge preservation. Specifically, too small weights or prototype numbers provide insufficient historical

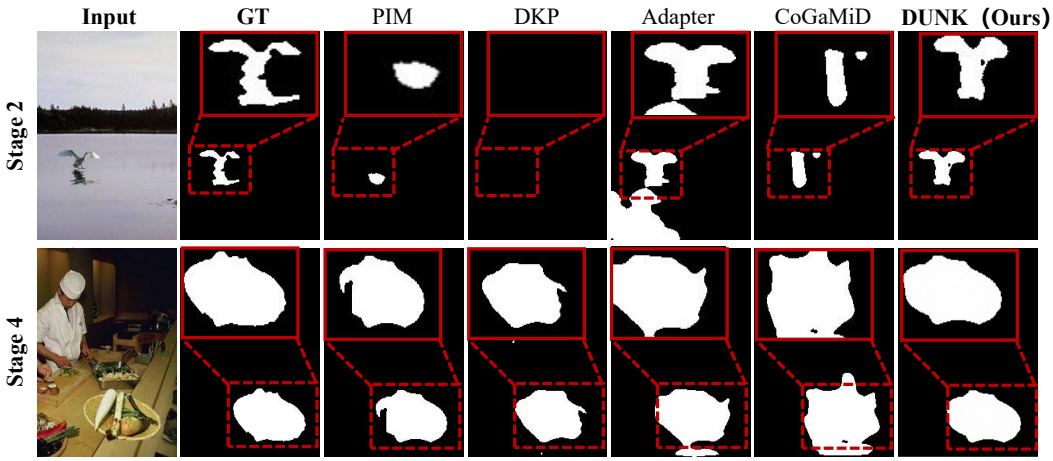


Figure 3: Comparisons between our DUNK framework and other competitive methods.

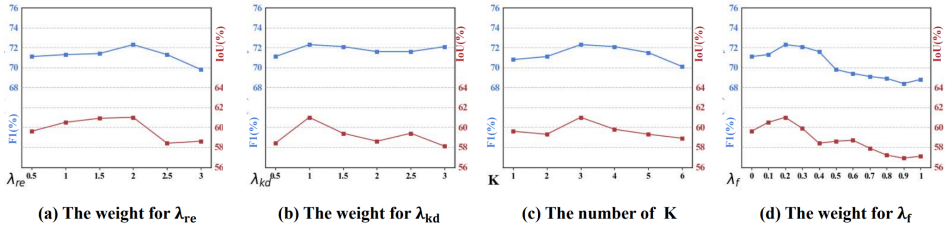


Figure 4: The influence of hyper-parameters in our DUNK.

authentic-forgery guidance, while excessively large values may over-constrain the model or introduce noisy prototype relations, thereby weakening the plasticity-stability balance in CIFL. Therefore, we adopt the above hyper-parameter settings in all experiments.

4.4 Visualization

To intuitively evaluate the effectiveness of our DUNK framework, we provide qualitative localization comparisons with several competitive methods in Fig. 3. It can be seen that existing methods usually generate incomplete, noisy, or over-segmented masks when facing diverse forgery patterns across different training stages. In contrast, our DUNK produces more accurate and complete localization results, which are more consistent with the ground-truth masks in both small manipulated regions and large forged objects. The above results intuitively illustrate the superior historical knowledge preservation and fine-grained localization ability of our DUNK framework during continual adaptation.

5 Conclusion

In this paper, we investigate Continual Image Forgery Localization (CIFL), which requires adapting to emerging manipulation patterns while preserving historical localization ability. To tackle this issue, we propose a Dual-Space Non-semantic Knowledge consolidation (DUNK) framework to maintain authentic-forgery discrepancy knowledge in both spatial and frequency spaces. Specifically, a Dual-space authentic-forgery Prototype Memory (DPM) module is designed to compactly preserve historical forgery knowledge, while a Dual-space Adaptive Prototype Replay (DAPR) module adaptively replays complementary spatial and frequency cues according to their reliability. In addition, a Multi-Perspective Knowledge Distillation (MPKD) module further alleviates the degradation of fine-grained localization ability from mask, boundary, and feature perspectives. Extensive experimental results demonstrate that our DUNK outperforms existing methods in both localization accuracy and anti-forgetting ability.

REPRODUCIBILITY STATEMENT

The main paper provides the CIFL setting, model formulation, training objectives, implementation details, and key hyper-parameter settings. Appendix B documents the dataset statistics and evaluation metrics, while the other appendices provide theoretical derivations, baseline adaptation details, further hyper-parameter analyses, and additional experiments. These materials are intended to support reproduction of our reported results.

AI USE STATEMENT

During the preparation of this manuscript, we used AI-assisted tools for language refinement and for facilitating the search of relevant literature. AI tools were not involved in developing the research ideas, formulating hypotheses, designing the proposed methodology or experiments, analyzing the experimental results, or drawing scientific conclusions. All technical content, claims, and conclusions were reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9233–9242, 2020.

Sungmin Cha, YoungJoon Yoo, Taesup Moon, et al. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. *Advances in neural information processing systems*, 34:10919–10930, 2021.

Jinpeng Chen, Runmin Cong, Yuxuan Luo, Horace Ip, and Sam Kwong. Saving 100x storage: Prototype replay for reconstructing training sample distribution in class-incremental semantic segmentation. *Advances in Neural Information Processing Systems*, 36:35988–35999, 2023.

Jinpeng Chen, Runmin Cong, Yuxuan Luo, Horace Ho Shing Ip, and Sam Kwong. Replay without saving: Prototype derivation and distribution rebalance for class-incremental semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2025.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2022.

Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China summit and international conference on signal and information processing*, pp. 422–426. IEEE, 2013.

Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4040–4050, 2021.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20606–20615, 2023.

Ruidong Han, Xiaofeng Wang, Ningning Bai, Jianpeng Hou, Wenshuo Zhang, and Jianghua Li. Hsff-net: Hierarchical spectral-feature fusion network for deepfake detection and localization. *Neural Networks*, pp. 107967, 2025.

Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *European conference on computer vision*, pp. 312–328. Springer, 2020.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28831–28841, 2025.

Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Autosplice: A text-prompt manipulated image dataset for media forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 893–903, 2023.

Jian Jiang, Yan Tian, Yongchuan Xu, Zhaocheng Xu, and Xun Wang. Sdreplay: diffusion model for continual semantic segmentation in traffic scenarios. *Multimedia Systems*, 31(6):463, 2025.

Chenqi Kong, Anwei Luo, Shiqi Wang, Haoliang Li, Anderson Rocha, and Alex C Kot. Pixel-inconsistency modeling for image manipulation localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4455–4472, 2025.

Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022.

Dong Li, Jiaying Zhu, Yidi Liu, Xin Lu, Xueyang Fu, Jiawei Liu, Aiping Liu, and Zheng-Jun Zha. Learnable frequency decomposition for image forgery detection and localization. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1359–1367, 2025.

Yuxi Li, Fuyuan Cheng, Wangbo Yu, Guangshuo Wang, Guibo Luo, and Yuesheng Zhu. Adaifl: Adaptive image forgery localization via a dynamic and importance-aware transformer network. In *European conference on computer vision*, pp. 477–493. Springer, 2024.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Zihan Lin, Zilei Wang, and Xu Wang. Towards continual universal segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29417–29427, 2025.

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

Andrea Maracani, Umberto Michieli, Marco Toldo, and Pietro Zanuttigh. Recall: Replay-based continual learning in semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7026–7035, 2021.

Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

Ju-Hyeon Nam, Dong-Hyun Moon, and Sang-Chul Lee. M2sformer: Multi-spectral and multi-scale attention with edge-aware difficulty guidance for image forgery localization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15927–15938. IEEE, 2025.

Adam Novozamsky, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops*, pp. 71–80, 2020.

Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–11, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Zhihao Sun, Haoran Jiang, Danding Wang, Xirong Li, and Juan Cao. Safl-net: Semantic-agnostic feature learning network with auxiliary plugins for image manipulation detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22424–22433, 2023.

Zhihao Sun, Haipeng Fang, Juan Cao, Xinying Zhao, and Danding Wang. Rethinking image editing detection in the era of generative ai revolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3538–3547, 2024.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024.

Kahim Wong, Jicheng Zhou, Haiwei Wu, Yain-Whar Si, and Jiantao Zhou. Adcd-net: Robust document image forgery localization via adaptive dct feature and hierarchical content disentanglement. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19280–19289. IEEE, 2025.

Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9543–9552, 2019.

Dengyun Xu, Xuanjing Shen, Yongping Huang, and Zenan Shi. Rb-net: integrating region and boundary features for image manipulation localization. *Multimedia Systems*, 29(5):3055–3067, 2023.

Kunlun Xu, Xu Zou, Yuxin Peng, and Jiahuan Zhou. Distribution-aware knowledge prototyping for non-exemplar lifelong person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16604–16613, 2024.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Hongmei Yin, Tingliang Feng, Fan Lyu, Fanhua Shang, Hongying Liu, Wei Feng, and Liang Wan. Beyond background shift: Rethinking instance replay in continual semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9839–9848, 2025.

Zekang Zhang, Guangyu Gao, Zhiyuan Fang, Jianbo Jiao, and Yunchao Wei. Mining unseen classes via regional objectness: A simple baseline for incremental segmentation. *Advances in neural information processing systems*, 35:24340–24353, 2022.

Guilin Zhu, Runmin Wang, Yuanjie Shao, Wei dong Yang, Nong Sang, and Changxin Gao. Continual gaussian mixture distribution modeling for class incremental semantic segmentation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Guilin Zhu, Dongyue Wu, Changxin Gao, Runmin Wang, Weidong Yang, and Nong Sang. Adaptive prototype replay for class incremental semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10932–10940, 2025b.

Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, and Jun Liu. Continual semantic segmentation with automatic memory sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3082–3092, 2023.

Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 11022–11030, 2025c.

A THEORETICAL JUSTIFICATION OF DUNK

In this section, we provide theoretical justification for the proposed Dual-Space Non-semantic Knowledge Consolidation (DUNK) framework. Specifically, we analyze how DPM approximates historical authentic-forgery distributions through prototype statistics, how DAPR preserves historical prototype-relation structures in both spatial and frequency spaces, and how the overall objective balances stability and plasticity during continual adaptation.

A.1 DUAL-SPACE AUTHENTIC-FORGERY PROTOTYPE MEMORY

We justify DPM by showing that the stored prototype statistics provide a maximum-likelihood diagonal Gaussian approximation to each historical authentic-forgery feature cluster. Specifically, each spatial or frequency cluster is regarded as an empirical distribution, and DPM approximates it with a diagonal Gaussian prototype. By minimizing the negative log-likelihood of the historical features, the optimal prototype center and variance are derived as the empirical mean and variance, respectively.

For each historical task, DPM stores authentic-forgery discrepancy in both spatial and frequency spaces. Let

$$q \in \{S, F\} \tag{15}$$

denote the representation space, where S and F correspond to the spatial and frequency spaces, respectively. For a given cluster $\mathcal{A}_k^{q,t}$ in space q, the empirical distribution of historical features can be written as

$$\hat{p}_k^{q,t}(v) = \frac{1}{|\mathcal{A}_k^{q,t}|} \sum_{v_i \in \mathcal{A}_k^{q,t}} \delta(v - v_i), \tag{16}$$

where v_i denotes a spatial feature from z_i when $q = S$, and a frequency descriptor d_i when $q = F$. $\delta(\cdot)$ is the Dirac delta function. We approximate this empirical distribution with a diagonal Gaussian prototype:

$$p_k^{q,t}(v) = \mathcal{N}\left(v; \mu_k^{q,t}, \mathrm{diag}(\sigma_k^{q,t})\right). \tag{17}$$

The optimal prototype parameters can be obtained by minimizing the negative log-likelihood of the empirical samples:

$$\mathcal{J}(\mu_k^{q,t}, \sigma_k^{q,t}) = -\sum_{v_i \in \mathcal{A}_k^{q,t}} \log p_k^{q,t}(v_i). \tag{18}$$

Ignoring constants independent of optimization, we have

$$\mathcal{J} = \frac{1}{2} \sum_{v_i \in \mathcal{A}_k^{q,t}} \sum_{c=1}^{C_q} \left[\log \sigma_{k,c}^{q,t} + \frac{(v_{i,c} - \mu_{k,c}^{q,t})^2}{\sigma_{k,c}^{q,t}} \right], \tag{19}$$

where C_q is the feature dimension in space q. Taking the derivative with respect to $\mu_{k,c}^{q,t}$ gives

$$\frac{\partial \mathcal{J}}{\partial \mu_{k,c}^{q,t}} = \sum_{v_i \in \mathcal{A}_k^{q,t}} \frac{\mu_{k,c}^{q,t} - v_{i,c}}{\sigma_{k,c}^{q,t}}. \tag{20}$$

Setting it to zero yields

$$\mu_{k,c}^{q,t} = \frac{1}{|\mathcal{A}_k^{q,t}|} \sum_{v_i \in \mathcal{A}_k^{q,t}} v_{i,c}. \tag{21}$$

Similarly, taking the derivative with respect to $\sigma_{k,c}^{q,t}$ gives

$$\frac{\partial \mathcal{J}}{\partial \sigma_{k,c}^{q,t}} = \frac{1}{2} \sum_{v_i \in \mathcal{A}_k^{q,t}} \left[\frac{1}{\sigma_{k,c}^{q,t}} - \frac{(v_{i,c} - \mu_{k,c}^{q,t})^2}{(\sigma_{k,c}^{q,t})^2} \right]. \tag{22}$$

Setting it to zero leads to

$$\sigma_{k,c}^{q,t} = \frac{1}{|\mathcal{A}_k^{q,t}|} \sum_{v_i \in \mathcal{A}_k^{q,t}} (v_{i,c} - \mu_{k,c}^{q,t})^2. \tag{23}$$

Therefore, the prototype statistics used in DPM are

$$\mu_k^{q,t} = \frac{1}{|\mathcal{A}_k^{q,t}|} \sum_{v \in \mathcal{A}_k^{q,t}} v, \qquad \sigma_k^{q,t} = \frac{1}{|\mathcal{A}_k^{q,t}|} \sum_{v \in \mathcal{A}_k^{q,t}} (v - \mu_k^{q,t})^2. \tag{24}$$

This shows that DPM preserves the first- and second-order statistics of historical authentic-forgery knowledge while avoiding raw image replay.

A.2 Dual-space Adaptive Prototype Replay

We justify DAPR from two aspects. First, we show that the diagonal Mahalanobis distance used for representation-to-prototype affinity can be motivated by the Gaussian likelihood of historical prototypes, where the prototype variance acts as a dimension-wise uncertainty measure. Second, we show that minimizing the KL divergence between old and current prototype-relation distributions bounds their distributional discrepancy, which encourages the current model to preserve historical prototype-relation structures during continual adaptation.

Given a forged-region representation $v_n \in \{z_i, d_i\}$ and a historical prototype (μ_k, σ_k), DAPR adopts the diagonal Mahalanobis distance:

$$d_{n,k} = \frac{1}{C} \sum_{c=1}^{C} \frac{(v_{n,c} - \mu_{k,c})^2}{\sigma_{k,c}}. \tag{25}$$

This distance naturally arises from the Gaussian likelihood. Specifically, under the prototype distribution

$$p_k(v_n) = \mathcal{N}(v_n; \mu_k, \mathrm{diag}(\sigma_k)). \tag{26}$$

the negative log-likelihood is

$$-\log p_k(v_n) = \frac{C}{2} \log(2\pi) + \frac{1}{2} \sum_{c=1}^{C} \log \sigma_{k,c} + \frac{1}{2} \sum_{c=1}^{C} \frac{(v_{n,c} - \mu_{k,c})^2}{\sigma_{k,c}}. \tag{27}$$

The last quadratic term corresponds to the diagonal Mahalanobis distance used in DAPR. Therefore, the Gaussian likelihood view motivates the variance-normalized affinity measure, where dimensions with large variance contribute less, while stable dimensions with small variance receive larger weights.

Based on this distance, the prototype relation distributions of the old model and the current model are defined as

$$r_n^{t-1} = \rho(-d_{n,\mathcal{K}}^{old}/\tau), \qquad r_n^{t} = \rho(-d_{n,\mathcal{K}}^{cur}/\tau), \tag{28}$$

where $\mathcal{K}$ denotes the index set of selected top-K nearest prototypes, τ is the temperature, and $\rho(\cdot)$ is the softmax function. This relation distribution can be interpreted as the posterior probability of assigning v_n to historical prototypes:

$$p(k|v_n) = \frac{\exp(-d_{n,k}/\tau)}{\sum_{j \in \mathcal{K}} \exp(-d_{n,j}/\tau)}. \tag{29}$$

DAPR minimizes the discrepancy between the prototype relation distributions produced by the frozen old model and the current model:

$$\mathcal{L}_{\mathrm{rep}}^{q} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{KL}\left(r_n^{t-1} \| r_n^{t}\right), \qquad q \in \{S, F\}. \tag{30}$$

By Pinsker's inequality, we have

$$\|r_n^{t-1} - r_n^{t}\|_1 \le \sqrt{2\,\mathrm{KL}\left(r_n^{t-1} \| r_n^{t}\right)}. \tag{31}$$

Therefore, minimizing the KL divergence between old and current prototype relations also bounds the total variation distance between their historical-prototype assignment distributions.

For any bounded relation-dependent function $\psi(\cdot)$ satisfying $|\psi(k)| \leq M$, we further have

$$\left|\mathbb{E}_{k\sim r_n^{t-1}}[\psi(k)] - \mathbb{E}_{k\sim r_n^{t}}[\psi(k)]\right| \leq M\|r_n^{t-1} - r_n^{t}\|_1. \tag{32}$$

Combining the above inequality gives

$$\left|\mathbb{E}_{k\sim r_n^{t-1}}[\psi(k)] - \mathbb{E}_{k\sim r_n^{t}}[\psi(k)]\right| \leq M\sqrt{2\,\mathrm{KL}\left(r_n^{t-1}\|r_n^{t}\right)}. \tag{33}$$

This indicates that DAPR explicitly constrains the current model to preserve the historical prototype-relation structure learned by the old model, thereby alleviating representation drift and catastrophic forgetting.

In addition to preserving prototype relations, DAPR needs to balance spatial and frequency replay strength. We therefore interpret the adaptive dual-space weighting as a reliability-based minimum-variance fusion problem, where a space with lower relation entropy and more forged-region representations is assigned a larger replay weight.

Spatial and frequency spaces provide complementary forgery cues. However, their reliability may vary across different manipulation types. DAPR therefore computes the reliability of each space according to the entropy of the old relation distribution. For space $q \in \{S, F\}$, let H_q be the average normalized entropy of the old relation distributions:

$$H_q = -\frac{1}{N_q \log K}\sum_{n=1}^{N_q}\sum_{k=1}^{K} r_{n,k}^{q,t-1}\log r_{n,k}^{q,t-1}. \tag{34}$$

A lower entropy indicates a more confident relation assignment. The reliability score is defined as

$$R_q = \frac{N_q}{B}\exp(-H_q), \tag{35}$$

where N_q is the number of forged-region representations in space q, and B is the batch size. The factor N_q/B measures the amount of available replay evidence, while $\exp(-H_q)$ measures the confidence of the old relations.

The final dual-space adaptive replay loss is defined as

$$\mathcal{L}_{\mathrm{rep}} = w_S\mathcal{L}_{\mathrm{rep}}^{S} + w_F\mathcal{L}_{\mathrm{rep}}^{F}, \tag{36}$$

where

$$w_S = \frac{R_S}{R_S + R_F}, \qquad w_F = \frac{R_F}{R_S + R_F}. \tag{37}$$

This weighting rule can be derived from a minimum-variance fusion view. Suppose the replay loss estimated from space q is an unbiased estimator of historical relation preservation, and its variance is inversely proportional to the reliability score:

$$\mathrm{Var}(\mathcal{L}_{\mathrm{rep}}^{q}) \propto \frac{1}{R_q}. \tag{38}$$

We seek weights w_S, w_F that minimize the variance of the fused estimator:

$$\min_{w_S,w_F} \quad \frac{w_S^2}{R_S} + \frac{w_F^2}{R_F}, \qquad \text{s.t.} \quad w_S + w_F = 1. \tag{39}$$

The Lagrangian is

$$\mathcal{G} = \frac{w_S^2}{R_S} + \frac{w_F^2}{R_F} + \lambda(w_S + w_F - 1). \tag{40}$$

Taking derivatives gives

$$\frac{\partial\mathcal{G}}{\partial w_S} = \frac{2w_S}{R_S} + \lambda = 0, \qquad \frac{\partial\mathcal{G}}{\partial w_F} = \frac{2w_F}{R_F} + \lambda = 0. \tag{41}$$

Thus,

$$\frac{w_S}{R_S} = \frac{w_F}{R_F}. \tag{42}$$

Together with $w_S + w_F = 1$, we obtain

$$w_S = \frac{R_S}{R_S + R_F}, \qquad w_F = \frac{R_F}{R_S + R_F}. \tag{43}$$

Therefore, the adaptive weighting strategy in DAPR can be interpreted as an optimal reliability-aware combination of spatial and frequency replay signals.

A.3 Overall Objective

We analyze the overall objective from the stability-plasticity perspective. The base loss promotes plasticity by adapting the model to the current manipulation patterns, while the replay and distillation terms promote stability by constraining historical prototype relations and localization-related outputs. Under a Lipschitz-continuity assumption on the old-task risk, the increase of historical localization risk can be bounded by the proposed replay and distillation losses. At incremental step t, the final objective of DUNK is

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{kd}\mathcal{L}_{\text{mpkd}} + \lambda_{re}\mathcal{L}_{\text{rep}}. \tag{44}$$

The base loss $\mathcal{L}_{\text{base}}$ encourages plasticity by learning the current manipulation pattern. The replay term preserves historical authentic-forgery prototype relations, and MPKD constrains mask, boundary, and feature drift.

Let $\mathcal{R}_{\text{old}}(\theta)$ denote the expected localization risk on previous tasks. Suppose the old-task loss is Lipschitz-continuous with respect to prototype-relation distributions and localization outputs. Then for some constants $C_r, C_m, C_b, C_f > 0$, the old-task risk increase can be bounded as

$$\mathcal{R}_{\text{old}}(\theta_t) - \mathcal{R}_{\text{old}}(\theta_{t-1}) \le C_r \left(w_S\mathcal{L}^S_{\text{rep}} + w_F\mathcal{L}^F_{\text{rep}}\right) + C_m\mathcal{L}_m + C_b\mathcal{L}_{bd} + C_f\mathcal{L}_f. \tag{45}$$

Equivalently,

$$\mathcal{R}_{\text{old}}(\theta_t) - \mathcal{R}_{\text{old}}(\theta_{t-1}) \le C_r\mathcal{L}_{\text{rep}} + C_{kd}\mathcal{L}_{\text{mpkd}}, \tag{46}$$

where C_{kd} absorbs the constants associated with mask, boundary, and feature distillation. Therefore, reducing the proposed replay and distillation losses directly controls the increase of historical localization risk. This explains why DUNK can maintain historical authentic-forgery discrepancy knowledge while still adapting to newly emerging manipulation patterns.

Overall, the proposed DPM, DAPR, and MPKD jointly provide a principled stability-plasticity mechanism: DPM compactly preserves historical distributions through moment-matched prototypes, DAPR maintains historical prototype relations in both spatial and frequency spaces, and MPKD bounds fine-grained localization drift from complementary prediction and feature perspectives. Consequently, DUNK mitigates catastrophic forgetting while preserving the ability to learn new forgery patterns.

B Details of the datasets and evaluation metrics

B.1 Datasets.

To evaluate the proposed DUNK under the Continual Image Forgery Localization (CIFL) setting, we construct a sequential benchmark with five representative datasets, including CASIAv2, IMD2020, AutoSplice, MMTD-Set, and SID-Set. These datasets cover diverse forgery patterns, ranging from traditional copy-move and splicing manipulations to recent AI-assisted image editing forgeries. The detailed numbers of training and testing samples are summarized in Table 4.

Table 4: Statistics of the datasets used in our CIFL benchmark.

Datasets	**Pub.**	**#Real**	**#Fake**
CASIAv2	*ChinaSIP'13*	7491	4873
IMD2020	*WACVW'20*	10000	10000
AutoSplice	*CVPR'23*	2273	3621
MMTD-Set	*ICLR'25*	10000	10000
SID-Set	*CVPR'25*	10000	10000

CASIAv2. CASIAv2 mainly contains traditional image forgery patterns, including copy-move and splicing manipulations. The forged regions are usually produced by copying contents from the same image or external images and inserting them into a target image, which often introduces boundary artifacts, texture inconsistency, and illumination mismatch.

IMD2020. IMD2020 is a large-scale image manipulation benchmark containing diverse realistic tampering cases. Its manipulated images are generated using a wide range of operations, including conventional image processing, inpainting, and GAN-based manipulation methods, leading to more natural object-level and region-level forgeries. Therefore, the tampered regions are often well blended with the surrounding background, making IMD2020 suitable for evaluating the robustness of forgery localization methods under realistic and advanced manipulation scenarios.

AutoSplice. AutoSplice introduces generative-model-assisted manipulation patterns. Different from conventional copy-move or manual splicing datasets, its forged images are created with the assistance of text-guided image editing models, resulting in visually coherent manipulations with more subtle forensic traces.

MMTD-Set. We use the AIGC-Editing subset of MMTD-Set, which is introduced by FakeShield for explainable image forgery detection and localization. This subset mainly focuses on generative editing forgeries, where tampered regions are produced by AIGC-based image editing. The resulting manipulations are usually more visually coherent and semantically plausible, making them suitable for evaluating localization performance under recent AI-assisted forgery patterns.

SID-Set. SID-Set mainly represents recent AI-based image editing forgeries. The manipulated regions are often highly realistic, semantically consistent, and visually smooth, making them difficult to distinguish from authentic regions using only RGB-domain cues.

Overall, these five datasets form a progressive CIFL benchmark from conventional image manipulations to recent generative and AI-editing forgeries, which is suitable for evaluating both continual adaptation and anti-forgetting ability.

B.2 Evaluation Metrics.

We evaluate different methods using Intersection over Union (IoU), F1-score (F1), and Average Forgetting (AF). IoU and F1 are used to measure pixel-level forgery localization accuracy, while AF is adopted to evaluate the degree of catastrophic forgetting during continual learning.

For each test image, let $\hat{Y}$ and Y denote the predicted forgery mask and the ground-truth mask, respectively. The Intersection over Union is computed as

$$\mathrm{IoU} = \frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|}. \tag{47}$$

Equivalently, using True Positives(TP), False Positives(FP), and False Negatives(FN), IoU can be written as

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{48}$$

The F1-score is defined as the harmonic mean of precision and recall:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{49}$$

where

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{50}$$

Thus, F1 can also be expressed as

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{51}$$

At stage t, IoU/F1 is averaged over all seen tasks $\{D_1, \ldots, D_t\}$:

$$A_t^m = \frac{1}{t} \sum_{i=1}^{t} a_{i,t}^m, \quad m \in \{\mathrm{IoU}, \mathrm{F1}\}. \tag{52}$$

To evaluate anti-forgetting ability, we further report Average Forgetting. Let $a_{i,j}^m$ denote the performance on the i-th dataset after the model has been trained on the j-th dataset, where $m \in \{\mathrm{IoU}, \mathrm{F1}\}$. Given a sequence of T datasets, the Average Forgetting under metric m is defined as

$$\mathrm{AF}^m = \frac{1}{T-1} \sum_{i=1}^{T-1} \left(\max_{j \in \{i,\dots,T-1\}} a_{i,j}^m - a_{i,T}^m \right). \tag{53}$$

Here, $\max_{j \in \{i,\dots,T-1\}} a_{i,j}^m$ denotes the best historical performance achieved on dataset $\mathcal{D}_i$ before the final step, while $a_{i,T}^m$ denotes its final performance after learning all datasets. A lower AF^m indicates stronger anti-forgetting ability.

C Adaptation of Continual Learning Methods

To ensure a fair comparison in the proposed CIFL setting, we adapt several continual learning methods originally designed for different tasks to the same PIM-based image forgery localization backbone. For each method, we retain its task-agnostic anti-forgetting mechanism and remove task-specific components that are not applicable to pixel-level forgery localization, such as semantic-class classifiers or identity-specific prediction heads. All adapted baselines follow the same dataset order, optimization settings, input resolution, and evaluation protocol as DUNK.

For all adapted continual learning baselines, the underlying localization network is instantiated with the same PIM architecture. Specifically, each model uses the PIM manipulation-mask branch, boundary branch, and reconstruction branch, and is trained with the same PIM base objective:

$$\mathcal{L}_{\mathrm{base}} = \mathcal{L}_M + \lambda_B \mathcal{L}_B + \lambda_C \mathcal{L}_C + \lambda_R \mathcal{L}_R. \tag{54}$$

Unless otherwise specified, we keep the original PIM loss weights unchanged for all adapted baselines, i.e., $\lambda_B = 1.0$, $\lambda_C = 0.001$, and $\lambda_R = 0.1$. In this way, the performance differences among the compared continual learning methods mainly come from their anti-forgetting strategies rather than from different localization backbones or base supervision settings.

LwF. Learning without Forgetting (LwF) mitigates catastrophic forgetting by distilling the softened outputs of the previous model during new-task learning. To adapt LwF to CIFL, both the old and new model are instantiated with the same PIM-based localization network. At stage t, the model trained after stage $t-1$ is frozen as the old knowledge, while the new model is initialized from this checkpoint and optimized on the current dataset $\mathcal{D}_t$. Since CIFL is a dense binary localization task, we apply LwF to pixel-wise manipulation-mask and boundary logits. The objective is

$$\mathcal{L}_{\mathrm{LwF}} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{kd}} \left(\mathcal{L}_{\mathrm{kd}}^m + \mathcal{L}_{\mathrm{kd}}^b \right). \tag{55}$$

For $q \in \{m, b\}$, the distillation loss is defined as

$$\mathcal{L}_{\mathrm{kd}}^q = T^2 \frac{1}{|\Omega_q|} \sum_{u \in \Omega_q} \mathrm{KL} \left(\rho \left(\frac{Z_{t-1}^q(u)}{T} \right) \middle\| \rho \left(\frac{Z_t^q(u)}{T} \right) \right), \tag{56}$$

where Z_{t-1}^q and Z_t^q denote the old and new logits, respectively, and $\rho(\cdot)$ denotes the softmax function. Ω_q denotes high-confidence foreground pixels predicted by the old model with threshold $\theta = 0.3$. We set $T = 2.0$ and $\lambda_{\mathrm{kd}} = 1.0$.

PLOP. PLOP is originally designed for continual semantic segmentation, where it combines entropy-based pseudo-labeling with Local POD distillation to alleviate background shift and catastrophic forgetting. To adapt PLOP to CIFL, we keep the PIM localization objective as the current-task supervision and apply Local POD between the frozen model from stage $t-1$ and the current model at stage t. Given intermediate features F_t^l and F_{t-1}^l from layer l, the adapted Local POD loss is

$$\mathcal{L}_{\mathrm{pod}} = \frac{1}{L} \sum_{l=1}^{L} \left\| \Psi(F_t^l) - \Psi(F_{t-1}^l) \right\|_2^2, \tag{57}$$

where $\Psi(\cdot)$ denotes the multi-scale horizontal and vertical pooling operator. Following PLOP, we also use the previous model to generate entropy-filtered pseudo masks on current images, retaining only confident forged predictions for auxiliary supervision. The final objective is

$$\mathcal{L}_{\mathrm{PLOP}} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{pod}} \mathcal{L}_{\mathrm{pod}} + \lambda_{\mathrm{ps}} \mathcal{L}_{\mathrm{ps}}. \tag{58}$$

We use three POD scales $\{1, 1/2, 1/4\}$ and set $\lambda_{\mathrm{pod}} = 0.01$ and $\lambda_{\mathrm{ps}} = 1.0$.

DKP. DKP is originally designed for non-exemplar lifelong person re-identification, where identity-level Gaussian prototypes are generated from instance-level distributions to preserve historical fine-grained knowledge. To adapt DKP to CIFL, we replace identity prototypes with forged-region prototypes extracted from the PIM mask branch. After stage $t-1$, forged-region features are summarized as a Gaussian prototype set $\mathcal{P}_{t-1} = \{(\mu_k, \sigma_k^2)\}_{k=1}^{K}$, where μ_k and σ_k^2 denote the prototype mean and diagonal variance. At stage t, we sample prototype features $G = \{g_k\}_{k=1}^{K}$ from $\mathcal{P}_{t-1}$ and compute the prototype-aware coordinate matrix:

$$C_p = \rho\left(F_t G^\top / \lambda_1\right), \tag{59}$$

where F_t denotes current forged-region features and $\rho(\cdot)$ denotes the softmax function. Following DKP, historical knowledge is transferred by aligning the prototype-aware affinity with the current feature affinity:

$$\mathcal{L}_{\mathrm{pkt}}^{\mathrm{DKP}} = \mathrm{KL}\left(\rho(C_p C_p^\top / \lambda_2) \middle\| \rho(F_t F_t^\top / \lambda_2)\right). \tag{60}$$

The final adapted objective is

$$\mathcal{L}_{\mathrm{DKP}} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{pkt}} \mathcal{L}_{\mathrm{pkt}}^{\mathrm{DKP}}. \tag{61}$$

Following the original DKP setting, we set $\lambda_1 = \lambda_2 = 0.1$ and use only the prototype set from the previous stage. The transfer loss weight is set to $\lambda_{\mathrm{pkt}} = 0.1$.

Adapter. Adapter is originally designed for class-incremental semantic segmentation, where it improves prototype replay by compensating for representation drift across incremental stages. To adapt Adapter to CIFL, we replace semantic class prototypes with forged-region prototypes extracted from the PIM mask branch. After each historical stage $j < t$, forged pixels are used to compute a task-level prototype P_j and its statistics. At stage t, the previous model and the current model are applied to high-confidence forged regions in the current batch to estimate the representation shift:

$$\Delta_j^t = \hat{P}_{j,t} - \hat{P}_{j,t-1}, \tag{62}$$

where $\hat{P}_{j,t}$ and $\hat{P}_{j,t-1}$ denote forged-region feature centers extracted by the current and previous models, respectively. The historical prototype is then compensated as

$$\bar{P}_j^t = \rho_j P_j + (1-\rho_j)(P_j + \Delta_j^t), \tag{63}$$

where ρ_j balances the stored prototype and the estimated shifted prototype according to the number of reliable forged pixels.

The compensated prototypes are replayed through the PIM mask classifier to preserve historical forged-region responses:

$$\mathcal{L}_{\mathrm{rep}}^{\mathrm{Ada}} = \frac{1}{|\mathcal{P}_{<t}|} \sum_{\bar{P}_j^t \in \mathcal{P}_{<t}} \mathrm{CE}\left(h_m(\bar{P}_j^t), 1\right), \tag{64}$$

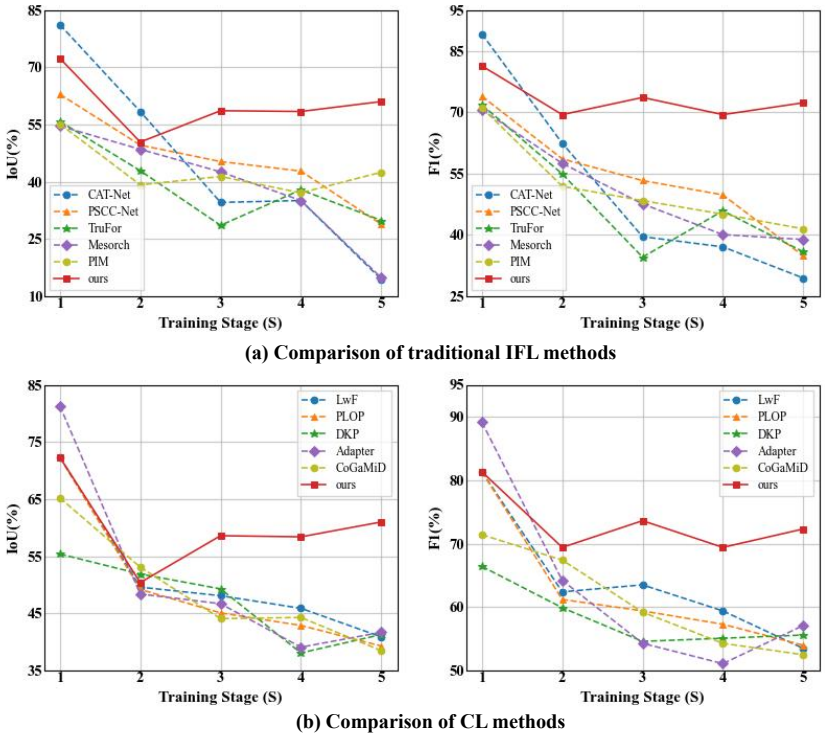
where $h_m(\cdot)$ denotes the PIM mask classifier and $\mathcal{P}_{<t}$ denotes the set of compensated prototypes from historical stages. Following Adapter, we further introduce an uncertainty-aware constraint on current predictions and a prototype discriminative constraint between current forged features and historical compensated prototypes. The final adapted objective is

$$\mathcal{L}_{\mathrm{Ada}} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{rep}} \mathcal{L}_{\mathrm{rep}}^{\mathrm{Ada}} + \lambda_{\mathrm{uac}} \mathcal{L}_{\mathrm{uac}} + \lambda_{\mathrm{cpd}} \mathcal{L}_{\mathrm{cpd}}. \tag{65}$$

Here, $\mathcal{L}_{\mathrm{uac}}$ reduces uncertain mask predictions, and $\mathcal{L}_{\mathrm{cpd}}$ separates current forged-region features from historical compensated prototypes. We set the confidence threshold to $\tau = 0.7$, following the original Adapter. The loss weights are set to $\lambda_{\mathrm{rep}} = 1.0$, $\lambda_{\mathrm{uac}} = 0.1$, and $\lambda_{\mathrm{cpd}} = 0.05$.

CoGaMiD. CoGaMiD is originally designed for class-incremental semantic segmentation, where it models old-class feature distributions with Gaussian mixtures and replays them during incremental learning. To adapt it to CIFL, we remove semantic-class-specific components and construct a task-wise forged-region feature bank on the PIM mask branch. After learning stage t, foreground mask features are collected from forged pixels and fitted by a diagonal-covariance Gaussian mixture:

$$\mathcal{G}_t(v) = \sum_{k=1}^{K_g} \pi_{t,k} \mathcal{N}\left(v; \mu_{t,k}, \mathrm{diag}(\sigma_{t,k})\right), \tag{66}$$



(a) Comparison of traditional IFL methods

(b) Comparison of CL methods

Figure 5: Performance tendency at different training stages.

where v denotes the mask-branch feature of PIM. At a later stage t, replay features are sampled from all historical mixtures $\{\mathcal{G}_j\}_{j<t}$ and fed into the PIM mask classifier as forged-region features:

$$\mathcal{L}_{\text{rep}}^{\text{CoGa}} = \frac{1}{|\tilde{\mathcal{V}}|} \sum_{\tilde{v}\in\tilde{\mathcal{V}}} \text{CE}\left(h_m(\tilde{v}), 1\right), \tag{67}$$

where $h_m(\cdot)$ is the mask classifier and $\tilde{\mathcal{V}}$ denotes sampled historical features. In addition, we preserve the mask-branch representation of the previous model through prototype-aware knowledge distillation:

$$\mathcal{L}_{\text{pkd}}^{\text{CoGa}} = \frac{1}{|\Omega|} \sum_{u\in\Omega} \left\|F_t^m(u) - F_{t-1}^m(u)\right\|_2^2 + \left\|s_t(u) - s_{t-1}(u)\right\|_2^2, \tag{68}$$

where F_t^m and F_{t-1}^m denote new and old mask features, respectively. Ω is the high-confidence forged region predicted by the old model, and $s_t(u)$ denotes the normalized likelihood vector of $F_t^m(u)$ over historical Gaussian banks. The final adapted objective is

$$\mathcal{L}_{\text{CoGa}} = \mathcal{L}_{\text{base}} + \lambda_{\text{rep}}\mathcal{L}_{\text{rep}}^{\text{CoGa}} + \lambda_{\text{pkd}}\mathcal{L}_{\text{pkd}}^{\text{CoGa}}. \tag{69}$$

We set $K_g = 3$, store at most 20,000 foreground features for each task, and sample 128 replay features from each historical task at every iteration. The confidence threshold is set to $\theta = 0.3$, and the loss weights are $\lambda_{\text{rep}} = 1.0$ and $\lambda_{\text{pkd}} = 5.0$.

D Performance Tendency at Different Training Stages

Compared with traditional IFL methods: As shown in Fig. 5(a), most traditional IFL methods exhibit obvious performance degradation as the training stage increases. Although several methods achieve competitive IoU/F1 scores at the early stages, their performance drops sharply in later stages, indicating severe catastrophic forgetting under sequential forgery distributions. In contrast, our method maintains a more stable performance tendency across different stages and consistently

achieves higher IoU and F1 scores in the later stages. This demonstrates that preserving historical authentic–forgery discrepancy knowledge is crucial for continual image forgery localization.

Compared with CL methods: As shown in Fig. 5(b), existing CL methods alleviate forgetting to some extent, but their performance still fluctuates significantly across training stages. This suggests that directly transferring semantic-oriented continual learning strategies to CIFL is insufficient, since CIFL requires preserving non-semantic forgery cues rather than category-level semantic knowledge. By contrast, our method shows a more robust and stable tendency in both IoU and F1, benefiting from dual-space prototype memory, adaptive prototype replay, and multi-perspective knowledge distillation. These results further verify the effectiveness of our framework in balancing new-task adaptation and historical-knowledge preservation.

E Distance Metric Analysis

In Sec. 3.5, we adopt the diagonal Mahalanobis distance in the Dual-space Adaptive Prototype Replay (DAPR) module to measure the relation between each current token and historical prototypes. Since each prototype stored in DPM contains both the prototype center and the dimension-wise variance, the distance metric is expected to account for the uncertainty of different feature dimensions. Based on this distance, DAPR selects the top-K nearest historical prototypes and constructs the prototype relation distribution by applying softmax over negative distances.

Given a token feature z_n, a historical prototype center μ_k, and the prototype variance σ_k, we compare three distance metrics. The Euclidean distance is formulated as:

$$d_E(z_n, \mu_k) = \frac{1}{C}\sum_{c=1}^{C}(z_{n,c} - \mu_{k,c})^2. \tag{70}$$

The Cosine distance is formulated as:

$$d_C(z_n, \mu_k) = 1 - \frac{z_n^\top \mu_k}{\|z_n\|_2 \|\mu_k\|_2}. \tag{71}$$

As shown in Fig. 6, the Diag. Mahalanobis distance achieves the best performance among the three distance metrics, demonstrating its effectiveness for prototype relation replay. Specifically, it reaches 61.0%/72.3% IoU/F1, outperforming Euclidean distance by 2.4%/0.7% and Cosine distance by 3.8%/2.9%. This is because Diag. Mahalanobis further considers the dimension-wise variance stored in DPM, thus suppressing unstable high-variance dimensions and producing more reliable prototype relations for DAPR.

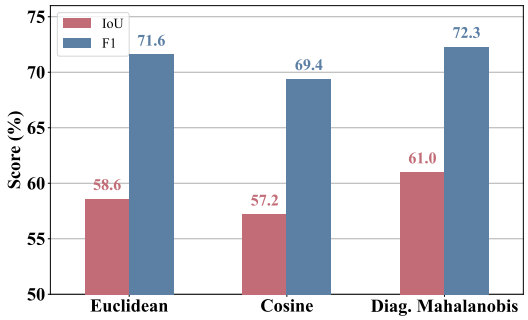


Figure 6: Comparison of different distance metrics for prototype relation replay in DAPR.

F Additional Visualization Results

To further intuitively evaluate the effectiveness of our DUNK framework, we provide more qualitative localization results across all training stages in Fig. 7. It can be seen that existing methods, including PIM, LwF, DKP, Adapter, and CoGaMiD, usually suffer from incomplete localization, fragmented predictions, or noisy false-positive regions when facing diverse forgery patterns. For example, some methods miss small manipulated regions in Stage 2 and Stage 5, while others generate over-segmented masks or introduce obvious background noise in Stage 3 and Stage 4. In contrast, our DUNK produces more complete and accurate localization masks across different stages. Specifically, our method can better preserve the shape of manipulated regions, suppress irrelevant background responses, and maintain stable localization ability for both small forged objects and

large manipulated areas. The above results further illustrate that DUNK can effectively preserve historical authentic-forgery discrepancy knowledge during continual adaptation, achieving stronger fine-grained localization performance under diverse forgery distributions.

G Effect of Training Order.

Table 5: Performance on training *Order-2*: *IMD2020→AutoSplice→CASIAv2→SID-Set→MMTD-Set*

Method	Pub.	IMD2020		AutoSplice		CASIAv2		SID-Set		MMTD-Set		AVG↑	
		IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
Joint-Train	-	67.2	76.9	86.5	92.6	64.8	74.7	79.2	90.6	79.8	91.8	75.5	85.3
PIM+LwF Li & Hoiem (2017)	*TPAMI'17*	18.9	29.6	42.5	51.6	22.5	39.2	69.4	84.1	68.4	81.3	44.3	57.2
+PLOP Douillard et al. (2021)	*CVPR'21*	15.9	28.4	48.5	66.1	26.2	36.1	49.1	64.5	79.4	85.2	43.8	56.1
+DKP Xu et al. (2024)	*CVPR'24*	18.7	31.5	45.3	63.1	33.2	37.5	55.4	72.4	77.4	88.4	46.0	58.6
+Adapter Zhu et al. (2025b)	*AAAI'25*	15.1	28.6	55.4	72.8	22.4	37.2	51.4	67.9	89.1	91.4	46.7	59.6
+CoGaMiD Zhu et al. (2025a)	*NeurIPS'25*	21.4	41.6	29.6	48.1	25.4	36.1	65.2	77.7	88.1	93.2	45.9	59.3
DUNK(Ours)	This paper	32.5	42.3	77.4	85.1	45.1	59.4	69.5	82.1	88.9	93.5	**62.7**	**72.5**

Table 6: Performance on training *Order-3*: *SID-Set→MMTD-Set→AutoSplice→IMD2020→CASIAv2*

Method	Pub.	SID-Set		MMTD-Set		AutoSplice		IMD2020		CASIAv2		AVG↑	
		IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
Joint-Train	-	79.2	90.6	79.8	91.8	86.5	92.6	67.2	76.9	64.8	74.7	75.5	85.3
PIM+LwF Li & Hoiem (2017)	*TPAMI'17*	27.8	41.2	38.6	52.3	44.7	59.8	42.1	56.4	59.8	70.8	42.6	56.1
+PLOP Douillard et al. (2021)	*CVPR'21*	25.4	39.1	35.8	49.3	47.6	62.5	41.3	55.6	58.4	71.0	41.7	55.5
+DKP Xu et al. (2024)	*CVPR'24*	29.6	43.5	39.7	53.7	49.8	64.1	44.6	58.2	59.8	69.0	44.7	57.7
+Adapter Zhu et al. (2025b)	*AAAI'25*	30.1	44.7	41.2	55.9	50.4	65.8	43.8	58.4	58.5	70.7	44.8	59.1
+CoGaMiD Zhu et al. (2025a)	*NeurIPS'25*	26.8	40.4	37.5	51.8	46.9	61.2	43.2	56.7	59.6	70.9	42.8	56.2
DUNK(Ours)	This paper	**48.9**	**63.0**	**58.7**	**70.7**	**68.4**	**78.3**	**61.5**	**73.5**	**71.0**	**79.5**	**61.7**	**73.0**

To further verify the robustness of DUNK under different task sequences, we conduct additional experiments under *Order-2* and *Order-3*. As shown in Table 5 and Table 6, DUNK consistently outperforms existing CL methods, achieving the best average IoU/F1 scores of **62.7%/72.5%**(*Order-2*) and **61.7%/73.0%**(*Order-3*). Compared with the strongest baseline Adapter, DUNK improves the average IoU/F1 by **16.0%/12.9%** under *Order-2* and **16.9%/13.9%** under *Order-3*, respectively. This is because DUNK preserves historical authentic-forgery discrepancy knowledge in both spatial and frequency spaces, which alleviates order-sensitive forgetting during continual adaptation.

H Limitation Discussion

Although DUNK shows strong performance under the evaluated CIFL settings, several aspects remain open for future exploration. First, we conduct experiments under multiple task orders, but evaluating more diverse and longer task sequences would further verify the robustness of DUNK in real-world continual forgery localization scenarios. Second, while prototype replay avoids storing massive historical raw samples, the dual-space memory and adaptive replay still introduce additional computation during continual training. Third, the effectiveness of prototype-based knowledge consolidation may be influenced by noisy annotations or highly ambiguous forgery traces. Future work will explore more scalable prototype updating and selection strategies for CIFL.

I Broader Impacts

Our method tackles a practical Continual Image Forgery Localization (CIFL) problem and introduces a Dual-space Non-semantic Knowledge consolidation (DUNK) framework to improve the long-term

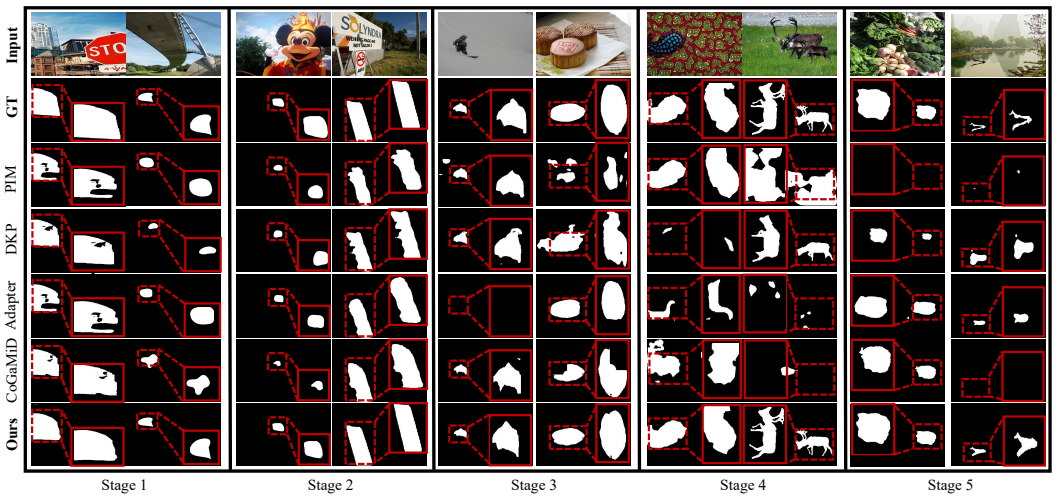


Figure 7: Additional qualitative localization results across different training stages.

adaptability of forgery localization models. The proposed method can continuously learn from newly emerging manipulation patterns while preserving the localization ability for historical forgery types.

The potential positive societal impacts of this research include:

1. Improving Digital Media Trustworthiness. With the rapid development of image editing tools and generative artificial intelligence, manipulated images are becoming increasingly realistic and diverse. Our method can help detect and localize forged regions at the pixel level, which is beneficial for media forensics, misinformation verification, copyright protection, and trustworthy visual content analysis.

2. Enhancing Adaptability to Emerging Forgery Techniques. Traditional image forgery localization methods are usually trained on fixed datasets and may fail when encountering new manipulation patterns. By addressing catastrophic forgetting in sequential forgery learning, our framework enables models to adapt to evolving forgery distributions while retaining historical knowledge, improving the long-term reliability of forensic systems in dynamic real-world environments.

3. Reducing Dependence on Historical Raw Data. Our framework preserves historical authentic-forgery discrepancy knowledge through compact spatial and frequency prototypes instead of storing massive raw historical samples. This design can reduce storage costs and alleviate potential privacy concerns caused by retaining sensitive manipulated or authentic images during continual learning.

The potential negative societal impacts and risks include:

1. Possible Misuse by Malicious Actors. Although the goal of this work is to improve forgery localization, a stronger forensic model may also provide feedback to malicious users about which forged regions are easier to detect. Such information could potentially be used to improve image manipulation techniques and evade detection systems.

2. Over-reliance on Automated Forensics. Forgery localization models may still produce false positives or false negatives, especially under unseen manipulation types, low-quality images, or complex post-processing operations. Therefore, predictions from our method should be used as supportive evidence rather than the sole basis for high-stakes decisions in journalism, law enforcement, or legal investigations.

Overall, this work is expected to contribute positively to trustworthy media analysis and robust image forensics. Meanwhile, careful deployment, human verification, and continuous evaluation are necessary to reduce the potential risks associated with misuse and over-reliance.