MAP: MAsk-Pruning for Source-Free Model Intellectual Property Protection —Supplementary Material

Boyang Peng[1]*, Sanqing Qu[1]*, Yong Wu[1], Tianpei Zou[1], Lianghua He[1],
Alois Knoll[2], Guang Chen[1]†, Changjun Jiang[1]
[1]Tongji University, [2] Technical University of Munich

A. Theoretical Analysis

Formally, we consider a source network $f_s : \mathcal{X}_s \to \mathcal{Y}_s$ trained on source domain $\mathcal{D}_s = \{(x_s, y_s)||x_s \sim \mathcal{P}^s_{\mathcal{X}}, y_s \sim \mathcal{P}^s_{\mathcal{Y}}\}$, a target network $f_t : \mathcal{X}_t \to \mathcal{Y}_t$, and target domain $\mathcal{D}_t = \{(x_t, y_t)||x_t \sim \mathcal{P}^t_{\mathcal{X}}, y_t \sim \mathcal{P}^t_{\mathcal{Y}}\}$. $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$ are the distribution of $\mathcal{X}$ and $\mathcal{Y}$, respectively. The goal of ***IP protection*** is to fine-tune f_t while minimizing the generalization region of f_t on target domain $\mathcal{D}_t$, in other words, degrade the performance of f_t on unauthorized target domain $\mathcal{D}_t$ while preserving performance on authorized source $\mathcal{D}_s$.

A.1. Definitions

Proposition 1 ([8]). *Let n be a nuisance for input x. Let z be a representation of x, and the label is y. The Shannon Mutual Information (SMI) is presented as $I(\cdot)$. For the information flow in representation learning, we have*

$$I(z; x) - I(z; y|n) \geq I(z; n) \tag{1}$$

Lemma 1 ([8]). *Let p be the predicted label outputted by a representation model when feeding with input x, and suppose that p is a scalar random variable and x is balanced on the ground truth label y. And $\mathcal{P}(\cdot)$ is the distribution. If the KL divergence loss $KL(\mathcal{P}(p)\|\mathcal{P}(y))$ increases, the mutual information $I(z; y)$ will decrease.*

A.2. Details of Optimization Objective Design

In the context of intellectual property (IP) protection, the objective is to maximize $I(z; n)$ on the unauthorized domain, and Proposition 1 provides guidance by aiming to minimize $I(z; y|n)$. According to Lemma 1, if the Kullback-Leibler (KL) divergence loss $KL(\mathcal{P}(p)\|\mathcal{P}(y))$ increases, the mutual information $I(z; y)$ will decrease. Since $I(z; y|n) = I(z; y) - I(z; n)$, the $I(z; y|n)$ will consistently decrease with $I(z; y)$. Please note that Proposition 1 and Lemma 1 have been proved in [8]. Therefore, the $\mathcal{L}_O$ in Eq. 2, $\mathcal{L}_{SA}$, and $\mathcal{L}_{SF}$ in the main paper are designed in the form of $\mathcal{L}_1 + \mathcal{L}_2$, where $\mathcal{L}_1 = KL(p_s\|y_s)$ and $\mathcal{L}_2 = -KL(p_t\|y_t)$. This design allows us to maximize $I(z; n)$ on the unauthorized (target) domain and minimize $I(z; n)$ on the authorized (source) domain.

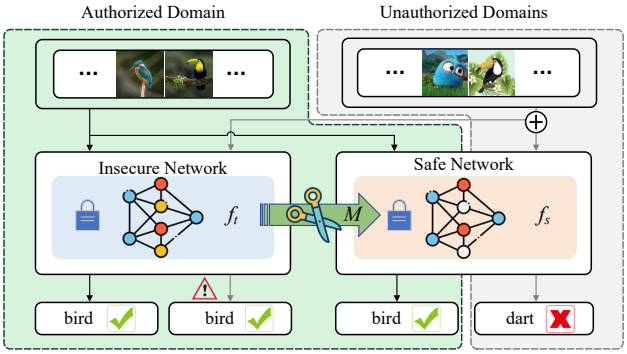

Figure 1. The total architecture of SA-MAP. A well-trained original source network f_s distills knowledge into the target network f_t, which shares the same architecture. We initialize and fix them with the same checkpoint, then update a *Learnable Binary Mask* (M) with consistency loss calculated from synthetic samples. The MAP limits a target domain generalization region while retaining source domain performance, leading to a beneficial outcome.

B. Details of MAP Architecture

As elaborated in the main text, we present the comprehensive architectural depiction of DF-MAP. In this supplementary, Fig. 1 and Fig. 2 showcase the exhaustive architectures of SA-MAP and SF-MAP, respectively. In the source-available scenario, access is available to labeled source samples $\{x^i_s, y^i_s\}^{N_s}_{i=1}$ and target samples $\{x^i_t, y^i_t\}^{N_t}_{i=1}$. We designate the source domain as the authorized domain, anticipating good performance, and the target domain as the unauthorized domain, expecting the opposite. As depicted in Fig. 1, the classification example illustrates that the secure network should correctly classify results on the authorized domain while producing erroneous results on unauthorized ones. We iteratively update a binary mask $M(\theta_M)$ for the

*Equal Contribution
†Corresponding author: guangchen@tongji.edu.cn

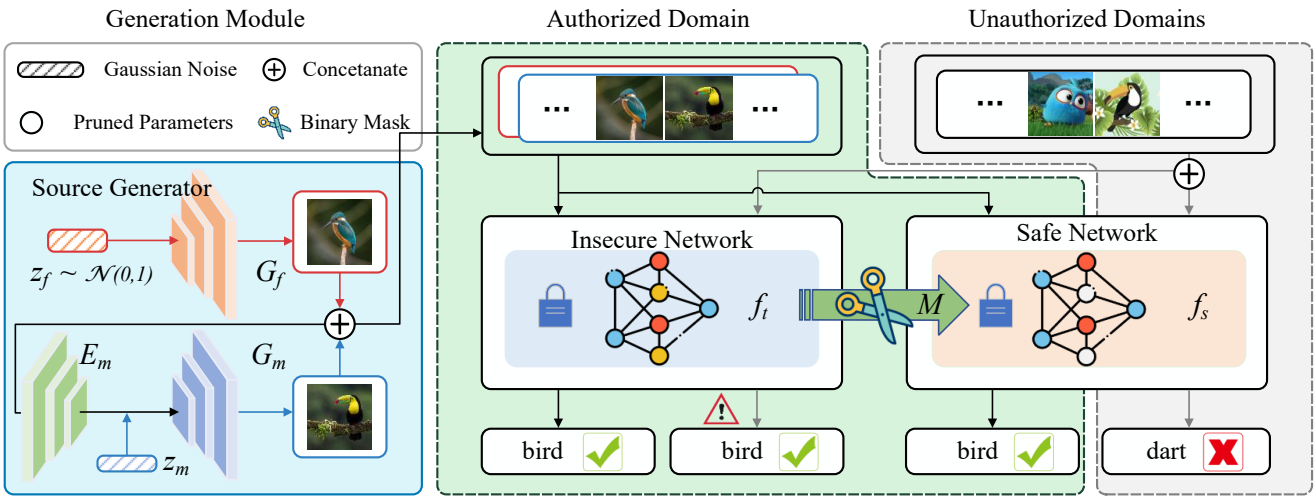

Figure 2. The total architecture of SF-MAP. **(a)** The Source Generator, displayed in the left part, consists of two generators. The *Fresh Generator (G_f)* generates synthetic novel featured samples, while *Memory Generator (G_m)* replays samples with features from previous images. **(b)** The right part illustrates the mask-pruning process. A well-trained original source network f_s distills knowledge into the target network f_t, which shares the same architecture. We initialize and fix them with the same checkpoint, then update a *Learnable Binary Mask (M)* with consistency loss from synthetic samples. The MAP limits the target generalization region, leading to a beneficial outcome.

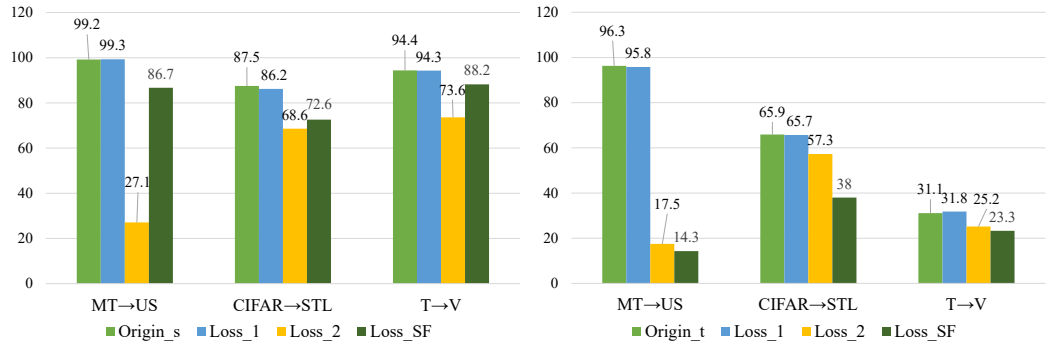

Figure 3. The accuracy (%) of SF-MAP with different losses on the target domain of MT → US, CIFAR10 → STL10, and VisDA-2017 (T → V). The **left** sub-figure is the performance of the source domain, and the **right** is the performance of the target domain. The light green bar, blue bar, yellow bar, and the dark green bar present the accuracy of origin model, the accuracy of $\mathcal{L}_1$-trained model, the accuracy of $\mathcal{L}_2$-trained model, and the accuracy of $\mathcal{L}_{SF}$-trained model, respectively.

insecure model to prune redundant parameters, which contributes to the generalization to unauthorized domains.

In the preceding discussion, f_t represents the target network, and f_s represents the source network, both sharing the same architecture and well-trained weight initialization. In the right part of Fig. 1, the pruned source network f_s can be regarded as a sub-network, aligning with our *Inverse Transfer Parameter Hypothesis* outlined in the main paper. More precisely, we gradually update $M(\theta_M)$ to prune redundant target-featured parameters of f_t, enabling the pruned f_t to progressively approximate the sub-network of f_s. The complete source network f_s is employed in the *Generation Module* in Fig. 2 to generate the source-style samples.

We further explore the SF-MAP architecture for the source-free model IP protection task, as illustrated in Fig. 2. In the source-free scenario, access is limited to unlabeled target samples $\{x_t^i\}_{i=1}^{N_t}$ and the well-trained weights (θ_s) of f_s trained on the source domain $\mathcal{D}_s$. Consequently, we generate pseudo source domain samples $\{x_s^{i\prime}, y_s^{i\prime}\}_{i=1}^{N_s'}$ and pseudo labels $\{y_{psd}^i\}_{i=1}^{N_t}$ for the target domain, as aforementioned. Please note that we are trying to improve the quality of y_{psd}, rather than creating a uniform distribution label space for the target domain. Because we want to precisely increase $I(z,n)$ on the target domain, the latter will have a negative influence on the feature z on other domains or even the source domain. Overall, the y_{psd} and *Generation Module* in Fig. 2 play a pivotal component within SF-MAP, particularly considering the challenge of synthesizing source-featured samples with only f_s and θ_s.

C. Details of Experiments

C.1. Details of Datasets

We implemented our methods and baseline on seven popular benchmarks widely used in domain adaptation and domain generalization. The digit benchmarks include MNIST [2], USPS [4], SVHN [6], and MNIST-M [3]. These benchmarks aim to classify each digit into one of ten classes (0-9). MNIST consists of 28x28 pixel grayscale images of handwritten digits, with a training set of 60,000 examples and a test set of 10,000 examples. USPS is a dataset scanned from envelopes, comprising 9,298 16x16 pixel grayscale samples. SVHN contains 600,000 32x32 RGB images of printed digits cropped from pictures of house number plates. MNIST-M is created by combining the MNIST with randomly drawn color photos from the BSDS500 [10] dataset as a background, consisting of 59,001 training images and 90,001 test images.

Additionally, we employed CIFAR10 [5], STL10 [1], and VisDA2017 [7] for image classification tasks. CIFAR10 is a subset of the Tiny Images dataset, featuring 60,000 32x32 color images with 10 object classes. STL-10 is an image dataset processed from ImageNet, comprising 13,000 96x96 pixel RGB images with 10 object classes. VisDA2017 is a simulation-to-real dataset for domain adaptation, containing 12 categories and over 280,000 images.

C.2. Details of SA-MAP result

Due to space constraints in the main paper, only a portion of the SA-MAP results are presented. In this supplementary section, we provide the detailed results in Table 1. The detailed version primarily includes the drop rate for each experiment group. The results of MAP in the source-available setting exhibit better performance. The true advantage of SA-MAP lies in the elimination of the need for retraining from scratch. The obtained sub-network effectively demonstrates our *Inverse Transfer Parameter Hypothesis*.

C.3. Details of Data-Free Model IP Protection

We provide the Algorithm 1 in the main paper. The optimization procedure adheres to the gradient as it represents the most effective path towards achieving the specified objective. In this particular instance, all produced domains exhibit alignment with a consistent gradient direction [8]. To introduce diversity in directional perspectives within the generated domains, we impose constraints on the gradient. Specifically, we decompose the generator network G_d into n_{dir} segments. To restrict the direction indexed by i, we employ a freezing strategy by fixing the initial i parameters of convolutional layers. This approach entails the immobilization of the gradient with respect to the convolutional layer parameters during training, thereby constraining the model's learning capacity along that particular direction.

Algorithm 1 Diversity Neighborhood Domains Generation

Require: The input samples $\mathcal{X}$ and label $\mathcal{Y}$ diversity generator network $G_d(x;\theta_\mu,\theta_\sigma)$, direction number n_{dir}, neighborhood samples $\mathcal{X}_{nbh} = [\,]$

while not converged **do**
  for d in n_{dir} **do**
    Generate sample $x_g = G_d(x)$
    Freeze the first d parts of G_d's each layers
    Build MI loss $\mathcal{L}_{MI}$ of x_g and x as Eq. (**??**)
    Build semantic loss $\mathcal{L}_{sem}$ of x_g, x as Eq. (**??**)
    Update θ_μ and θ_σ by $\mathcal{L}_{MI} + \mathcal{L}_{sem}$
  end for
  Append x_g to $\mathcal{X}_{nbh}$
end while
return Neighborhood samples $\mathcal{X}_{nbh}$

C.4. Details of Ownership Verification

We adopt the method from [8] to introduce a model watermark to the source domain data, creating a new auxiliary domain $\mathcal{D}_a = \{(x_a, y_a)||x_a \sim \mathcal{P}^a_{\mathcal{X}}, y_a \sim \mathcal{P}^a_{\mathcal{Y}}\}$. In this experiment, we utilize the watermarked auxiliary samples $\{x^i_a, y^i_a\}^{N_a}_{i=1}$ as the unauthorized samples, exhibiting poor performance when evaluated by model f_t. The original source samples $\{x^i_s, y^i_s\}^{N_s}_{i=1}$ without watermarks represent the authorized samples, intended to yield good performance. The details of the ownership verification experiments are outlined in Algorithm 2.

$$\begin{aligned}\mathcal{L}_O(f_t;\mathcal{X}_s,\mathcal{Y}_s,\mathcal{X}_a,\mathcal{Y}_a) = \frac{1}{N_s}\sum_{i=1}^{N_s} KL(p^S_t\|y_s) \\ -min\{\lambda\cdot\frac{1}{N_a}\sum_{i=1}^{N_a} KL(p^A_t\|y_a),\gamma\}\end{aligned} \tag{2}$$

where $KL(\cdot)$ presents the Kullback-Leibler divergence. $p^S_t = f_t(x_s)$ and $p^A_t = f_t(x_a)$ mean the prediction of target model f_t. λ means a scaling factor and γ means an upper bound. We set $\lambda = 0.1$ and $\gamma = 1$

The detailed results in Table 2 reveal that the original model in supervised learning (SL) struggles to differentiate between the source domain $\mathcal{D}_s$ and the watermarked authorized domain $\mathcal{D}_a$, achieving similar results on both. In contrast, established model intellectual property (IP) protection methods such as NTL [8], CUTI [9], and our SA-MAP exhibit distinct advantages. These methods showcase superior performance on $\mathcal{D}_s$ compared to $\mathcal{D}_a$. Particularly noteworthy is the performance of SA-MAP, which outperforms the other methods, showcasing a 1.9% improvement over the second-best approach.

C.5. Details of Ablation Study

The detailed ablation figure, highlighting various loss functions, is depicted in Fig. 3. As stated before, the model

Model	Source/Target	MT	US	SN	MM	Source Drop↓	Target Drop↑	ST-D↓
NTL	MT	98.9 / 97.4	96.3 / 14.0	36.3 / 19.0	64.9 / 11.2	1.5 (1.5%)	50.9 (77.6%)	0.019
	US	90.0 / 10.8	99.7 / 99.9	32.8 / 7.1	42.5 / 8.5	**-0.2 (-0.2%)**	**46.3 (84.0%)**	**-0.024**
	SN	68.4 / 9.0	74.9 / 8.1	91.9 / 91.1	32.8 / 9.0	0.8 (0.9%)	**50.0 (85.2%)**	0.011
	MM	97.6 / 11.3	88.2 / 16.4	40.1 / 19.2	96.8 / 95.1	2.0 (2.1%)	59.7 (79.2%)	0.027
	Mean	/	/	/	/	1.0 (1.1%)	**51.7 (81.5%)**	0.013
CUTI	MT	98.9 / 98.9	96.3 / 7.8	36.3 / 19.1	64.9 / 12.7	0 (0%)	**52.7 (80.0%)**	0
	US	90.0 / 16.7	99.7 / 99.8	32.8 / 10.1	42.5 / 8.5	-0.1 (-0.1%)	42.3 (78.6%)	-0.013
	SN	68.4 / 9.3	74.9 / 12.6	91.9 / 91.6	32.8 / 9.2	0.3 (0.3%)	48.3 (82.3%)	0.036
	MM	97.6 / 11.6	88.2 / 14.1	40.1 / 19.8	97.1/96.3	0.8 (0.8%)	60.1 (80.0%)	0.010
	Mean	/	/	/	/	0.3 (0.3%)	50.9 (80.2%)	0.004
MAP (ours)	MT	98.9 / 99.0	96.3 / 14.3	36.3 / 18.9	64.9 / 10.7	**-0.1 (-0.1%)**	51.0 (77.8%)	**-0.013**
	US	90.0 / 11.0	99.7 / 99.7	32.8 / 7.8	42.5 / 10.8	0 (0%)	45.2 (82.1%)	0
	SN	68.4 / 9.5	74.9 / 8.5	91.9 / 92.7	32.8 / 9.4	**-0.8 (-0.9%)**	49.6 (84.4%)	**-0.012**
	MM	97.6 / 11.2	88.2 / 14.3	40.1 / 19.3	97.1 / 97.2	**-0.1 (-0.1%)**	**60.4 (80.2%)**	**-0.012**
	Mean	/	/	/	/	**-0.3 (-0.3%)**	51.6 (81.1%)	**-0.004**

Table 1. Results of SA-MAP in source-available situation. In the table, MNIST, USPS, SVHN, and MNIST-M datasets are abbreviated as MN, US, SN, and MM, separately. The left of '/' represents the accuracy of the model trained on the source domain with **SL**, and the right of '/' means the accuracy of NTL, CUTI, and MAP, which are trained on the **SL** setting. The 'Source/Target Drop' means the average degradation (relative degradation) of the above models. The '↓' means a smaller number gives a better result, and the '↑' means the opposite. The data of the NTL and CUTI are obtained by their open-source code. Finally, we bold the number with the best performance.

Source	Methods				Avg Drop			
	SL	NTL	CUTI	MAP	SL	NTL	CUTI	MAP
MT	99.2 / 99.4	11.2 / 99.1	11.4 / 99.1	9.7 / 98.3	0.2	87.9	87.7	88.6
US	99.5 / 99.6	14.0 / 99.7	6.8 / 99.8	6.8 / 99.3	0.1	85.7	93.0	92.5
SN	91.1 / 90.3	24.0 / 90.4	43.5 / 90.4	34.8 / 82.3	-0.8	66.4	46.9	47.5
MM	92.1 / 96.5	12.8 / 96.7	17.1 / 96.7	16.7 / 95.9	4.4	83.9	79.6	79.2
CIFAR	85.7 / 85.7	56.8 / 84.2	45.3 / 83.7	23.5 / 79.7	0	27.4	38.4	56.2
STL	93.2 / 86.5	26.4 / 81.2	22.7 / 84.7	22.1 / 82.2	-6.7	54.8	62.0	60.1
VisDA	92.4 / 92.5	93.1 / 93.2	73.8 / 92.9	70.6 / 89.7	0.1	0.1	22.4	19.1
Mean	/	/	/	/	-0.3	58.0	61.4	63.3

Table 2. Ownership verification. The left of '/' denotes NTL, CUTI, and MAP's results on watermarked auxiliary domains, while the right on source domains. The average drop (Avg Drop) presents the drop between source domains and auxiliary domains, the higher, the better.

Algorithm 2 Ownership Verification with MAP

Require: The source dataset $\mathcal{X}_s$, target model $f_t(x;\theta_t)$, pre-trained model parameters θ_0, mask $M(\theta_M)$.

1: Initialize θ_t with θ_0 and fix θ_t
2: **while** not converged **do**
3: Add watermark to x_s to build x_a as [8].
4: Update θ_M by x_s and x_a as Eq. (2)
5: **end while**
6: **return** Learned mask parameters θ_M

trained with $\mathcal{L}_{SF}$ demonstrates superior performance on both the source and target domains. In the case of $\mathcal{L}_1$, the outcome closely resembles that of the original model, indicating a lack of effective model intellectual property (IP) protection. On the other hand, with $\mathcal{L}_2$, although there is a reduction in accuracy on the target domain, there is a noteworthy decline on the source domain, signifying a failure to adequately preserve the source performance.

References

[1] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 3

[2] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 3

[3] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 3

[4] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE TPAMI*, 16(5):550–554, 1994. 3

[5] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[6] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 3

[7] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 3

[8] Lixu Wang, Shichao Xu, Ruiqi Xu, Xiao Wang, and Qi Zhu. Non-transferable learning: A new approach for model ownership verification and applicability authorization. *arXiv preprint arXiv:2106.06916*, 2021. 1, 3, 4

[9] Lianyu Wang, Meng Wang, Daoqiang Zhang, and Huazhu Fu. Model barrier: A compact un-transferable isolation domain for model intellectual property protection. In *CVPR*, 2023. 3

[10] Jimei Yang, Brian Price, Scott Cohen, Honglak Lee, and Ming-Hsuan Yang. Object contour detection with a fully convolutional encoder-decoder network. In *CVPR*, 2016. 3