applied sciences

MDPI

Article

SupCon-MPL-DP: Supervised Contrastive Learning with Meta Pseudo Labels for Deepfake Image Detection

Kyeong-Hwan Moon [1,2], Soo-Yol Ok [1] and Suk-Hwan Lee [1,*]

[1] Department of Computer Engineering, Dong-A University, Busan 49315, Republic of Korea; drmoon_1st@naver.com (K.-H.M.); sooyol@dau.ac.kr (S.-Y.O.)
[2] Department of Information Convergence Engineering, Pusan National University, Busan 46241, Republic of Korea
* Correspondence: skylee@dau.ac.kr

Abstract: Recently, there has been considerable research on deepfake detection. However, most existing methods face challenges in adapting to the advancements in new generative models within unknown domains. In addition, the emergence of new generative models capable of producing and editing high-quality images, such as diffusion, consistency, and LCM, poses a challenge for traditional deepfake training models. These advancements highlight the need for adapting and evolving existing deepfake detection techniques to effectively counter the threats posed by sophisticated image manipulation technologies. In this paper, our objective is to detect deepfake videos in unknown domains using unlabeled data. Specifically, our proposed approach employs Meta Pseudo Labels (MPL) with supervised contrastive learning, so-called SupCon-MPL, allowing the model to be trained on unlabeled images. MPL involves the simultaneous training of both a teacher model and a student model, where the teacher model generates pseudo labels utilized to train the student model. This method aims to enhance the adaptability and robustness of deepfake detection systems against emerging unknown domains. Supervised contrastive learning utilizes labels to compare samples within similar classes more intensively, while encouraging greater distinction from samples in dissimilar classes. This facilitates the learning of features in a diverse set of deepfake images by the model, consequently contributing to the performance of deepfake detection in unknown domains. When utilizing the ResNet50 model as the backbone, SupCon-MPL exhibited an improvement of 1.58% in accuracy compared with traditional MPL in known domain detection. Moreover, in the same generation of unknown domain detection, there was a 1.32% accuracy enhancement, while in the detection of post-generation unknown domains, there was an 8.74% increase in accuracy.

Keywords: deepfake detection; deepfake unknown domain; meta pseudo labels; supervised contrastive learning; generative misuse

Citation: Moon, K.-H.; Ok, S.-Y.; Lee, S.-H. SupCon-MPL-DP: Supervised Contrastive Learning with Meta Pseudo Labels for Deepfake Image Detection. *Appl. Sci.* **2024**, *14*, 3249. https://doi.org/10.3390/app14083249

Academic Editor: David Megías

Received: 8 March 2024
Revised: 6 April 2024
Accepted: 6 April 2024
Published: 12 April 2024

1. Introduction

Recently, with the advancement of generative artificial intelligence models [1–15], deepfakes have become increasingly similar to real images/videos, making them difficult to distinguish. Deepfakes can be broadly categorized into three generations based on the evolution of generative models. First-generation deepfake generative models [1–5] typically attempt to generate simple and low-resolution images/videos based on probability distributions or synthesize multiple images/videos by exploiting features in tasks such as Face2Face and FaceSwap. In particular, generative adversarial network (GAN)-based models such as CGAN [2], WGAN [4], and WGAN-GP [5], as well as autoencoder-based models like VAE [6] and conditional VAE [7] have enabled the generation of various deepfake images/videos. However, first-generation deepfakes often exhibit noticeable artifacts that can be discerned by the human eye. With the transition to second-generation deepfake generative models [6–8], there has been progress in generating high-resolution deepfake

images that are more difficult to distinguish compared with first-generation ones, along with performance improvements in various tasks. In particular, second-generation deepfake generative models like StyleGAN, proposed by T. Karras et al. [8], produce deepfake images that are difficult for the human eye to distinguish, excluding some flaws such as artifacts in hair. Second-generation deepfakes can be generated using deepfake generation tools such as DeepfaceLab [10], DeepSwap [11], Synthesia [12], and others. Finally, third-generation deepfake generative models [9,13–15] produce images/videos that are even more flexible and difficult to distinguish than those generated by second-generation models, across various tasks. In particular, Stable Diffusion, proposed by R. Rombach et al. [13], is currently being used for the generation of various human and artwork images, raising concerns related to copyright and human rights issues. Furthermore, the consistency model proposed by Y. Song et al. [15] has enabled state-of-the-art deepfake generative model training at a lower cost by reducing the extensive iteration process required by previous diffusion models for restoring original images from noise. As deepfakes increasingly become difficult to distinguish from real images/videos, they are being utilized in various criminal activities.

To address issues caused by deepfakes, methods have been proposed to identify flaws in landmarks that occur when deepfake generative models create images, aiding in the detection of deepfakes [16–19]. Meanwhile, recent advancements in deepfake detection for single models have shown improvement in detecting deepfake videos and images. D.A. Coccomini et al. [20] enhanced the detection performance of deepfake videos and images by combining EfficientNet [21] and Vision Transformer [22] when training a single model. In other words, existing deepfake detection models [16–23] verify flaws in facial landmarks during the preprocessing stage and construct large models for flexible predictions.

Previous studies have primarily focused on the detection performance of labeled known domain (known domain) tasks in deepfake detection. However, deepfake generation models are rapidly evolving, and similar generations of deepfake generation models are also being developed diversely. Therefore, detecting deepfake images in unknown domains (unknown domain) is also crucial. A few studies have proposed generalized deepfake detection models using techniques such as contrastive learning, meta learning, and others [24–36].

In this paper, we propose SupCon-MPL, a combination of the Meta Pseudo Labels (MPL) [37] with supervised contrastive learning (SupCon) [38] to further train the model with unlabeled images/videos, simultaneously enhancing the model's generalization ability to distinguish deepfakes in unknown domains. The proposed SupCon-MPL utilizes the basic structure of MPL, where two models, namely teacher and student, are simultaneously trained. Each model influences the other during training. The teacher model constructs pseudo labels for unlabeled images and transfers them to the student model. Through this approach, the student model learns from unlabeled data, providing the potential to train effectively with limited labeled data. Furthermore, during the training process, we apply the supervised contrastive loss (SupConLoss) [38] to the encoder of each model, enabling contrastive representation learning, thereby inducing generalized model training.

The performance evaluation experiments were conducted in two parts: model validation experiments and deepfake detection experiment based on scenario. In the model validation experiments, we utilized the data from five domains within FaceForensics++ [39]. We evaluated the detection performance in labeled known domains by combining the data in various ways and assessed the generalized detection performance in unknown domains. The deepfake detection experiment based on scenario involves training the model with first-generation deepfake datasets (FaceForensics++ [39], DFDC [40], Celeb-DF [41]) and evaluating the detection performance on first- and second-generation unknown deepfake datasets (StyleGAN [8], NeuralTextures [39]). The experimental results showed that SupCon-MPL achieved performance improvements of 1.58%, 1.32%, and 8.74% over the baseline MPL model in the proposed evaluation scenario, respectively. The main contributions of this paper are as follows:

(1) The proposed method enables additional training through unlabeled data. Especially, while two models are trained simultaneously, the student model infers information about unlabeled data from the other model and provides feedback, allowing for additional training to be conducted with less bias towards a specific model. Ultimately, the proposed method enhances the performance of deepfake detection by enabling additional training with a large amount of unlabeled data.
(2) Our model enables generalized deepfake detection model training through contrastive learning. We improved the generalized deepfake detection performance on unknown data, which was previously low in the Meta Pseudo Labels-based deepfake detection model [24], through contrastive learning.
(3) Our model exhibited higher deepfake detection performance compared with all other models in the comparison with various generalized deepfake detection models [24,31,34,35]. The experimental results demonstrate that our model outperforms existing deepfake detection models, showing robust detection capability across diverse labeled datasets and even unknown generational deepfakes.

2. Related Works

As deepfake generation models advance, various detection methods have also been researched. A common approach in deepfake detection is to explore flaws in facial images [16–19]. However, the continual development of new generative models has led to the problem of being unable to train detection models using data from all generative models. To address this problem, a few studies have explored training generalized deepfake detection models [24–36].

2.1. Generalized Deepfake Detection

The generalization of deepfake detection implies the ability to detect deepfake videos generated not only by the models used during training but also by unseen or new generative models. In other word, as generative models progress from GANs and VAEs to diffusion and consistency models, achieving the ability to detect deepfakes generated by various and new generative models simultaneously is the main goal of generalized deepfake detection techniques. Recently, research has been conducted on detecting deepfake videos that are unknown from both the data and training perspectives.

On the data perspective, SBL [29] and OST [30] enhanced the generalization of deepfake detection by synthesizing additional training data by combining original images from each generative model with various other images and selectively using them. On the training perspective, A. Jain et al. [25] utilized datasets from Google, Jigsaw, FaceForensics++ [39], Celeb-DF [41], Deepfake-TIMIT [42], and their own database DF-Mobio to train a generalized deepfake detection model using contrastive representation learning across various domains. A. Nadimpalli et al. [26] proposed a hybrid learning technique combining supervised learning and reinforcement learning. In particular, during the training process, the reinforcement learning agent selects the top k augmentations that have the most significant impact on performance improvement when training through convolutional neural networks (CNN) and uses them for testing, enabling the training of a generalized deepfake detection model. We employed the meta-learning technique, Meta Pseudo Labels, in the deepfake training process, applying it after domain splitting for each data, resulting in training a model with higher performance in the same model training [24], and conducted experiments using various CNN-based models specialized in extracting features from images, including EfficientNet [21], ResNet [43], ResNext [44], and WideResNet [45], which are image classification models.

2.2. Contrastive Representation Learning in Deepfake Detection

Currently, research applying contrastive representation learning (CRL) for training generalized deepfake detection models is conducted. CRL enables learning similar features in the feature space between a specific image and from the same domain (positive images),

while also learning features that differentiate from different domains (negative images). H. Chih-Chung et al. [27] utilized contrastive loss [43] to train the encoder, following which they trained the classifier to generalize the discriminative performance on deepfake images generated by various GAN-based models. S. Fung et al. [36] trained the encoder using unsupervised CRL with image pairs that include random augmentation applied to the same image during the training process. Following this, they trained the classifier using labeled images to develop a generalized deepfake detection model. X. Ying et al. [35] addressed the issue of conventional CRL techniques not utilizing label information of deepfake images by applying supervised contrastive learning (SupCon) [38], which uses label information. However, CRL requires a large amount of data, especially a significant number of negative samples. In this paper, we performed CRL using SupCon [38], while simultaneously combining the SupCon model with meta-learning method, MPL [37]. This approach also allows for additional CRL training on unlabeled data, even when using the same labeled data, enabling the training of a more generalized deepfake detection model compared with conventional Meta Pseudo Labels.

2.3. Meta Pseudo Labels

Meta Pseudo Labels (MPL) [37] trains the model using unlabeled images, and when the same model is trained on an image classification task, it has shown improved performance compared with conventional models. MPL gained significant attention by achieving over 90% Top 1 score on the ImageNet [44] classification task, marking a significant milestone. Figure 1a shows the MPL facilitates the learning of the teacher model through feedback from the student model, thus enhancing the conventional learning techniques such as knowledge distillation [45] or noisy student [46], where the teacher model passes on information to the student model. This improvement addresses the issue of inadequate learning of the student model when the performance of the teacher model is subpar. The training process of the MPL is shown in Figure 1b. The student model in MPL learns through the Pseudo Labels inferred by the teacher model. Subsequently, it imparts the feedback value regarding the learning to the teacher model. The teacher model learns through the labeled loss from the labeled data, UDA loss [47], feedback from the student model, and MPL loss from the unlabeled data. However, it consumes substantial computing resources due to the simultaneous training of the two models.

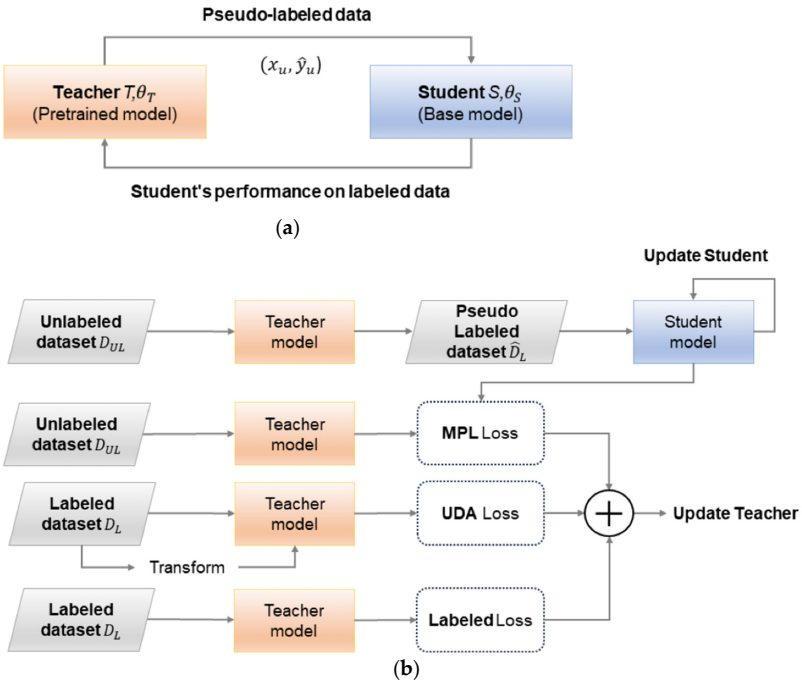

Figure 1. (**a**) Training overview and (**b**) process with label/unlabeled datasets of Meta Pseudo Labels [37].

From the perspective of training a generalized deepfake detection model, MPL can enhance the performance of generalized deepfake detection by enabling additional training through unlabeled data, compared with models trained solely with labeled data. In this paper, we experiment with the enhancement of detection capabilities for unknown domains and post-generation deepfakes, using both MPL [37] and SupCon [38].

3. Proposed SupCon-MPL-Based Deepfake Detection

To detect deepfake videos in the deepfake unknown domain, the proposed method introduces SupCon-MPL, a meta-learning model based on contrastive learning, utilizing unlabeled images from the deepfake known domain. You can find notations used in the rest of paper summarized in Table A1.

3.1. Proposed Training Strategy

3.1.1. Known Domain and Unknown Domain in Deepfake

A deepfake domain can be defined as a collection of images and their features, generated from a "specific deepfake generative model". In this paper, we distinguish deepfake domains into the known domain ($\boldsymbol{K}$) and unknown domain ($\boldsymbol{U}$). The known domain ($\boldsymbol{K}$) refers to a collection of deepfake images that are labeled when training models. The data in $\boldsymbol{K}$ is labeled and therefore can be directly used for training. Meanwhile, the unknown domain ($\boldsymbol{U}$) refers to data created by unknown deepfake generative models. The data in $\boldsymbol{U}$ is not labeled, hence it is not possible to determine whether the image is real or fake. Also, as they are created from various generative models, they can involve various features. The known domain $\boldsymbol{K}$ can be defined as $\boldsymbol{K} = \{\mathrm{K}_1, \mathrm{K}_2, \cdots\}$ where K_i is i-th known deepfake generative model, and the deepfake dataset $\boldsymbol{D_K} = \{D_{K_1}, D_{K_2}, \cdots\}$ consists of a dataset $\boldsymbol{D_{K_i}} = \{(x_i, y_i)\}$ composed with a set of deepfake images x_i and labels y_i generated by K_i. On the other hand, the unknown domain $\boldsymbol{U}$ can be defined as $\boldsymbol{U} = \{U_1, U_2, \cdots\}$ where U_i is i-th unknown deepfake generative model, and the deepfake dataset $\boldsymbol{D_U} = \{D_{U_1}, D_{U_2}, \cdots\}$ consists of a dataset $D_{U_i} = \{(x_i)\}$ composed with a set of deepfake images x_i generated by U_i.

In this paper, to address $\boldsymbol{U}$, we first experiment by distinguishing $\boldsymbol{D_K}$ into a labeled dataset (D_L) and an unlabeled dataset (D_{UL}), as shown in Figure 2a. Subsequently, to verify the influence of $\boldsymbol{D_U}$ on the training process, we assume $\boldsymbol{D_U}$ as D_{UL} and perform experiments, as shown in Figure 2b. In the deepfake training scenario, from the perspective of generative models by generation, both D_L and D_{UL} constitute with first-generation $\boldsymbol{D_K}$, and evaluation is conducted using the first- and second-generation $\boldsymbol{D_U}$.

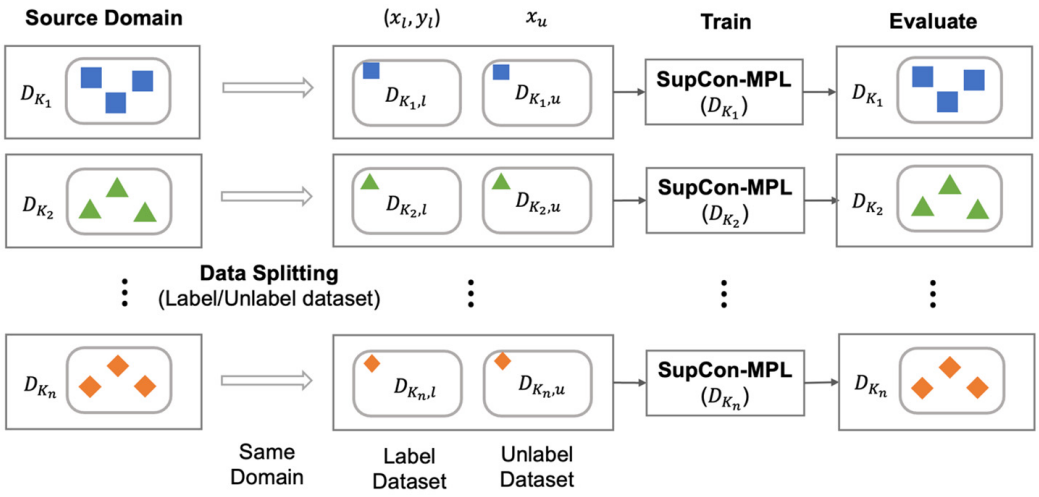

Figure 2. *Cont.*

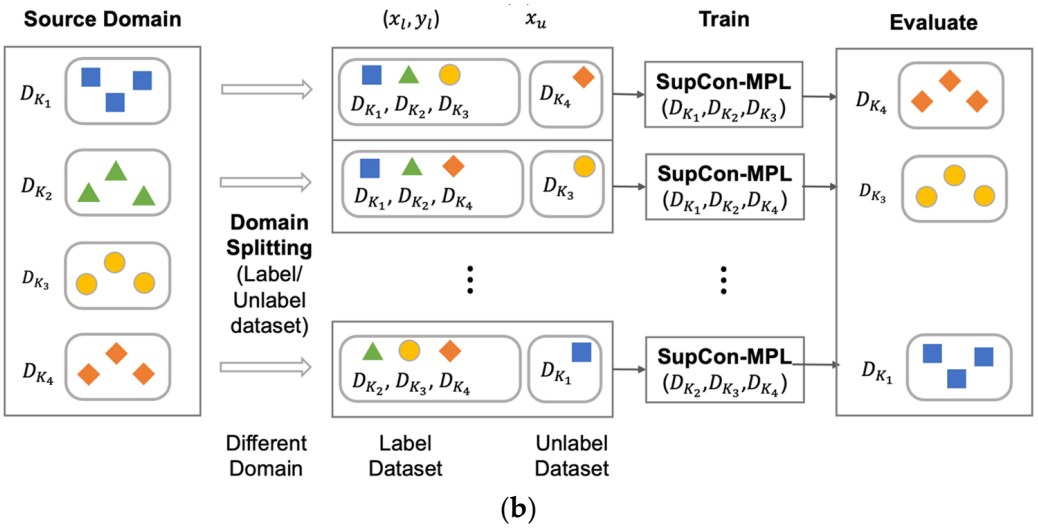

Figure 2. Deepfake image discrimination strategy targeting for the (**a**) known domain and (**b**) unknown domain.

3.1.2. Training Strategy for Deepfake Unknown Domain Detection

The training process is employed based on a comparison between the base model and the student model of the MPL (SupCon-MPL). Upon completion of training the base model with the entire dataset $\boldsymbol{D_K}$, the model is subsequently employed as the teacher model to train the student model. In other words, we aim to verify performance improvement when training the model under the same conditions. If performance enhancement is validated at this method, it suggests that superior performing models can be trained under identical training conditions, even when employing larger or state-of-the-art (SOTA) models.

The training images are constructed considering the problems of existing deepfake detection. While deepfakes by known generative models exist in $\boldsymbol{K}$, deepfake images by unknown generative models also exist in $\boldsymbol{U}$. Therefore, during the training phase, we enhance the deepfake detection performance in $\boldsymbol{K}$ using labeled data and contribute to the generalization of the learning model by using data $\boldsymbol{D_K}$ and $\boldsymbol{D_U}$ from $\boldsymbol{K}$ and $\boldsymbol{U}$ as unlabeled data, respectively. Consistent with this approach, the data D_L and D_{UL} are structured into $\boldsymbol{D_K}$, with images from dataset $\boldsymbol{D_U}$ serving as D_{UL}.

In the proposed method, we combine data in three strategies to detect the unknown domain dataset $\boldsymbol{D_U}$. The first strategy is to use the data from D_L and D_{UL} as the same domain, aiming to verify whether unlabeled data from a specific $\boldsymbol{K}$ contributes to the improvement of model performance. Figure 2a illustrates the training strategy of using $\boldsymbol{D_K}$ as unlabeled data. The second strategy aims to solve the realistic deepfake problem by experimenting with the impact of unlabeled data on the detection performance of the corresponding domain. Figure 2b illustrates the feasibility of improving model performance by employing dataset $\boldsymbol{D_K}$ as labeled data D_L and dataset $\boldsymbol{D_U}$ as unlabeled data D_{UL}. Finally, in the deepfake scenario experiment, after training the model using the first-generation deepfake dataset as D_L and D_{UL}, the generalized deepfake detection model learning is assessed through the first-generation $\boldsymbol{D_K}$ and the first- and second-generation $\boldsymbol{D_U}$.

3.2. SupCon-MPL: Supervised Contrastive Learning with Meta Pseudo Labels

In the proposed method, following the strategy in Figure 2b, the MPL model is trained for the detection of deepfakes in the unknown domain $\boldsymbol{U}$. SupCon-MPL allows supplementary training utilizing unlabeled videos, and with the aid of CRL, it enhances the deepfake detection in feature space. Furthermore, it affords the flexibility to employ diverse encoder models during the training phase and enables the fine tuning of the SupCon-MPL-trained model.

In particular, the limitations of deepfake detection with limited labeled data can be mitigated by using unlabeled data, and a generalized detection model can be trained through CRL. Another notable advantage lies in the capability to conduct concurrent learning via feedback from the student model, even if the performance of the T model is low. The details of the proposed method are elucidated in Figure 3. The most significant distinction from the conventional MPL and SupCon model training is that learning through unlabeled data not only resolves the training issue of CRL due to limited data but also enhances detection capabilities in both $\boldsymbol{K}$ and $\boldsymbol{U}$. Ultimately, the final goal is to enhance the detection capabilities of deepfakes in domains that are not targeted, especially in a situation where new deepfake models in $\boldsymbol{U}$ continue to be developed.

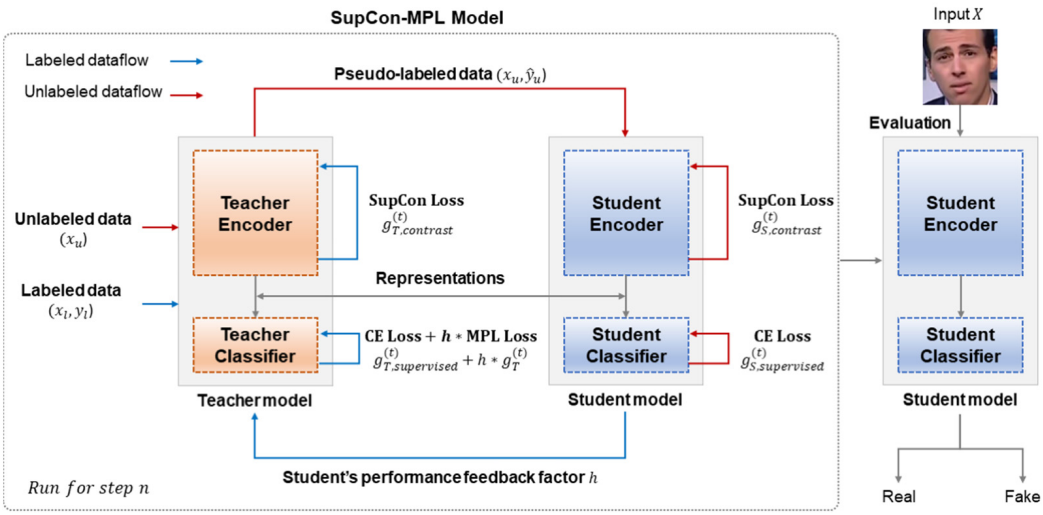

Figure 3. Modified Meta Pseudo Labels and loss functions for deepfake detection.

Supcon-MPL consists of a teacher model T and a student model S. Each model has the same structure but different parameter values. The teacher model (T) utilizes a pre-trained model, while the student model (S) starts its training from the initial state before being trained. Each model consists of an identical structure of an encoder and a classifier. The encoder is modified to extract a 128-dimensional feature by removing the classifier layer of a specific model, allowing it to primarily learn the representations of labeled and unlabeled images. The classifier is composed of a single linear layer, which performs classification based on the representation values received from the encoder.

The training of the SupCon-MPL is conducted by first having the student model T use unlabeled data to perform CRL, followed by fine tuning with labeled data. In this process, the teacher model T's classifier learns through the feedback from S, while S learns dependently on T.

3.3. SupCon-MPL Loss Function

SupCon-MPL, as shown in Figure 3, is composed of a teacher model (T) and a student model (S), each of which consists of an encoder and a linear classifier. SupCon-MPL has two loss functions in order to sequentially train each model. One involves the teacher model T distilling knowledge to the student model S, while the other entails the teacher model T training from the feedback factor provided by S on the labeled data. The knowledge distilled by T includes previously learned content about deepfakes.

In SupCon-MPL, let the parameters of T classifier and S classifier be θ_T, θ_S, respectively, and denote the batch of images and labels on the labeled data as $(x_l, y_l) \in D_K$, and the

batch of images on the unlabeled data as $x_u \in D_U$. The goal of SupCon-MPL is to minimize the parameters θ_S^{PL} of the generalized deepfake detection model S:

$$\theta_S^{PL} = \underset{\theta_S}{argmin} \underbrace{E_{x_u}[CE(T(x_u;\theta_T), S(x_u;\theta_S))]}_{L_u := (\theta_T, \theta_S)} \tag{1}$$

Hence, the objective function of SupCon-MPL is defined as follows.

$$\begin{gathered} L_l \text{ with respective to } \theta_T: \\ \min_{\theta_T} L_l\left(\theta_S^{PL}(\theta_T)\right), \\ \text{where } \theta_S^{PL}(\theta_T) = \underset{\theta_S}{\operatorname{argmin}} L_u(\theta_T, \theta_S). \end{gathered} \tag{2}$$

For optimization, SupCon-MPL approximates $\theta_S^{PL}(\theta_T)$ by the learning rate ηs, and then,

$$\theta_S^{PL}(\theta_T) \approx \theta_S - \eta s \cdot \nabla_{\theta_S} L_u(\theta_T, \theta_S) \tag{3}$$

defines the final objective function as follows:

$$\begin{gathered} L_l \text{ with respective to } \theta_T: \\ \min_{\theta_T} L_l\left(\theta_S - \eta s \cdot \nabla_{\theta_S} L_u(\theta_T, \theta_S)\right), \\ \text{where } \theta_S^{PL}(\theta_T) = \underset{\theta_S}{\operatorname{argmin}} L_u(\theta_T, \theta_S). \end{gathered} \tag{4}$$

Both T and S consist o-f an encoder and a classifier and are trained according to their respective loss functions. The loss function of ENC_T, the encoder of T, is composed of SupConLoss [38], and the loss function of CLF_T, the classifier, is composed of labeled loss for the labeled data x_l and MPL loss, reflecting the feedback from S. First and foremost, $g_{T,contrast}^{(t)}$ the loss function of ENC_T, receives image pairs $(RandAugment_a(x_l), RandAugment_b(x_l), y_l)$ as inputs that reflect different random augmentations on the same image x_l and the label y_l. Subsequently, the loss value is obtained by passing image pairs through SupConLoss [38]. At this juncture, given the similarity between the current training process and that of the original MPL's UDA loss, the utilization of the UDA loss is no longer used:

$$g_{T,contrast}^{(t)} = \nabla_{\theta_T} SupConLoss(RandAugment_a(x_l), RandAugment_b(x_l), y_l)\big|_{\theta_T = \theta_T^{(t)}} \tag{5}$$

ENC_T is promptly updated following the computation of the $g_{T,contrast}^{(t)}$:

$$\theta_{T,ENC}^{(t+1)} = \theta_{T,ENC}^{(t)} - \eta_S \cdot g_{T,contrast}^{(t)} \tag{6}$$

The labeled loss of ClF_T, $g_{T,supervised}^{(t)}$, measures the difference between y_l and the label predicted by T through cross-entropy loss (CE Loss). Here, emb_l^T denotes the embedding value derived by passing the labeled data x_l through ENC_T:

$$g_{T,supervised}^{(t)} = \nabla_{\theta_T} CE\left(y_l, CLF_T\left(emb_l^T; \theta_T\right)\right)\big|_{\theta_T = \theta_T^{(t)}} \tag{7}$$

The MPL loss $g_T^{(t)}$ calculates the difference between the hard pseudo label y_u, which is the maximum value extracted from the pseudo labels generated by T through x_u, and the logit. Here, emb_u^T denotes the embedding value derived by passing the labeled data x_u through ENC_T:

$$g_T^{(t)} = h \cdot \nabla_{\theta_T} CE\left(\hat{y}_u, CLF_T\left(emb_u^T; \theta_T\right)\right)|_{\theta_T=\theta_T^{(t)}} \tag{8}$$

The feedback factor h of S was calculated in the same way as the original Meta Pseudo Labels [37], using Taylor expansion to calculate the difference before and after the training of S. In the proposed method, we approximated h using the difference from the CE loss value for the labeled data after S was trained to the value before training. This allows the final loss value to converge as the training progresses.

$$h = CE\left(y_l, S\left(x_l; \theta_S^{(t+1)}\right)\right) - CE\left(y_l, S\left(x_l; \theta_S^{(t)}\right)\right) \tag{9}$$

The final loss function of ClF_T is composed of the sum of each loss function value:

$$\theta_T^{(t+1)} = \theta_T^{(t)} - \eta s \cdot \left(g_T^{(t)} + g_{T,supervised}^{(t)}\right) \tag{10}$$

S is trained through unlabeled data. The loss function of the student model's encoder ENC_S, denoted as $g_{S,contrast}^{(t)}$, is trained utilizing SupConLoss [38], akin to ENC_T. It leverages $(x_u, \hat{y}_u)$, comprising an unlabeled image x_u paired with pseudo labels $\hat{y}_u$, generated by T:

$$\begin{aligned} &g_{S,contrast}^{(t)} \\ &= \nabla_{\theta_T} SupConLoss(RandAugment_a(x_u), RandAugment_b(x_u), \hat{y}_u)|_{\theta_S=\theta_S^{(t)}} \end{aligned} \tag{11}$$

ENC_S is also promptly updated following the computation of the $g_{S,contrast}^{(t)}$:

$$\theta_{S,ENC}^{(t+1)} = \theta_{S,ENC}^{(t)} - \eta_S \cdot g_{S,contrast}^{(t)} \tag{12}$$

The loss function of CLF_S is calculated using CE loss for the hard pseudo label $\hat{y}_u$ of T for x_u and the prediction of S. Here, emb_u^S denotes the embedding value derived by passing the labeled data x_u through ENC_S:

$$\theta_S^{(t+1)} = \theta_S^{(t)} - \eta s \cdot \nabla_{\theta_S} CE\left(\hat{y}_u, CLF_S\left(emb_u^S; \theta_S\right)\right)|_{\theta_S=\theta_S^{(t)}} \tag{13}$$

The SupConLoss in $g_{T,contrast}^{(t)}$, $g_{S,contrast}^{(t)}$ on the teacher model T and student model S are as follows:

$$SupConLoss = \sum_{i \in I} SupConLoss_i = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp\left(z_i \cdot z_p / \tau\right)}{\sum_{a \in A(i)} exp(z_i \cdot z_a / \tau)} \tag{14}$$

Here, $i \in I \equiv \{1, \cdots, 2N\}$ is the index of the randomly augmented data, and $P(i) \equiv \{p \in A(i) : y_p = y_i\}$ is the set of indices for all positives in the batch (since y_p and y_i are labels of images that have been randomly augmented from y_l, they are the same as y_l). z_i and z_p are the embedding values of each randomly augmented image in $(RandAugment_a(x_l), RandAugment_b(x_l))$ passed through the encoder ENC, and $A(i) \equiv I \backslash \{i\}$, and τ is the temperature parameter. In other words, the inner product between positive pairs (i and p are the same class but different samples) is maximized through $exp\left(z_i \cdot z_p / \tau\right)$, and the inner product between negative pairs is minimized through $exp(z_i \cdot z_a / \tau)$, so that the SupConLoss is minimized.

The training process of the proposed SupCon-MPL model for deepfake detection is shown in Algorithm 1.

Algorithm 1 The Deepfake detection method based on SupCon-MPL (Pseudo code)

Set Labeled data, Unlabeled data with domain splitting [37].
Input: Labeled data x_l, y_l and unlabeled data x_u.
Initialize $\theta_T^{(0)}$ and $\theta_S^{(0)}$.
Pretrain Teacher model with x_l, y_l.
For $t = 0$ to $N - 1$ **do**
Sample an unlabeled example x_u and a labeled example x_l, y_l.
Sample a pseudo label $\hat{y}_u \sim P(\cdot|x_u; \theta_T)$.
Compute contrastive loss of student encoder ENC_S using the pseudo label $\hat{y}_u$:

$$g_{S,contrast}^{(t)} = \nabla_{\theta_T} SupConLoss(RandAugment_a(x_u), RandAugment_b(x_u), \hat{y}_u)|_{\theta_S=\theta_S^{(t)}}$$

Update the student encoder $\mathrm{ENC_S}$ using the pseudo label $\hat{y}$:

$$\theta_{S,ENC}^{(t+1)} = \theta_{S,ENC}^{(t)} - \eta_S \cdot g_{S,\,contrast}^{(t)}$$

Update the student classifier CLF_S using the pseudo label $\hat{y}_u$:

$$\theta_S^{(t+1)} = \theta_S^{(t)} - \eta s \cdot \nabla_{\theta_S} CE(\hat{y}_u, CLF_S(emb_u^S; \theta_S))|_{\theta_S=\theta_S^{(t)}}$$

Compute contrastive loss of teacher encoder ENC_T using the labeled data (x_l, y_l):

$$g_{T,contrast}^{(t)} = \nabla_{\theta_T} SupConLoss(RandAugment_a(x_l), RandAugment_b(x_l), y_l)|_{\theta_T=\theta_T^{(t)}}$$

Update the teacher encoder ENC_T using the labeled data (x_l, y_l):

$$\theta_{T,ENC}^{(t+1)} = \theta_{T,ENC}^{(t)} - \eta_S \cdot g_{T,\,contrast}^{(t)}$$

Compute gradient on labeled data (x_l, y_l):

$$g_{T,\,supervised}^{(t)} = \nabla_{\theta_T} CE(y_l, CLF_T(emb_l^T; \theta_T))|_{\theta_T=\theta_T^{(t)}}$$

Compute feedback factor h from student:

$$h = CE\left(y_l, S\left(x_l; \theta_S^{(t+1)}\right)\right) - CE\left(y_l, S\left(x_l; \theta_S^{(t)}\right)\right)$$

Compute MPL loss from unlabeled data x_l:

$$g_T^{(t)} = h \cdot \nabla_{\theta_T} CE(\hat{y}_u, CLF_T(emb_u^T; \theta_T))|_{\theta_T=\theta_T^{(t)}}$$

Update the teacher classifier CLF_T:

$$\theta_T^{(t+1)} = \theta_T^{(t)} - \eta s \cdot \left(g_T^{(t)} + g_{T,\,supervised}^{(t)}\right)$$

end for
return $\theta_{S,ENC}^{(N)}, \theta_S^{(N)}$ ▷ Only the student encoder and classifier are returned for evaluations.

4. Experiment

In this chapter, we first describe the experimental setup. Subsequently, we present the experimental results for Figure 2a,b in Sections 4.2 and 4.3, respectively. Our main results consist of comparisons with the pretrained model and SupCon model in Section 4.4 and comparisons with state-of-the-art models in Section 4.5.

4.1. Experiment Setup

The experiments are conducted using NVIDIA Tesla V100 32 and NVIDIA RTX-3090 (NVIDIA, Santa Clara, CA, USA) with ubuntu 20.04 environment for reproducibility and stability. The single-domain experiment and the multi-domain experiment are existing outputs of the Meta Learning-based Deepfake Detection Project [24]. The pretrained model, Meta Pseudo Labels model (MPL model), SupCon model, and SupCon-MPL model are experimented for their training performance under the same conditions and hyperparameters. The training dataset uses the videos of the deepfakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT), and real videos in FaceForensics++ [39], with DFDC [40], and Celeb-DF [41]. In scenario evaluation, deepfake videos of first-generation's unknown domain are NeuralTextures(NT) [39] with real videos, and for post-generation's unknown domain, we selected StyleGAN [8] images with CelebA [48] videos. We used MTCNN [49] to extract face images frame by frame of each video.

In the single-domain experiment, 260,000 real and 340,000 fake data from each domain in the FaceForensics++ [39] are used. During training, the amount of validation data used is 20% of the training data, and the evaluation dataset uses 150,000 per each data domain. In the multi-domain experiment, 200,000 labeled data are randomly extracted from four domains, and 180,000 unlabeled data are extracted from a single domain for use.

The generational deepfake scenario trains using 170,000 each of the first-generation known domain's FaceForensics++ (DF, F2F, FS, Real) [39], DFDC [40], Celeb-DF [41] data, and then evaluates using 51,200 each of first- and second-generation unknown domain data. In the backbone model in scenario evaluation, we used ResNet50 [50] due to lack of computational resources.

The hyperparameters used in the experiment are a learning rate of 10^{-4}, an image size of 64, and a batch size of 512. In the MPL, SupCon-MPL models, the batch sizes of labeled and unlabeled images are 64 and 448, respectively. Finally, the threshold is set at 0.95. The training models used were ResNet50 [50], ResNet101 [50], ResNext50 [51], EfficientNet-b5 [21], and WideResNet50 [52].

Experiment data and evaluation data use a mix of fake and real data. In the experiment in Section 4.2, video data from one domain is used as labeled and unlabeled data, and in the experiment in Section 4.3, videos from multiple domains are used as labeled data, and one domain is used as unlabeled data.

4.2. Single-Domain Experiment

In this section, single-domain experiments evaluate the detection performance for a single deepfake generation model, while, in Section 4.3, the performance on new deepfakes is evaluated through training on various deepfake generation models. Sections 4.4 and 4.5 assess the performance of deepfake detection across diverse deepfake datasets and scenario-based deepfake detection, respectively.

In the experiment using only one domain, as shown in Table 1, the performance of the known domain increased in most of the situation. Furthermore, it was confirmed that the performance in unknown domains also increased in most of the situation. Based on Efficient-Net-b5 [21] in Table 1, ACC and AUC in K improved by an average of 4.47% and 4.53%, respectively, and ACC in U improved by an average of 0.20%, but AUC decreased by an average of 0.13%. However, overall, out of a total of 64 ACC and AUC validations recorded in Tables 1–8, 44 and 41 cases improved, respectively. Through this, it was confirmed that when using the same data, the performance of the MPL model is higher than the pretrained model.

Table 1. The performance of pretrained and MPL model on the known domain (EfficientNetb5 [21]).

Baseline Model	Train Dataset	Test Dataset	Pretrained Model			MPL Model		
			ACC	AUC	F1 Score	ACC	AUC	F1 Score
EfficientNetb5 [21]	DF	DF	89.35	89.35	89.19	**90.08**	**90.08**	**90.36**
	F2F	F2F	77.21	77.21	77.61	**80.35**	**80.35**	**79.45**
	FS	FS	84.52	84.31	83.20	**87.90**	**87.43**	**85.98**
	NT	NT	64.13	63.33	58.18	**74.79**	**74.47**	**71.36**

Table 2. The performance of pretrained and MPL model on the unknown domain (EfficientNetb5 [21]).

Baseline Model	Train Dataset	Test Dataset	Pretrained Model			MPL Model		
			ACC	AUC	F1 Score	ACC	AUC	F1 Score
EfficientNetb5 [21]	DF	F2F	51.59	51.60	22.09	**51.70**	**51.70**	**23.01**
		FS	55.97	52.06	23.23	**56.45**	**52.65**	**25.01**
		NT	55.37	51.26	20.06	**55.41**	**51.37**	**21.28**
	F2F	DF	**57.56**	**57.55**	**47.78**	53.87	53.84	32.11
		FS	**56.47**	**54.05**	**39.66**	55.48	51.92	26.57
		NT	54.63	**51.91**	**33.96**	**54.76**	50.99	23.26
	FS	DF	57.25	57.22	**39.84**	**57.61**	**57.88**	36.54
		F2F	51.76	51.77	**25.78**	**51.98**	**51.99**	20.16
		NT	53.79	**49.89**	**20.80**	**54.32**	49.80	12.35
	NT	DF	58.72	58.71	52.73	**61.03**	**61.00**	**52.78**
		F2F	54.52	54.53	**45.23**	**55.88**	**55.89**	43.68
		FS	51.65	**49.42**	**34.03**	**53.29**	49.34	27.25

Table 3. The performance of pretrained and MPL model on the known domain (ResNet50 [50]).

Baseline Model	Train Dataset	Test Dataset	Pretrained Model			MPL Model		
			ACC	AUC	F1 Score	ACC	AUC	F1 Score
ResNet50 [50]	DF	DF	91.16	91.16	90.91	**91.29**	**91.30**	**90.97**
	F2F	F2F	82.47	82.48	81.84	**83.98**	**83.97**	**82.75**
	FS	FS	87.55	87.21	86.01	**88.31**	**88.09**	**87.06**
	NT	NT	74.96	74.28	71.20	**76.22**	**75.84**	**73.15**

Table 4. The performance of pretrained and MPL model on the unknown domain (ResNet50 [50]).

Baseline Model	Train Dataset	Test Dataset	Pretrained Model			MPL Model		
			ACC	AUC	F1 Score	ACC	AUC	F1 Score
ResNet50 [50]	DF	F2F	**52.91**	**52.89**	**22.41**	51.92	51.86	20.36
		FS	56.80	52.68	21.41	**57.02**	**53.00**	**22.63**
		NT	**55.91**	**51.77**	**18.54**	55.45	51.33	18.53
	F2F	DF	**53.94**	**53.90**	**31.72**	53.79	53.75	28.28
		FS	54.62	50.97	**24.47**	**55.32**	**51.35**	21.84
		NT	54.76	**51.20**	**24.07**	**54.96**	51.04	21.00
	FS	DF	56.34	56.30	33.89	**59.57**	**59.53**	**41.50**
		F2F	51.16	51.10	20.08	**51.57**	**51.52**	**21.01**
		NT	53.67	49.44	14.32	**53.81**	**49.61**	**15.49**
	NT	DF	**60.45**	**60.43**	49.93	60.17	60.15	**50.85**
		F2F	**54.74**	**54.70**	**38.88**	53.56	53.51	36.77
		FS	51.59	48.15	22.30	**51.84**	**48.51**	**25.20**

Table 5. The performance of pretrained and MPL model on the known domain (ResNet101 [50]).

Baseline Model	Train Dataset	Test Dataset	Pretrained Model			MPL Model		
			ACC	AUC	F1 Score	ACC	AUC	F1 Score
ResNet101 [50]	DF	DF	**91.16**	**91.16**	90.84	91.13	91.13	**90.85**
	F2F	F2F	81.41	81.41	80.27	**83.50**	**83.49**	**82.73**
	FS	FS	87.59	87.37	86.13	**87.75**	**87.66**	**86.39**
	NT	NT	74.17	74.56	73.16	**76.03**	**75.53**	**73.22**

Table 6. The performance of pretrained and MPL model on the unknown domain (ResNet101 [50]).

Baseline Model	Train Dataset	Test Dataset	Pretrained Model			MPL Model		
			ACC	AUC	F1 Score	ACC	AUC	F1 Score
ResNet101 [50]	DF	F2F	**52.69**	**52.63**	**21.22**	51.26	51.20	18.01
		FS	**56.90**	**52.73**	**20.42**	56.02	51.84	18.71
		NT	**55.93**	**51.74**	**18.33**	54.93	50.73	16.53
	F2F	DF	**53.63**	**53.59**	**30.44**	53.50	53.46	27.42
		FS	54.33	50.08	**22.82**	**54.86**	**50.85**	19.67
		NT	54.62	51.01	**23.30**	**55.17**	**51.28**	20.98
	FS	DF	57.04	57.70	37.00	**59.55**	**59.51**	**42.55**
		F2F	51.43	51.37	20.63	**51.84**	**51.79**	**24.60**
		NT	**53.66**	49.55	15.04	53.61	**49.59**	**17.59**
	NT	DF	**62.66**	**62.65**	**59.08**	60.16	60.14	49.99
		F2F	**52.69**	**52.63**	**47.55**	51.26	51.20	37.97
		FS	56.90	52.73	**29.98**	**56.02**	**51.84**	21.29

Table 7. The performance of pretrained and MPL model on the known domain (ResNext50 [51]).

Baseline Model	Train Dataset	Test Dataset	Pretrained Model			MPL Model		
			ACC	AUC	F1 Score	ACC	AUC	F1 Sore
ResNext50 [51]	DF	DF	90.29	90.29	90.01	**91.14**	**91.14**	**91.02**
	F2F	F2F	81.19	81.17	80.07	**82.87**	**82.85**	**82.11**
	FS	FS	87.04	86.77	85.44	**87.36**	**87.36**	**86.36**
	NT	NT	74.26	74.43	73.21	**76.32**	**75.78**	**73.30**

Table 8. The performance of pretrained and MPL model on the unknown domain (ResNext50 [51]).

Baseline Model	Train Dataset	Test Dataset	Pretrained Model			MPL Model		
			ACC	AUC	F1 Score	ACC	AUC	F1 Score
ResNext50 [51]	DF	F2F	**51.83**	**51.73**	**20.17**	51.66	51.56	18.92
		FS	56.57	52.53	21.48	**57.24**	**53.21**	**23.08**
		NT	**55.25**	**51.57**	**18.50**	54.72	50.50	15.47
	F2F	DF	**54.57**	**54.58**	**31.56**	54.46	54.47	31.25
		FS	54.31	50.38	18.21	**55.28**	**51.45**	**22.28**
		NT	54.67	50.89	21.12	**54.80**	**51.02**	**21.58**
	FS	DF	59.23	59.23	41.25	**60.33**	**60.33**	**46.22**
		F2F	52.02	51.93	23.43	**52.98**	**52.89**	**27.91**
		NT	53.31	49.22	15.05	**53.41**	**49.55**	**17.30**
	NT	DF	**60.58**	**60.58**	43.46	58.50	58.51	**44.82**
		F2F	**54.16**	**54.11**	30.69	52.46	52.38	**31.00**
		FS	49.17	46.78	19.36	**51.26**	**47.80**	**21.71**

4.3. Multi-Domain Experiment

In the multi-domain experiment, the combination of data is configured considering the actual deepfake situation. The situation is assumed to have $\boldsymbol{K}$ deepfake data from multiple domains and D_{UL} data from $\boldsymbol{U}$. Afterwards, the evaluation is conducted through the $\boldsymbol{U}$ data used as D_{UL}. Therefore, MPL trains with deepfake videos from multiple domains, and, after training, MPL experiments the ability to detect unlabeled data using unlabeled data. As a result of the experiment, ACC and AUC increased by an average of 1.59% and 1.26% in two models in Tables 9 and 10. Through this, it was confirmed that the MPL model improved the deepfake detection ability of the unknown model by learning with unlabeled data.

Table 9. The performance of pretrained and MPL model on the targeted unknown domain (ResNext50 [51]).

Baseline Model	Train Dataset	Unlabeled Dataset	Pretrained Model			MPL Model		
			ACC	AUC	F1 Score	ACC	AUC	F1 Score
ResNext50 [51]	F2F, FS, NT	DF	63.33	63.32	70.02	**66.23**	**66.22**	**70.35**
	DF, FS, NT	F2F	57.35	57.39	60.69	**58.04**	**58.08**	**61.64**
	DF, F2F, NT	FS	50.69	50.53	46.69	**51.81**	**51.54**	**47.22**
	DF, F2F, FS	NT	53.63	53.05	**47.65**	**53.89**	**53.06**	46.28

Table 10. The performance of pretrained and MPL model on the targeted unknown domain (WideResNet50 [52]).

Baseline Model	Train Dataset	Unlabeled Dataset	Pretrained Model			MPL Model		
			ACC	AUC	F1 Score	ACC	AUC	F1 Score
WideResNet50 [52]	F2F, FS, NT	DF	63.17	63.14	69.19	**66.03**	**66.00**	**70.15**
	DF, FS, NT	F2F	56.86	56.87	**60.76**	**57.60**	**57.60**	58.91
	DF, F2F, NT	FS	48.37	47.74	39.32	**52.86**	**51.13**	**41.60**
	DF, F2F, FS	NT	**54.22**	**53.43**	**47.37**	53.92	51.96	39.87

4.4. SupCon-MPL Experiment

The experiment uses the labeled data identically to the multi-domain experiment and evaluates according to each combination of known domain and unknown domain. At this time, Celeb-DF [41] is used as unlabeled data in the training of the SupCon-MPL model. As a result of the experiment, shown in Table 11, when evaluating FS data as an unknown domain compared with the SupCon model [35], ACC and AUC decreased, but in other validations, the performance of the SupCon-MPL model was similar or higher than the performance of the two models being compared. This shows that the SupCon-MPL model has been trained as a generalized detection model compared with the existing deepfake detection model.

Table 11. The performance of SupCon-MPL compared with the baseline model and SupCon model [35].

Baseline Model	Train Dataset	Unlabeled Dataset	Pretrained Model			SupCon Model [35]			SupCon-MPL(Ours) Model		
			ACC	AUC	F1 Score	ACC	AUC	F1 Score	ACC	AUC	F1 Score
ResNet50 [50]	FF (without DF)	DF (unknown)	64.24	64.27	**65.49**	62.88	62.84	58.55	**64.60**	**64.55**	61.60
		F2F + FS + NT (known)	70.56	71.36	83.36	75.44	75.52	83.01	**75.84**	**76.00**	**83.50**
	FF (without F2F)	F2F (unknown)	55.76	55.64	40.63	56.61	56.64	49.47	**58.74**	**58.77**	**51.52**
		DF + FS + NT (known)	77.26	76.89	81.66	75.54	75.54	84.13	**78.11**	**78.28**	**85.84**
	FF (without FS)	FS (unknown)	54.47	52.07	79.18	**55.75**	**53.68**	77.20	55.72	53.41	**79.29**
		DF + F2F + NT (known)	75.99	75.87	81.66	76.02	75.88	84.84	**77.22**	**77.12**	**85.84**
	FF (without NT)	NT (unknown)	56.71	**54.95**	62.66	56.58	54.31	66.27	**56.76**	54.02	**69.43**
		DF + F2F + FS (known)	77.39	77.41	81.66	79.09	79.26	84.13	**81.22**	**81.32**	**85.84**

4.5. Deepfake Scenario Experiment

In this section, we construct a training scenario for a deepfake detection model in the real world and train the model. The scenario involves training a deepfake model with first-generation deepfake data, then experimenting with the detection of first-generation deepfakes (NT) that were not participated while training, and post-generation (second-generation) deepfakes (StyleGAN) that are newly developed and unknown. Table 12 shows the training results of various models according to the scenario. As a result of the experiment, among various models, the SupCon-MPL model achieved highest performance in all scenario evaluations compared with other models.

Table 12. The performance of SupCon-MPL compared with other deepfake detection methods.

Model	Scenario Deepfakes (Known)	Current-Generation Deepfakes (Unknown)	Post-Generation Deepfakes (Unknown)
Tar [34]	52.40	44.62	49.96
DDT [31]	80.41	44.62	49.49
MPL [24]	79.82	56.53	43.16
SupCon [35]	79.01	56.66	47.77
SupCon-MPL (ours)	**81.40**	**57.85**	**51.90**

4.6. Limitations

The main goal of the SupCon-MPL is to enhance the deepfake detection performance in unknown domains using meta-learning. Therefore, in this paper, we conducted experiments by reconfiguring a limited deepfake dataset into scenarios.

The main limitation is related to computing resources. As the training in Section 4.3. and Section 4.4. was conducted using NVIDIA RTX-3090, only ResNet50 [50] could be used as the backbone model in SupCon-MPL, which uses two models. The ResNet50 [50] model currently stands out among various image classification models for its stability and consistently decent performance scores. Additionally, within the given resources, it can be utilized for training the SupCon-MPL model. Subsequent experiments are needed with various backbone models and larger image sizes through more computing resources.

The next limitation is that we could not find a verified deepfake dataset for the third generation and higher. Further experiments are needed through the corresponding dataset in the future.

5. Conclusions

With the development of various deepfake generative models, it has become important to develop a generalized deepfake detection model that guarantees the detection performance of unknown domain deepfakes, not just the data of the domain used for training. The proposed SupCon-MPL allows for training from unlabeled data and performs contrastive learning to enhance the detection performance of unknown deepfakes. This enables contrastive learning with a diverse range of data, and, during training, the teacher model and the student model infer information and provide feedback to each other, facilitating the student model to surpass the performance of the teacher model.

Our model improved all detection performances over other generalized deepfake detection models' known and unknown domains in deepfake scenario evaluation. Indeed, one of the significant features of SupCon-MPL is its ability to train models using a large amount of unlabeled data. This provides a method to enhance the model's performance utilizing the countless images and videos available on the internet and so on. Through this, our model enables the training of a generalized deepfake detection model and provides robust detection capability for real-world deepfakes. SupCon-MPL can contribute to the detection of newly developed deepfake generative models as they become increasingly difficult to distinguish from real images, especially with the development of various deepfake generative models.

Future research will focus on improving the detection performance of higher generation deepfake images/videos using these methods. Additionally, studies on reducing the training cost of SupCon-MPL will be conducted. The goal is to develop a more efficient and economical deepfake detection model through these efforts.

Author Contributions: Conceptualization, K.-H.M. and S.-H.L.; Methodology, K.-H.M., S.-Y.O. and S.-H.L.; Software, K.-H.M.; Validation, K.-H.M.; Resources, S.-Y.O.; Writing—original draft, K.-H.M.; Writing—review & editing, S.-H.L.; Supervision, S.-Y.O. and S.-H.L.; Project administration, K.-H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Dong-A University research fund.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Summary of notations.

Symbol	Definition
CRL	Contrastive Representation Learning.
MPL	Meta Pseudo Labels.
GAN	Generative Adversarial Networks.
VAE	Variational Auto Encoder.
$\boldsymbol{K}$	Known domain which involves a set of known deepfake generative models.
$\boldsymbol{U}$	Unknown domain which involves a set of unknown deepfake generative models.
$\boldsymbol{D_K}$	Deepfake datasets of the known domain which are labeled.
$\boldsymbol{D_U}$	Deepfake datasets of the unknown domain which are unlabeled.
K_i	i-th known deepfake generative model in known domain.
U_i	i-th unknown deepfake generative model in unknown domain.
x_i	Images of real and deepfake.
y_i	Labels of real and deepfake.
D_L	Labeled dataset used while training.
D_{UL}	Unlabeled dataset used while training.
T	Teacher model of SupCon-MPL.
S	Student model of SupCon-MPL.
θ_T	Parameters of the teacher model's classifier.
θ_S	Parameters of the student model's classifier.
θ_S^{PL}	Minimized parameters of model in SupCon-MPL.
$\theta_{T,ENC}^{(t)}$	Parameters of teacher encoder.
$\theta_{S,ENC}^{(t)}$	Parameters of student encoder.
ηs	Learning rate used while training.
ClF_T	Classifier in the teacher model.
ENC_T	Encoder in the teacher model.
CLF_S	Classifier in the student model.
ENC_S	Encoder in the student model.
x_l	Labeled images used while training.
x_u	Unlabeled images used while training.
y_l	Labels used while training.
$\hat{y}_u$	Pseudo label generated by the teacher model.
emb_l^T	Embedding value derived by passing the labeled data through the teacher encoder.
emb_u^S	Embedding value derived by passing the unlabeled data through the student encoder.
h	Student model's feedback factor.
I	Index of the randomly augmented data.
$P(i)$	Set of indices for all positives in the batch.
z_i, z_p	Embedding values of each randomly augmented image.

References

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9. [CrossRef]
2. Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784.
3. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797. [CrossRef]
4. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 214–223.
5. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of wasserstein gans. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.

6. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
7. Sohn, K.; Lee, H.; Yan, X. Learning structured output representation using deep conditional generative models. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9.
8. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410. [CrossRef]
9. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the ICML 2021 Workshop on Unsupervised Reinforcement Learning, Virtual, 18–24 July 2021; pp. 8821–8831.
10. DeepFaceLab. Available online: https://github.com/iperov/DeepFaceLab (accessed on 5 March 2024).
11. Deepswap. Available online: https://deepfaceswap.ai/ (accessed on 5 March 2024).
12. Synthesia. Available online: https://www.synthesia.io (accessed on 5 March 2024).
13. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695. [CrossRef]
14. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.L.; Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36479–36494.
15. Song, Y.; Dhariwal, P.; Chen, M.; Sutskever, I. Consistency models. *arXiv* **2023**, arXiv:2303.01469.
16. Li, Y.; Lyu, S. Exposing deepfake videos by detecting face warping artifacts. *arXiv* **2018**, arXiv:1811.00656.
17. Matern, F.; Riess, C.; Stamminger, M. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops, Waikoloa, HI, USA, 7–11 June 2019; pp. 83–92. [CrossRef]
18. Li, Y.; Chang, M.C.; Lyu, S. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *arXiv* **2018**, arXiv:1806.02877.
19. Ciftci, U.A.; Demir, I.; Yin, L. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Trans. Pattern Anal. Mach. Intell.* 2020, *early access*. [CrossRef]
20. Coccomini, D.A.; Messina, N.; Gennaro, C.; Falchi, F. Combining efficientnet and vision transformers for video deepfake detection. In Proceedings of the International Conference on Image Analysis and Processing, Lecce, Italy, 23–27 May 2022; pp. 219–229. [CrossRef]
21. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Los Angeles, CA, USA, 9–15 June 2019; pp. 6105–6114.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
23. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258. [CrossRef]
24. Moon, K.-H.; Ok, S.-Y.; Seo, J.; Lee, S.-H. Meta Pseudo Labels Based Deepfake Video Detection. *J. Korea Multimed. Soc.* **2024**, *27*, 9–21. [CrossRef]
25. Jain, A.; Korshunov, P.; Marcel, S. Improving generalization of deepfake detection by training for attribution. In Proceedings of the 2021 IEEE 23rd International Workshop on Multimedia Signal Processing, Tampere, Finland, 6–8 October 2021; pp. 1–6. [CrossRef]
26. Nadimpalli, A.V.; Rattani, A. On improving cross-dataset generalization of deepfake detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 91–99. [CrossRef]
27. Hsu, C.C.; Lee, C.Y.; Zhuang, Y.X. Learning to detect fake face images in the wild. In Proceedings of the 2018 International Symposium on Computer, Consumer and Control, Taichung, Taiwan, 6–8 December 2018; pp. 388–391. [CrossRef]
28. Dong, F.; Zou, X.; Wang, J.; Liu, X. Contrastive learning-based general Deepfake detection with multi-scale RGB frequency clues. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 90–99. [CrossRef]
29. Shiohara, K.; Yamasaki, T. Detecting deepfakes with self-blended images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18720–18729. [CrossRef]
30. Chen, L.; Zhang, Y.; Song, Y.; Wang, J.; Liu, L. Ost: Improving generalization of deepfake detection via one-shot test-time training. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24597–24610.
31. Aneja, S.; Nießner, M. Generalized zero and few-shot transfer for facial forgery detection. *arXiv* **2020**, arXiv:2006.11863.
32. Kim, M.; Tariq, S.; Woo, S.S. Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1001–1012. [CrossRef]
33. Qi, H.; Guo, Q.; Juefei-Xu, F.; Xie, X.; Ma, L.; Feng, W.; Zhao, J. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 4318–4327. [CrossRef]
34. Lee, S.; Tariq, S.; Kim, J.; Woo, S.S. Tar: Generalized forensic framework to detect deepfakes using weakly supervised learning. In Proceedings of the IFIP International Conference on ICT Systems Security and Privacy Protection, Oslo, Norway, 22–24 June 2021; pp. 351–366. [CrossRef]

35. Xu, Y.; Raja, K.; Pedersen, M. Supervised contrastive learning for generalizable and explainable deepfakes detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 379–389.
36. Fung, S.; Lu, X.; Zhang, C.; Li, C.T. DeepfakeUCL: Deepfake Detection via Unsupervised Contrastive Learning. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8. [CrossRef]
37. Pham, H.; Dai, Z.; Xie, Q.; Le, Q.V. Meta pseudo labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11557–11568. [CrossRef]
38. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
39. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Niessner, M. Faceforensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1–11. [CrossRef]
40. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The deepfake detection challenge (dfdc) dataset. *arXiv* **2020**, arXiv:2006.07397.
41. Li, Y.Z.; Yang, X.; Sun, P.; Qi, H.G.; Lyu, S. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In Proceedings of the Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3207–3216. [CrossRef]
42. Korshunov, P.; Marcel, S. Deepfakes: A new threat to face recognition? assessment and detection. *arXiv* **2018**, arXiv:1812.08685.
43. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; pp. 1735–1742. [CrossRef]
44. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
45. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
46. Xie, Q.; Luong, M.T.; Hovy, E.; Le, Q.V. Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10687–10698. [CrossRef]
47. Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; Le, Q. Unsupervised data augmentation for consistency training. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 6256–6268.
48. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3730–3738. [CrossRef]
49. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
51. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500. [CrossRef]
52. Zagoruyko, S.; Komodakis, N. Wide residual networks. *arXiv* **2016**, arXiv:1605.07146.