

## Article

# Separable Confident Transductive Learning for Dairy Cows Teat-End Condition Classification

Youshan Zhang <sup>1,\*</sup> , Ian R. Porter <sup>1</sup> , Matthias Wieland <sup>2</sup>  and Parminder S. Basran <sup>1</sup> 

<sup>1</sup> Department of Clinical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA; irp5@cornell.edu (I.R.P.); psb92@cornell.edu (P.S.B.)

<sup>2</sup> Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA; mjw248@cornell.edu

\* Correspondence: yz945@cornell.edu

**Simple Summary:** The health of dairy cows is important for milk quality and the health of the mammary gland. Traditionally, teat-end health has been assessed manually through visual inspection of teat-end callosity thickness and roughness (i.e., hyperkeratosis), which is a risk-factor for mastitis. Here, we describe a computer-vision approach to replace the time-consuming and expensive manual assessment of teat-end hyperkeratosis. Using separable confident transductive learning, a convolutional neural network is trained with the goal of increasing the feature differences in the images of teat-ends with different classifications of hyperkeratosis. When compared with the traditional approach of transfer learning of a convolution neural network for classifying the extent of hyperkeratosis, the overall accuracy of our model increased from 61.8 to 77.6%. This substantial improvement in accuracy renders the possibility of using image-based machine learning to routinely monitor hyperkeratosis on commercial dairy farm settings.

**Abstract:** Teat-end health assessments are crucial to maintain milk quality and dairy cow health. One approach to automate teat-end health assessments is by using a convolutional neural network to classify the magnitude of teat-end alterations based on digital images. This approach has been demonstrated as feasible with GoogLeNet but there remains a number of challenges, such as low performance and comparing performance with different ImageNet models. In this paper, we present a separable confident transductive learning (SCTL) model to improve the performance of teat-end image classification. First, we propose a separation loss to ameliorate the inter-class dispersion. Second, we generate high confident pseudo labels to optimize the network. We further employ transductive learning to narrow the gap between training and test datasets with categorical maximum mean discrepancy loss. Experimental results demonstrate that the proposed SCTL model consistently achieves higher accuracy across all seventeen different ImageNet models when compared with retraining of original approaches.

**Keywords:** transductive learning; dairy cows; teat-end health assessments



**Citation:** Zhang, Y.; Porter, I.R.; Wieland, M.; Basran, P.S. Separable Confident Transductive Learning for Dairy Cows Teat-End Condition Classification. *Animals* **2022**, *12*, 886. <https://doi.org/10.3390/ani12070886>

Academic Editor: Kiro Petrovski

Received: 1 March 2022

Accepted: 26 March 2022

Published: 31 March 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

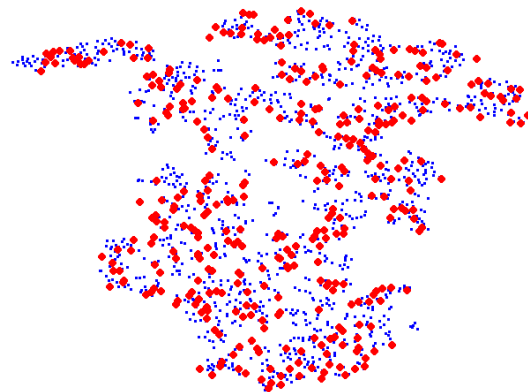
Mastitis remains one of the most frequently occurring diseases in dairy cows, often arising from intramammary infections by way of the teat canal. Machine milking can affect teat canal integrity and lead to increased teat-end callosity, which can increase the risk of bacterial infections of the mammary gland [1]. Frequent monitoring of teat-end callosity is critical for a mastitis prevention program [2]. However, cow-side manual assessments of teat-health, which is the current best practice, is time-consuming and suffers from inter- and intra-rater variability [3]. Another challenge is the inability to assess the entire herd in large dairy farms. To address some of these challenges, deep learning (DL) has been proposed where GoogLeNet transfer learning was used to classify the extent of hyperkeratosis using

a four-level classification scheme [4]. The overall accuracy of this approach was 46.7–61.8%, suggesting feasibility but, as of yet, insufficient accuracy to be useful as a clinical decision tool. As shown in Figure 1 with a t-SNE map, the training (red) and test (blue) data are observed as mixed together after retraining GoogLeNet and unable to discriminate the four classes. The indistinct boundaries of the classes lead to another challenge to improve the performance of the teat-end condition classification problem.

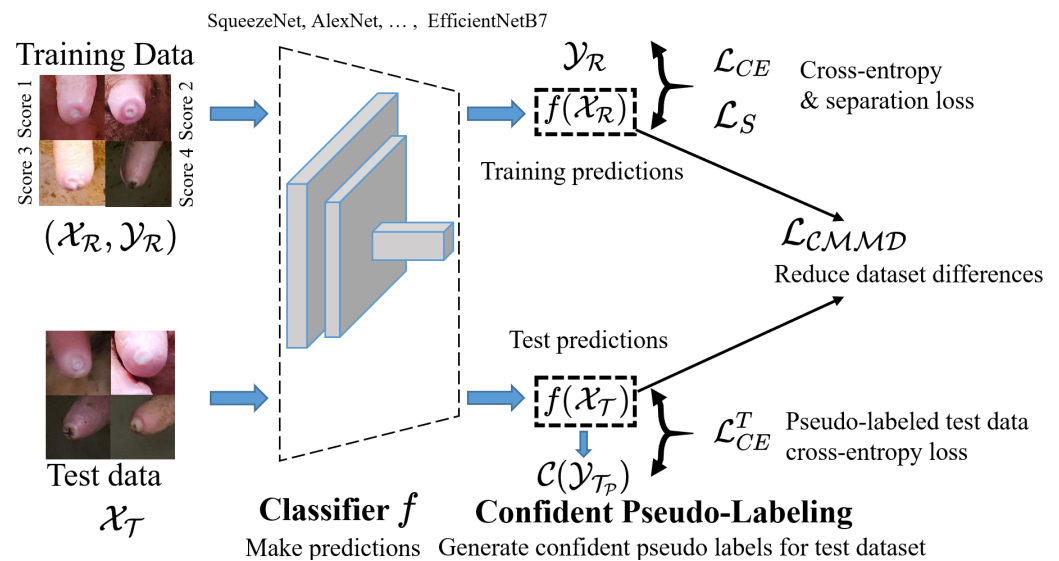
In this paper, we propose a new paradigm that yields a substantial improvement in accuracy of teat-end image classification while retaining the flexibility and accessibility of commonly used ImageNet classifiers such as AlexNet [5], GoogLeNet [6], Xception [7], and NasNetLarge [8]. To address the aforementioned challenges, we aggregate four different loss functions in one framework: classification loss, separation loss, pseudo labeled test data classification loss, and categorical maximum mean discrepancy (MMD) loss. As shown in Figure 2, using these proposed novel loss functions, our model can realize the inter-class dispersion and intra-class compactness. This paper provides three specific contributions:

1. We propose a novel separable confident transductive learning model (SCTL) to improve accuracy for the teat-end image classification. To improve the discrimination of different classes, we first propose a separation loss to enlarge the dissimilarity between different categories.
2. We develop a pseudo labeling adjustment learning paradigm to continuously generate high confidence examples for the test data and further optimize the network with test data information.
3. We narrow the gap between intra-class differences between training and test data with transductive learning by minimizing categorical MMD loss and further align the condition distribution between training and test data.

In this study, we performed experiments with seventeen benchmark ImageNet models by optimizing these loss functions and increasing the way that differences are detected between the images. The accuracy of our SCTL model with GoogLeNet is increased from 61.8 to 77.6%. This substantial increase in accuracy may render image-based hyperkeratosis classifications feasible on commercial dairy farm settings.



**Figure 1.** t-SNE [9] view of training (blue) and test (red) dataset from a retrained GoogLeNet [4]. Different categories are mixed together after training.



**Figure 2.** The learning scheme of our proposed SCTL model. We first fine-tune the classifier  $f$  from seventeen well-known ImageNet models and then make predictions for the training ( $\mathcal{X}_R$ ) and test datasets ( $\mathcal{X}_T$ ). For the training data, we minimize the typical cross-entropy loss ( $\mathcal{L}_{CE}$ ) and the separation loss ( $\mathcal{L}_S$ ) to improve the inter-class dispersion. For the test data, we generate confident pseudo labeled examples ( $\{\mathcal{C}(\mathcal{X}_T), \mathcal{C}(\mathcal{Y}_{T_p}^t)\}$ ) in the  $t$  adjustment learning, and then we minimize the pseudo labeled test data cross-entropy loss ( $\mathcal{L}_{CE}^T$ ). To reduce the dataset differences, we also develop a categorical maximum mean discrepancy loss  $\mathcal{L}_{CMMD}$  to improve intra-class compactness.

## 2. Related Work

### 2.1. Teat-End Classification

In the dairy industry, mastitis, which is an inflammation of one or more of the cow's mammary glands, is a frequently occurring disease that affects dairy cow health and milk quality. Mechanical stresses on the cow's teat-end can evoke circulatory changes to it and, over the course of several weeks, can result in increased teat-end callosity thickness and roughness [10]. These changes to the teat-end can increase the risk of pathogenic bacteria infiltrating the cow's udders. To monitor and reduce these risks, regular inspections of the dairy cow's teat-end health is recommended [11]. This is achieved by manually inspecting the teat-ends of at least 20% of the cows in the herd [12]; however, herd-level assessments are time-consuming, expensive, imprecise, and subjective [3]. Four classes scoring of hyperkeratosis is a standard classification that is usually used in cow teat-end classification (Score 1: no ring; Score 2: smooth ring; Score 3: rough ring; Score 4: very rough ring) [13].

Transfer learning has been applied in various computer vision tasks whose performance relies on the diversity of image data and fine-tuning of network parameters. With the invention of different ImageNet models, transfer learning has been widely adapted in image classification, objection detection, and segmentation problems. Porter et al. [4] recently described a machine-learning approach where images of teat-ends could be used to train a convolution neural network (CNN), such as GoogleNet, to classify the extent of teat-end hyperkeratosis using a four-point scoring system. Although the overall accuracy of our original approach on test data showed promise of automatic teat-end assessment, the accuracy was relatively low, which highlighted the need for improvement.

### 2.2. Transductive Learning

Transductive learning (TL) is a process that trains both labeled training data and unlabeled test data [14,15]. It is generally used in semi-supervised learning scenarios. Different from frequently used supervised inductive classification, which aims to train a classification model based on the labeled training data to approximate test data class distribution, the goal of transductive learning is to find an admissible function using the unlabeled data to improve classification performance [16]. The key idea of TL is that the

predicted labels for the test samples are viewed as optimization variables, which can be iteratively updated in the training process [17]. When TL is applied, it is often assumed that the training and test sets share a similar distribution [18]. As shown in Figure 1, the training and test sets are not separated, suggesting they are sampled from the same distribution. This observation has motivated us to employ TL for teat-end image classification and, as of yet, has not yet been explored in our specific problem.

### 2.3. Pseudo Labeling

The purpose of pseudo labeling is to seek the generation of labels or pseudo labels for unlabeled data to guide the learning process [19]. Pseudo labeling typically generates pseudo labels for the unlabeled data either based on hard assigned labels (the predictions from neural network [17,20]) or the predicted class probability [21–23]. Under such a regime, label information from unlabeled data can be included during training. In deep networks, the classifier from the training data is usually treated as an initial pseudo labeler to generate the pseudo labels for the test data (and use them as if they were real labels). There are several algorithms for obtaining pseudo labels and promote the performance of unlabeled data. Xie et al. [22] proposed a Moving Semantic Transfer Network (MSTN) to develop semantic matching and domain adversary losses to obtain pseudo labels. Iscen et al. [24] assigned pseudo labels to unlabeled samples based on neighborhood graphs. Zhang et al. [25] offer a label propagation with augmented anchors method to improve label propagation via the generation of unlabeled virtual samples with label prediction. Haase et al. [26] trained re-initialized networks and unlabeled datasets on each partition. The trained networks were used to filter the labels for training the newer networks. However, most of their experiments are conducted based on noisy data. Although previous pseudo labeling approaches are general and domain-agnostic, they tend to underperform since noisy pseudo labeled samples degrade model performance. In addition, most pseudo labeling methods employ a two-stage paradigm. The pseudo labels in the first stage (using the trained training data classifier) are generated and then used to train the model along with the labeled training data in the second stage. Our work differs from these approaches by generating high confidence examples with adjustment learning using a novel scheme, which allows for competitive results for teat-end image classification.

## 3. Methods

### 3.1. Motivation

The scientific goal of our paper is to develop a fully automated deep learning model that can accurately identify different categories of dairy cow teat-end conditions. The utilitarian goal is to detect hyperkeratosis of the teat-end area (Score 3 and Score 4) in the commercial dairy farm setting. Our research problem is the teat-end image classification task, and we aim to improve the classification accuracy using transductive learning and pseudo labeling.

### 3.2. Problem

Let  $\mathcal{D}$  be a dataset and subscripts  $\mathcal{R}$  and  $\mathcal{T}$  refer to training or testing subsets of the data. Image classification can be formulated as the problem of learning a classifier  $f$  from a set of training data,  $\mathcal{D}_{\mathcal{R}} = \{(\mathcal{X}_{\mathcal{R}}^i, \mathcal{Y}_{\mathcal{R}}^i)\}_{i=1}^{\mathcal{N}_{\mathcal{R}}}$ , where  $y_i$  is the ground-truth label in  $C$  categories corresponding to  $x_i$ , and  $\mathcal{N}_{\mathcal{R}}$  is the number of samples in the training dataset. In our setting,  $f$  is a classifier from the CNN. The goal of a vanilla image classification problem is to improve the accuracy of the unlabeled  $\mathcal{T}$  dataset examples:  $\mathcal{D}_{\mathcal{T}} = \{(\mathcal{X}_{\mathcal{T}}^j, \mathcal{Y}_{\mathcal{T}}^j)\}_{j=1}^{\mathcal{N}_{\mathcal{T}}}$ . However, due to the diversity of the datasets and fuzzy differences between different categories, the accuracy of test samples remains difficult to improve.

### 3.3. Transfer Learning

With the emergence of different ImageNet models, fine-tuning one of the ImageNet models with transfer learning is often applied in classifying new datasets. The parameters

of these different ImageNet models are fit by optimizing a typical categorical cross-entropy (CE) loss function  $\mathcal{L}$

$$\mathcal{L}_{CE} = -\frac{1}{N_{\mathcal{R}}} \sum_{i=1}^{N_{\mathcal{R}}} \sum_{c=1}^C \mathcal{Y}_{\mathcal{R}_c}^i \log(f_c(\mathcal{X}_{\mathcal{R}}^i)), \quad (1)$$

where  $\mathcal{Y}_{\mathcal{R}_c}^i \in [0, 1]^C$  is the binary indicator of each class  $c$  in true label for observation  $\Phi(\mathcal{X}_{\mathcal{S}}^i)$ , and  $f_c(\mathcal{X}_{\mathcal{R}}^i)$  is the predicted probability of class  $c$  using classifier  $f$ .

### 3.4. Separation Loss

As shown in Figure 1, different classes of training and test datasets are mixed together. The decision boundary for the trained network remains fuzzy, leading to poor model performance and low accuracy of test-end classification. Hence, it is necessary to improve the discrimination between different classes.

The purpose of a new separation loss function,  $\mathcal{L}_S$ , is to improve the inter-class dispersion so that the boundaries between different categories can be separable, and samples in the same categories can be more closely associated with each other. The core part of separation loss is to reduce the similarity between different classes. Since the network is trained using batch-wise samples, we inevitably encounter situations where the number of samples in different classes are imbalanced. We hence calculate the covariance matrix of the output of each categories' samples and then minimize the structural similarity [27] between each two categories' covariance matrix as follows.

$$\mathcal{L}_S = \frac{1}{\sum_{i=1}^{C-1} \sum_{j=i+1}^C} \sum_{i=1}^{C-1} \sum_{j=i+1}^C |SSIM(COV(c_i(f(B(\mathcal{X}_{\mathcal{R}}))))), COV(c_j(f(B(\mathcal{X}_{\mathcal{R}}))))|, \quad (2)$$

where  $B$  represents batch-wise data,  $c_{i/j}$  generates the categorical output by  $c_{i/j}(f(B(\mathcal{X}_{\mathcal{R}}))) = f(B(\mathcal{X}_{\mathcal{R}}^{\mathcal{Y}_{\mathcal{R}}=i/j}))$ .  $COV$  calculates the covariance matrix of categorical features as in Equation (3) and  $|\cdot|$  takes the absolute value to accelerate the convergence.

$$COV = \frac{1}{N_B} \sum_{z=1}^{N_B} (B_{\mathcal{Z}}^z - \mu_{\mathcal{Z}})(B_{\mathcal{Z}}^z - \mu_{\mathcal{Z}})^T, \quad \text{where } \mu_{\mathcal{Z}} = \frac{1}{N_B} \sum_{z=1}^{N_B} B_{\mathcal{Z}}^z \quad (3)$$

where  $\mu_{\mathcal{Z}}$  is the data mean and  $B_{\mathcal{Z}}$  is either  $c_i(f(B(\mathcal{X}_{\mathcal{R}})))$  or  $c_j(f(B(\mathcal{X}_{\mathcal{R}})))$ . The SSIM can be computed in Equation (4).

$$SSIM(B_1, B_2) = \frac{(2\mu_{B_1}\mu_{B_2} + C_1)(2\sigma_{B_1B_2} + C_2)}{(\mu_{B_1}^2 + \mu_{B_2}^2 + C_1)(\sigma_{B_1}^2 + \sigma_{B_2}^2 + C_2)} \quad (4)$$

where  $B_1 = COV(c_i(f(B(\mathcal{X}_{\mathcal{R}}))))$  and  $B_2 = COV(c_j(f(B(\mathcal{X}_{\mathcal{R}}))))$  are batch-wise features,  $\mu_{B_1}, \mu_{B_2}, \sigma_{B_1}, \sigma_{B_2}$ , and  $\sigma_{B_1B_2}$  are mean, standard deviations of domain invariant and specific features batch, and cross-covariance for  $(B_1, B_2)$ .  $C_1$  and  $C_2$  are two variables to stabilize the division with weak denominator. This loss function is derived from structural similarity index measure (SSIM) [27]. It has the advantages of measuring luminance, contrast, and structural difference between  $B_1$  and  $B_2$ . Therefore,  $\mathcal{L}_S$  has more capability of measuring the similarity between any two different categorical samples. In addition, the range of the  $\mathcal{L}_S$  is from 0 to 1, where 1 indicates high similarity between batch features and 0 means they are not similar.

During the training, minimizing  $\mathcal{L}_S$  can lead to the minimal similarity between each of the two categories. Hence, it can achieve the inter-class dispersion.

### 3.5. Confident Pseudo Labeling

By combining separation loss with cross-entropy loss, we can improve the discrimination of classifier  $f$  using training dataset. To improve the performance of the test dataset, we leverage transductive learning to mitigate the difference between the training and

test datasets. Transductive learning can train both labeled training data and test samples (without true labels); hence, the difference between them can be minimized [15].

To obtain knowledge from the test dataset, we first generate confident pseudo labels. Previous work either utilized hard pseudo labels or predicted class probability. In contrast to previous approaches, we aim to continuously train the new confident pseudo labeled test data. In this stage, we also take advantage of the initial training classifier  $f$  to generate initial pseudo labels and examples for the test data. We define a confident pseudo label in the following equation.

$$\mathcal{C}(\mathcal{Y}_{\mathcal{T}_p}^j) = \arg \max_{c \in \mathcal{C}} \{f_c(\mathcal{X}_{\mathcal{T}}^j)\} \text{ if } \max(f_c(\mathcal{X}_{\mathcal{T}}^j)) > p, \quad (5)$$

where  $\mathcal{C}$  represents confidence.  $\mathcal{C}(\mathcal{Y}_{\mathcal{T}_p}^j)$  is the confident label and  $\mathcal{C}(\mathcal{X}_{\mathcal{T}}^j)$  is its corresponding confident sample. Here,  $f_c(\mathcal{X}_{\mathcal{T}}^j)$  is the predicted probability in class  $c$  given the observation  $\mathcal{X}_{\mathcal{T}}^j$ .  $\max(\cdot)$  takes the dominant class probability, and it is higher than the threshold  $p$ , and  $p$  is between 0 and 1. The confident samples and their confident labels are able to push the decision boundary of classifier  $f$  toward the test dataset.

We can construct a pseudo label test domain  $\mathcal{D}_p = \{\mathcal{X}_{\mathcal{P}}^n, \mathcal{Y}_{\mathcal{P}}^n\}_{n=1}^{\mathcal{N}_p}$ , which consists of confident test examples with its confident pseudo labels, where  $\mathcal{N}_p \leq \mathcal{N}_{\mathcal{T}}$ ,  $\mathcal{X}_{\mathcal{P}} = \mathcal{C}(\mathcal{X}_{\mathcal{T}})$  and  $\mathcal{Y}_{\mathcal{P}} = \mathcal{C}(\mathcal{Y}_{\mathcal{T}_p})$ , and  $\mathcal{N}_p$  is controlled by  $p$ .  $\mathcal{N}_p = 0$  if  $p = 1$ , and  $\mathcal{N}_p = \mathcal{N}_{\mathcal{T}}$  if  $p = 0$ .

However, this pseudo labeling method generates confident pseudo labels with only a single high probability. The classifier  $f$  can be updated in the early stages of training but may not be able to train more examples on successive iterations since all high probability samples are treated as confident samples. Therefore, we propose to continuously generate confident examples in  $T$  times adjustment learning so that the classifier  $f$  could be updated in each adjustment learning. In adjustment learning, the pseudo label test domain becomes:  $\mathcal{D}_{\mathcal{P}^t}^t = \{\mathcal{X}_{\mathcal{P}^t}^n, \mathcal{Y}_{\mathcal{P}^t}^n\}_{n=1}^{\mathcal{N}_{\mathcal{P}^t}}$ , where  $\mathcal{X}_{\mathcal{P}^t} = \mathcal{C}(\mathcal{X}_{\mathcal{T}^t})$  and  $\mathcal{Y}_{\mathcal{P}^t} = \mathcal{C}(\mathcal{Y}_{\mathcal{T}_p^t})$ , and  $t$  is between  $[1, T]$ . To remove noisy pseudo labels of the predicted target domain in every  $t$ , we set the number of  $t$ -th updated domain  $\mathcal{N}_{\mathcal{P}^t}$  is not larger than the target domain sample size  $\mathcal{N}_{\mathcal{T}}$ , which means  $\mathcal{N}_{\mathcal{P}^t} \leq \mathcal{N}_{\mathcal{T}}$ .

$$\mathcal{C}(\mathcal{Y}_{\mathcal{T}_p^t}^j) = \arg \max_{c \in \mathcal{C}} \{f_c(\mathcal{X}_{\mathcal{T}^t}^j)\} \text{ if } \max(f_c(\mathcal{X}_{\mathcal{T}^t}^j)) > p_t. \quad (6)$$

In addition,  $\mathcal{C}(\mathcal{Y}_{\mathcal{T}_p^t}^j)$  is updated using Equation (6) with probability threshold  $p_t$  of every  $t$ , it also meets the requirements ( $p_{t+1} \leq p_t$  and  $0 \leq p_t \leq 1$ ), and we could obtain confident examples and pseudo labels during each  $t$ -th iteration and the classifier  $f$  will lean toward the test data. In  $T$  times iterations, we then form a set of probability threshold as  $p_T = \{p_t\}_{t=1}^T$ . This approach produces confident examples and pseudo labels in each recurrent training interval.

During training, the constructed pseudo labeled test data domain  $\mathcal{D}_{\mathcal{P}}^t$  will keep optimizing the trained classifier  $f$  after minimizing cross-entropy loss and separation loss functions. The pseudo labeled test data are also minimized by the cross-entropy loss. Therefore, the loss function for  $\mathcal{D}_{\mathcal{P}}^t$  in each training iteration is given by:

$$\mathcal{L}_{CE}^T = -\frac{1}{\mathcal{N}_{\mathcal{P}^t}} \sum_{n=1}^{\mathcal{N}_{\mathcal{P}^t}} \sum_{c=1}^{\mathcal{C}} \mathcal{Y}_{\mathcal{P}^t}^n \log(f_c(\mathcal{X}_{\mathcal{P}^t}^n)), \quad (7)$$

where  $\mathcal{N}_{\mathcal{P}^t}$  is the number of confident samples of  $t$ -th adjustment learning.  $\mathcal{Y}_{\mathcal{P}^t}^n \in [0, 1]^{\mathcal{C}}$  is the confident pseudo labeled binary indicator of each class  $c$  for the confident sample  $\mathcal{X}_{\mathcal{P}^t}^n$  in the  $t$ -th adjustment training, and  $f_c(\mathcal{X}_{\mathcal{P}^t}^n)$  is also the predicted probability of each class  $c$  given the input of confident sample  $\mathcal{X}_{\mathcal{P}^t}^n$ .

### 3.6. Categorical Maximum Mean Discrepancy

The proposed confident pseudo labeling process can optimize the network parameters, and it is not necessary to minimize the differences between the training and test data. To reduce the discrepancy between training and test data, we also compute the maximum mean discrepancy (MMD) loss [28], which is a frequently used distance-based loss function that reduces the divergence between the training and test data. However, MMD loss in conventional form focuses on only the marginal distribution alignment, which is more suitable for large domain divergence problems. As shown in Figure 1, the training and test data overlap, suggesting the marginal distribution alignment is not important for these cow teat images. Due to the fuzzy boundaries between different categories, conditional distribution alignment is required. Hence, we propose a categorical MMD (CMMD) loss, which attempts to align the conditional distribution of each category of training and test data.

$$\mathcal{L}_{CMMD} = \frac{1}{C} \sum_{c=1}^C \left( \frac{1}{N_{\mathcal{R}^c}^2} \sum_{i,j}^{N_{\mathcal{R}^c}} \kappa(L_{\mathcal{R}^c}^i, L_{\mathcal{R}^c}^j) + \frac{1}{N_{\mathcal{P}^c}^2} \sum_{i,j}^{N_{\mathcal{P}^c}} \kappa(L_{\mathcal{T}^c}^i, L_{\mathcal{T}^c}^j) - \frac{2}{N_{\mathcal{R}^c} \cdot N_{\mathcal{P}^c}} \sum_{i,j}^{N_{\mathcal{R}^c}, N_{\mathcal{P}^c}} \kappa(L_{\mathcal{R}^c}^i, L_{\mathcal{P}^c}^j) \right), \tag{8}$$

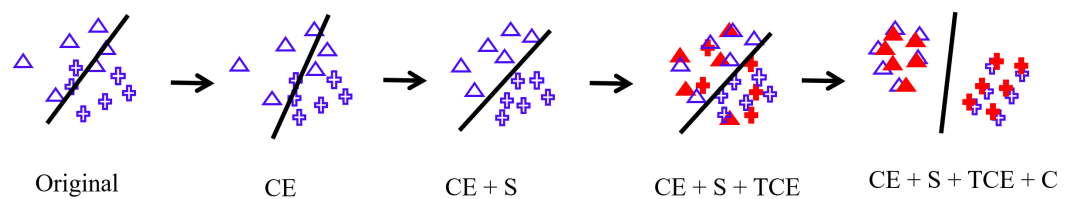
where  $N_{\mathcal{R}^c}$  and  $N_{\mathcal{P}^c}$  are the number of samples in each class of training and confident pseudo labeled test data,  $L_{\mathcal{R}^c} = f(\mathcal{X}_{\mathcal{R}^c})$ , and  $L_{\mathcal{T}^c} = f(\mathcal{X}_{\mathcal{P}^c})$ .  $\mathcal{X}_{\mathcal{R}^c}$  and  $\mathcal{X}_{\mathcal{P}^c}$  are categorical samples. This proposed CMMD loss measures the discrepancy between training and test datasets.

### 3.7. SCTL Model

The framework of our proposed SCTL model is depicted in Figure 2. Combining all loss functions, our model minimizes the following objective function:

$$\mathcal{L}(\mathcal{X}_{\mathcal{R}}, \mathcal{Y}_{\mathcal{R}}, \mathcal{X}_{\mathcal{T}}) = \arg \min(\mathcal{L}_{CE} + \alpha \mathcal{L}_S + \beta \mathcal{L}_{CE}^T + \gamma \mathcal{L}_{CMMD}) \tag{9}$$

where  $\mathcal{L}_{CE}$  is the source classification loss,  $\mathcal{L}_S$  is the separation loss, and  $\mathcal{L}_{CMMD}$  minimizes the categorical distance between training and test data.  $\mathcal{L}_{CE}^T$  is cross-entropy loss for confident pseudo labeled test data.  $\alpha$ ,  $\beta$ , and  $\gamma$  are three trade-off parameters. Figure 3 shows a toy example of our SCTL model. The overall training algorithm is shown in Algorithm 1.



**Figure 3.** A toy example of our SCTL learning paradigm. The blue color is the training data, and the red color is the test data. “Original” is the binary classification problem, “CE” refers to performing cross-entropy loss in training data, “CE + S” minimizes the proposed separation to enlarge differences between the two classes. “CE + S + TCE” can additionally minimize the pseudo labeled test data using the cross-entropy loss. “CE + S + TCE + C” adds another categorical maximum mean discrepancy loss to reduce the divergence between training and test data and form our SCTL model.

**Algorithm 1** Separable Confident Transductive Learning Network.  $B(\cdot)$  denotes the mini-batch sets,  $I$  is the number of iterations.  $p_T = \{p_t\}_{t=1}^T$ , and  $T$  is the number of adjustment learning.  $iter_t$  is the  $t$ -th adjustment learning.

- 1: **Input:** labeled training data  $\mathcal{D}_{\mathcal{R}} = \{(\mathcal{X}_{\mathcal{R}}^i, \mathcal{Y}_{\mathcal{R}}^i)\}_{i=1}^{\mathcal{N}_{\mathcal{R}}}$  and unlabeled test data  $\mathcal{D}_{\mathcal{T}} = \{\mathcal{X}_{\mathcal{T}}^j\}_{j=1}^{\mathcal{N}_{\mathcal{T}}}$
- 2: **Output:** predicted test labels
- 3: **repeat**
- 4:   Derive batch-wise data  $(B(\mathcal{X}_{\mathcal{R}}), B(\mathcal{Y}_{\mathcal{R}}))$  and  $B(\mathcal{X}_{\mathcal{T}})$  from  $\mathcal{D}_{\mathcal{R}}$  and  $\mathcal{D}_{\mathcal{T}}$
- 5:   **for**  $iter = 1$  to  $I$  **do**
- 6:     Train classifier  $f$  using Equations (1) and (2)
- 7:     **if**  $iter = iter_t$  **then**
- 8:       Get  $p_t$
- 9:     **end if**
- 10:    Generate confident pseudo test labels  $\mathcal{C}(\mathcal{Y}_{\mathcal{T}_p}^t)$  using Equation (6)
- 11:    Optimize  $f$  using Equation (7)
- 12:    Minimize the differences between training and test data using Equation (8)
- 13:    Minimize overall loss with Equation (9)
- 14:   **end for**
- 15: **until** converged
- 16: Make prediction for test samples based on trained classifier  $f$

## 4. Experiments

### 4.1. Datasets

We utilize the dataset from [4]. A total of 398 digital images of dairy cows on two commercial New York dairy farms were obtained: farm A milked approximately 1600 Holstein cows in a 60-stall rotary parlor, and farm B milked approximately 4000 Holstein cows in a 100-stall rotary parlor. Thus, our dataset includes 398 dairy cows, and a total of 1529 teat images were extracted in four categories (Score 1, Score 2, Score 3, and Score 4). A total of 380 teat images (around 70 cows) were utilized for the test dataset. For a fair comparison of different algorithms, we split the dataset into training (75%, 1149 images, around 288 cows) and test (25%, 380 images, around 70 cows) datasets. All results are reported based on the test dataset. Table 1 shows the statistics of the teat-end images dataset. Scores 3 and 4 are not common compared with Scores 1 and 2.

**Table 1.** Statistics of training and test data.

Label	Training		Test	
	Number	Percentage	Number	Percentage
Score 1	450	39.2	149	39.2
Score 2	491	42.7	163	42.9
Score 3	187	16.3	62	16.3
Score 4	21	1.8	6	1.6

### 4.2. Implementation Details

As shown in Figure 2, we utilize seventeen different ImageNet models as the backbone network during the training. The parameters during the training are epochs (100), batch size (16), learning rate ( $3 \times 10^{-5}$ ), ( $\alpha = 0.3$ ), ( $\beta = 1$ ), ( $\gamma = 0.5$ ), ( $T = 3$ ), and  $p_t = \{0.9, 0.8, 0.5\}$ . We report the accuracy of test dataset by:  $Accuracy = \frac{1}{\mathcal{N}_{\mathcal{T}}} \sum_{i=1}^{\mathcal{N}_{\mathcal{T}}} (\mathcal{Y}_{\mathcal{T}_p}^i == \mathcal{Y}_{\mathcal{T}}^i) \times 100$ . We also compare our results with [4] and conduct an ablation study to show the effect of different loss functions on classification accuracy. Since four categories are unbalanced, we also assigned the weight to each class according to the number of each category in the training dataset, and the assigned weights are [0.71, 0.65, 1.70, 15.17] for four categories, respectively. We implement our approach using PyTorch (version 1.7.1, CUDA version: 11.1). The model



has trained on a Dell Latitude 7420 laptop (Windows 10) with 16 GB RAM using GeForce 1080 Ti GPU.

#### 4.3. Results

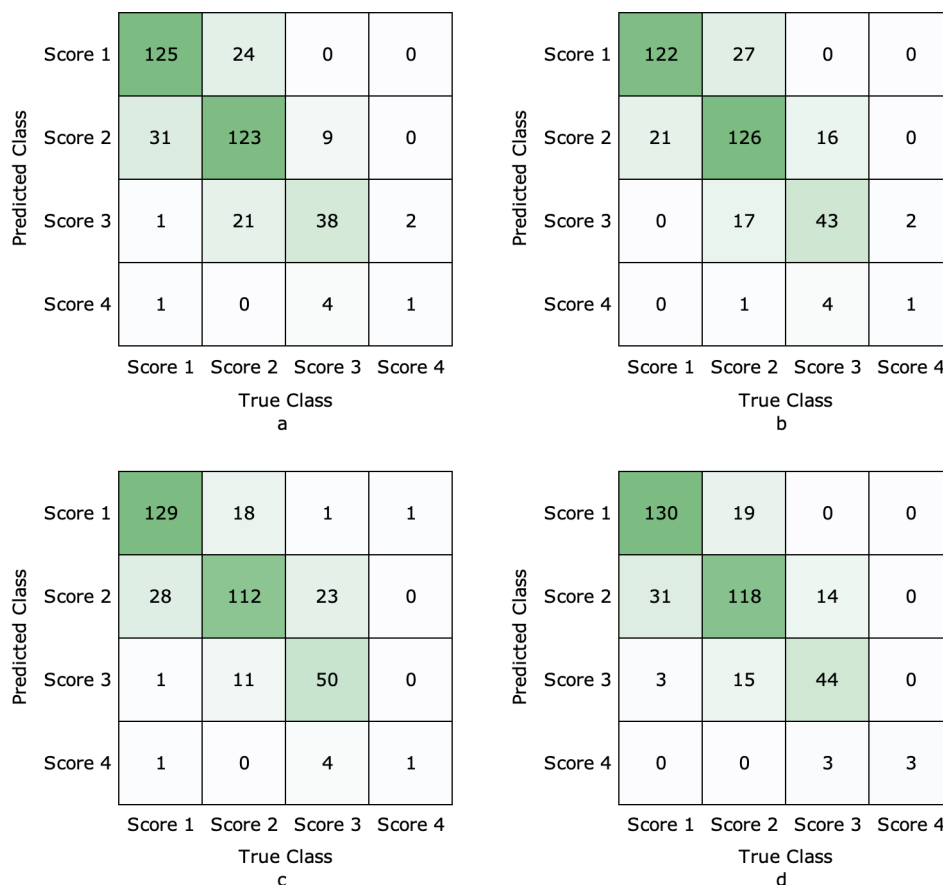
As shown in Table 2, we compared the accuracy of seventeen different ImageNet models. We observe that our proposed SCTL with GoogLeNet achieves the highest accuracy when compared with all other models. Moreover, there is consistent improvement across seventeen different ImageNet models, and we achieve 4.9% average improvement. We conclude that our proposed SCTL model improves the performance of ImageNet models. The accuracy of each category of the top-4 highest accuracy model is shown in Table 3. Our model with GoogLeNet has the highest accuracy in the Score 4 category, suggesting our SCTL paradigm is able to handle the unbalanced class problem. Score 4 corresponds to the teat-end condition with the highest degree of hyperkeratosis. This result further indicates the SCTL model can improve inter-class dispersion since SCTL can improve the accuracy in detecting severe teat-end affection even with a small number of training samples. Although the performance of Score 3 is slightly lower than NasNetLarge-SCTL, it is still much higher than DenseNet161-SCTL and DenseNet201-SCTL. The confusion matrixes of these four models are shown in Figure 4. We find that GoogLeNet-SCTL achieves the highest performance, and it has better performance than the other three models in Scores 1 and 4. When compared with earlier work [4], our model improves performance by 15.8%; our SCTL model substantially enhances the accuracy of teat-end image classification datasets. We also notice that the accuracy of “Original” with GoogLeNet, which only minimizes the cross-entropy loss, is still higher than the result from [4]. One possible explanation for the difference is that Porter et al. [4] trained GoogLeNet using MATLAB, while our model uses PyTorch. We also compare results from one transductive learning model (GSM) and three domain adaptation methods (DAN, DCORAL, and CAN). Experimental results show that our GoogLeNet-SCTL still achieves the highest performance. Tables 4 and 5 display findings from the ablation studies.

**Table 2.** Accuracy of different ImageNet models and the ablation study. GoogLeNet demonstrated the highest accuracy and greatest improvement when adopting SCTL.

Networks	SCTL	Original	Improvement
SqueezeNet [29]	68.2	67.4	0.8
AlexNet [5]	69.5	67.1	2.5
<b>GoogLeNet [6]</b>	<b>77.6</b>	64.2	13.4
ShuffleNet [30]	66.3	64.5	1.8
ResNet18 [31]	75.3	<b>72.6</b>	2.7
VGG16 [32]	73.4	66.8	6.6
VGG19 [32]	74.5	67.9	6.6
MobileNetv2 [33]	72.6	70.5	2.1
ResNet50 [31]	72.9	70.8	2.1
ResNet101 [31]	74.0	68.4	5.6
DenseNet161 [34]	75.5	72.9	2.6
DenseNet201 [34]	76.8	70.5	6.3
InceptionV3 [7]	73.7	71.6	2.1
Xception [35]	75.0	70.5	4.5
InceptionResNetV2 [36]	72.1	66.3	5.8
NasNetLarge [8]	76.8	68.2	8.6
EfficientNetB7 [8]	74.7	71.3	3.4
Average	<b>73.5</b>	<b>68.6</b>	4.9
GoogLeNet [4]	-	61.8	<b>15.8</b>
GSM [15]	-	65.2	12.4
DAN [37]	-	62.1	15.5
DCORAL [38]	-	63.8	13.8
CAN [39]	-	67.4	10.2

**Table 3.** Accuracy of each class in the test dataset. Shown in bold are the highest performance of a network for each score. Note that no single network achieves the highest accuracy for all scores.

Networks	Score 1	Score 2	Score 3	Score 4	Ave.
DenseNet161-SCTL	83.9	75.5	61.3	16.7	75.5
DenseNet201-SCTL	85.2	<b>77.3</b>	62.9	0.0	76.8
NasNetLarge-SCTL	86.6	71.2	<b>74.2</b>	16.7	76.8
<b>GoogLeNet-SCTL</b>	<b>87.3</b>	72.4	71.0	<b>50.0</b>	<b>77.6</b>



**Figure 4.** Confusion matrix of DenseNet161-SCTL, DenseNet201-SCTL, NasNetLarge-SCTL, and GoogLeNet-SCTL models. (a) DenseNet161-SCTL, (b) DenseNet201-SCTL, (c) NasNetLarge-SCTL, (d) GoogLeNet-SCTL.

To demonstrate the effects of different loss functions ( $\mathcal{L}_S$ : “S” (separation loss),  $\mathcal{L}_{CE}^T$ : “T” (cross-entropy loss of confident pseudo labeled test data), and  $\mathcal{L}_{CMMD}$ : “C” (categorical MMD loss) an ablation study is shown in Table 5. Notice that cross-entropy loss is required for the training data. “SCTL-T-S-C” is implemented without  $\mathcal{L}_{CE}^T$ ,  $\mathcal{L}_S$ , and  $\mathcal{L}_{CMMD}$  loss. It only reduces training data cross-entropy loss. “SCTL-T-S” minimizes the cross-entropy loss and categorical MMD loss. “SCTL-C” reports results without performing categorical MMD loss. Based on the average accuracy, we find  $\mathcal{L}_S > \mathcal{L}_{CMMD} > \mathcal{L}_{CE}^T$ . Therefore, the proposed separation loss, categorical MMD loss, and confident pseudo labeling approaches are effective in improving the performance of the test dataset. To show the effectiveness of our proposed  $\mathcal{L}_S$ ,  $\mathcal{L}_{CE}^T$ , and  $\mathcal{L}_{CMMD}$ , we also conducted an ablation study to show different variants of them. In Section 3.4, we utilize the SSIM to measure the similarity between the training and test data, and we take an absolute value to accelerate the convergence. As shown in Table 4, we report the accuracy and the number of convergence of different variants of our proposed separation loss. We find that Jaccard similarity has a lower

accuracy than cosine similarity, although it has a longer convergence number. Furthermore, compared with  $\mathcal{L}_S$  without taking the absolute value (w/o abs.), our model achieves high accuracy with the fastest convergence times. When comparing  $\mathcal{L}_{CMMMD}$  with original MMD loss, our proposed loss function still achieves better accuracy and requires fewer training iterations. Our proposed confident pseudo labeling with adjustment learning is again better than the pseudo label strategy in [20]. Therefore, our proposed loss functions can fast and accurately improve classification accuracy.

**Table 4.** Ablation study of  $\mathcal{L}_S$  and  $\mathcal{L}_{CMMMD}$ .

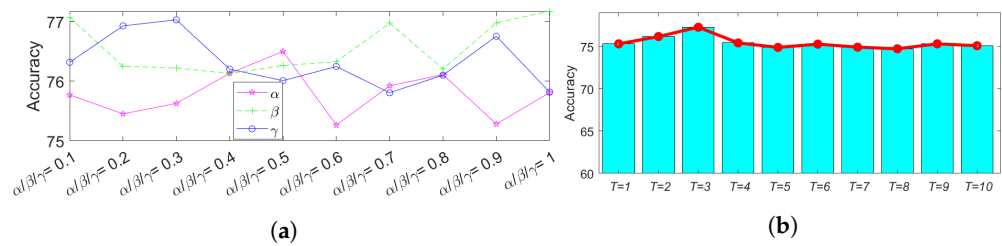
Method	Accuracy	# Convergence
Cosine similarity	75.3	92
Jaccard similarity w/o abs.	74.5 76.8	83 78
MMD	76.3	75
PL [20]	75.5	83
<b>SCTL</b>	<b>77.6</b>	<b>70</b>

**Table 5.** Ablation study of different loss functions on test accuracy.

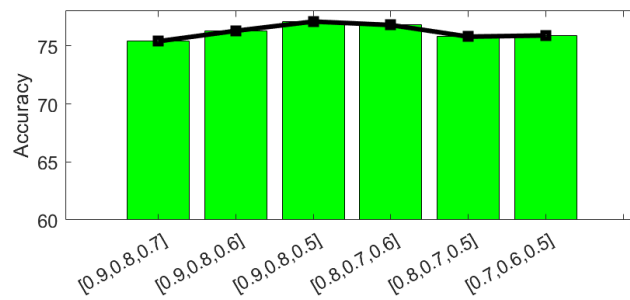
Loss	Accuracy
SCTL-T-S-C	64.2
SCTL-T-S	66.6
SCTL-T-C	71.1
SCTL-S-C	72.9
SCTL-T	74.7
SCTL-S	72.1
SCTL-C	74.4
<b>SCTL</b>	<b>77.6</b>

#### 4.4. Parameter Analysis

There are five hyperparameters  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $T$ , and  $p_t$  in our SCTL model.  $\alpha$ ,  $\beta$ , and  $\gamma$  are three trade-off parameters to balance the weight between separation loss, pseudo labeled test cross-entropy loss, and categorical MMD loss.  $T$  and  $p_t$  control the number of adjustment learning and the probability of selecting the confident examples, respectively. To obtain the optimal parameters, we use GoogLeNet as the backbone network. We first show the influence of  $\alpha$ ,  $\beta$ , and  $\gamma$  on test data accuracy.  $\alpha$ ,  $\beta$ , and  $\gamma$  are selected from  $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$  and fix one parameter while varying the others. As shown in Figure 5a, the x-axis represents that different values of  $\alpha$ ,  $\beta$ , and  $\gamma$ . We observed that the test data accuracy achieves the highest value when  $\alpha = 0.5$ ,  $\beta = 1$ , and  $\gamma = 0.3$ , respectively.  $T$  is selected from  $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ , and  $p_t$  is selected from  $\{0.9, 0.8, 0.7, 0.6, 0.5\}$ . Since we need to obtain confident examples,  $p_1$  should be a very high probability. Thus, we set  $p_1 \geq 0.5$ . For  $t > 2$ ,  $p_t$  is selected from  $\{0.9, 0.8, 0.7, 0.6, 0.5\}$  and require  $p_t \geq p_{t+1}$ . As shown in Figure 5b, we observed that our model achieves the highest test data accuracy when  $T = 3$ . We then examined how different  $p_t$  values affect the accuracy in Figure 6. We observed that when  $p_t = \{0.9, 0.8, 0.5\}$ , the highest accuracy in the test data is achieved. By carefully examining these parameters and their influence on overall performance, we find the best hyperparameters for our SCTL model are:  $\alpha = 0.5$ ,  $\beta = 1$ ,  $\gamma = 0.3$ ,  $T = 3$ , and  $p_t = \{0.9, 0.8, 0.5\}$ .



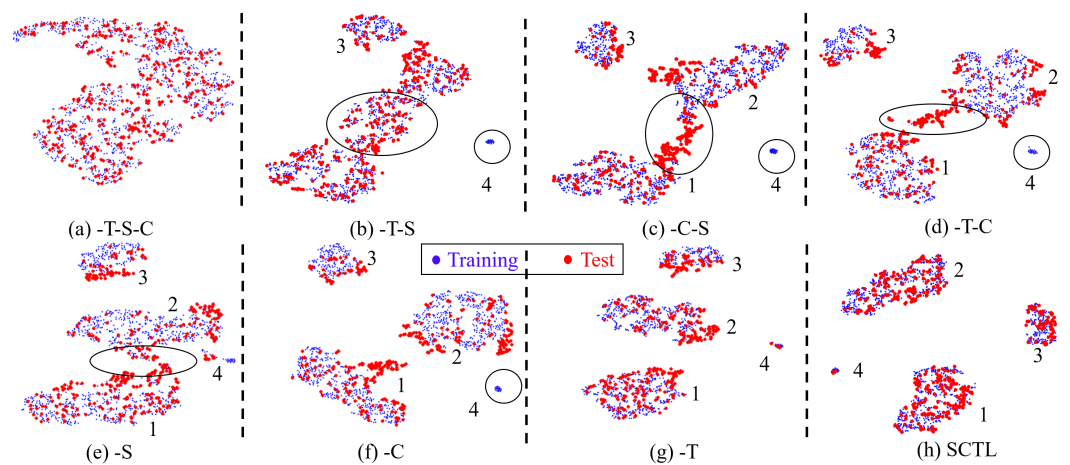
**Figure 5.** Parameter analysis for  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $T$ . The effect of different  $\alpha$ ,  $\beta$ , and  $\gamma$  on test accuracy is shown in (a), and effect of different  $T$  on test accuracy is shown in (b). When  $\alpha = 0.5$ ,  $\beta = 1$ , and  $\gamma = 0.3$ , and  $T = 3$ , accuracy is highest.



**Figure 6.** Effect of different  $p_t$  on test accuracy.

4.5. Feature Visualization

To further demonstrate the effectiveness of different loss functions, we utilize t-SNE [9] to visualize the deep features of network activations in 2D space. As shown in Figure 7a, we cannot observe four distinctive classes if we only minimize the cross-entropy loss. From Figure 7b to Figure 7g, the four classes become more distinctive after adding separation loss, pseudo labeled test cross-entropy loss, and categorical MMD loss. Comparing Figure 7b,c with Figure 7d, the four categories cannot be correctly classified if we only train the network with a single loss (especially Score 4 in the test data which are missing). There is also contamination between classes 1 and 2 among these three figures. These two issues are ameliorated if we train the model with two losses (from Figure 7e to Figure 7g). Figure 7g has a similar trend as Figure 7h with less class divergence. Finally, with SCTL (Figure 7h), we see inter-class dispersion and intra-class compactness of the test dataset.

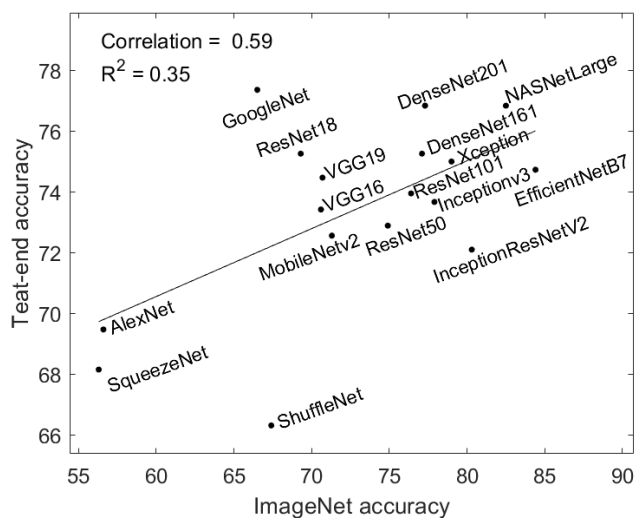


**Figure 7.** t-SNE view of training and test dataset with different loss functions. Blue color represents training dataset and red represents test dataset. Numbers 1 to 4 mean the location of four classes (Score 1 to Score 4).

## 5. Discussion

### 5.1. Relationship between ImageNet Accuracy and Teat-End Accuracy

Previous work [40] noted that ResNet and DenseNet are usually the better neural networks for transfer learning, and a better ImageNet model can produce better features for domain adaptation [41], which is one special case of transductive learning. We explore how different ImageNet models affect the teat-end classification accuracy, their correlation score, and the  $R^2$  value as per [41]. As shown in Figure 8, both correlation score and  $R^2$  value are low, which suggests no strong relationship between ImageNet model accuracies and teat-end classification accuracies. This result differs from [40,41], suggesting the optimal ImageNet model for teat-end classification may not be based on the ImageNet model with the highest accuracy.



**Figure 8.** ImageNet models' accuracy and teat-end classification accuracy. There is no strong relationship between them since both correlation and  $R^2$  have a low value.

### 5.2. What Can We Draw from Our Experiments?

Fine-tuning different ImageNet models for transfer learning has been one of the most popular methods for image classification problems. However, choosing the optimal ImageNet for a given dataset remains a challenge. For our dataset, the teat images vary from the ImageNet datasets and thus there is no strong relationship between ImageNet model accuracy and teat-end classification accuracy. As shown in Table 2, GoogLeNet unexpectedly achieved the highest performance. This suggests there is value in using ImageNet models with lower memory size when first exploring such techniques, such as (SqueezeNet and GoogLeNet) for transfer learning if the image data are very different from the pre-trained ImageNet dataset. If image data are very similar to the ImageNet images, there may be value in using more accurate networks such as Xception and EfficientNet.

### 5.3. Advantages and Limitations

There are several advantages of our proposed SCTL model. First, our proposed separation loss enlarges the difference between different categories and leads to greater inter-class dispersion. Second, we generate high confident pseudo labels for test data in three times adjustment learning to optimize the network with pseudo labels information. Last, we propose a categorical MMD loss to reduce the divergence between training and test data. By aggregating all three of these novel loss functions, our SCTL model can enhance the performance of the teat-end image classification problem.

One limitation of our work is that we have a small sample size of teat-end images (1529 images). Especially, the category Score 4 is unbalanced. However, Score 4 corresponds to the severe hyperkeratosis of the teat-end, which is less prevalent in the study population when compared with the other three categories. As for future work, aside from collecting

more data, improving the pseudo label quality of the test dataset can be a useful technique to further improve performance. Our SCTL model can be applied to other image classification tasks (e.g., teat skin condition assessments). However, five hyperparameters,  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $T$ , and  $p_t$ , should be adjusted according to different datasets.

## 6. Conclusions

In this paper, we propose a separation confident transductive learning model for teat-end image classification. We first propose a separation loss to enlarge the differences between different categories. We then generate confident labels for the test data using adjustment learning to optimize the network. Finally, we employ transductive learning to minimize the divergence between the training and test data with a categorical MMD loss. Although the level of affection of cows' teats can influence the performance of our SCTL model, we demonstrate that the proposed SCTL model can achieve higher accuracy when compared with ImageNet transfer learning models. We believe that through the aid of SCTL, the detection of hyperkeratosis is feasible in the commercial dairy farm setting. Our approach offers the opportunity for more frequent and automated teat-end condition assessments. Such an automated hyperkeratosis detection method may help farmers mitigate the risks of intramammary infections, decrease the use of antimicrobials, control the costs associated with detecting and managing mastitis, and improve the quality of life of dairy cows and farmers.

**Author Contributions:** Conceptualization, Y.Z. and P.S.B.; methodology, Y.Z.; writing—original draft preparation, Y.Z. and P.S.B.; writing—review and editing, M.W. and I.R.P.; supervision, P.S.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Cornell Initiative for Digital Agriculture (CIDA).

**Institutional Review Board Statement:** The study was conducted according to a Cornell University Institutional Animal Care and Use Committee reviewed study (protocol no. 2013-0064).

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** We evaluate our model using a dataset from previous work [4]. Source code, accessed 25 March 2022, is available at <https://github.com/YoushanZhang/SCTL>.

**Acknowledgments:** The authors thank the farm owners and their employees for their willingness to participate in the study and their support during the data collection.

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

## References

1. Neijenhuis, F.; Barkema, H.; Hogeveen, H.; Noordhuizen, J. Relationship between teat-end callosity and occurrence of clinical mastitis. *J. Dairy Sci.* **2001**, *84*, 2664–2672. [[CrossRef](#)]
2. Williams, D.; Mein, G. Physical and physiological factors affecting milk flowrate from the bovine teat during machine milking. *Occas. Publ. N. Z. Soc. Anim. Prod. (N. Z.)* **1982**, *8*, 42–74.
3. Basran, P.; Wieland, M.; Porter, I. A digital technique and platform for assessing dairy cow teat-end condition. *J. Dairy Sci.* **2020**, *103*, 10703–10708. [[CrossRef](#)]
4. Porter, I.; Wieland, M.; Basran, P. Feasibility of the use of deep learning classification of teat-end condition in Holstein cattle. *J. Dairy Sci.* **2021**, *104*, 4529–4536. [[CrossRef](#)] [[PubMed](#)]
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
6. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
7. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 2818–2826.
8. Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q.V. Learning transferable architectures for scalable image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8697–8710.
9. Maaten, L.v.d.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

10. Ruegg, P.L. A 100-Year Review: Mastitis detection, management, and prevention. *J. Dairy Sci.* **2017**, *100*, 10381–10397. [[CrossRef](#)] [[PubMed](#)]
11. Neijenhuis, F.; Mein, G.; Britt, J.; Reinemann, D.; Hillerton, J.; Farnsworth, R.; Baines, J.; Hemling, T.; Ohnstad, I.; Cook, N.; et al. Evaluation of bovine teat condition in commercial dairy herds: 4. Relationship between teat-end callosity or hyperkeratosis and mastitis. In Proceedings of the 2nd International Symposium on Mastitis and Milk Quality, Vancouver, BC, Canada, 12–14 September 2001; pp. 362–366.
12. Reinemann, D.; Rasmussen, M.; LeMire, S.; Neijenhuis, F.; Mein, G.; Hillerton, J.; Morgan, W.; Timms, L.; Cook, N.; Farnsworth, R.; et al. Evaluation of bovine teat condition in commercial dairy herds: 3. Getting the numbers right. In Proceedings of the 2nd International Symposium on Mastitis and Milk Quality, NMC/AABP, Vancouver, BC, Canada, 12–14 September 2001; pp. 357–361.
13. Mein, G.; Neijenhuis, F.; Morgan, W.; Reinemann, D.; Hillerton, J.; Baines, J.; Ohnstad, I.; Rasmussen, M.; Timms, L.; Britt, J.; et al. Evaluation of bovine teat condition in commercial dairy herds: 1. Non-infectious factors. In Proceedings of the 2nd International Symposium on Mastitis and Milk Quality, Vancouver, BC, Canada, 12–14 September 2001; pp. 347–351.
14. Joachims, T. Transductive learning via spectral graph partitioning. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington, DC, USA, 21–24 August 2003; pp. 290–297.
15. Zhang, Y.; Xie, S.; Davison, B.D. Transductive Learning Via Improved Geodesic Sampling. In Proceedings of the 30th British Machine Vision Conference, Cardiff, UK, 9–12 September 2019.
16. Zhou, D.; Bousquet, O.; Lal, T.N.; Weston, J.; Schölkopf, B. Learning with local and global consistency. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 13–18 December 2004; pp. 321–328.
17. Shi, W.; Gong, Y.; Ding, C.; Tao, Z.M.; Zheng, N. Transductive semi-supervised deep learning using min-max features. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 299–315.
18. Quadrianto, N.; Petterson, J.; Smola, A. Distribution matching for transduction. *Adv. Neural Inf. Process. Syst.* **2009**, *22*, 1500–1508.
19. Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N.E.; McGuinness, K. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
20. Lee, D.H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop Chall. Represent. Learn. ICML* **2013**, *3*, 896.
21. Saito, K.; Ushiku, Y.; Harada, T. Asymmetric tri-training for unsupervised domain adaptation. *arXiv* **2017**, arXiv:1702.08400.
22. Xie, S.; Zheng, Z.; Chen, L.; Chen, C. Learning semantic representations for unsupervised domain adaptation. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 5423–5432.
23. Zhang, W.; Ouyang, W.; Li, W.; Xu, D. Collaborative and Adversarial Network for Unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3801–3809.
24. Iscen, A.; Tolia, G.; Avrithis, Y.; Chum, O. Label propagation for deep semi-supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5070–5079.
25. Zhang, Y.; Deng, B.; Jia, K.; Zhang, L. Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 781–797.
26. Haase-Schütz, C.; Stal, R.; Hertlein, H.; Sick, B. Iterative Label Improvement: Robust Training by Confidence Based Filtering and Dataset Partitioning. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 9483–9490.
27. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)]
28. Long, M.; Zhu, H.; Wang, J.; Jordan, M.I. Unsupervised domain adaptation with residual transfer networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 136–144.
29. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
30. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
31. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
32. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
33. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
34. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.

35. 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.
36. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
37. Long, M.; Cao, Y.; Wang, J.; Jordan, M.I. Learning transferable features with deep adaptation networks. *arXiv* **2015**, arXiv:1502.02791.
38. Sun, B.; Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 443–450.
39. Kang, G.; Jiang, L.; Yang, Y.; Hauptmann, A.G. Contrastive adaptation network for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4893–4902.
40. Kornblith, S.; Shlens, J.; Le, Q.V. Do better ImageNet models transfer better? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2661–2671.
41. Zhang, Y.; Davison, B.D. Impact of ImageNet Model Selection on Domain Adaptation. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops, Snowmass Village, CO, USA, 1–5 March 2020; pp. 173–182.