Animal Fiber Identification under the Open Set Condition

Oliver Rippel[1][a], Sergen Gülçelik[1], Khosrow Rahimi[2], Juliana Kurniadi[2], Andreas Herrmann[2] and Dorit Merhof[1][b]

[1]*Institute of Imaging & Computer Vision, RWTH Aachen University, Aachen, Germany*

[2]*DWI – Leibniz-Institut für Interaktive Materialien, Aachen, Germany*

Keywords: Out-of-Distribution Detection, Natural Fiber Identification, Classification, Open Set Recognition, Machine Learning.

Abstract: Animal fiber identification is an essential aspect of fabric production, since specialty fibers such as cashmere are often targeted by adulteration attempts. Proposed, automated solutions can furthermore not be applied in practice (i.e. under the open set condition), as they are trained on a small subset of all existing fiber types only and simultaneously lack the ability to reject fiber types unseen during training at test time. In our work, we overcome this limitation by applying out-of-distribution (OOD)-detection techniques to the natural fiber identification task. Specifically, we propose to jointly model the probability density function of in-distribution data across feature levels of the trained classification network by means of Gaussian mixture models. Moreover, we extend the open set F-measure to the so-called area under the open set precision-recall curve ($AUPR_{os}$), a threshold-independent measure of joint in-distribution classification & OOD-detection performance for OOD-detection methods with continuous OOD scores. Exhaustive comparison to the state of the art reveals that our proposed approach performs best overall, achieving highest area under the class-averaged, open set precision-recall curve ($AUPR_{os,avg}$). We thus show that the application of automated fiber identification solutions under the open set condition is feasible via OOD detection.

1 INTRODUCTION

Animal fibers possess desirable characteristics such as thermal insulation, moisture wicking and softness, making them an important material for fabric production (McGregor, 2018). Since specialty fibers (e.g. cashmere) excel at one or more of the above properties they achieve premium prices on the market (International Wool Textile Organisation, 2018). Said prices, however, render specialty fibers an attractive target for adulteration, and adulteration rates between 15-60% have been reported for cashmere products (Waldron et al., 2014; Phan and Wortmann, 2001).

Various fiber identification methods have been proposed to counteract adulteration (Rane and Barve, 2011; Kim et al., 2013; Zoccola et al., 2013; International Wool Textile Organisation, 2000; American Society for Testing and Materials, 1993). Out of those, optical identification methods are the most widely applicable. Here, fibers are identified based on their surface morphology using either optical microscopy (American Society for Testing and Materials, 1993) or scanning electron microscopy (SEM) (International Wool Textile Organisation, 2000). Despite being subjective in nature and requiring extensively trained experts to achieve reliable results (Zhang and Ainsworth, 2005; Wortmann, 1991), optical identification methods are still the ones predominantly used in industry.

In order to overcome the limitations of human experts, it has been proposed to automate the fiber identification by means of pattern recognition (Yildiz, 2020; Robson, 1997; Robson, 2000; Xing et al., 2020a; Xing et al., 2020b; Rippel et al., 2021a). While prior work has shown that accurate fiber identification is possible, the application of the developed solutions in practice is hindered by the following two facts: (I) Developed solutions train & evaluate their algorithms only on a small subset of all existing fiber types. Even though up to 11 fiber types (10 specialty fibers + wool) can be identified by a human expert (International Wool Textile Organisation, 2000), research typically focusses on binary classification, e.g.

[a] https://orcid.org/0000-0002-4556-5094

[b] https://orcid.org/0000-0002-1672-2185

Rippel, O., Gülçelik, S., Rahimi, K., Kurniadi, J., Herrmann, A. and Merhof, D.
Animal Fiber Identification under the Open Set Condition.
DOI: 10.5220/0010769800003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 36-47
ISBN: 978-989-758-555-5; ISSN: 2184-4321

distinguishing between cashmere and wool (Robson, 1997; Robson, 2000; Xing et al., 2020a; Xing et al., 2020b) or mohair and wool (Yildiz, 2020). (II) Developed solutions lack the ability to reject fiber types unseen during training at test time. Such a rejection can be achieved in principle by out-of-distribution (OOD)-detection techniques (Geng et al., 2020), facilitating the application of algorithms under the open set condition (i.e. when training and test data do not originate from the same data distribution). However, the applicability of OOD-detection techniques to the task at hand has not yet been demonstrated, and is the main goal of our work.

Our contributions are as follows:

- We set up an exhaustive dataset comprising SEM-images of 4 animal fiber types from 10 different sources. In total, the dataset contains 6500 images and covers the major axes of variation in natural fiber surface morphology.
- We expand the open set F-measure (Mendes Júnior et al., 2017) to facilitate the evaluation of joint in-distribution classification & OOD-detection performance for OOD-detection methods that output continuous OOD scores. The proposed metric can be used to assess joint performance across all possible OOD thresholds.
- We propose to perform OOD detection by modeling the joint distribution of in-distribution data across feature levels of converged, convolutional neural network (CNN)-based classifiers via Gaussian mixture models (GMMs). We compare this approach to state-of-the-art OOD-detection algorithms.
- We also investigate effects of outlier exposure (OE) on animal fiber identification under the open set condition.

2 RELATED WORK

So far, the performance of animal fiber identification algorithms has not yet been investigated under the open set condition. We will therefore give a short definition of open set recognition (OSR) and related terminology first. This will be followed by a brief overview of proposed OOD-detection methods as well as OE. Last, we will also present the open set F-measure, which can be used to evaluate algorithms under the open set condition.

2.1 Open Set Recognition

In general, OSR is concerned with problems that arise specifically when training and test data do not originate from the same data distribution (Geng et al., 2020). In context of classification, this means one is first tasked with distinguishing between in-distribution and OOD data followed by the subsequent d-class classification of the presumed in-distribution data. The in-distribution is furthermore composed of the d target classes, which are referred to as known known classes (KKCs) (e.g. specialty fibers such as cashmere). Opposed to the in-distribution, the OOD is composed of known unknown classes (KUCs), i.e. negative OOD samples available during training/validation, and unknown unknown classes (UUCs), i.e. samples not available for training/validation.

The task of OOD detection can now be formulated as "developing a measure which achieves lower values for in-distribution data compared to OOD data, allowing for their separation".

2.2 Methods for OOD Detection

OOD-detection methods commonly make use either of the classifier's unnormalized predictions (also referred to as "logits" in literature) or of the underlying feature representations to formulate their OOD scores.

2.2.1 OOD Detection based on Unnormalized Predictions

The most straightforward method is to take the maximum softmax probability (MSP) of the unnormalized predictions as the OOD score (Hendrycks and Gimpel, 2017). Specifically, let $\phi : \mathbb{R}^{c\times h\times w} \rightarrow \mathbb{R}^d$ be a pre-trained CNN classifying into d classes. Then, for a given input $\boldsymbol{x}$, the softmax probability s_i for class i with $i \in \{1,\ldots,d\}$ is calculated as

$$s_i(\boldsymbol{x}) = \frac{e^{\phi_i(\boldsymbol{x})}}{\sum_d e^{\phi_d(\boldsymbol{x})}}. \quad (1)$$

The probability that a given input sample belongs to class i is given by $s_i(\boldsymbol{x})$. The MSP now takes $-\max_i s_i(\boldsymbol{x})$ as the OOD score, since the classifier should be uncertain for samples not originating from the in-distribution. Since it has been shown that classifiers suffer from so-called overconfident predictions (Nguyen et al., 2015), i.e. from assigning high probabilities to UUCs, modifications have been developed. For example, temperature scaling and input-preprocessing have been proposed to improve MSP

performance in outlier detection using in-degree number (ODIN) (Liang et al., 2018). Additionally, the classifier's output has been used to define the energy-based score (EBS)

$$\text{EBS} = -\log \sum_i e^{\phi_i(\boldsymbol{x})} \tag{2}$$

based on the ties between energy-based modeling and machine learning (Liu et al., 2020). Alternatively, it has been proposed to use the maximum of the unnormalized predictions (MaxLogit) as OOD score, arguing that putting the class-predictions in relation to each other via the softmax operator may be detrimental when semantically similar classes are present in the in-distribution (Hendrycks et al., 2019a).

2.2.2 OOD Detection based on Intermediate Features

Complementary to the classifier's predictions, the intermediate feature representations of a CNN can also be used for OOD detection. Here, algorithms commonly try to estimate the probability density function (PDF) of the in-distribution, often using the Gaussian assumption (Rippel et al., 2021b; Kamoi and Kobayashi, 2020) and its mixture models (Lee et al., 2018b; Ahuja et al., 2019). Alternatively, deep generative models have also been used to fit unconstrained PDFs to in-distribution data in intermediate features (Kirichenko et al., 2020; Zisselman and Tamar, 2020; Zhang et al., 2020; Blum et al., 2021). The OOD score is then defined as the negative log-likelihood (NLL) of a given input image $\boldsymbol{x}$ under the estimated PDF.

Apart from the PDF-estimation, it has also been proposed to use the distance of an input image $\boldsymbol{x}$ to a fixed reference UUC point inside the intermediate features, resulting in the feature space singularity distance (FSSD) (Huang et al., 2020). Here, it is proposed to use uniformly distributed noise samples as the reference UUC point, arguing that uniform noise possesses the highest degree of OOD-ness.

OOD detection has also been performed by autoencoders (AEs), where the encoder's features are simultaneously used to classify images into d KKCs and to reconstruct the input images via the decoder (Oza and Patel, 2019; Sun et al., 2020; Neal et al., 2018). The OOD score is then defined based on the residual of the image reconstruction, which arguably should be higher for OOD than in-distribution data. It should be noted that these approaches roughly double the computational complexity of the CNN and fail to consistently achieve state-of-the-art results, and are therefore not further regarded in this work.

2.3 Outlier Exposure

The OOD-detection methods from subsection 2.2 can be applied to any converged CNN and require no knowledge about OOD data. To now include available information about OOD data, it has been proposed to use KUCs during training by means of OE (Hendrycks et al., 2019c). In principle, an additional loss term is introduced for the proposed OOD score that is maximized for OOD samples (and optionally minimized for in-distribution samples), e.g. Equation 2. While it has been shown that sampling the OOD introduces a bias in the resulting OOD detector, i.e. OE may actually hurt OOD detection for some UUCs (Ye et al., 2021), OE has also been shown to work for most UUCs (Hendrycks et al., 2019c). As an alternative to real OOD images, approaches facilitating OE by means of synthetic OOD images have also been proposed (Neal et al., 2018; Lee et al., 2018a; Grcić. et al., 2021).

2.4 Open Set F-measure

In order to jointly assess OOD-detection and in-distribution classification performance, the open set F-measure has been proposed (Mendes Júnior et al., 2017). Specifically, for a d-class classification problem, the open set F-measure for class i is defined as the harmonic mean between its open set precision and recall, defined as

$$P_i = \frac{TP_i}{TP_i + FP_i}, \quad FP_i = \sum_{j=1}^{d} FP_{i,j} + FP_{i,UUC} \tag{3}$$

and

$$R_i = \frac{TP_i}{TP_i + FN_i}, \quad FN_i = \sum_{j=1}^{d} FN_{i,j} + FN_{i,UUC} \tag{4}$$

respectively. Compared to the closed set precision and recall variants, it can be seen that false positives may now also be incurred by failing to reject an OOD image followed by its misclassification to class i Equation 3, and false negatives may be incurred by incorrectly labeling in-distribution data as OOD Equation 4 (refer also Figure 1a).

As can be inferred from Equation 3 and Equation 4, the open set F-measure requires that every image is labeled either as OOD and rejected or labeled as in-distribution & subsequently classified. It is therefore ill-suited for OOD-detection methods that yield continuous OOD scores and thus require thresholding to achieve the aforementioned partitioning of images into in-distribution and OOD.

3 NATURAL FIBER IDENTIFICATION UNDER THE OPEN SET CONDITION

We specify both our proposed OOD-detection method as well as the threshold-independent evaluation of joint in-distribution classification & OOD-detection performance in the following.

3.1 OOD Detection via Modeling the Joint PDF Across Feature Levels

Similar to (Ahuja et al., 2019), we model the PDF of in-distribution data of a converged CNN by means of GMMs. However, as a significant extension, we propose to model the joint PDF of in-distribution data across feature levels instead of modeling PDFs layer-wise and summing their individual OOD scores. We motivate this by the fact that consistency of an input image $\boldsymbol{x}$ across feature levels of a network has been shown to be an indicator of model generalization, i.e. models with consistent representations generalize well (Natekar and Sharma, 2020). We argue that inconsistent representations may be an indicator of OOD, and show that joint PDF-estimation is beneficial for OOD detection empirically in subsubsection 4.3.2.

Specifically, let $\phi : \mathbb{R}^{c \times h \times w} \rightarrow \mathbb{R}^d$ again be a pre-trained CNN classifying into d classes. Each intermediate layer's output $\psi_m := \phi_m(\psi_{m-1})$ for an input image $\psi_0 := \mathbf{x}_i$ has c_m features and spatial dimension h_m by w_m. We now reduce over the spatial dimensions h_w and w_m by means of averaging, resulting in a feature vector $\boldsymbol{c}$ extracted per intermediate layer m. By concatenating $\boldsymbol{c}$ of all intermediate layers, a larger feature vector $\boldsymbol{c}_{cat}$ is generated. To model the PDF of in-distribution data in $\boldsymbol{c}_{cat}$, we make use of GMMs, defined as

$$p(\boldsymbol{x}) = \sum_{i=1}^{k} \theta_i \mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{5}$$

with $\sum_{i=1}^{k} \theta_i = 1$, k being the number of Gaussian mixture components and $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ denoting the mean vector and covariance matrix of mixture component i. We approximate the parameters of the GMM by using the expectation maximization (EM) algorithm, and determine the number of Gaussians k by means of the Bayesian information criterion (BIC) (Bishop, 2006). Similar to other PDF-estimation approaches, we use the NLL of an input image $\boldsymbol{x}$ under the estimated PDF as the OOD score.

3.2 Threshold-independent Evaluation of Joint In-distribution Classification & OOD-detection Performance

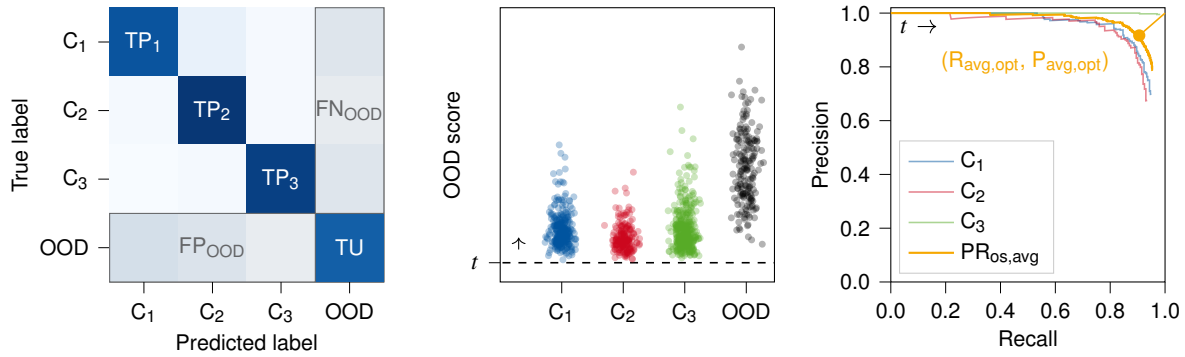
When looking at recent work on OOD detection, it becomes apparent that OOD-detection and in-distribution classification performances are reported individually (Liu et al., 2020; Lee et al., 2018b; Hendrycks and Gimpel, 2017; Liang et al., 2018). We argue that reporting OOD-detection and in-distribution classification performances separately oversimplifies the open set classification problem, and verify this claim experimentally in subsubsection 4.3.1.

Since most OOD-detection approaches proposed in literature yield continous OOD scores, we extend the open set F-measure (refer subsection 2.4) to be threshold-independent. Specifically, we propose to generate open set precision-recall curves (Davis and Goadrich, 2006) by plotting the open set precision & recall values (Equation 3 and Equation 4) across all potential OOD thresholds t (refer Figure 1). Such curves are generated for each class individually, and the area under the open set precision-recall curve (AUPR_{os}) can now be used to measure the joint in-distribution classification & OOD-detection performance of a single class.

As we want to quantify the overall performance across all classes, we furthermore propose to compute the average of the class-wise open set precision & recall values for each threshold t. The resulting values can then again be plotted for all t to yield a class-averaged, open set precision-recall curve (refer the yellow curve in Figure 1c). The area under the class-averaged, open set precision-recall curve ($\text{AUPR}_{\text{os,avg}}$) can now be used to jointly assess the overall in-distribution classification & OOD-detection performance for multi-class classification problems. Similar to the macro-averaged F1-score, $\text{AUPR}_{\text{os,avg}}$ regards all KKCs as equally important. Since not all OOD-detection methods provide continuous OOD scores, we also report the Euclidean distance of the class-averaged, open set precision-recall curve to the optimal point (1, 1), yielding PR_{dist} for comparison.

4 EXPERIMENTS

In the following, we conduct experiments to assess the applicability of natural fiber identification algorithms under the open set condition.



(a) (b) (c)

Figure 1: Threshold-independent evaluation of joint OOD-detection and in-distribution classification performance for multi-class classification problems. The open set confusion matrix at a single OOD threshold t is shown in (a), whereas (b) shows a scatter plot of OOD scores for a hypothetical 3-class open set classifier. By iterating over all possible OOD thresholds t, class-wise as well as a class-averaged open set precision-recall curves can be computed, shown in (c). The areas under the open set precision-recall curves can now be used to assess joint OOD-detection and in-distribution classification performance.

4.1 Dataset

We set up an exhaustive dataset to facilitate the performed experiments. We employ SEM-imaging over optical microscopy as it provides higher-resolution images of the fiber surface morphology, and is the only imaging technique shown to be reliable in combination with human operators (Wortmann and Wortmann, 1992). Our dataset contains 4 animal fiber types from 10 different sources, and 6500 images in total. Furthermore, all specimen samples were checked for purity & identity by a certified laboratory prior to image acquisition.

When composing the dataset, focus was put on sampling the three major axes of variation in natural fiber surface morphology:

1. *Inter-species* variation. These are variations present between different species, e.g. cashmere & yak.
2. *Intra-species* variation. These are variations present between different races of the same species, e.g. merino wool and typical wool.
3. *Treatment status* variation. These are variations introduced by the mechanical & chemical processes applied to the fibers during fabric production (e.g. they are dyed & bleached (d&b)). Variations incurred by *treatment status* are important for the industrial application of animal fiber identification since adulteration attempts often involve treatment of non-specialty fibers to make their surface morphology more similar to the one of specialty fibers.

Reference images for all three axes of variation are shown in Figure 2.

Table 1: Characteristics of the animal fiber dataset used in this work.

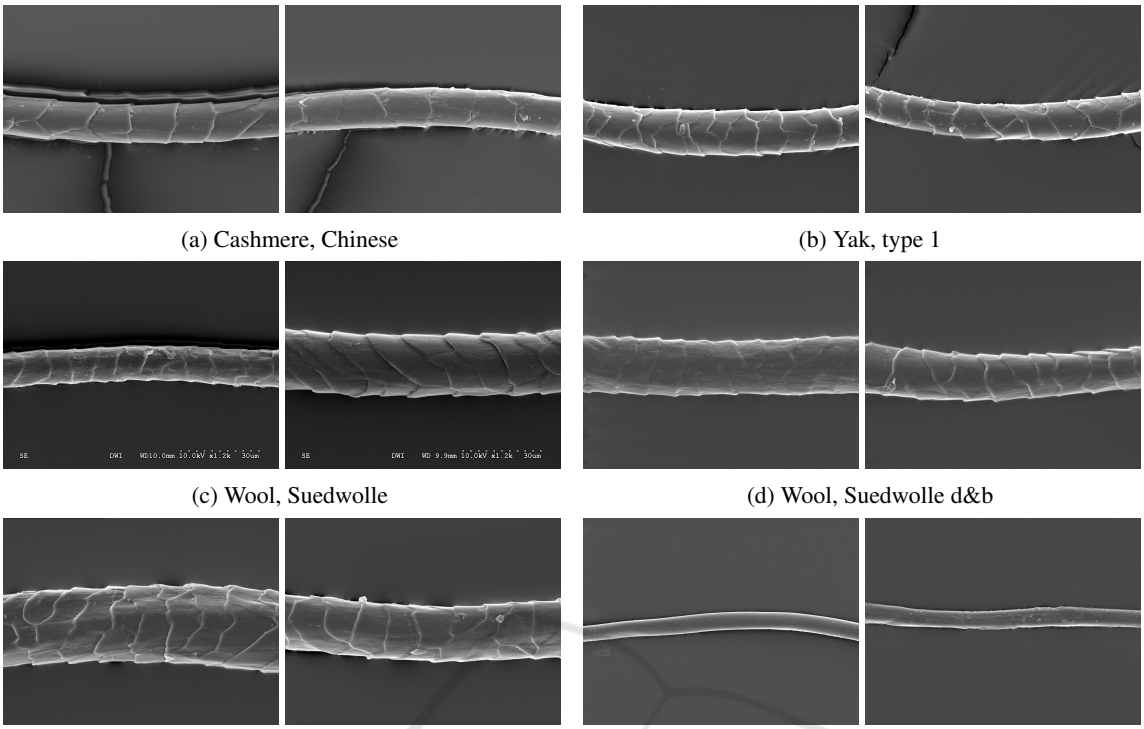
Fiber Type	#Samples	Use
Cashmere, Iranian	500	KKC
Cashmere, Chinese	500	KKC
Cashmere, brown	500	KKC
Yak, type 1	500	KKC
Yak, type 2	500	KKC
Wool, Suedwolle	1000	KKC
Wool, Interwool	1000	KKC
Wool, Suedwolle d&b	1000	UUC/KUC
Wool, merino	500	UUC/KUC
Silk	500	UUC/KUC

4.2 Experimental Setup

In the following, we describe the experimental setup required to assess the applicability of natural fiber identification algorithms under the open set condition.

4.2.1 Dataset Composition & Splits

In all experiments, the KKCs are cashmere, yak and wool (refer Table 1). We pool all sources for the three KKCs as it is not necessary to distinguish between subtypes of animal fibers under current regulations (Council of European Union, 2011; Freer, 1946). In order to represent all three possible axes of variation the UUCs/KUCs are: silk (inter-species), merino wool (intra-species) and d&b wool (treatment status). Since all axes are equally important, we will report the evaluation scores per axis and give their mean and standard deviation to denote overall algorithmic performance.



(a) Cashmere, Chinese

(b) Yak, type 1

(c) Wool, Suedwolle

(d) Wool, Suedwolle d&b

(e) Wool, merino

(f) Silk

Figure 2: Reference sample images from the animal fiber dataset.

Following good scientific practice, we split our dataset into training, validation and test splits. Since KKCs, UUCs and KUCs serve different purposes during model training, validation and testing, they are split independently.

The KKCs are used to train the in-distribution classifier for classifying into $d = 3$ classes. Since the dataset is rather small, we perform a 5-fold cross-validation to improve the robustness of our evaluations. To this end, the data is split in a 3-1-1 split, meaning that in each fold 60% of the data is used for training, and 20% is used each for validation and testing. Splits are furthermore stratified according to the prevalence of the KKCs.

Assessing algorithmic performance under the open set condition requires UUCs. Since UUCs are per-definition unknown during model training/evaluation, they are only used at test time. As is common in literature, the size of the UUC dataset is set to a fifth of the KKCs' testing set (Hendrycks et al., 2019c; Liu et al., 2020). In order to cover most of the variation of the UUC data, the model is tested with a randomly sampled UUC set for each fold.

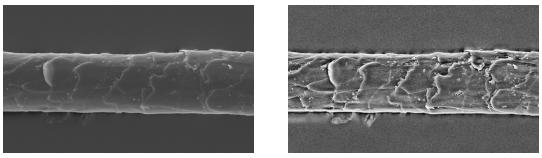
For the experiments on OE (subsubsection 4.3.3), KUCs are also used during training. Identical to the UUCs, KUCs also compose a fifth of the KKCs' training set and are also sampled randomly for each fold. In addition to the performance on the KUC class, we will also report the OOD-detection performance on the two remaining UUCs to assess whether potential gains on the KUC also generalize to the UUCs.

4.2.2 Evaluation Details

Apart from the proposed joint-performance metric (subsection 3.2), we also report OOD-detection performance and in-distribution classification performance individually. For the OOD-detection performance, we compute the area under the receiver operating characteristic (ROC) curve (AUROC) of the binary in-distribution/OOD classification problem, where in-distribution data is the positive and OOD data is the negative class. For the in-distribution classification performance, we compute the macro-averaged F1-score of the in-distribution test set.

4.2.3 Model Architecture & Evaluated OOD-detection Methods

For all our experiments, we employ the EfficientNet-B0 (Tan and Le, 2019) model architecture, which consists of nine levels. Furthermore, we perform transfer learning with an ImageNet(Deng et al., 2009)-pretrained EfficientNet-B0, as pre-training on large-



(a) (b)

Figure 3: Iranian cashmere sample before (a) and after (b) the application of CLAHE.

scale datasets improves model robustness and uncertainty (Hendrycks et al., 2019b). For our proposed method, we extract features from levels 2, 3, 4, 6 and 9.

We compare our proposed method with the Mahalanobis (Maha) score (Lee et al., 2018b), MSP, ODIN, EBS, MaxLogit and the FSSD. Note that for Maha, we extract features from the same levels as for our approach, and give the unweighted mean of level-wise scores as the overall OOD score.

4.2.4 Image Preprocessing & Training Details

To compensate for inhomogeneities in image contrast caused by the image acquisition procedure, contrast limited adaptive histogram equalization (CLAHE) (Pizer et al., 1987) is applied (refer Figure 3 for a reference image). Further, the images are normalized and scaled down to the expected input resolution of the pre-trained classifier (224×224 pixels).

We fine-tune the pre-trained EfficientNet-B0 using the cross-entropy loss (Goodfellow et al., 2016) in combination with the Adam optimizer (Kingma and Ba, 2015), a learning rate of 0.0001 and batch-size of 16. The training set is used to train the 3-class classifier, and the validation set is used to detect the best model by calculating the macro-averaged F1-score over in-distribution data only. Model training is furthermore stopped when no improvement of at least 2% is achieved within 10 epochs for the in-distribution F1-score. Moreover, the validation set is used to parametrize the OOD-detection methods after the model has converged (i.e. estimate the joint in-distribution PDF by means of GMM). The test set is subsequently used to test model performance.

For experiments that employ OE, we use samples from the KUC to minimize

$$\mathcal{L}_{OE}(\phi(\boldsymbol{x}_{KUC})) = \log \sum_{i=1}^{d} e^{\phi_i(\boldsymbol{x}_{KUC})} - \frac{1}{d} \sum_{i=1}^{d} \phi_i(\boldsymbol{x}_{KUC}) \tag{6}$$

in addition to the supervised cross-entropy loss, which is similar to (Hendrycks et al., 2019c). The

Fold	F-score
0.0	96.2
1.0	95.5
2.0	96.4
3.0	95.3
4.0	96.4
μ	96.0
σ	0.5

(a)

(b)

Figure 4: In-distribution classification performance without application of OE. (a) shows per-fold F1-scores whereas (b) shows the corresponding confusion matrix of fold 1.

overall loss for training with OE is thus given as

$$\mathcal{L} = \mathcal{L}_{CE}(\phi(\boldsymbol{x}_{KKC})) + \lambda \mathcal{L}_{OE}(\phi(\boldsymbol{x}_{KUC})). \tag{7}$$

Based on preliminary experiments, we set λ to 0.5.

4.3 Results

We first compare with state-of-the-art OOD-detection methods in subsubsection 4.3.1. Afterwards, we perform an ablation study to investigate the influence of the proposed PDF-modeling across feature levels on joint performance in subsubsection 4.3.2. Next, we assess influence of OE on open set classification performance in subsubsection 4.3.3.

4.3.1 Natural Fiber Identification Performance under the Open Set

Table 2 shows that joint performance of the individual algorithms varies with the axes of biological variation that needs to be detected as OOD. Here, it can be seen that our proposed method is best-suited for detecting modifications incurred by chemical treatments (wool d&b), making it especially suitable for detecting adulteration attempts. Furthermore, our proposed method performs best and most consistent over all axes of variation, achieving an $\text{AUPR}_{\text{os,avg}}$ of 91.9 ± 2.1.

Regarding in-distribution classification performance, it can be seen that classification rates comparable to human raters are achieved with a macro-averaged in-distribution F1-score of 96.0 ± 0.5. Note that all methods share the same F1-scores given in Figure 4a.

Assessing pure OOD-detection performance, it can be seen that our method again performs the most consistent across all axes of biological variation, achieving an AUROC of 84.8 ± 8.4 (Table 3). Moreover, Maha surprisingly performs best with respect to OOD detection for two of the three UUCs while

Table 2: $AUPR_{os,avg}$ (%) for different UUCs. Best value per row is boldfaced.

	prediction-based				feature-based		
UUC	MSP	ODIN	EBS	MaxLogit	Maha	FSSD	Ours
Wool d&b	86.6	85.5	85.5	85.5	90.5	79.9	**93.8**
Merino	93.0	93.0	93.0	**93.1**	74.5	74.3	89.7
Silk	**93.5**	92.2	91.6	91.8	90.3	80.7	92.1
μ	91.0	90.3	90.0	90.1	85.1	78.3	**91.9**
σ	3.9	4.1	4.0	4.1	9.2	3.5	**2.1**

Table 3: AUROC (%) for different UUCs. Best value per row is boldfaced.

	prediction-based				feature-based		
UUC	MSP	ODIN	EBS	MaxLogit	Maha	FSSD	Ours
Wool d&b	32.4	31.6	31.8	31.8	**94.7**	35.2	92.8
Merino	82.6	83.4	**84.1**	84.0	38.6	32.4	76.0
Silk	89.9	84.9	83.4	84.1	**96.2**	59.1	85.5
μ	68.3	66.6	66.4	66.6	76.5	42.2	**84.8**
σ	31.3	30.4	30.0	30.2	32.8	14.7	**8.4**

achieving subpar $AUPR_{os,avg}$ values for them (refer Table 2).

We investigate the reasons behind this next by assessing the class-averaged, open set precision-recall curve for both Maha and our method on the UUC silk in Figure 5. Here, it can be seen that the curve of our proposed method possesses desirable characteristics, as no major dips in precision can be observed for low recall values, indicating that correctly classified in-distribution samples of all classes achieve lowest OOD scores. Conversely, Maha achieves higher precision & recall values later on, but exhibits dips in precision for low recall values. In combination with high OOD-detection performance, this indicates that lowest OOD scores are assigned to misclassified in-distribution data. Therefore, it is important to assess the joint-performance of OOD-detection & in-distribution classification when evaluating classifiers under the open set condition.

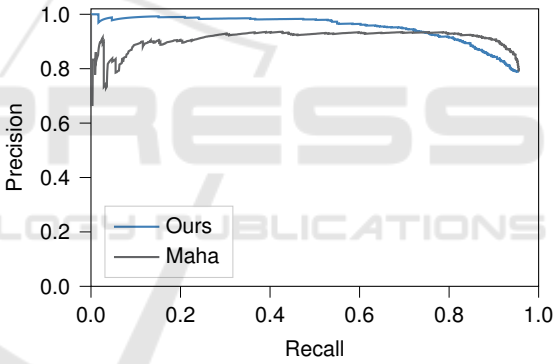


Figure 5: $PR_{os,avg}$ curve of the proposed OOD-method as well as Maha. UUC is silk.

4.3.2 Ablation Study

We perform an ablation experiment to assess the importance of modeling the joint PDF of in-distribution data across feature levels of a classifier. To this end, we compare our approach to fitting GMMs to the in-distribution data in every feature level individually, giving their average NLL as the overall OOD score.

Assessing results in Table 4, it can be seen that the proposed joint-modeling boosts OOD-detection performance measured by AUROC as well as joint performance as measured by $AUPR_{os,avg}$ for all evaluated UUCs. While lower standard deviations are achieved for the level-wise PDF-estimation, this is explained by the fact that the improvement of joint PDF-estimation varies with respect to the chosen UUC class.

4.3.3 Influence of OE

We assess the influence of OE on natural fiber identification under the open set condition next. To this end, we iterate over the three UUCs, using them as KUC for OE. In addition to OE, we also evaluate a simple k+1 classifier trained with the KUC as the *reject* class.

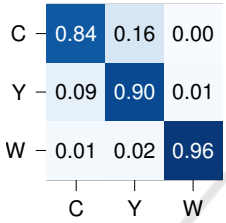
The results in Table 5 show that OE on average decreases joint performance, as indicated by larger PR_{dist}-values for all methods. In fact, OE does im-

Table 4: AUROC (%) and $AUPR_{os,avg}$ (%) for joint PDF-estimation across feature levels vs. level-wise PDF-estimation followed by averaging of level-wise OOD scores. Best values per metric and row are boldfaced.

UUC	AUROC		$AUPR_{os,avg}$	
	level-wise	joint	level-wise	joint
Wool d&b	91.9	**92.8**	93.3	**93.8**
Merino	75.4	**76.0**	89.5	**89.7**
Silk	83.6	**85.5**	91.6	**92.1**
μ	83.6	**84.8**	91.5	**91.9**
σ	**8.3**	8.4	**1.9**	2.1

Class	μ	σ
Wool d&b	93.0	1.4
Merino	95.5	1.0
Silk	92.6	1.7
no OE	96.0	0.5

(a)



(b)

Figure 6: In-distribution classification performance under application of OE. (a) shows per-fold F1-scores whereas (b) shows the corresponding confusion matrix of fold 1 with KUC = Wool d&b.

prove joint performance only for the KUC, and significantly reduces the performance for the UUCs not used for OE. The aforementioned effect is furthermore strongest for the k+1 classifier, and its inverse can be observed for FSSD.

When investigating the mechanisms underlying this phenomenon by assessing OOD detection and in-distribution classification individually, it can be seen that OE does improve OOD-detection performance on average (Table 6) at the cost of a reduced in-distribution classification performance (Figure 6a). Moreover, the OOD-detection results improve substantially only for the KUC class. In fact, they actually degrade for the UUCs across all assessed algorithms with the exception of Maha. Thus, OOD-detection improvements achieved by OE for KUCs do not propagate to UUCs. Coupled with an overall decrease in in-distribution classification performance, this translates to overall benefits as measured by PR_{dist} only for the KUCs.

5 DISCUSSION

In our work, we have investigated the performance of natural fiber identification algorithms under the open set condition. To this end, we identified the three main axes of variation that classifiers need to be robust against & set up a dataset that reflects these variations.

Experiments revealed that the proposed joint PDF-modeling across feature levels of a CNN performs best overall. Moreover, the obtained in-distribution classification rates were high enough to warrant a potential model deployment (>96% agreement with the nominal value is required for human operators (Zhang and Ainsworth, 2005)). The ablation study in subsubsection 4.3.2 further showed that the joint PDF-modeling was beneficial for all UUCs. This shows that consistency of model representations is not only predictive of model generalization (Natekar and Sharma, 2020), but can furthermore be used to boost OOD-detection performance of feature-based OOD-detection methods. Note that the joint modeling across feature levels was also shown to be beneficial for transfer-learning anomaly detection (AD), which is concerned with performing OOD detection under the one-class-classification setting (Defard et al., 2020).

Moreover, the importance of assessing joint OOD-detection & in-distribution classification performance became evident in subsubsection 4.3.1, where Maha achieved subpar joint performance as measured by $AUPR_{os,avg}$ despite achieving strong OOD-detection and in-distribution classification results. Therefore, we argue that one should report joint performance in future when evaluating algorithms under the open set condition rather than assessing in-distribution performance and OOD-detection performance separately, as was best practice so far (Liu et al., 2020; Lee et al., 2018b; Hendrycks and Gimpel, 2017; Liang et al., 2018). Here, the next step is to expand the proposed $AUPR_{os,avg}$ to the binary classification as well as to the object detection task.

We also assessed the influence of OE on natural fiber identification under the open set condition. Interestingly, it was found that performance improvements for the KUC used for OE did not propagate to the UUCs. These results are in line with the finding that a bias is introduced by sampling OOD data for OE, which may negatively impact OOD-detection performance of UUCs(Ye et al., 2021). Furthermore, OE reduced in-distribution classification performance in our experiments. Therefore, natural fiber identification under the open set condition does not benefit from OE, especially since many UUCs, e.g. adulteration procedures, are present.

Table 5: PR_{dist} (%) for open set classification with OE. μ and σ are calculated over all possible UUC – KUC combinations (All), over all combinations where UUC $\neq$ KUC, over all combinations where UUC = KUC and, for comparison, when no OE is applied.

			prediction-based				feature-based			
			MSP	ODIN	EBS	MaxLogit	Maha	FSSD	Ours	k+1
OE	All	μ	17.1	17.9	17.9	17.6	15.5	20.0	15.6	14.7
		σ	2.1	2.6	3.1	2.9	4.3	3.9	3.6	6.5
	UUC = KUC	μ	16.5	17.3	17.0	16.9	15.2	21.4	13.8	6.5
		σ	2.4	3.2	3.4	3.2	3.3	1.8	3.5	0.3
	UUC $\neq$ KUC	μ	17.3	18.1	18.3	18.0	15.6	19.2	16.5	18.8
		σ	1.8	2.5	2.7	2.6	4.8	4.5	3.6	2.8
No OE		μ	14.5	15.7	16.2	15.9	**13.8**	17.9	15.0	—
		σ	**0.5**	1.9	3.0	2.6	4.7	3.1	2.9	—

Table 6: AUROC (%) for OOD-detection with OE. μ and σ are calculated over all possible UUC – KUC combinations (All), over all combinations where UUC $\neq$ KUC, over all combinations where UUC = KUC and, for comparison, when no OE is applied.

			prediction-based				feature-based		
			MSP	ODIN	EBS	MaxLogit	Maha	FSSD	Ours
OE	All	μ	70.7	69.9	69.8	70.4	82.5	43.5	87.1
		σ	21.4	20.7	20.9	20.8	24.3	21.8	13.3
	UUC = KUC	μ	79.0	80.5	82.2	81.7	90.8	50.1	94.0
		σ	17.8	14.9	14.3	14.9	8.7	25.2	5.5
	UUC $\neq$ KUC	μ	66.5	64.6	63.7	64.7	78.4	40.2	83.7
		σ	22.0	21.2	21.0	21.1	29.0	18.4	15.1
No OE		μ	68.3	66.6	66.4	66.6	76.5	42.2	84.8
		σ	31.3	30.4	30.0	30.2	32.8	14.7	8.4

5.1 Limitations

While, to the best of our knowledge, we have used the largest and most diverse dataset so far, human operators are capable of distinguishing between at least 11 fiber types (10 specialty fibers + wool) (International Wool Textile Organisation, 2000). We will therefore expand the dataset, focussing on covering as many fiber types from as diverse sources as possible. Moreover, while our proposed method performed best overall, it did not achieve highest values for every single UUC. In fact, the best performance for two of the three UUCs was achieved by methods that are based on the classifier's unnormalized predictions (refer Table 2). In our future work, we will therefore develop OOD-detection methods that leverage both intermediate feature representations as well as the classifier's output. Last, we did not assess the performance of the model when distribution shifts occur for the KKCs. Ideally, the classifier would be robust to this under the open set condition, i.e. it would accept & correctly classify KKCs which have undergone input-distribution shifts rather than rejecting them. This would require the OOD-detection method to assign lower OOD scores to KKCs that have undergone distribution shifts compared to UUCs. Advances from the field of domain adaptation can be used as a starting point here (Saito and Saenko, 2021; Bashkirova et al., 2021).

6 CONCLUSION

In our work, we have thoroughly investigated the performance of natural fiber identification algorithms under the open set condition. To this end, we identified the three main axes of variation that natural fiber identification algorithms need to be robust against, and have created a dataset that is able to reflect them. Our experiments revealed that the proposed joint PDF-modeling across feature levels of a CNN performs best, achieving highest $AUPR_{os,avg}$ amongst all evaluated methods. Furthermore, we demonstrated that metrics of joint performance are necessary to fully re-

capitulate the behavior of a classifier under the open set condition. Our work thus shows that natural fiber identification algorithms provide promising results in real-world scenarios, i.e. under the open set condition. Our future work will focus on improving the OOD-detection performance and investigating the behavior of the classifier when simultaneously challenged with distribution shifts of KKCs and OOD-detection of UUCs.

ACKNOWLEDGEMENTS

This work was supported by the German Federation of Industrial Research Associations (AiF) under the grant number 21376 N.

REFERENCES

Ahuja, N. A., Ndiour, I., Kalyanpur, T., and Tickoo, O. (2019). Probabilistic modeling of deep features for out-of-distribution and adversarial detection. *arXiv preprint arXiv:1909.11786*.

American Society for Testing and Materials (1993). Standard methods for quantitative analysis of textiles, method d629-88.

Bashkirova, D., Hendrycks, D., Kim, D., Mishra, S., Saenko, K., Saito, K., Teterwak, P., and Usman, B. (2021). Visda-2021 competition universal domain adaptation to improve performance on out-of-distribution data. *arXiv preprint arXiv:2107.11011*.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blum, H., Sarlin, P.-E., Nieto, J., Siegwart, R., and Cadena, C. (2021). The fishyscapes benchmark: Measuring blind spots in semantic segmentation. *International Journal of Computer Vision*, 129(11):3119–3135.

Council of European Union (2011). Council regulation (EU) no 1007/2011.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Defard, T., Setkov, A., Loesch, A., and Audigier, R. (2020). Padim: a patch distribution modeling framework for anomaly detection and localization. *arXiv preprint arXiv:2011.08785*.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Freer, R. E. (1946). The wool products labeling act of 1939. *Temp. LQ*, 20:42.

Geng, C., Huang, S.-J., and Chen, S. (2020). Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Grcić., M., Bevandić., P., and Segvić., S. (2021). Dense open-set recognition with synthetic outliers generated by real nvp. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP,*, pages 133–143. INSTICC, SciTePress.

Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., and Song, D. (2019a). Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*.

Hendrycks, D. and Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Hendrycks, D., Lee, K., and Mazeika, M. (2019b). Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR.

Hendrycks, D., Mazeika, M., and Dietterich, T. (2019c). Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*.

Huang, H., Li, Z., Wang, L., Chen, S., Dong, B., and Zhou, X. (2020). Feature space singularity for out-of-distribution detection. *arXiv preprint arXiv:2011.14654*.

International Wool Textile Organisation (2000). Test method no. iwto-58-00: Scanning electron microscopic analysis of speciality fibres and sheep's wool and their blends.

International Wool Textile Organisation (2018). Statistics for the global wool production and textile industry,.

Kamoi, R. and Kobayashi, K. (2020). Why is the mahalanobis distance effective for anomaly detection? *arXiv preprint arXiv:2003.00402*.

Kim, Y., Kim, T., and Choi, H.-M. (2013). Qualitative identification of cashmere and yak fibers by protein fingerprint analysis using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Industrial & Engineering Chemistry Research*, 52(16):5563–5571.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kirichenko, P., Izmailov, P., and Wilson, A. G. (2020). Why normalizing flows fail to detect out-of-distribution data. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20578–20589. Curran Associates, Inc.

Lee, K., Lee, H., Lee, K., and Shin, J. (2018a). Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*.

Lee, K., Lee, K., Lee, H., and Shin, J. (2018b). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in*

Neural Information Processing Systems, pages 7167–7177.

Liang, S., Li, Y., and Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*.

Liu, W., Wang, X., Owens, J., and Li, S. Y. (2020). Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33.

McGregor, B. A. (2018). 4 - physical, chemical, and tensile properties of cashmere, mohair, alpaca, and other rare animal fibers. In Bunsell, A. R., editor, *Handbook of Properties of Textile and Technical Fibres (Second Edition)*, The Textile Institute Book Series, pages 105 – 136. Woodhead Publishing, second edition edition.

Mendes Júnior, P. R., de Souza, R. M., Werneck, R. d. O., Stein, B. V., Pazinato, D. V., de Almeida, W. R., Penatti, O. A. B., Torres, R. d. S., and Rocha, A. (2017). Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106(3):359–386.

Natekar, P. and Sharma, M. (2020). Representation based complexity measures for predicting generalization in deep learning. *arXiv preprint arXiv:2012.02775*.

Neal, L., Olson, M., Fern, X., Wong, W.-K., and Li, F. (2018). Open set learning with counterfactual images. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

Oza, P. and Patel, V. M. (2019). C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Phan, K.-H. and Wortmann, F. (2001). Quality assessment of goat hair for textile use. In *Silk, Mohair, Cashmere and Other Luxury Fibres*, pages 227–233. Elsevier.

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368.

Rane, P. P. and Barve, S. (2011). Standardization and optimization of mtdna isolation and molecular genetic analysis of d-loop region in animal natural fibres. *International Journal of Zoological Research*, 7(2):190.

Rippel, O., Bilitewski, N., Rahimi, K., Kurniadi, J., Herrmann, A., and Merhof, D. (2021a). Identifying pristine and processed animal fibers using machine learning. In *2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6.

Rippel, O., Mertens, P., and Merhof, D. (2021b). Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6726–6733.

Robson, D. (1997). Animal fiber analysis using imaging techniques: Part i: Scale pattern data. *Textile Research Journal*, 67(10):747–752.

Robson, D. (2000). Animal fiber analysis using imaging techniques: Part ii: Addition of scale height data. *Textile Research Journal*, 70(2):116–120.

Saito, K. and Saenko, K. (2021). Ovanet: One-vs-all network for universal domain adaptation. *arXiv preprint arXiv:2104.03344*.

Sun, X., Yang, Z., Zhang, C., Ling, K.-V., and Peng, G. (2020). Conditional gaussian distribution learning for open set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tan, M. and Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Waldron, S., Brown, C., and Komarek, A. M. (2014). The chinese cashmere industry: a global value chain analysis. *Development Policy Review*, 32(5):589–610.

Wortmann, F.-J. (1991). Quantitative fiber mixture analysis by scanning electron microscopy: Part iii: Round trial results on mohair / wool blends. *Textile Research Journal*, 61(7):371–374.

Wortmann, F.-J. and Wortmann, G. (1992). Quantitative fiber mixture analysis by scanning electron microscopy: Part iv: Assessment of light microscopy as an alternative tool for analyzing wool/specialty fiber blends. *Textile Research Journal*, 62(7):423–431.

Xing, W., Liu, Y., Deng, N., Xin, B., Wang, W., and Chen, Y. (2020a). Automatic identification of cashmere and wool fibers based on the morphological features analysis. *Micron*, 128:102768.

Xing, W., Liu, Y., Xin, B., Zang, L., and Deng, N. (2020b). The application of deep and transfer learning for identifying cashmere and wool fibers. *Journal of Natural Fibers*, 0(0):1–17.

Ye, Z., Chen, Y., and Zheng, H. (2021). Understanding the effect of bias in deep anomaly detection. *arXiv preprint arXiv:2105.07346*.

Yildiz, K. (2020). Identification of wool and mohair fibres with texture feature extraction and deep learning. *IET Image Processing*, 14(2):348–353.

Zhang, H., Li, A., Guo, J., and Guo, Y. (2020). Hybrid models for open set recognition. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M., editors, *Computer Vision - ECCV 2020*, pages 102–117, Cham. Springer International Publishing.

Zhang, L. and Ainsworth, W. (2005). Microscope analysis of animal fiber blends - training of operators. Technical report.

Zisselman, E. and Tamar, A. (2020). Deep residual flow for out of distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zoccola, M., Lu, N., Mossotti, R., Innocenti, R., and Montarsolo, A. (2013). Identification of wool, cashmere, yak, and angora rabbit fibers and quantitative determination of wool and cashmere in blend: a near infrared spectroscopy study. *Fibers and Polymers*, 14(8):1283–1289.