entropy

MDPI

Article

An Asymmetric Contrastive Loss for Handling Imbalanced Datasets

Valentino Vito * and **Lim Yohanes Stefanus**

Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia
* Correspondence: valentino.vito11@ui.ac.id

Abstract: Contrastive learning is a representation learning method performed by contrasting a sample to other similar samples so that they are brought closely together, forming clusters in the feature space. The learning process is typically conducted using a two-stage training architecture, and it utilizes the contrastive loss (CL) for its feature learning. Contrastive learning has been shown to be quite successful in handling imbalanced datasets, in which some classes are overrepresented while some others are underrepresented. However, previous studies have not specifically modified CL for imbalanced datasets. In this work, we introduce an asymmetric version of CL, referred to as ACL, in order to directly address the problem of class imbalance. In addition, we propose the asymmetric focal contrastive loss (AFCL) as a further generalization of both ACL and focal contrastive loss (FCL). The results on the imbalanced FMNIST and ISIC 2018 datasets show that the AFCL is capable of outperforming the CL and FCL in terms of both weighted and unweighted classification accuracies.

Keywords: asymmetric loss; class imbalance; contrastive loss; entropy; focal loss

check for updates

Citation: Vito, V.; Stefanus, L.Y. An Asymmetric Contrastive Loss for Handling Imbalanced Datasets. *Entropy* **2022**, *24*, 1303. https://doi.org/10.3390/e24091303

Academic Editors: Andrea Prati, Luis Javier García Villalba and Vincent A. Cicirello

Received: 10 August 2022
Accepted: 12 September 2022
Published: 15 September 2022

1. Introduction

Class imbalance is a major obstacle occurring within a dataset when certain classes in the dataset are overrepresented (referred to as majority classes), while some are underrepresented (referred to as minority classes). This can be problematic for a large number of classification models. A deep learning model such as a convolutional neural network (CNN) might not be able to properly learn from the minority classes. Consequently, the model would be less likely to correctly identify minority samples as they occur. This is especially crucial in medical imaging, since a model that cannot identify rare diseases would not be effective for diagnostic purposes. For example, the ISIC 2018 dataset [1,2] is an imbalanced medical dataset that consists of images of skin lesions that appear in various frequencies during screening.

To produce a less imbalanced dataset, it is possible to resample the dataset by either increasing the number of minority samples [3–6] or decreasing the number of majority samples [7–10]. Other methods for handling class imbalance include substituting the standard cross-entropy (CE) loss for a more suitable loss, such as the focal loss (FL). Lin et al. [11] modified the CE loss into FL so that minority classes can be prioritized. This is done by ensuring that the model focuses on samples that are harder to classify during model training. Recent studies also unveiled the potential of contrastive learning as a way to combat imbalanced datasets [12–15].

Contrastive learning is performed by contrasting a sample (called an *anchor*) to other similar samples (called positive samples) so that they are mapped closely together in the feature space. As a consequence, dissimilar samples (called negative samples) are pushed away from the anchor, forming clusters in the feature space based on similarity. In this research, contrastive learning is done using a two-stage training architecture, which utilizes the contrastive loss (CL) formulated by Khosla et al. [16]. This formulation of CL is supervised, and it can contrast the anchor to multiple positive samples belonging to the same class. This is unlike self-supervised contrastive learning [17–20], which contrasts the anchor to only one positive sample in the mini-batch.

In this work, we propose a modification of supervised CL that is referred to as the asymmetric contrastive loss (ACL). Unlike CL, the ACL is able to directly contrast the anchor to its negative samples so that they are pushed apart in the feature space. This becomes important when a rare sample has no other positive samples in the mini-batch. To our knowledge, we are the first to modify the supervised version of CL in order to address class imbalance, effectively augmenting several studies performed previously in [12,13]. The proposed ACL is aimed toward improving the effectiveness of the two-stage architecture originally presented in [12,13], especially in the feature learning aspect. In addition, the ACL is designed as a generalization of CL, and thus, it provides more flexibility and tuning opportunities as a loss function.

We also consider the asymmetric variant of the focal contrastive loss (FCL) [21], which is called the asymmetric focal contrastive loss (AFCL). Using FMNIST and ISIC 2018 as datasets, experiments were performed to test the performance of both the ACL and AFCL in binary classification tasks. It was observed that the AFCL was superior to the CL and FCL in multiple class imbalance scenarios, provided that suitable hyperparameters were used. In addition, this work provides a streamlined survey of the literature related to entropy and loss functions.

2. Related Work

Several studies have been conducted in recent years on the application of contrastive losses to imbalanced datasets. On Siamese networks, for example, Wang et al. [14] and Alenezi et al. [15] proposed the novel focal CL and W-shaped CL, respectively. Their methods managed to achieve state-of-the-art performance in handling the class imbalance problem, wherein Wang et al. used satellite images and Alenezi et al. used skin lesion images as datasets. Their CL functions had a different form from that of the supervised CL of Khosla et al. [16], which is the CL that upon which our study is based.

Marrakchi et al. [12] and Chen et al. [13] independently adopted supervised CL to combat class imbalance in the medical domain. They both used a two-stage architecture consisting of (1) feature learning using CL, followed by (2) fine-tuning using classification loss. Their architectures were almost identical; they differed only in the type of loss function during fine-tuning (Marrakchi et al. used cross-entropy loss, while Chen et al. used focal loss). One limitation present in these studies was that CL was not modified further to deal with imbalance and was implemented as is. Therefore, our aim is to generalize CL in order to effectively learn from imbalanced datasets using the aforementioned two-stage architecture.

In this paper, we present a novel CL referred to as the ACL, and we include its focal-based variant, AFCL. Our motivation for introducing the losses comes from both the asymmetric loss due to Ben-Baruch et al. [22] and the focal contrastive loss due to Zhang et al. [21], whose explanations are provided in Section 3. Although these losses were proposed for different applications (fine-tuning and multi-label classification, respectively), it turns out that these ideas can be applied to our goal of modifying CL so as to handle imbalance.

3. Background on Entropy and Loss Functions

In this section, we provide a literature review on the basics of information theory and loss functions for easy reference.

3.1. Entropy, Information, and Divergence

Introduced by Shannon [23], entropy provides a measure of the amount of information contained in a random variable, usually in bits. The *entropy* $H(X)$ of a random variable X is given by the formula

$$H(X) = \mathbb{E}_{P_X}[-\log(P_X(X))]. \tag{1}$$

Given two random variables X and Y, their *joint entropy* $H(X,Y)$ is the entropy of the joint random variable (X,Y):

$$H(X,Y) = \mathbb{E}_{P_{(X,Y)}}\Big[-\log(P_{(X,Y)}(X,Y))\Big]. \tag{2}$$

In addition, the *conditional entropy* $H(Y \mid X)$ is defined as

$$H(Y \mid X) = \mathbb{E}_{P_{(Y,X)}}\Big[-\log(P_{Y|X}(Y \mid X)\Big]. \tag{3}$$

Conditional entropy is used to measure the average amount of information contained in Y when the value of X is given. Conditional entropy is bounded above by the original entropy; that is, $H(Y \mid X) \leq H(Y)$, with equality if and only if X and Y are independent [24]. The formulas for entropy, joint entropy, and conditional entropy can be derived via an axiomatic approach [25,26].

The *mutual information* $I(X;Y)$ is a measure of dependence between random variables X and Y [27]. It provides the amount of information about one random variable provided by the other random variable, and it is defined by

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X). \tag{4}$$

Mutual information is symmetric. In other words, $I(X;Y) = I(Y;X)$. Mutual information is also nonnegative ($I(X;Y) \geq 0$), and $I(X;Y) = 0$ if and only if X and Y are independent [24].

The dissimilarity between random variables X and X' on the same space $\mathcal{X}$ can be measured using the notion of *KL-divergence*:

$$D_{\mathrm{KL}}(X \,\|\, X') = \mathbb{E}_{P_X}\left[\log\left(\frac{P_X(X)}{P_{X'}(X)}\right)\right]. \tag{5}$$

Similarly to mutual information, KL-divergence is nonnegative ($D_{\mathrm{KL}}(X \,\|\, X') \geq 0$), and $D_{\mathrm{KL}}(X \,\|\, X') = 0$ if and only if $X = X'$ [24]. Unlike mutual information, KL-divergence is asymmetric, so $D_{\mathrm{KL}}(X \,\|\, X')$ and $D_{\mathrm{KL}}(X' \,\|\, X)$ are not necessarily equal.

3.2. Cross-Entropy and Focal Loss

Given random variables X and $\hat{X}$ on the same space $\mathcal{X}$, their *cross-entropy* $H(X;\hat{X})$ is defined as [28]:

$$H(X;\hat{X}) = \mathbb{E}_{P_X}[-\log(P_{\hat{X}}(X)]. \tag{6}$$

Cross-entropy is the average number of bits needed to encode the true distribution X when its estimate $\hat{X}$ is provided [29]. A small value of $H(X;\hat{X})$ implies that $\hat{X}$ is a good estimate for X. Cross-entropy is connected to KL-divergence via the following identity:

$$H(X;\hat{X}) = H(X) + D_{\mathrm{KL}}(X \,\|\, \hat{X}). \tag{7}$$

When $\hat{X} = X$, the equality $H(X;\hat{X}) = H(X)$ holds.

Now, the cross-entropy loss and focal loss are provided within the context of a binary classification task consisting of two classes labeled 0 and 1. Suppose that $y \in \{0,1\}$ denotes the ground-truth class and $p \in [0,1]$ denotes the estimated probability for the class labeled 1. The value of $1-p$ is then the estimated probability for the class labeled 0. The *cross-entropy (CE) loss* is given by

$$\begin{aligned}\mathcal{L}_{\mathrm{CE}} &= -y\log(p) - (1-y)\log(1-p) \\ &= \begin{cases} -\log(p) & y = 1, \\ -\log(1-p) & y = 0. \end{cases}\end{aligned}$$

If $y = 1$, then the loss $\mathcal{L}_{\text{CE}}$ is zero when $p = 1$. On the other hand, if $y = 0$, then the loss is zero when $1 - p = 1$. In either case, the CE loss is minimized when the estimated probability of the true class is maximized, which is the desired property of a good classification model.

The *focal loss* (FL) [11] is a modification of the CE loss introduced to put more focus on hard-to-classify examples. It is given by the following formula:

$$\mathcal{L}_{\text{foc}} = -y(1-p)^{\gamma}\log(p) - (1-y)p^{\gamma}\log(1-p). \tag{8}$$

The parameter γ in $\mathcal{L}_{\text{foc}}$ is known as the *focusing parameter*. Choosing a larger value of γ would push the model to focus on training from the misclassified examples. For instance, suppose that $\gamma = 4$ and denote the estimated probability of the true class by p_t. The graph in Figure 1 shows that when $p_t > 0.5$, the FL is quite small. Hence, the model would be less concerned about learning from an example when p_t is already sufficiently large. FL is a useful choice when class imbalance exists, as it can help the model focus on the less represented samples within the dataset.

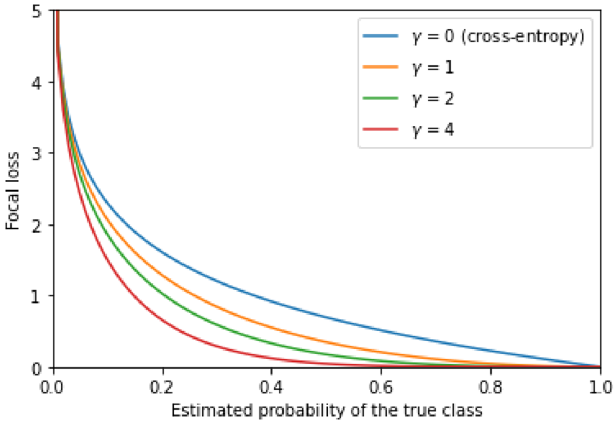

Figure 1. A graph illustrating the focal loss given the predicted probability of the ground-truth class, with varying values of γ.

3.3. Asymmetric Loss

For multi-label classification with K labels, let $y_i \in \{0,1\}$ be the ground truth for class i and let $p_i \in [0,1]$ be its estimated probability obtained by the model. The aggregate classification loss is then

$$\mathcal{L} = \sum_{i=1}^{K} \mathcal{L}_i, \tag{9}$$

where

$$\mathcal{L}_i = -y_i\mathcal{L}_i^{+} - (1-y_i)\mathcal{L}_i^{-}. \tag{10}$$

If FL is the chosen type of loss, $\mathcal{L}_i^{+}$ and $\mathcal{L}_i^{-}$ are set as follows:

$$\mathcal{L}_i^{+} = (1-p_i)^{\gamma}\log(p_i) \quad \text{and} \quad \mathcal{L}_i^{-} = p_i^{\gamma}\log(1-p_i). \tag{11}$$

In a typical multi-label dataset, the ground truth y_i has value 0 for the majority of classes i. Consequently, the negative terms $\mathcal{L}_i^{-}$ dominate in the calculation of the aggregate loss $\mathcal{L}$. *Asymmetric loss* (ASL) [22] is a proposed solution to this problem. ASL emphasizes the contribution of the positive terms by modifying the losses of (11) to

$$\mathcal{L}_i^{+} = (1-p_i)^{\gamma^{+}}\log(p_i) \tag{12}$$

and

$$\mathcal{L}_i^{-} = (p_i^{(m)})^{\gamma^{-}}\log(1-p_i^{(m)}), \tag{13}$$

where γ^+, γ^- are hyperparameters and $p_i^{(m)}$ is the *shifted probability* of p_i obtained from the *probability margin* $m \geq 0$ via the formula

$$p_i^{(m)} = \max(p_i - m, 0). \tag{14}$$

This shift helps decrease the contribution of $\mathcal{L}_i^-$. Indeed, if we set $m = 1$, then $\mathcal{L}_i^- = 0$.

3.4. Contrastive Loss

Contrastive learning is a learning method for learning representations from data. A supervised approach of contrastive learning was introduced by Khosla et al. [16] to learn from a set of sample–label pairs $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ in a mini-batch of size N. The samples $\boldsymbol{x}_i$ are fed through a feature encoder $\mathrm{Enc}(\cdot)$ and a projection head $\mathrm{Proj}(\cdot)$ in succession to obtain features $\boldsymbol{z}_i = \mathrm{Proj}(\mathrm{Enc}(\boldsymbol{x}_i))$. The feature encoder extracts features from $\boldsymbol{x}_i$, whereas the projection head projects the features into a lower dimension and applies ℓ_2-normalization so that $\boldsymbol{z}_i$ lies in the unit hypersphere. In other words, $\|\boldsymbol{z}_i\|_2 = 1$.

A pair $(\boldsymbol{z}_i, \boldsymbol{z}_j)$, where $i \neq j$, is referred to as a *positive pair* if the features share the same class label ($\boldsymbol{y}_i = \boldsymbol{y}_j$), and it is a *negative pair* if the features have different class labels ($\boldsymbol{y}_i \neq \boldsymbol{y}_j$). Contrastive learning aims to maximize the similarity between $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ whenever they form a positive pair and minimize their similarity whenever they form a negative pair. This similarity is measured with cosine similarity [29]:

$$\kappa(\boldsymbol{z}_i, \boldsymbol{z}_j) = \frac{\boldsymbol{z}_i \cdot \boldsymbol{z}_j}{\|\boldsymbol{z}_i\|_2 \|\boldsymbol{z}_j\|_2} = \boldsymbol{z}_i \cdot \boldsymbol{z}_j. \tag{15}$$

From the above equation, we have $\kappa(\boldsymbol{z}_i, \boldsymbol{z}_j) \in [-1, 1]$. In addition, $\kappa(\boldsymbol{z}_i, \boldsymbol{z}_j) = 1$ when $\boldsymbol{z}_i = \boldsymbol{z}_j$, and $\kappa(\boldsymbol{z}_i, \boldsymbol{z}_j) = -1$ when $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ form a 180° angle.

Fixing $\boldsymbol{z}_i$ as the anchor, let $A_i = \{\boldsymbol{z}_k \mid k \neq i\}$ be the set of features other than $\boldsymbol{z}_i$ and let $P_i = \{\boldsymbol{z}_k \in A_i \mid \boldsymbol{y}_k = \boldsymbol{y}_i\}$ be the set of $\boldsymbol{z}_k$ such that $(\boldsymbol{z}_i, \boldsymbol{z}_k)$ is a positive pair. The predicted probability p_{ij} that $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ belong to the same class is obtained by applying the softmax function to the the set of similarities between $\boldsymbol{z}_i$ and $\boldsymbol{z}_k \in A_i$:

$$p_{ij} = \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_j / \tau)}{\sum_{\boldsymbol{z}_k \in A_i} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_k / \tau)}, \tag{16}$$

where τ is referred to as the *temperature parameter*. Since our goal is to maximize p_{ij} whenever $\boldsymbol{z}_j \in P_i$, the *contrastive loss* that is to be minimized is formulated as

$$\mathcal{L}_{\mathrm{con}} = -\sum_{i=1}^{n} \frac{1}{|P_i|} \sum_{\boldsymbol{z}_j \in P_i} \log(p_{ij}). \tag{17}$$

Information-theoretical properties of $\mathcal{L}_{\mathrm{con}}$ are given in [21], for which we provide a summary. Let X, Y, and Z denote random variables of the samples, labels, and features, respectively. The following theorem states that $\mathcal{L}_{\mathrm{con}}$ is positively proportional to $H(Z \mid Y) - H(Z)$ under the assumption that no class imbalance exists.

Theorem 1 (Zhang et al. [21])**.** *Assuming that features are ℓ_2-normalized and the dataset is balanced,*

$$\mathcal{L}_{con} \propto H(Z \mid Y) - H(Z). \tag{18}$$

Theorem 1 implies that minimizing $\mathcal{L}_{\mathrm{con}}$ is equivalent to minimizing the conditional entropy $H(Z \mid Y)$ and maximizing the feature entropy $H(Z)$. Since $I(Z;Y) = H(Z) - H(Z \mid Y)$, minimizing $\mathcal{L}_{\mathrm{con}}$ is equivalent to maximizing the mutual information $I(Z;Y)$ between features Z and class labels Y. In other words, contrastive learning aims to extract the maximum amount of information from class labels and encode it in the form of features.

After the features are extracted, a classifier $\mathrm{Clas}(\cdot)$ is assigned to convert z_i into a prediction $\hat{y}_i = \mathrm{Clas}(z_i)$ of the class label. The random variable of predicted class labels is denoted by $\hat{Y}$.

For the next theorem, the definition of *conditional cross-entropy* $H(Y; \hat{Y} \mid Z)$ is given as follows:

$$H(Y; \hat{Y} \mid Z) = \mathbb{E}_{P_{(Y,Z)}}\Big[-\log(P_{(\hat{Y},Z)}(Y, Z)\Big]. \tag{19}$$

Conditional CE measures the average amount of information needed to encode the true distribution Y using its estimate $\hat{Y}$ given the value of Z. A small value of $H(Y; \hat{Y} \mid Z)$ implies that $\hat{Y}$ is a good estimate for Y given Z.

Theorem 2 (Zhang et al. [21])**.** *Assuming that features are ℓ_2-normalized and the dataset is balanced,*

$$\mathcal{L}_{con} \propto \inf H(Y; \hat{Y} \mid Z) - H(Y), \tag{20}$$

where the infimum is taken over classifiers.

Theorem 2 implies that minimizing $\mathcal{L}_{\mathrm{con}}$ will minimize the infimum of conditional cross-entropy $H(Y; \hat{Y} \mid Z)$ taken over classifiers. As a consequence, contrastive learning is able to encode features in Z such that the best classifier can produce a good estimate of Y given the information provided by the feature encoder.

The formula for $\mathcal{L}_{\mathrm{con}}$ can be modified so as to resemble the focal loss, resulting in a loss function known as the *focal contrastive loss* (FCL) [21]:

$$\mathcal{L}_{\mathrm{FC}} = -\sum_{i=1}^{n} \frac{1}{|P_i|} \sum_{z_j \in P_i} (1 - p_{ij}) \log(p_{ij}). \tag{21}$$

4. Proposed Loss Functions and Architecture

In this section, our proposed modification of the contrastive loss, which is called the asymmetric contrastive loss, is introduced. In addition, the architecture of the model in which the contrastive losses are implemented is explained. Our proposed asymmetric loss function is novel, while the architecture is obtained from [12,13] with no changes made. Thus, our contribution lies simply in the change of the loss function.

4.1. Asymmetric Contrastive Loss

In (17), the inside summation of the contrastive loss is evaluated over P_i. Consequently, according to (16), each anchor z_i is contrasted with vectors z_j that belong to the same class. This does not present a problem when the mini-batch contains plenty of examples from each class. However, the calculated loss may not give each class a fair contribution when some classes are less represented in the mini-batch.

In Figure 2, a sampled mini-batch consists of 11 examples with a blue-colored class label and one example with a red-colored class label. When the anchor z_i is the representation of the red-colored sample, z_i does not directly contribute to the calculation of $\mathcal{L}_{\mathrm{con}}$, since P_i is empty. In other words, z_i cannot be contrasted to any other sample in the mini-batch. This scenario is likely to happen when the dataset is imbalanced, and it motivates us to modify CL so that each anchor z_i can also be contrasted with z_j not belonging to the same class.

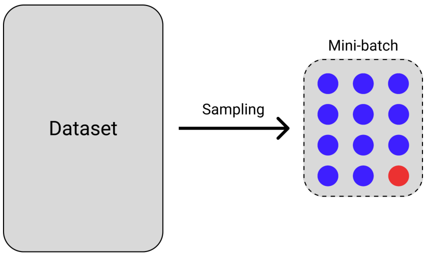

Figure 2. A mini-batch consisting of 11 examples with a blue-colored class label and one example with a red-colored class label.

Let $N_i = A_i \setminus P_i$ be the set of vectors z_k such that (z_i, z_k) is a negative pair. Motivated by the $\mathcal{L}_i^+$ and $\mathcal{L}_i^-$ of (10), we define

$$\mathcal{L}_i^+ = \frac{1}{|P_i|} \sum_{z_j \in P_i} \log(p_{ij}) \tag{22}$$

and

$$\mathcal{L}_i^- = \frac{1}{|N_i|} \sum_{z_j \in N_i} \log(1 - p_{ij}), \tag{23}$$

where $p_{ij} = \exp(z_i \cdot z_j / \tau) / \sum_{z_k \in A_i} \exp(z_i \cdot z_k / \tau)$. The loss function $\mathcal{L}_i^+$ contrasts z_i to vectors in P_i, whereas $\mathcal{L}_i^-$ contrasts z_i to vectors in N_i. The resulting *asymmetric contrastive loss* (ACL) is given by the formula

$$\mathcal{L}_{\text{AC}} = -\sum_{i=1}^{n} (\mathcal{L}_i^+ + \eta \mathcal{L}_i^-), \tag{24}$$

where $\eta \geq 0$ is a fixed hyperparameter. If $\eta = 0$, then $\mathcal{L}_{\text{AC}} = \mathcal{L}_{\text{con}}$. Hence, ACL is a generalization of CL.

When the batch size is set to a large number (over 100, for example), the value p_{ij} tends to be very small. This causes $\mathcal{L}_i^-$ to be much smaller than $\mathcal{L}_i^+$. In order to balance their contribution to the total loss $\mathcal{L}_{\text{AC}}$, a large value for η is usually chosen (between 60 and 300 in our experiment).

In summary, we propose ACL in order to (1) generalize the CL via the addition of a summation over negative samples and (2) specifically address the problem of class imbalance. ACL is intended to be both more flexible and robust to imbalances than the vanilla CL.

4.2. Asymmetric Focal Contrastive Loss

Following the formulation of $\mathcal{L}_{\text{FC}}$ in (21), $\mathcal{L}_i^+$ can be modified to have the following formula:

$$\mathcal{L}_i^+ = \frac{1}{|P_i|} \sum_{z_j \in P_i} (1 - p_{ij})^\gamma \log(p_{ij}). \tag{25}$$

Using this loss, the *asymmetric focal contrastive loss* (AFCL) is then given by

$$\mathcal{L}_{\text{AFC}} = -\sum_{i=1}^{n} (\mathcal{L}_i^+ + \eta \mathcal{L}_i^-), \tag{26}$$

where $\mathcal{L}_i^- = \frac{1}{|N_i|} \sum_{z_j \in N_i} \log(1 - p_{ij})$. We do not modify $\mathcal{L}_i^-$ by adding the multiplicative term $(p_{ij})^\gamma$, since p_{ij} is usually too small and would make $\mathcal{L}_i^-$ vanish if the term is added.

We have $\mathcal{L}_{\text{AFC}} = \mathcal{L}_{\text{FC}}$ when $\gamma = 1$. Thus, AFCL generalizes the FCL. Unlike with the FCL, we add the hyperparameter $\gamma \geq 0$ to the loss function so as to provide some flexibility to the loss function.

4.3. Model Architecture

This section explains the inner workings of the classification model used for the implementation of the contrastive losses. The architecture of the model is taken from [12,13]. The training strategy for the model, as shown in Figure 3, comprises two stages: the feature-learning stage and the fine-tuning stage.

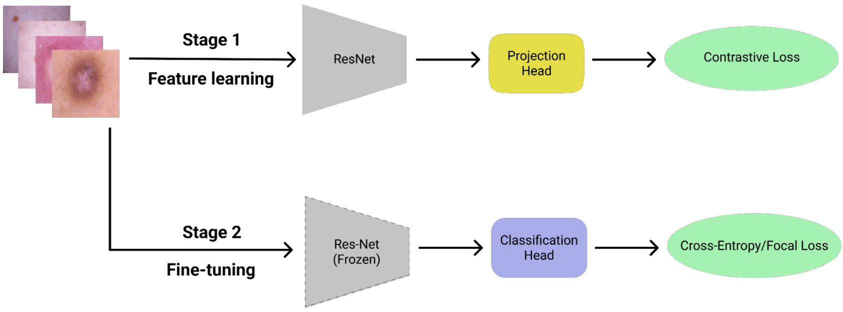

Figure 3. A two-stage training strategy consisting of: (1) feature learning using contrastive loss and (2) classifier fine-tuning using either FL or CE loss.

In the first stage, each mini-batch is fed through a feature encoder. We consider either ResNet-18 or ResNet-50 [30] for the architecture of the feature encoder. The output of the feature encoder is projected by the projection head to generate a vector z of length 128. If ResNet-18 is used for the feature encoder, then the projection head consists of two layers of lengths 512 and 128. If ResNet-50 is used, then the two layers are of lengths 2048 and 128. Afterwards, z is ℓ_2-normalized, and the model parameters are updated using some version of the contrastive loss (either CL, FCL, ACL, or AFCL).

After the first stage is complete, the feature encoder is frozen and the projection head is removed. In its place, we have a one-layer classification head that generates the estimated probability that the training sample belongs to a certain class. The parameters of the classification head are updated using either the FL or CE loss. The final classification model is the feature encoder trained during the first stage, together with the classification head trained during the second stage. Since the classification head is a significantly smaller architecture than the feature encoder, the training is mostly focused on the first stage. As a consequence, we typically need a larger number of epochs for the feature-learning stage compared to the fine-tuning stage.

5. Experiments

The datasets and settings of our experiments are outlined in this section. We provide and discuss the results of the experiments on the FMNIST and ISIC 2018 datasets. The PyTorch implementation is available on GitHub (https://github.com/valentinovito/Asymmetric-CL, accessed on 8 September 2022).

5.1. Datasets

In our experiments, the training strategy outlined in Section 4.3 was applied to two imbalanced datasets. The first was a modified version of the Fashion-MNIST (FMNIST) dataset [31], and the second was the International Skin Imaging Collaboration (ISIC) 2018 medical dataset [1,2].

The FMNIST dataset consisted of low-resolution (28 × 28 pixels) grayscale images of ten classes of clothing. In this study, we took only two classes to form a binary classification task: the T-shirt and shirt classes. The samples were taken such that the proportion between the T-shirt and shirt images could be imbalanced, depending on the scenario. On the other hand, the ISIC 2018 dataset consisted of high-resolution RGB images of seven classes of skin lesions. As with FMNIST, we used only two classes for the experiments: the melanoma and dermatofibroma classes. Illustrations of the sample images of both datasets are provided in Figure 4.

FMNIST was chosen as our dataset, since, although simple, it is a benchmark dataset for testing deep learning models for computer vision. On the other hand, ISIC 2018 was chosen since it is a domain-appropriate imbalanced dataset for our model. We first applied the model (using AFCL as the loss function) to the more lightweight FMNIST dataset under various class imbalance scenarios. This was conducted to check the appropriate values of the η and γ parameters of the AFCL under different imbalance conditions. Afterwards, the model was applied to the ISIC 2018 dataset using the optimal parameter values obtained during the FMNIST experiments.

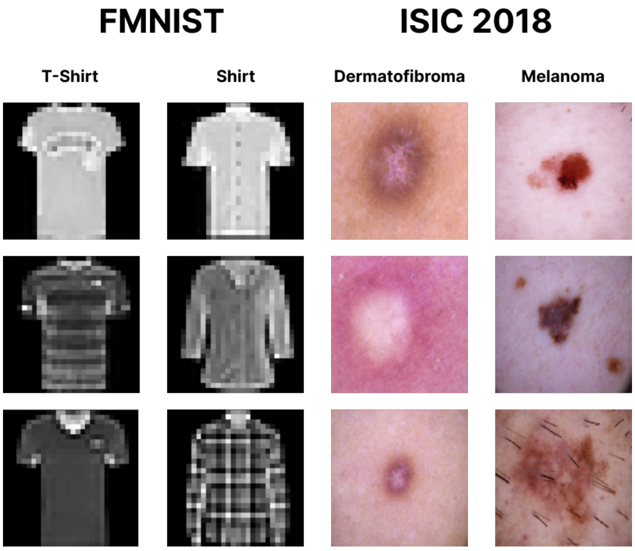

Figure 4. Sample images of the FMNIST and ISIC 2018 datasets.

5.2. Experimental Details

The experiments were conducted using the NVIDIA Tesla P100-PCIE GPU allocated by the Google Colaboratory Pro platform. The models and loss functions were implemented using PyTorch. To process the FMNIST dataset, we used the simpler ResNet-18 architecture as the feature encoder and trained it for 20 epochs. On the other hand, to process the ISIC 2018 dataset, we used the deeper ResNet-50 as the feature encoder and trained it for 40 epochs. For both the FMNIST and ISIC 2018 datasets, the learning rate and batch size were set to 10^{-2} and 128, respectively. In addition, the classification head was trained for 10 epochs. The encoder and the classification head were both trained using the Adam optimizer. Finally, the temperature parameter τ of the contrastive loss was set to its default value of 0.07.

The evaluation metrics utilized in the experiment were (weighted) accuracy and unweighted accuracy (UWA), both of which could be calculated from the number of true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP) using the formulas

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{27}$$

and

$$\text{UWA} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right), \tag{28}$$

respectively. Unlike accuracy, the UWA provided the average of the individual class accuracies regardless of the number of samples in the test set of each class. UWA is an appropriate metric when a dataset is significantly imbalanced [32].

For heavily imbalanced datasets, a high accuracy and low UWA may mean that the model is biased towards classifying samples as part of the majority class. This indicates that the model does not properly learn from the minority samples. In contrast, a lower accuracy with a high UWA indicates that the model takes significant risks to classify some samples as part of the minority class. Our aim was to construct a model that maximized both metrics simultaneously; that is, a model that could learn unbiasedly from both the majority and minority samples with minimal misclassification error.

5.3. Experiments Using FMNIST

The data used in the FMNIST experiment comprised 1000 images classified as either a T-shirt or a shirt. The dataset was split 70/30 for model training and testing. The images were augmented using random rotations and random flips. We deployed 11 class imbalance scenarios on the dataset, which controlled the proportion between the T-shirt class and the shirt class. For example, if the proportion was 60:40, then 600 T-shirt images and 400 shirt images were sampled to form the experimental dataset. Our proportions ranged from 50:50 to 98:2.

During the first stage, the ResNet-18 encoder was trained using the AFCL. Afterwards, the classification head was trained using the CE loss during the second stage. As AFCL contains two parameters, η and γ, our goal was to tune each of these parameters independently, keeping the other parameter fixed. First, η was tuned as we set $\gamma = 0$, followed by the tuning of γ as we set $\eta = 0$. Each experiment was performed four times in total. The average accuracy and UWA of these four runs are provided in Table 1 (for the tuning of η) and Table 2 (for the tuning of γ).

Table 1. The accuracy and UWA (averaged over four independent runs) of 11 class imbalance scenarios using various values of η for the AFCL. The parameter γ was consistently set to 0.

Scenario	**Metric**	η					
		0	**60**	**120**	**180**	**240**	**300**
50:50	Accuracy	78.92	77.83	**79.75**	71.08	77.17	78.83
	UWA	79.00	78.28	**80.32**	72.53	77.87	79.42
55:45	Accuracy	**79.50**	**79.50**	79.33	77.83	77.67	77.75
	UWA	78.70	**79.34**	79.15	77.17	78.21	76.50
60:40	Accuracy	**84.50**	82.92	82.42	81.33	82.08	83.17
	UWA	**83.09**	81.82	81.27	79.71	81.74	81.66
65:35	Accuracy	81.50	**83.42**	83.25	81.59	82.58	79.25
	UWA	79.19	**80.91**	80.73	77.92	79.43	75.42
70:30	Accuracy	82.50	84.33	**85.08**	82.08	83.42	83.00
	UWA	78.41	78.26	**80.91**	77.78	79.14	75.11
75:25	Accuracy	86.75	85.17	85.58	85.17	**86.92**	86.58
	UWA	77.87	76.48	77.74	77.03	**78.63**	77.57
80:20	Accuracy	86.00	87.25	87.33	87.92	87.00	**88.25**
	UWA	76.16	74.65	76.94	76.28	**77.49**	76.97
85:15	Accuracy	87.33	87.08	86.75	87.42	87.33	**87.67**
	UWA	**70.08**	66.34	55.77	68.33	69.83	62.83
90:10	Accuracy	90.83	91.00	90.83	90.67	89.50	**91.67**
	UWA	64.91	68.61	66.11	64.02	61.77	**72.58**
95:5	Accuracy	**94.42**	93.33	93.42	94.00	92.83	93.25
	UWA	54.77	**60.70**	54.24	50.00	49.38	54.80
98:2	Accuracy	97.42	97.83	98.08	98.08	**98.33**	98.08
	UWA	52.45	52.66	**55.87**	**55.87**	49.83	52.79

Table 2. The accuracy and UWA (averaged over four independent runs) of 11 class imbalance scenarios using various values of γ for the AFCL. The parameter η was consistently set to 0.

Scenario	**Metric**	γ					
		0	**1**	**2**	**4**	**7**	**10**
50:50	Accuracy	**78.08**	74.83	77.08	77.58	76.58	77.50
	UWA	**77.70**	74.84	76.77	77.55	76.55	77.25
55:45	Accuracy	80.17	81.25	80.75	80.00	**81.75**	76.83
	UWA	80.14	81.19	80.69	79.96	**81.70**	76.82
60:40	Accuracy	79.42	78.50	77.92	80.17	**80.67**	80.08
	UWA	**84.42**	83.42	80.00	83.00	82.42	82.92
65:35	Accuracy	**84.42**	83.42	80.00	83.00	82.42	82.92
	UWA	**81.98**	81.22	77.87	80.39	80.68	80.16
70:30	Accuracy	83.75	83.83	82.17	82.58	**84.83**	82.25
	UWA	79.64	79.18	77.82	77.51	**79.67**	78.71
75:25	Accuracy	85.42	**86.17**	84.42	84.83	85.75	86.00
	UWA	76.27	**79.85**	77.08	76.41	77.34	78.47
80:20	Accuracy	89.33	**89.58**	87.67	89.42	87.33	88.00
	UWA	77.59	78.67	78.43	**79.31**	78.97	70.12
85:15	Accuracy	87.42	89.00	88.17	88.33	89.08	**90.08**
	UWA	64.97	72.08	71.99	71.47	71.95	**77.04**
90:10	Accuracy	92.42	92.33	**93.42**	93.25	92.58	91.25
	UWA	64.00	67.94	66.04	74.42	**80.54**	68.35
95:5	Accuracy	94.17	93.17	**95.33**	95.00	94.00	95.09
	UWA	**62.13**	53.11	57.64	59.17	55.22	55.82
98:2	Accuracy	**96.92**	**96.92**	95.00	96.00	**96.92**	96.67
	UWA	**56.59**	51.56	55.61	52.63	53.10	52.98

For the tuning of η, six values of η were experimented on: $\eta \in \{0, 60, 120, 180, 240, 300\}$. When $\eta = 0$, the loss function was reduced to the ordinary CL. As observed in Table 1, the optimal value of η tended to be larger when the dataset was moderately imbalanced. As the scenario went from 60:40 to 90:10, the parameter η that maximized accuracy increased in value, from $\eta = 0$ when the proportion was 60:40 to $\eta = 300$ when the proportion was 90:10. In general, this indicated that the $\mathcal{L}_i^-$ term of the ACL became more essential to the overall loss as the dataset got more imbalanced, confirming the reasoning contained in Section 4.1.

As seen in Table 2, we experimented on $\gamma \in \{0, 1, 2, 4, 7, 10\}$, where choosing $\gamma = 0$ meant that we were using the CL. Although the overall pattern of the optimal γ was less apparent than η of the previous experiment, some insights could still be obtained. When the scenario was between 70:30 and 90:10, the focusing parameter γ was optimally chosen when it was larger than zero. This was in direct contrast to when the proportion was perfectly balanced (50:50), where $\gamma = 0$ was the most optimal parameter. This suggests that a larger value of γ should be considered when class imbalance is significantly present within a dataset.

When the dataset was balanced, however, our experiments suggested that neither asymmetry nor focality was markedly helpful. Indeed, in the 50:50 scenario, CL already provided the second-best accuracy in Table 1 and the best accuracy in Table 2. In Table 1, the CL was the case where $\eta = 0$ was chosen. In Table 2, on the other hand, the CL was used when $\gamma = 0$. Therefore, our proposed loss function works best with imbalanced datasets.

5.4. Experiments Using ISIC 2018

From the ISIC 2018 dataset, a total of 1113 melanoma images and 115 dermatofibroma images were combined to create the experimental dataset. As with the previous experiment, the dataset was split 70/30 for training and testing. The images were resized to 128 × 128 pixels. The ResNet-50 encoder was trained using one of the available contrastive losses, which included the CL/FCL as baselines and the ACL/AFCL as the proposed loss functions. The classification head was trained using FL as the loss function, with its focusing parameter set to $\gamma = 2$.

The proportion between the melanoma class and the dermatofibroma class in the experimental dataset was close to 90:10. Using the results from Tables 1 and 2 as a heuristic for determining the optimal parameter values, we set $\eta = 300$ and $\gamma = 2, 7$. It is worth mentioning that even though $\gamma = 2$ produced the best accuracy in the FMNIST experiment, the UWA of the resulting model was quite poor. However, we decided to include this value in this experiment for completeness.

The results of this experiment are given in Table 3. As in the previous section, each experiment was conducted four times, so the table lists the average accuracy and UWA of these four runs for each contrastive loss tested. Each run, which included both model training and testing, was completed in roughly 80 min using our computational setup.

From Table 3, CL and ACL performed the worst in terms of UWA and accuracy, respectively. However, ACL gave the best UWA among all losses. This may indicate that the ACL encouraged the model to take the risky approach of classifying some samples as part of the minority class at the expense of accuracy. Overall, AFCL with $\eta = 300$ and $\gamma = 7$ emerged as the best loss in this experiment, producing the best accuracy and the second-best UWA behind the ACL. This led us to conclude that the AFCL, with optimal hyperparameters chosen, is superior to the vanilla CL and FCL.

Table 3. The accuracy and UWA (averaged over four independent runs) of the model when trained using various contrastive losses.

Loss Function	Accuracy	UWA
CL [16]	93.00	72.25
FCL [21]	93.07	74.34
ACL (η = 300)	85.94	**75.54**
AFCL (η = 300, γ = 2)	92.39	74.36
AFCL (η = 300, γ = 7)	**93.75**	74.62

6. Conclusions and Future Work

In this work, we introduced an asymmetric version of both contrastive loss (CL) and focal contrastive loss (FCL), which are referred to as ACL and AFCL, respectively. These asymmetric variants of the contrastive loss were proposed to provide more focus on the minority class. The experimental model used was a two-stage architecture consisting of a feature-learning stage and a classifier fine-tuning stage. This model was applied to the imbalanced FMNIST and ISIC 2018 datasets using various contrastive losses. Our results show that the AFCL was able to outperform the CL and FCL in terms of both weighted and unweighted accuracies. On the ISIC 2018 binary classification task, AFCL, with $\eta = 300$ and $\gamma = 7$ as hyperparameters, achieved an accuracy of 93.75% and an unweighted accuracy of 74.62%. This is in contrast to the FCL, which achieved 93.07% and 74.34% on both metrics, respectively.

The experiments in this research were conducted using datasets consisting of approximately 1000 images in total. In the future, the experimental model may be applied to larger-scale datasets in order to test its scalability. In addition, other models based on the ACL and AFCL can also be developed for specific datasets, ideally within the realm of multi-class classification.

Author Contributions: Conceptualization, V.V. and L.Y.S.; methodology, V.V.; software, V.V.; validation, V.V. and L.Y.S.; formal analysis, V.V.; investigation, V.V.; resources, V.V. and L.Y.S.; data curation, V.V.; writing—original draft preparation, V.V.; writing—review and editing, V.V. and L.Y.S.;

visualization, V.V.; supervision, L.Y.S.; project administration, L.Y.S.; funding acquisition, L.Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Faculty of Computer Science, Universitas Indonesia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv* **2019**, arXiv:1902.03368.
2. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **2018**, *5*, 180161. [CrossRef] [PubMed]
3. Bej, S.; Davtyan, N.; Wolfien, M.; Nassar, M.; Wolkenhauer, O. LoRAS: An oversampling approach for imbalanced datasets. *Mach. Learn.* **2021**, *110*, 279–301. [CrossRef]
4. Fajardo, V.A.; Findlay, D.; Houmanfar, R.; Jaiswal, C.; Liang, J.; Xie, H. Vos: A method for variational oversampling of imbalanced data. *arXiv* **2018**, arXiv:1809.02596.
5. Karia, V.; Zhang, W.; Naeim, A.; Ramezani, R. Gensample: A genetic algorithm for oversampling in imbalanced datasets. *arXiv* **2019**, arXiv:1910.10806.
6. Tripathi, A.; Chakraborty, R.; Kopparapu, S.K. A novel adaptive minority oversampling technique for improved classification in data imbalanced scenarios. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 10650–10657.
7. Arefeen, M.A.; Nimi, S.T.; Rahman, M.S. Neural network-based undersampling techniques. *IEEE Trans. Syst. Man, Cybern. Syst.* **2020**, *52*, 1111–1120. [CrossRef]
8. Dai, Q.; Liu, J.w.; Liu, Y. Multi-granularity relabeled under-sampling algorithm for imbalanced data. *Appl. Soft Comput.* **2022**, *124*, 109083. [CrossRef]
9. Koziarski, M. Radial-based undersampling for imbalanced data classification. *Pattern Recognit.* **2020**, *102*, 107262. [CrossRef]
10. Rayhan, F.; Ahmed, S.; Mahbub, A.; Jani, R.; Shatabda, S.; Farid, D.M. Cusboost: Cluster-based under-sampling with boosting for imbalanced classification. In Proceedings of the 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, 21–23 December 2017; pp. 1–5.
11. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
12. Marrakchi, Y.; Makansi, O.; Brox, T. Fighting class imbalance with contrastive learning. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2021; pp. 466–476.
13. Chen, K.; Zhuang, D.; Chang, J.M. SuperCon: Supervised contrastive learning for imbalanced skin lesion classification. *arXiv* **2022**, arXiv:2202.05685.
14. Wang, Z.; Peng, C.; Zhang, Y.; Wang, N.; Luo, L. Fully convolutional siamese networks based change detection for optical aerial images with focal contrastive loss. *Neurocomputing* **2021**, *457*, 155–167. [CrossRef]
15. Alenezi, F.; Öztürk, Ş.; Armghan, A.; Polat, K. An Effective Hashing Method using W-Shaped Contrastive Loss for Imbalanced Datasets. *Expert Syst. Appl.* **2022**, *204*, 117612. [CrossRef]
16. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
17. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Online, 13–18 July 2020; pp. 1597–1607.
18. Henaff, O. Data-efficient image recognition with contrastive predictive coding. In Proceedings of the International Conference on Machine Learning, Online, 13–18 July 2020; pp. 4182–4192.
19. Hjelm, R.D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv* **2018**, arXiv:1808.06670.
20. Tian, Y.; Krishnan, D.; Isola, P. Contrastive multiview coding. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 776–794.
21. Zhang, Y.; Hooi, B.; Hu, D.; Liang, J.; Feng, J. Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29848–29860.
22. Ben-Baruch, E.; Ridnik, T.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; Zelnik-Manor, L. Asymmetric loss for multi-label classification. *arXiv* **2020**, arXiv:2009.14119.
23. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]

24. Ajjanagadde, G.; Makur, A.; Klusowski, J.; Xu, S. *Lecture Notes on Information Theory*; Laboratory for Information and Decision Systems, Massachusetts Institute of Technology: Cambridge, MA, USA, 2017.
25. Gowers, W. Topics in Combinatorics. 2020. Available online: https://drive.google.com/file/d/1V778zHQTx4XE8FxDgznt2jTshZzxAFot/view (accessed on 13 May 2022).
26. Khinchin, A.Y. *Mathematical Foundations of Information Theory*; Dover Publications: Mignola, NY, USA, 1957.
27. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2006.
28. Boudiaf, M.; Rony, J.; Ziko, I.M.; Granger, E.; Pedersoli, M.; Piantanida, P.; Ayed, I.B. A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 548–564.
29. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
31. Xiao, H.; Rasul, K.; Vollgraf, R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv* **2017**, arXiv:1708.07747.
32. Fahad, M.S.; Ranjan, A.; Yadav, J.; Deepak, A. A survey of speech emotion recognition in natural environment. *Digit. Signal Process.* **2021**, *110*, 102951. [CrossRef]