

# Triplet Loss-less Center Loss Sampling Strategies in Facial Expression Recognition Scenarios

Hossein Rajoli\*, Fatemeh Lotfi\*, Adham Atyabi<sup>†</sup>, Fatemeh Afghah\*

\*Holcombe Department of Electrical and Computer Engineering, Clemson University, Clemson, SC, USA

Emails: {hrajoli, flotfi, fafghah}@clemson.edu,

<sup>†</sup>Department of Computer Science, University of Colorado Colorado Springs, Colorado Springs, USA

Email: aatyabi@uccs.edu

**Abstract**—Facial expressions convey massive information and play a crucial role in emotional expression. Deep neural network (DNN) accompanied by deep metric learning (DML) techniques boost the discriminative ability of the model in facial expression recognition (FER) applications. DNN, equipped with only classification loss functions such as Cross-Entropy cannot compact intra-class feature variation or separate inter-class feature distance as well as when it gets fortified by a DML supporting loss item. The triplet center loss (TCL) function is applied on all dimensions of the sample's embedding in the embedding space. In our work, we developed three strategies: fully-synthesized, semi-synthesized, and prediction-based negative sample selection strategies. To achieve better results, we introduce a selective attention module that provides a combination of pixel-wise and element-wise attention coefficients using high-semantic deep features of input samples. We evaluated the proposed method on the RAF-DB, a highly imbalanced dataset. The experimental results reveal significant improvements in comparison to the baseline for all three negative sample selection strategies.

**Index Terms**—Deep metric learning, Triplet loss, Facial expression recognition, Negative-samples selection.

## I. INTRODUCTION

Facial expression recognition (FER) has always played an essential role in the development and flourishing of numerous daily-used applications. Autonomous cars, human-computer interaction (HCI), quality of experiment measurement, forensic section, healthcare, and treatment, detecting deception, etc. are fields in which automatic FER is widely used as monitoring or assessment tools [1]. Increasing computational capacity, emerging sophisticated deep neural networks (DNN), and providing large-scale datasets have made great progress in this hot spot field of computer vision. Expressed emotions by face depend on the main three categories of signals proposed by the face, namely, static, slow, and fast signals. All discriminative features fall into these three signals, for instance, skin color, and facial bone shape of individuals belong to the static group, while muscle tone and skin texture, and eyebrow rapid rising respectively belong to slow and fast signals. Therefore, in the same category of facial emotional expression, there are high diversity of samples that vary due to age, level of intelligence, race, gender and so on [1] [2].

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0090 and the National Science Foundation under Grant Numbers CNS-2232048, and CNS-2204445.

In general, there are three main steps in a standard FER task with the DNN; extracting spatial deep features of the input sample, mapping highly semantic feature space to embedding space, and then estimating the probability distribution over all categories using softmax activation. Meanwhile, deep metric learning (DML) can play a crucial role. By imposing embedding space toward the point that intra-class samples are grouped tightly and inter-class ones placed as far as possible, DML can improve our supervised classification task [3]. One of the effective DML-based loss functions is loss-less triplet loss [4] which has been designed to take the information of all samples in backpropagation.

In an ordinary triplet loss, if the distance between the embedding space's sample representation and its corresponding center, which represents a positive sample, is less than its distance to the negative ones with the amount of defined margin, then that sample would not contribute to backpropagation. To boost the performance, a particular form of TC3L, namely, *adaptive margin T3L (AMTC3L)*, is introduced. Compared with ordinary TC3L, which considers a constant margin for all inputs, AMTC3L applies adaptive margins for each sample. The proposed loss function focuses on highlighting the samples with features that improve the discriminative ability of the network and let them contribute more to the backpropagation process.

Two different attention modules are introduced, pixel-wise and element-wise inclusion/exclusion modules. The core idea is, only distinguishing features from discriminative spots of the face get filtered out for recognizing the expressed emotion. In this way, the model learns how to discriminate noisy features and areas based on the deep features that have been extracted.

The main contributions of this work can be summarized as follows;

- We proposed a new form of triplet loss (TL), namely, AMTC3L, where the network learns the appropriate margin that well suits the network and lets it learn from some samples more than others by assigning appropriate margin to each sample attentively.
- To emphasize more on discriminative features we proposed an attention network, yields exclusion/inclusion weights from the deep features.
- To apply AMTC3L to the FER scenario, we proposed three different strategies offering the most negative sam-



ple. They are full-synthesized, semi-synthesized, and prediction-based negative sample selection strategies.

- The source code will publicly be available at the following github repository: [AMTC3L sampling strategies](#).

## II. RELATED WORK

This paper mainly focuses on introducing three strategies to customize and leverage triplet loss in FER scenarios fortified by a DML approach. Hence, we consider conducted works from two perspectives of DML usage in FER and triplet loss sampling strategies in classification scenarios.

### A. Triplet Loss sampling strategies in classification

Triplet loss enhances both inter-class compactness and intra-class separability. The authors in [5] developed a class-paired margin identifier mechanism in which the association with an outlier detector improves the performance of FER application. They proposed an outlier detector preventing occluded and confusing samples from having contributions in selected negative samples. To classify multi-labeled samples, [6] proposed a two-step sample selection mechanism. In the first step, a subgroup of the most informative diverse samples have been selected from the mini-batches as anchors and then in the second step, the positive and negative samples would be chosen considering the relevancy, hardness, and diversity of the images during the selection. To address the 3D retrieval challenge, [7] leveraged triplet center loss (TCL) and introduced a class center as the anchor sample, and adopted the nearest negative sampling as the negative sample selection strategy. Since the triplet loss increases the chance of contribution in back-propagation for minority classes in an imbalance dataset as the negative sample, Lei et al. [8] takes this advantage to classify the imbalance dataset with triplet and triplet center loss functions. The authors adopt all class centers as the negative sample, not only the nearest ones in their classification scenario. Huang et al. [9] conducted research to address the challenges of classifying hyperspectral images that are prone to inter-class inseparability due to occupying a wide range in feature space. To leverage the triplet loss function, in each mini-batch they randomly select a subset of classes and for each selected class they randomly take a subset of their corresponding samples within the mini-batch. Selecting a set of the hardest positive and negative samples for each input prevents outliers to affect the training procedure. Triplet loss has been also used to address knowledge distillation challenges. Georgescu and Ionescu [10] distilled and transferred knowledge of a trained network on the full-face dataset to a network that has been trained on the occluded-face dataset. In the proposed scenario, the occluded sample in embedding space and the negative samples come from the student network while the full-face networks give the positive samples. In both the positive and negative sampling process, the distance between the anchor and one-tenth of the randomly selected available samples is considered. Then, the farthest and the nearest samples are chosen as the positive and negative samples. Zhou et al. [11] utilized TCL with a fixed margin

in a FER scenario. In this work, the input sample's center is considered as the positive sample, and the nearest class center (other than the sample's class center) as the negative one. Chi-squared distance has been applied to the triplet loss function in a facial emotion recognition scenario by [12]. The authors considered an identity annotation (if available) in negative sample selection to reduce the computation cost and the effect of identity features in the training process.

### B. DML in FER scenarios

Deep metric learning using a neural network, is an approach for projecting input samples into a manifold space where a distance metric such as euclidean, cosine, or chi-squared can measure the distance between two points [13]. For facial emotion recognition, Fard and Mahoor [14] proposed an adaptive correlation loss that guides the network within the training process to the point of being more discriminative. They define an adaptive loss function based on the mean of  $k$  embedding vectors that the network produces in the embedding space for every single class. The loss function penalizes less correlated intra-class mean vectors and correlated inter-class ones. Farzaneh and Qi [15] takes advantage of center loss with an attentive network to increase the intra-class compactness. Huang et al. [16] developed a Mahalanobis metric distance and implemented it by a fully connected network, applied to the higher semantic features. To reach the level of pose-aware and identity-invariant, authors of [17] proposed a five-tuple DML consisting of two sets of triplets including, pos triplet and identity triplet with a common anchor, then trained the network using this DML approach. Liu et al. [18] proposed  $(N + M)$ -tuple clusters' loss that selects  $N$  different expressions from the same identity as the negative samples and searches for  $M$  positive samples on the fly then optimizes the network jointly based on this loss plus softmax to alleviate attribute variations coming from different identities.

## III. PROPOSED MODEL

In this section, we elaborate on the proposed approach and explain all modules, namely, Resnet-50 as the backbone, the classifier, the multi-task loss function, purifying module that eliminates outlier features from the embedding vector while providing the adaptive margin for input samples. In our approach, using weights generated by purifying network the margin would be calculated as,  $0 \leq \{\alpha_i = \sum_{j=1}^{C_d} w_j\} \leq C_d$ . Given a mini batch of  $m$  samples-  $D_m = \{(x^{(i)}, y^{(i)}) \mid i = 1, 2, \dots, m\}$ , where  $x^{(i)} \in \mathbb{X}$  is the training sample from the training set and  $y^{(i)} \in \{1, 2, \dots, K\}$  is its corresponding categorical label. The backbone network,  $\mathcal{N}_{BB}$ , maps mini batch input samples to their feature space representation,  $x^* \in \mathbb{R}^{C_f \times H_f \times W_f}$ , where  $C_f$ ,  $H_f$ , and  $W_f$  respectively denote the number of channels, rows, and width of the tensor in the feature space. By leveraging two depth-wise convolutional layers, the classifier network,  $\mathcal{N}_C$ , takes the feature space representation of the sample and extracts deep high semantic context,  $x^d \in \mathbb{R}^{C_d \times H_f \times W_f}$ . Then, using a global average pooling layer,  $\mathcal{P}$ , the tensor gets converted to the embedding





Figure 1: A sample of RAF-DB facial expressions dataset.

$e \in \mathbb{R}^{C_d \times 1 \times 1}$ . Finally, another depth-wise convolutional layer maps the embedding to row score of  $x^z \in \mathbb{R}^{K \times 1 \times 1}$ . Then, a probability distribution over  $K$  channels,  $Pr(y = j | x^{(i)})$ , is taken using Softmax over channels. In the end, the discrepancy between prediction,  $\hat{y}$  and true labels,  $y^{(i)}$ , would be calculated using cross-entropy (CE) loss function,  $\mathcal{L}_{CE}$ , as follows:

$$\mathcal{L}_{CE} = -\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^K \log y^{(i)} Pr(y^{(i)} = j | x^{(i)}). \quad (1)$$

#### A. Multi-task Loss

To calculate discrepancy, for any training sample, the following loss function would be calculated:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{AMTC3L}, \quad (2)$$

where [1] calculates  $\mathcal{L}_{CE}$ , and  $\mathcal{L}_{AMTC3L}$  measures Adaptive margin triplet center loss.

1) *Adaptive margin triplet loss-less center Loss (AMTC3L)*: The functionality of AMTC3L relies on two sub-modules, the margin estimator, and the negative class-center selector. As explained in [19] and demonstrated in [3], in general, the marginal triplet loss has three main terms that should be chosen carefully, the positive sample, the negative sample, and the margin. To improve intra-class compactness and inter-class separability, the sample's class center in embedding space is considered as the positive sample and the negative one would be selected from the remaining  $k - 1$  class centers.

$$\mathcal{L}_{TL} = \max \left( 0, \frac{1}{2m} \sum_{i=0}^m \sum_{j=1}^{c_d} \| (e_j^{(i)}) - (c_{y^{(i)}j}^p) \|_2^2 - \| (e_j^{(i)}) - (c_{y^{(i)}j}^n) \|_2^2 + \alpha_i \right), \quad (3)$$

where,  $\sigma$  represents the sigmoid activation,  $e_j^{(i)} \in e^{(i)} = [e_1^{(i)}, e_2^{(i)}, \dots, e_d^{(i)}]^T$  represents  $j$ 'th element of the  $i$ 'th sample's embedding from the input mini batch. Corresponding class center,  $c_{y^{(i)}j}^p \in c_{y^{(i)}}^p = [c_{y^{(i)}1}^p, c_{y^{(i)}2}^p, \dots, c_{y^{(i)}d}^p]$  defines the most positive embedding, while  $c_{y^{(i)}j}^n = [c_{y^{(i)}1}^n, c_{y^{(i)}2}^n, \dots, c_{y^{(i)}d}^n]$  represents the class center of the most negative one. Note that  $\alpha_i$  is the margin of sample  $i$ th of minibatch which comes from the adaptive margin generator.

In cases where the embedding falls into the region close enough to their class prototype, and far well from the most negative prototype,  $\mathcal{L}_{TL} = 0$ . Consequently, the sample contribution in backpropagation would be nothing. During the training process, the AMG learns to assign an appropriate  $\alpha_i$  to samples in a way that it dismissed the samples with simple

#### Algorithm MS-NSS: Mathematically Synthesized Negative-Sample Selection

**Inputs:** mini batch's embedding tensors and corresponding labels,  $D_m = \{(e^{(i)}, y^{(i)}) | i = 0, 1, \dots, m; e^{(i)} \in \mathbb{R}^{C_d \times 1 \times 1}\}$ ; embedding tensors' centers,  $c = \{c_k | k = 1, 2, \dots, K; c_{(k)} \in \mathbb{R}^{1 \times C_d}\}$ .  
**Output:**  $c^n = \{c_i^n | i = 1, 2, \dots, m\}$ .

- 1: while not converged do
- 2: compute inter cluster euclidean distance:
 
$$\eta_{kj} = \{\| \sigma((e_j^{(i)}) - \sigma((c_{k,j})) \|_2^2 | i = 1, \dots, m; j = 1, \dots, C_d; \forall c_k | k \neq y^{(i)}, k = 1, 2, \dots, K\}.$$
- 3:  $c_j^n = \{ \underset{k \in \{1, 2, \dots, C_d\}}{\text{minimize}} (\eta_{kj}) | j = 1, 2, \dots, C_d \}$
- 4: compute the triplet center loss by [4].
- 5: end while

#### Algorithm NS-NSS: Network's Semantic Negative Sample Selection

**Inputs:** mini batch's embedding tensors and corresponding labels,  $D_m = \{(e^{(i)}, y^{(i)}) | i = 0, 1, \dots, m; e^{(i)} \in \mathbb{R}^{C_d \times 1 \times 1}\}$ ; mini batch's prediction,  $\hat{y}^{(i)}$ ; embedding tensors' centers,  $c = \{c_k | k = 1, 2, \dots, K; c_{(k)} \in \mathbb{R}^{1 \times C_d}\}$ .  
**Output:**  $c^n = \{c_i^n | i = 1, 2, \dots, m\}$ .

- 1: while not converged do
- 2: if  $\hat{y}^{(i)} \neq y^{(i)}$  then:
- 3:  $c^n = \{c_{\hat{y}^{(i)}} | i \in \{i | \hat{y}^{(i)} \neq y^{(i)}\}\}$
- 4: else  $\hat{y}^{(i)} = y^{(i)}$ :
- 5:  $\gamma^{(i)} = \text{argmax}\{v(\hat{y}^{(i)} = k) | k \neq y^{(i)}\}$   
here  $v()$  represents frequency function.
- 6:  $c^n = \{c_{\gamma^{(i)}} | i \in \{i | \hat{y}^{(i)} = y^{(i)}\}\}$
- 7: end if
- 8: compute the triplet center loss by [4].
- 9: end while

#### Algorithm MM-NSS: Mixed-Method Negative Sample Selection

**Inputs:** mini batch's embedding tensors and corresponding labels,  $D_m = \{(e^{(i)}, y^{(i)}) | i = 0, 1, \dots, m; e^{(i)} \in \mathbb{R}^{C_d \times 1 \times 1}\}$ ; embedding tensors' centers,  $c = \{c_k | k = 1, 2, \dots, K; c_{(k)} \in \mathbb{R}^{1 \times C_d}\}$ .  
**Output:**  $c^n = \{c_i^n | i = 1, 2, \dots, m\}$ .

- 1: while not converged do
- 2: if  $\hat{y}^{(i)} \neq y^{(i)}$  then:
- 3:  $c^n = \{c_{\hat{y}^{(i)}} | i = 1, \dots, m\}$
- 4: else  $\hat{y}^{(i)} = y^{(i)}$ :
- 5: compute inter-cluster euclidean distance:
 
$$\eta_{kj} = \{\| \sigma((e_j^{(i)}) - \sigma((c_{k,j})) \|_2^2 | i = 1, \dots, m; j = 1, \dots, C_d; \forall c_k | k \neq y^{(i)}, k = 1, 2, \dots, K\}.$$
- 6:  $c_j^n = \{ \underset{k \in \{1, 2, \dots, C_d\}}{\text{minimize}} (\eta_{kj}) | j = 1, 2, \dots, C_d \}$
- 7: end if
- 8: compute the triplet center loss by [4].
- 9: end while



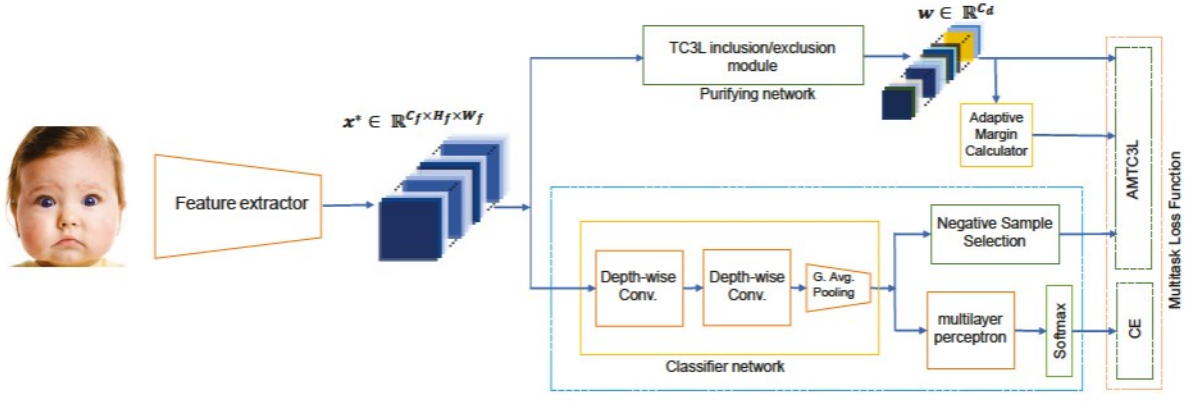


Figure 2: Detailed structure of the model.

features and performs hard-mining on samples with the most difficult discriminative features.

To boost the within-cluster compactness, in [4] similar to [3], we use the cluster centers as the most positive prototype while we synthesize the most negative embedding according to Algorithms **MS-NSS**, **NS-NSS**, and **MM-NSS**: that correspondingly named mathematically synthesized, network’s semantic, and mixed method negative sample selection algorithms.

The first algorithm synthesizes the most negative prototype totally based on the euclidean distance, the second one chooses the most negative one for each embedding based on the model output. In misclassified samples, the most negative class prototype would be exactly the class center of the incorrect prediction. However, for correct predictions, the algorithm looks at the records and selects the class center of the most challenging category that has defeated the model more. The third algorithm is a combination of the two first algorithms, it acts similar to algorithm **NS-NSS** for incorrect predictions and algorithm **MS-NSS** for correct predictions.

To elaborate, algorithms **MS-NSS** gets all mini batch’s embeddings, their corresponding labels, and class center vectors as its inputs and returns a synthesized class center vector as the most negative sample’s prototype  $c_{y(i)}^n$ . For the  $j$ th dimension of the  $i$ th embedding, the algorithm calculates the Euclidean distance to the  $j$ ’th dimension of all class centers. Then, it selects the smallest one. In this way, the synthesized most negative prototype on each dimension would be a linear combination of all class center vectors projection on that dimension. Note that the coefficients of the linear combination are all zero, and only there is a single one.

Algorithm **NS-NSS** acts based on the network predictions. It gets embeddings, their corresponding labels, and the class center matrix as inputs. Then, it compares the predictions and labels and finds the misclassified samples to create a new square matrix of misclassification statistic,  $S$ , that demonstrates the distribution of misclassified prediction classes for each class label. For incorrectly classified samples, the prediction class center would be considered as the negative sample. However, for the correctly classified ones, the most misclas-

sification of each class during training with that minibatch determines the negative class.

Algorithms **MM-NSS** works similarly to Algorithms **MS-NSS** for samples classified correctly, and the same as Algorithms **NS-NSS** for samples that are misclassified.

Farzaneh and Qi [15] argues that not all elements of the embedding have the same effect on the discrimination ability. Inspired by this fact, we believe not all samples’ embeddings have the same effect on the discrimination ability. As a result, the AM module provides appropriate  $\alpha_i$  to adjust the value of error for each sample as follows:

$$\mathcal{L}_{AMTC3L} = \frac{1}{2m} \sum_{i=0}^m \left[ \sum_{j=1}^{c_d} w_j \left( \left\| \sigma(e_j^{(i)}) - \sigma(c_{y(i)}^p) \right\|_2^2 \right) - w_j \left( \left\| \sigma(e_j^{(i)}) - \sigma(c_{y(i)}^n) \right\|_2^2 \right) \right] + \alpha_i, \quad (4)$$

$$\text{s.t., } 0 < \alpha_i \leq c_d, \quad (5)$$

where  $\sigma$  represents the Sigmoid function. In this way, the projection of the distance on every single dimension gets a value between zero and one [4]. Hence, if  $\alpha_i = \{\sum_{j=1}^{c_d} w_j \mid w_j \in \{0, 1\}\}$  then it assures that in all circumstances  $\mathcal{L}_{AMTC3L} > 0$  and the sample would contribute on the backpropagation. However, the number of outlier features that get multiplied by zero cause the margin varies sample by sample.

### B. Classifier

To have the advantage of being FCN, using the average pooling layer in the classifier network is inevitable,  $\mathcal{N}_C$ . As mentioned, the basic architecture of  $\mathcal{N}_C$  consists of two depth-wise  $1 \times 1$  kernel-size convolution layers that consecutively reduce the channel size and map the feature space representation,

Table I: Performance comparison of AMTC3L-aided CE with ordinary CE and state of the art methods on RAF-DB dataset. Note that our method’s results is based on five-fold. cross-validation.

Datasets	Models	Sur.	Fea.	Dis.	Hap.	Sad.	Ang.	Neu.	Average
RAF-DB	Baseline	0.85	0.60	0.57	0.94	0.82	0.74	0.86	0.7685
	MS-NSS	0.85	0.64	0.59	0.95	0.87	0.81	0.84	0.7929
	NS-NSS	0.86	0.66	0.62	<b>0.96</b>	0.82	0.78	0.86	0.7943
	MM-NSS	<b>0.87</b>	<b>0.66</b>	<b>0.64</b>	0.95	<b>0.89</b>	<b>0.80</b>	<b>0.87</b>	<b>0.8114</b>



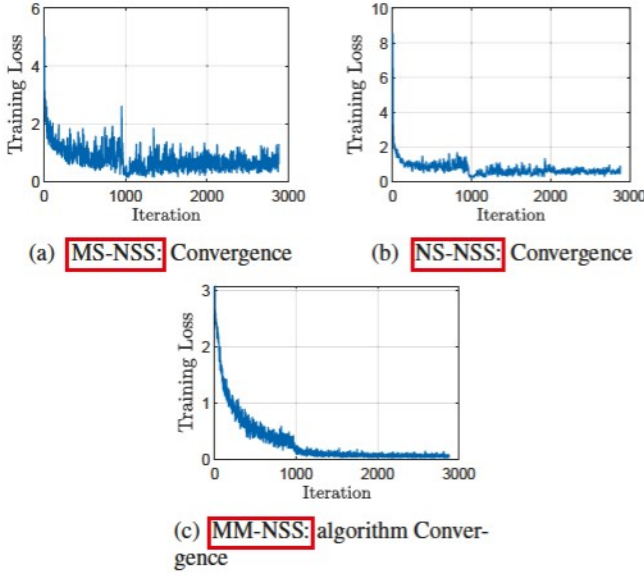


Figure 3: Training loss vs iterations in all three proposed algorithms.

$\mathbb{R}^{C_f \times H_f \times W_f}$ , to context space,  $\mathbb{R}^{C_a \times H_f \times W_f}$ . Then, a global average pooling layer produces embedding tensor,  $\mathbb{R}^{C_a \times 1 \times 1}$ , from context space representation of the input sample. Finally, a fully connected layer maps the embedding to the row score vector.

1) *Selective attention module*: In general, not all spatial features have the same contribution to the classifying process. As [20] indicates, every single feature map from channel-wise stacked does not necessarily demonstrate a unique semantic feature. Therefore, leveraging the advantage of the attention layer makes sense. Inspired by [20, 23], a selective inclusion/exclusion attention mechanism is applied to all projections of the embedding vector, anchor, and negative vectors to prevent outliers' contribution to the DML loss function. As mentioned before, this module also defines adaptive margins.

#### IV. EXPERIMENTAL RESULT

In this section, first, we introduce the dataset utilized in this study, the RAF-DB. Then, we compare visually the performance of all three strategies of selecting negative samples that we developed in our work. Some samples of RAF-DB has demonstrated in Fig. 1.

##### A. RAF-DB

The Real-world Affective Faces Database (RAF-DB) contains 29672 real-world images with two sets of simple and compound emotions and is considered as one of the richest facial expression databases [24]. The first set of labels includes basic emotions like Fearful, Happy, Surprised, Sad, Angry, Disgusted, and Natural. This dataset was gathered in 2017. The dataset's complex labeled emotions included 12 labels such as Fearfully Surprised, Angrily Surprised, and so on. The basic emotions were represented by 15,339 instances of all images, which were divided into 12,271 training samples and 3,068 testing samples.

##### B. Implementation details

We used pre-trained Resnet-50 as our backbone, it acts very well in extracting deep features. All proposed algorithms and baseline are trained with standard stochastic gradient descent (SGD) optimizer with the momentum of 0.9 and the weight decay of  $5 \times 10^{-4}$ . To augment the input samples we used multi-scale random cropping, random horizontal flipping, random erasing, and normalizing with regard to mean and standard deviation. In order to apply multi-scale random cropping, before feeding a mini-batch of samples to the network we scale them randomly then randomly crop patches with the size of  $224 \times 224$  and feed them to the network. In this way, the model can learn local high-discriminative features as well as global features. Note that applying correct augmentation has a significant effect to increase both the baseline and proposed technique. We trained our model on RAF-DB with 60 epochs and a learning rate of 0.01 (only for CE) for baseline and 0.05 (for either CE and AMTC3L) in all three algorithms. The learning rates decay with a factor of 10 every 20 epochs. In our works, the hyper-parameter of  $\lambda$  needs to get different values for different algorithms to help the model works at the optimum point. In our architecture, the deep feature is  $x^* \in \mathbb{R}^{2048 \times 4 \times 4}$ , and embedding is  $e \in \mathbb{R}^{64 \times 1 \times 1}$ .

In Fig. 3 the convergence ability of the proposed algorithms has been demonstrated. Furthermore, in Table I and Table II the average and overall accuracy have been shown for the proposed algorithms and the baseline. According to Table II, significant improvement has been achieved by using AMTC3L-aided CE for all proposed algorithms. However, it can get better if we apply the adaptive module scheme that excludes the outliers' effect from backpropagation. As Fig. 3a demonstrates, there is a huge noisy fluctuation in decreasing loss due to mathematically synthesized negative samples. MS-NSS explores the class centers and builds up single-by-single dimensions of negative samples from the closest elements of other classes. It is the hardest negative sampling scenario among the three proposed approaches. In Algorithm NS-NSS, smooth convergence comes from the fact that our selection strategy is exactly based on the performance that the model has presented before the current state. There is a tendency to compact/separate intra-/inter- class features but it is not as effective as algorithm MS-NSS. Algorithm MS-NSS that takes advantage of both methods shows smooth fluctuation and the fastest convergence.

#### V. CONCLUSION

In this paper, we have developed a DML attention-aided mechanism that leverages the benefits of the hinge-based loss function, namely TCL. To address the challenge of selecting the most negative samples, we introduced three novel approaches that consider either mathematically the nearest synthesized samples or statistically nearest samples, or a mix of both samples. Accordingly, we realized utilizing the mix of both samples (MM-NSS) results in a more robust approach. The simulation results have shown up to 3.12% improvement in accuracy compared to the baseline on the RAF-DB dataset.



Table II: Ablation study

Settings	MS-NSS			NS-NSS			MM-NSS		
	Classifier	TC3L	AMTC3L	Classifier	TC3L	AMTC3L	Classifier	TC3L	AMTC3L
baseline	✓	✗	✗	✓	✗	✗	✓	✗	✗
pipeline a (ablation)	✓	✓	✗	✓	✓	✗	✓	✓	✗
pipeline b (ablation)	✓	✓	✓	✓	✓	✓	✓	✓	✓
baseline (accuracy)		% 85.78			% 85.78			% 85.78	
pipeline a (accuracy)		% 87.12			% 87.29			% 88.53	
pipeline b (accuracy)		% 87.77			% 87.81			% 88.90	

Furthermore, the results have highlighted the importance of a negative sampling strategy in DML-based Loss-less triplet loss scenarios.

## REFERENCES

- [1] I Michael Revina and WR Sam Emmanuel. A survey on human face expression recognition techniques. *Journal of King Saud University-Computer and Information Sciences*, 33(6):619–628, 2021.
- [2] Niloufar Mehrabi and Sayed Pedram Haeri Boroujeni. Age estimation based on facial images using hybrid features and particle swarm optimization. In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, pages 412–418. IEEE, 2021.
- [3] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Embedding transfer with label relaxation for improved metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2021.
- [4] Gábor Kertész. Different triplet sampling techniques for loss-less triplet loss on metric similarity learning. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 000449–000454. IEEE, 2021.
- [5] Weicheng Xie, Haoqian Wu, Yi Tian, Mengchao Bai, and Linlin Shen. Triplet loss with multistage outlier suppression and class-pair margins for facial expression recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):690–703, 2021.
- [6] Gencer Sumbul, Mahdyar Ravanbakhsh, and Begüm Demir. Informative and representative triplet selection for multi-label remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.
- [7] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1945–1954, 2018.
- [8] Weixian Lei, Rong Zhang, Yang Yang, Ruixuan Wang, and Wei-Shi Zheng. Class-center involved triplet loss for skin disease classification on imbalanced data. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2020.
- [9] Ke-Kun Huang, Chuan-Xian Ren, Hui Liu, Zhao-Rong Lai, Yu-Feng Yu, and Dao-Qing Dai. Hyperspectral image classification via discriminative convolutional neural network with an improved triplet loss. *Pattern Recognition*, 112:107744, 2021.
- [10] Mariana-Iuliana Georgescu and Radu Tudor Ionescu. Teacher-student training and triplet loss for facial expression recognition under occlusion. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2288–2295. IEEE, 2021.
- [11] Linyi Zhou, Xijian Fan, Tardi Tjahjadi, and Sruti Das Choudhury. Discriminative attention-augmented feature learning for facial expression recognition in the wild. *Neural Computing and Applications*, 34(2):925–936, 2022.
- [12] Hamid Sadeghi and Abolghasem-A Raie. Histnet: Histogram-based convolutional neural network with chi-squared deep metric learning for facial expression recognition. *Information Sciences*, 608:472–488, 2022.
- [13] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–285, 2018.
- [14] Ali Pourramezan Fard and Mohammad H Mahoor. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, 10:26756–26768, 2022.
- [15] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2402–2411, 2021.
- [16] Wei Huang, Siyuan Zhang, Peng Zhang, Yufei Zha, Yuming Fang, and Yanning Zhang. Identity-aware facial expression recognition via deep metric learning based on synthesized images. *IEEE Transactions on Multimedia*, 2021.
- [17] Yuanyuan Liu, Wei Dai, Fang Fang, Yongquan Chen, Rui Huang, Run Wang, and Bo Wan. Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition. *Information Sciences*, 578:195–213, 2021.
- [18] Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–29, 2017.
- [19] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [20] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, pages 1–38, 2022.
- [21] Fatemeh Lotfi, Omid Semiari, and Walid Saad. Semantic-aware collaborative deep reinforcement learning over wireless cellular networks. In *ICC 2022-IEEE International Conference on Communications*, pages 5256–5261. IEEE, 2022.
- [22] Fatemeh Lotfi, Omid Semiari, and Fatemeh Afghah. Evolutionary deep reinforcement learning for dynamic slice management in o-ran. In *2022 IEEE Globecom Workshops (GC Wkshps)*, pages 227–232. IEEE, 2022.
- [23] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3601–3610, 2021.
- [24] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1): 356–370, 2019.