Hard Sample-aware Consistency for Low-resolution Facial Expression Recognition

Bokyeung Lee, Kyungdeuk Ko, Jonghwan Hong, Hanseok Ko
Department of Electrical and Computer Engineering, Korea University
{bksain,kdko,jhong2661,hsko}@korea.ac.kr

Abstract

Facial expression recognition (FER) plays a pivotal role in computer vision applications, encompassing video understanding and human-computer interaction. Despite notable advancements in FER, performance still falters when handling low-resolution facial images encountered in real-world scenarios and datasets. While consistency constraint techniques have garnered attention for generating robust convolutional neural network models that accommodate input variations through augmentation, their efficacy is diminished in the realm of low-resolution FER. This decline in performance can be attributed to augmented samples that networks struggle to extract expressive features. In this paper, we identify hard samples that cause an overfitting problem when considering various degrees of resolution and propose novel hard sample-aware consistency (HSAC) loss functions, which include combined attention consistency and label distribution learning. The combined attention consistency aligns an attention map from multi-scale low-resolution images with an appropriate target attention map by combining activation maps from high-resolution and flipped low-resolution images. We measure the classification difficulty for low-resolution face images and adaptively apply label distribution learning by combining the original target and predictions of high-resolution input. Our HSAC empowers the network to achieve generalization by effectively managing hard samples. Extensive experiments on various FER datasets demonstrate the superiority of our proposed method over existing approaches for multi-scale low-resolution images. Furthermore, we achieved a new state-of-the-art performance of 90.97% on the original RAF-DB dataset.

1. Introduction

Facial expression recognition (FER) is important in human-computer interaction and various computer vision tasks. Previous deep network-based FER studies have primarily focused on addressing noisy label problems in datasets [16, 19, 26], resulting in limited performance when dealing with low-resolution (LR) facial images commonly encountered in real-world environments. As shown in Figure 2 (a), multi-scale LR images arise from factors such as limitations in camera equipment quality and the distance between subjects and the camera lens. Compared to high-resolution (HR) images, LR facial images lack crucial details, such as muscle movement-induced wrinkles, which serve as vital cues for accurate facial expression recognition. Several FER methods for various LR images have been proposed [15, 23], but their performance remains limited. Consequently, recognizing facial expressions in multi-

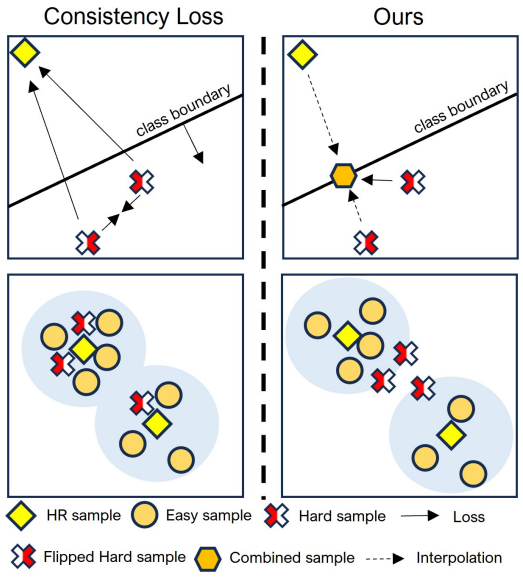

Figure 1. Given a hard sample, we intuitively describe how to work when employing previous consistency loss and our proposed loss function. Black arrows denote the effect of the loss function and dashed arrows mean interpolation between two samples. Blue circles indicate the distribution of samples corresponding to each class.

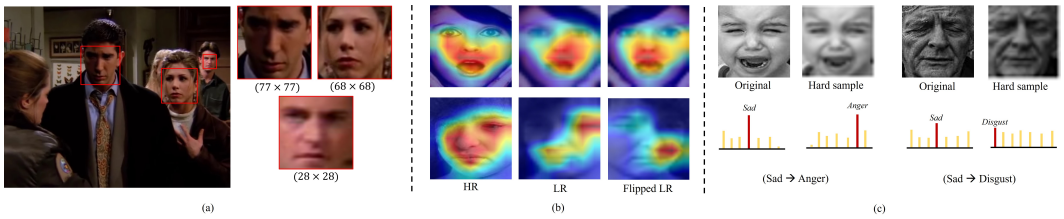

Figure 2. (a) The multi-scale LR examples of real environments. (b) The visualization of the attention maps from high-resolution (HR), low-resolution (LR), and flipped LR images (LR Flip). (c) The examples of the hard sample.

scale LR facial images presents a significant challenge, yet it is essential for real-life applications.

Figure 2 (b) shows class activation maps (CAM) [29] for the FER model trained with HR images when HR, LR, and flipped LR images are input, respectively. The attention maps derived from LR images cover smaller regions of the face compared to those from HR images. Then, the network generates different attention maps for the LR image and the flipped LR face image as confirmed in [27]. To generate a robust convolutional neural network (CNN) model that can be utilized in degraded images, the consistency constraint has been employed in many studies [3, 20, 21, 27]. They encourage results (attention map and prediction) from degraded input to be close to results from the original input.

Existing consistency constraints suffer from limitations in improving performance in low-resolution facial expression recognition (FER) when using down-sampling augmentation. As the facial details are distorted during down-sampling, augmentations can degrade the original image, potentially causing the network to be trained on face images from which the original target's emotional features cannot be effectively extracted. We define 'hard samples' as those for which the network cannot extract emotional features when considering various levels of resolution. Figure 2 (c) displays examples of hard samples. Due to down-sampling, tears and fine eye wrinkles disappear, resulting in an appearance resembling an angry face. Consequently, the prediction of the FER model is 'Anger' rather than the original label, 'Sad', as seen on the left side. The issue of overfitting arises when the representation of hard samples is compelled to closely resemble the high-resolution (HR) representation through the application of consistency constraints. This occurs because the network learns to achieve consistency based on various representations, such as face identity, rather than focusing on the emotion of the hard sample itself. As shown in Figure 1 (left), the hard sample is coerced to closely resemble the original sample when employing the previous consistency loss, thus being treated similarly to the original and easy samples from which expression features can be extracted. This impedes the learning of a generalized model.

To address these challenges, we present a solution called hard sample-aware consistency (HSAC) for multi-scale low-resolution (FER), which includes both combined attention consistency and label distribution learning. Our approach involves down-sampling the high-resolution image (I_H) to create a low-resolution image (I_L), along with generating a horizontally flipped low-resolution image (I_{LF}). These three images (I_H, I_L, I_{LF}) are then fed into a shared CNN, and predictions as well as attention maps are obtained using CAM.

We construct a combined attention map through linear interpolation of the attention maps from I_H and I_{LF}, and we aim to align the attention map of I_L closely with this combined attention map. Label distribution learning techniques are employed by merging the original target and predictions from I_H. We determine whether the input image is classified as a hard sample by assessing its classification difficulty. In the event that an input sample is categorized as a hard sample, label distribution learning for an LR image is not executed. Our HSAC exhibits linear behavior across multi-resolution samples while circumventing optimization for hard samples. After sufficient optimization, hard samples are predominantly positioned at the periphery of the class sample distribution, unlike the previous consistency loss function, as illustrated in Figure 1. In our visualization, we observe that hard samples are only positioned at the outskirts of the distribution of class samples in Section 4.6. Then, it encourages the network to learn more discriminative representation. Experimental results demonstrate that our HSAC outperforms models utilizing previous consistency constraints and augmentation techniques on both synthetic and real datasets. Moreover, HSAC attains a new state-of-the-art performance on original FER datasets.

2. Related Works

2.1. Facial expression recognition

Facial expression recognition (FER) aims to provide affective behavior information of humans for real-life applications in human-computer interaction systems. While FER on "in-the-lab" datasets has shown good results, existing

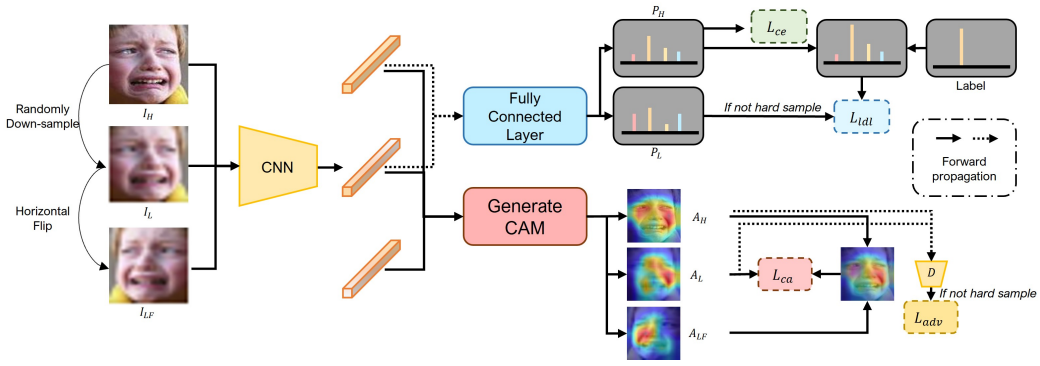

Figure 3. The overview of our proposed method. The black arrows denote forward propagation.

methods have limitations in terms of generalization. Consequently, researchers have focused on addressing challenges in "in-the-wild" scenarios, which involve label noise, occluded faces, and degraded images [16, 18, 23]. Various approaches have been proposed to overcome these challenges. For example, [19] addresses label noise by identifying uncertain samples and relabeling them in a low-importance group. [11] employs label distribution learning to handle noise labels by combining one-hot labels with distributions based on uncertainty values. To tackle occluded facial expression recognition, [22] propose step-wise and adversarial learning using unpaired non-occluded face data. They demonstrate robust performance under intra- and cross-database evaluation. Additionally, [15] employs feature super-resolution and adversarial learning to enhance FER for LR face inputs. However, there are still limitations in achieving accurate multi-scale LR facial expression recognition.

2.2. Attention Mechanism for Classification

CAM is an effective tool to visualize the predicted class scores and localize discriminative parts detected by the CNN. The attention maps from CAM can be also utilized to generate pseudo labels and are trained to capture more complete regions to design robust models in recent studies [2,3]. For the number of classes K, height H, and width W, the attention map $A \in \mathbb{R}^{K\times H\times W}$ is calculated as:

$$A_k = \sum_{c=1}^{C} \mathrm{W}_{k,c}\mathrm{F}_c, \tag{1}$$

where C is the number of channels for the feature map $\mathrm{F} \in \mathbb{R}^{C\times H\times W}$ of the last convolutional layer. $\mathrm{W} \in \mathbb{R}^{K\times C}$ denotes the weights of the fully connected layer.

[1, 2] generate pseudo labels by introducing a self-supervised task and perform weakly supervised learning to improve the attention maps. [9, 21] eliminate the most discriminative region of attention maps and encourage the network to learn classification features from other regions and capture expanded activations. [20] proposed a self-supervised equivariant attention mechanism to solve inconsistency problems on the generated attention maps when applying down-sampling with different factors in general image classification. [6, 27] achieve better visual perceptual plausibility and classification performance by employing attention consistency methods in multi-label classification and facial expression recognition.

3. Proposed Method

Figure 3 illustrates the framework of HSAC, comprising a CNN-based backbone network and a fully connected layer. Initially, we down-sample the HR facial image I_H using a randomly selected down-sampling factor to generate a multi-scale LR face image I_L. Subsequently, I_L is horizontally flipped to create the flipped LR version, I_{LF}. Through bicubic interpolation, both I_L and I_{LF} are up-sampled to match the dimension of I_H before being fed into the network. The network then generates feature maps from I_H, I_L, and I_{LF}. The feature maps originating from I_H and I_L are propagated into a fully connected layer, yielding two predictions (P_H and P_L) corresponding to I_H and I_L. For the generation of attention maps, the feature maps for each input and the weights of the fully connected layer are utilized, following the same approach as described in Equation 1. This process yields three attention maps (A_H, A_L, and A_{LF}) from I_H, I_L, and I_{LF}.

3.1. Preliminary

The existing methods employ attention consistency loss function [6, 27] between A_H and A_L to solve inconsistency problems as follows:

$$L_a = \frac{1}{K}\sum_{k=1}^{K} D_a(A_{H_k}, A_{L_k}), \tag{2}$$

where D_a is the function that can compute the distance between two attention maps, such as L1- or L2-norm. Similarly, they can try to use the consistency loss function for prediction from augmented input as:

$$L_p = D_p(P_L, P_T), \tag{3}$$

where D_p is a loss function that can capture the difference between two categorical distributions and P_T denotes the target probability distribution, such as a one-hot logical label and prediction from original input, P_H.

3.2. Combined Attention Consistency

The attention consistency loss function mentioned above in Eq. 2 has the potential to result in an overfitting issue. This occurs as the network is conditioned to treat multi-resolution samples equally in terms of attention maps. This uniform treatment is extended to hard samples, diverting the network's focus toward learning facial image identity features. As a consequence, the network's capacity to concentrate on acquiring discriminative facial expression features becomes compromised.

To mitigate the overfitting problem, we introduced a combined attention consistency loss function. The purpose of the combined attention consistency is to align attention maps from multi-scale low-resolution images with appropriate target attention maps. A_{LF} and A_H can serve as valuable references for A_L. We first calculate the L1 distance between the attention maps from LR (A_L) and the attention maps from HR and LRF (A_H and A_{LF}), and we can acquire distances $\delta_{A_{LH}} = ||A_L - A_H||_1$ and $\delta_{A_{LF}} = ||A_L - Flip(A_{LF})||_1$, respectively. $Flip(\cdot)$ is horizontal flip operation. $\delta_{A_{LH}}$ is the distance between A_L and A_H, and $\delta_{A_{LF}}$ denotes the distance between A_L and flipped A_{LF}. The similarity between A_L and A_H can be calculated as $s_A = \frac{\delta_{A_{LF}}}{\delta_{A_{LF}} + \delta_{A_{LH}}}$, and the similarity between A_L and A_{LF} is defined as $1 - s_A$. The combined attention consistency loss can be expressed as

$$L_{ca} = \frac{1}{K} \sum_{k=1}^{K} D_a(A_{C_k}, A_{L_k}). \tag{4}$$

$A_C = ((1 - s_A) \cdot A_H + s_A \cdot Flip(A_{LF}))$ denotes combined attention maps, and the combined attention consistency minimizes the difference between A_L and combined attention maps. We use L1-norm as $D_a(\cdot)$. The large s_A means that activation maps are imbalanced in the early training epoch because attention maps (A_L and flipped A_{LF}) are different despite images of the same resolution. So, as A_{LF} occupies a large proportion of combined attention maps, A_L and A_{LF} become close to each other. In contrast, the small s_A denotes that attention maps are balanced, and L_{ca} focuses on transferring the knowledge of discriminative representation into attention maps from LR. In contrast to the prior attention consistency loss, our proposed method offers a suitable target attention map from the perspective of label distribution learning. This is achieved due to the balanced nature of s_A, as opposed to being biased towards a single side, as elaborated in Section 4.7. As a result, attention maps derived from multi-scale low-resolution inputs retain a distance from the high-resolution attention map commensurate with the extent of distortion. The combined attention consistency enables the network to learn linear behavior on features from multi-resolution LR images in sample augmentation points of view [3, 24, 26].

3.3. Label Distribution Learning

Label distribution learning proves to be an effective solution for addressing noisy label issues stemming from ambiguous samples, low-quality images, inconsistent annotations, and incorrect annotations. It has already been incorporated into numerous FER models. Considering this, LR samples can be perceived as ambiguous and of lower quality. As a result, we utilize label distribution learning to tackle ambiguity and formulate target distributions for easy samples that do not fall under the category of hard samples. We calculate the L1 distance to acquire $\delta_{P_{LH}} = ||P_L - P_H||_1$ and $\delta_{P_{LT}} = ||P_L - T||_1$, where T is one-hot logical label. The similarity, s_P, between P_L and P_H is calculated as $\frac{\delta_{LT}}{\delta_{LT} + \delta_{LH}}$.

Our label distribution learning loss function is formulated as:

$$L_{ldl} = D_p(P_L, P_C), \tag{5}$$

where $D_p(\cdot)$ is the cross-entropy loss function. $P_C = (s_P \cdot P_H + (1 - s_P) \cdot T)$ denotes target label distributions for LR generated by combining P_H and T with s_P. When dealing with images exhibiting slight degradation, the minimization of L_{ldl} prompts P_L to converge towards a distribution that is more akin to the distribution shared between P_H and T.

3.4. Total loss function for Hard Sample

To learn discriminative features, we require a cross-entropy loss function using logical labels for HR images, which is defined as $L_{ce} = D_p(P_H, T)$. Furthermore, we also utilize adversarial learning to create an indistinguishable A_L from A_H, thereby compensating for the deficiency in representing A_L when solely utilizing L_{ca}. For discriminator $D(\cdot)$, adversarial loss function for discriminator is defined $L_{adv_d} = -logD(A_H) - log(1 - D(A_L))$ and adversarial loss function is expressed as

$$L_{adv_g} = log(1 - D(A_L)). \tag{6}$$

To circumvent overfitting issues, it becomes imperative to selectively apply the aforementioned loss function to hard samples. To distinguish hard samples, we compare the expression class outcomes among P_L, P_H, and logical label

Loss Functions							Accuracy of Down-sampling Factors (%)			
no.	L_{ce}	L_a	L_p	L_{ca}	L_{ldl}	L_{adv}	$\times$ 2	$\times$ 4	$\times$ 8	Avg
1	✓	×	×	×	×	×	88.75	87.03	76.96	84.24
2	✓	✓	×	×	×	×	90.55	89.90	84.35	88.26
3	✓	×	×	✓	×	×	90.48	89.93	85.33	88.58
4	✓	×	✓	×	×	×	88.75	88.10	83.83	86.89
5	✓	×	×	×	✓	×	88.75	88.07	85.06	87.29
6	✓	✓	✓	×	×	×	89.80	89.41	81.58	86.93
7	✓	×	✓	✓	×	×	90.61	90.06	85.21	88.62
8	✓	×	×	✓	✓	×	90.38	89.86	86.38	88.87
9	✓	×	×	✓	✓	✓	90.58	90.03	86.25	88.95

Table 1. Ablation study for the proposed loss functions on RAF-DB. Avg is the average accuracy for all downscale factors $\times 2$, $\times 4$, and $\times 8$.

T. If the maximum index (estimated result) of P_L differs from both the expression class outcomes of P_H and T, we consider an LR image as a hard sample. Subsequently, we define the classification difficulty indicator $\tau = 0$ when the LR image is identified as a hard sample. The classification difficulty indicator is computed as

$$\tau = \begin{cases} 0 & \text{if } C(P_L) \neq C(P_H) \text{ and } C(P_L) \neq C(T), \\ 1 & \text{otherwise,} \end{cases} \quad (7)$$

where $C(\cdot) = argmax_k(\cdot)$ that estimated class is extracted, where k is the class index. Our total loss function is formulated as:

$$L_T = \begin{cases} L_{ce} + \lambda_{ca} L_{ca} + \lambda_{ldl} L_{ldl} + \lambda_{adv} L_{adv_g}, & \text{if } \tau = 1 \\ L_{ce} + \lambda_{ca} L_{ca}, & \text{otherwise} \end{cases} \quad (8)$$

where $\lambda_{ca}, \lambda_{ldl}$ and λ_{adv} are hyper-parameters. We prevent the network from overfitting to hard samples by incorporating a classification difficulty indicator, and L_{ca} inherently establishes connections between each representation from multi-scale low-resolution images. Further training details for the algorithm and loss functions can be found in the Supplementary Materials.

4. Experiments

4.1. Implementation Details

Datasets To evaluate the facial expression recognition performance on synthetic multi-scale LR images, we use RAF-DB [13], AffectNet [14], and SFEW2.0 [4] in our experiments, which are "in-the-wild" datasets. They are created by collecting images from the internet, movies, and other real-life scenarios, and these datasets contain seven facial expression classes (surprise, fear, disgust, happy, sad, angry, and neutral). Our data split process is following previous works [11, 15]. To evaluate the facial expression recognition performance on real LR images, We employ CAER-S and FER+ because these are very close to real LR scenarios. We extracted 1223 images with less than 68 pixels in width and height for the real LR image test from the CAER-S dataset. The FER$^+$ dataset consists of 3588 48×48 grayscale LR images.

Preprocess In the training phase, we use the horizontal flip and the random erasing as augmentation techniques and then the original face image is resized to 224×224 pixels to make HR. To acquire I_L, I_H is down-sampled by non-integer factors $\times 1$ to $\times 8$ with bicubic down-sampling. We can obtain I_{LF} by flipping I_L horizontally. I_L and I_{LF} are upsampled by 224×224 pixels with bicubic upsampling. In the test phase, we resize the original image to 28×28 (×8), 56×56 (×4), 112×112 (×2), and 224×224 (×1) using bicubic regardless of the original image size for fair comparisons. Then, we interpolate resized images to 224×224 using bicubic for inference, as the input size of the networks is 224×224.

Hyper-parameter Setting For a fair comparison, we use the pre-trained ResNet50 (RN50) [8] with MS-Celeb-1M [7] dataset as a backbone network. During training, the mini-batch size is 32 and the learning rate is set to 10^{-4}. We empirically set $\lambda_{ca}, \lambda_{ldl}$ and λ_{adv} to 3, 10^{-2}, and 10^{-5}, respectively. We use Adam optimizer [10] and the training epoch is set to 60.

4.2. Ablation Study

In this section, we will demonstrate the impact of the proposed loss functions and validate the criteria for selecting hard samples, followed by an analysis of performance variations based on hyper-parameter settings. All experiments in the ablation study are conducted on the RAF-DB dataset. Table 1 illustrates the effectiveness of our proposed loss functions for multi-scale low-resolution Facial Expression Recognition (FER). In the first row, a model trained with L_{ce} exhibits poor performance at a down-sampling ratio of $\times 8$. In the second and third rows, the results using L_a are lower than those using our L_{ca} by 0.98% in the $\times 8$ down-sampling case. Similarly, in the fourth and fifth rows, our L_{ldl} outperforms L_p by 1.23% in the $\times 8$ down-sampling case. In the sixth row, where the previous consistency loss function is employed, significant performance

Dataset	Down-sampling Factors	RN50	RN50 + L_p	RN50 + RCAN	FSR-FER	EAC	EAC +L_a	HSAC
RAF-DB	×2 (%)	88.75	88.75	86.44	84.02	90.06	89.90	**90.58**
	×4 (%)	87.03	88.10	86.54	80.02	89.47	88.55	**90.03**
	×8 (%)	76.96	83.83	76.11	65.97	81.94	84.62	**86.25**
	Avg (%)	84.24	86.89	83.03	76.67	87.15	87.69	**88.95**
AffecNet-7	×2 (%)	62.80	63.09	59.03	-	65.20	65.26	**65.43**
	×4 (%)	61.26	62.77	58.63	-	64.37	64.31	**65.17**
	×8 (%)	49.89	57.83	42.26	-	61.31	63.03	**63.40**
	Avg (%)	57.98	61.23	53.30	-	63.62	64.20	**64.66**
SFEW2.0	×2 (%)	49.01	46.78	42.57	55.14	56.19	53.47	**59.16**
	×4 (%)	49.01	47.28	43.81	49.64	55.94	54.21	**58.91**
	×8 (%)	45.79	46.53	38.12	40.00	53.47	53.96	**55.20**
	Avg (%)	47.93	46.86	41.50	48.26	55.20	53.88	**57.75**

Table 2. Evaluation of our proposed HSAC for multi-scale LR FER on various FER datasets. The best results are highlighted.

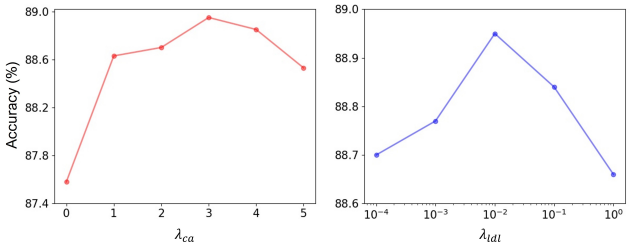

Figure 4. Evaluation results for the proposed method according to different regularization parameters, L_{ca} and L_{ldl}.

The hard sample selection criteria	Acc (%)
$C(P_L) \neq C(P_H)$	88.76
$C(P_L) \neq C(T)$	88.82
$C(P_L) \neq C(P_H)$ and $C(P_L) \neq C(T)$	88.95

Table 3. Comparison of HSAC according to the hard sample selection criteria.

degradation is observed when the network is trained with both L_a and L_p. In contrast, our HSAC consistently maintains high performance regardless of the down-sampling ratio, as demonstrated in the ninth row. These results underscore that our proposed loss functions, in conjunction with the consideration of hard samples, empower the network to surmount performance degradation stemming from overfitting.

Figure 4 shows evaluation results for our proposed method according to different regularization parameters, λ_{ca} and λ_{ldl}. The left and right figures show accuracy for multi-scale LR FER about λ_{ca} and λ_{ldl}, respectively. λ_{ca} ranges from 0 to 5. As λ_{ca} increases from 0 to 3, we can see a trend of improved performance and then accuracy decreases slightly. λ_{ldl} ranges from 10^{-4} to 1. As λ_{ldl} increases from 10^{-4} to 10^{-2}, we can see a trend of improved performance and then accuracy decreases rapidly. We acquire the best performance at $\lambda_{ca} = 3$ and $\lambda_{ldl} = 10^{-2}$.

As demonstrated in Table 3, a performance comparison of HSAC based on the hard sample selection criteria in Eq. 7 is presented. In the first row, the LR image is classified as a hard sample when the prediction outcome of the HR image diverges from that of the LR image. This implies that only the predicted value for the HR sample is considered, and the corresponding label is disregarded. This can reduce the number of trainable samples when predictions for HR fail. The criteria of the second row mean that the LR image is considered a hard sample when the target expression class is different from the prediction outcome of the LR image. This signifies that samples with label noise are excluded, even if they represent less distorted images. This approach also contributes to a reduction in the number of training samples. The third row introduces our proposed criterion for determining whether a facial sample qualifies as a hard sample. Our hard sample selection criterion was carefully designed to secure as many hard samples as feasible for inclusion in the learning process.

4.3. Evaluation of HSAC on Synthetic Multi-Scale Low-resolution Datasets

To compare our proposed method on three datasets, we employ several FER models and techniques for multi-scale LR FER. As shown in Table 2, RN50 and EAC [27] are trained with the FER dataset without a data augmentation technique for multi-scale LR. EAC is our baseline model, which uses the consistency method. $+L_p$ and $+L_a$ mean that the model is trained with previous consistency constraints. We also employ single-image super-resolution method RCAN [25] for ×2, ×4, ×8 to interpolate LR images. FER models without consistency constraints show limited performance especially as the down-sampling factor becomes large. RCAN performed poorly at all down-sampling factors, which, as reported in [15], adversely affects the FER that needs to capture detailed facial movements due to small and large artifacts. EAC+L_a performed worse in SFEW2.0 compared to EAC. This is be-

Dataset	DACL	MViT	SCN	EfficientFace	DMUE	RUL	EAC	HSAC
RAF-DB	87.78%	88.62%	88.14%	88.36%	89.42%	88.98%	90.35%	**90.97%**
AffectNet-7	65.20%	64.57%	-	63.70%	63.11%	-	65.23%	**65.29%**
SFEW2.0	-	-	-	-	58.34%	-	56.68%	**58.42%**

Table 4. Comparison with state-of-the-art methods on original face images. The best performance is highlighted.

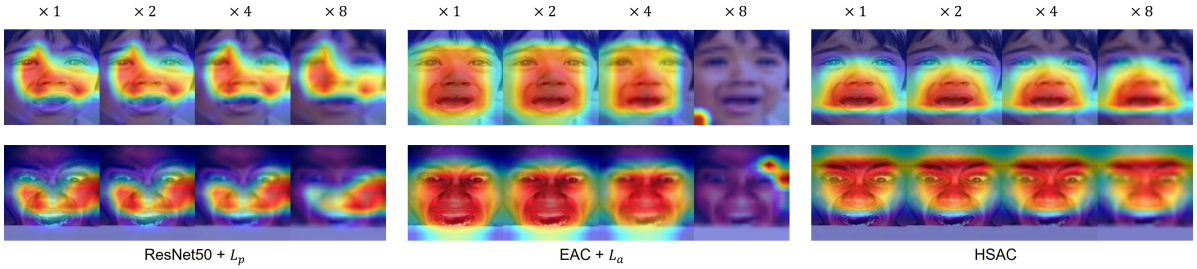

Figure 5. Visualization of the activation maps generated by CAM about RN50 + L_p, EAC + L_a, and HSAC on RAF-DB.

Datasets	RN50	RN50 $+L_p$	EAC	EAC $+L_a$	HSAC
CAER-S	62.71%	62.74%	63.19%	63.88%	**65.63%**
FER$^+$	64.34%	64.61%	64.55%	65.34%	**67.15%**

Table 5. Evaluation of HSAC on real LR datasets.

cause SFEW2.0 contains more ambiguous emotional expressions than other datasets, and hard samples are trained with original labels. In contrast, our HSAC shows consistently high performance on all datasets. Specifically, HSAC outperforms EAC$+L_a$ by 1.26%, 0.46%, and 3.87% in terms of the average accuracy on RAF-DB, AffectNet, and SFEW2.0, respectively. In addition, our method outperforms other methods by a large performance difference at $\times 8$ while maintaining performance at other down-sampling factors.

4.4. Comparison with State-of-the-art Methods

Our proposed method is compared with the state-of-the-art on original datasets. The comparison methods can be divided into architecture and solving noisy label problem-based methods. DACL [5] designed attention network and MViT [12] is based on vision transformer architecture for FER. EfficientFace [28] employed a teacher network to provide target distribution. SCN [19], DMUE [16], RUL [26], and EAC are methods to solve the noisy label problem in the FER dataset and they show state-of-the-art performance. Table 4 shows the performances of comparison methods on original FER datasets. Our HSAC outperforms state-of-the-art methods in original FER performance even though it was trained for multi-scale LR FER. This result suggests that our hard sample-aware strategy on ambiguous samples makes the model robust.

4.5. Evaluation of HSAC on Real Low-resolution Datasets

We trained RN50, RN50 $+L_p$, EAC, EAC $+L_a$, and our HSAC on both AffecNet and RAF-DB to evaluate CAER-S dataset. Additionally, we trained RN50, RN50 $+L_p$, EAC, EAC $+L_a$, and our HSAC on a grayscale AffectNet and RAF-DB to evaluate FER$^+$ dataset. As shown in Table 5, our HSAC still outperforms the state-of-the-art method and models with previous consistency constraints on real LR images. The results imply that our strategy of considering hard samples works well in real scenarios.

4.6. Visualization

To demonstrate that our proposed methods work as intended, we visualize the acquired feature maps. As depicted in Figure 5, we present activation maps generated through Class Activation Mapping (CAM) for RN50 $+L_p$, EAC $+L_a$, and our proposed HSAC. RN50 $+L_p$ failed to capture the distinctive regions across all down-sampling factors, and it is evident that the attention area gradually diminishes as the down-sampling factor's intensity increases. EAC $+L_a$ encompassed the entire facial area, encompassing not only the discriminative facial region but also irrelevant regions for Facial Expression Recognition (FER). Furthermore, EAC $+L_a$ completely faltered in classifying the facial expression of the low-resolution image at a $\times 8$ down-sampling ratio. In contrast, our proposed HSAC precisely captures the most discriminative facial area, specifically the facial wrinkles, while effectively excluding irrelevant regions.

To validate the superiority of our proposed methods, we further visualize feature distributions on RAF-DB for all down-sampling factors ($\times 1$, $\times 2$, $\times 4$, and $\times 8$) using t-SNE [17]. Figure 7 illustrates the feature distributions of

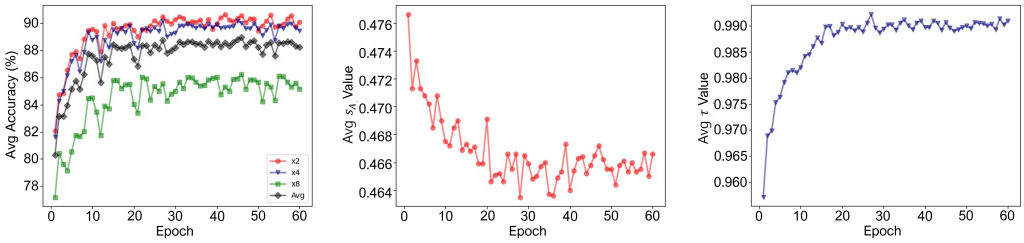

Figure 6. Evaluation results for down-sampling factors according to training epoch (Left). The value of similarity s_A and average τ according to training epoch (Middle and Right).

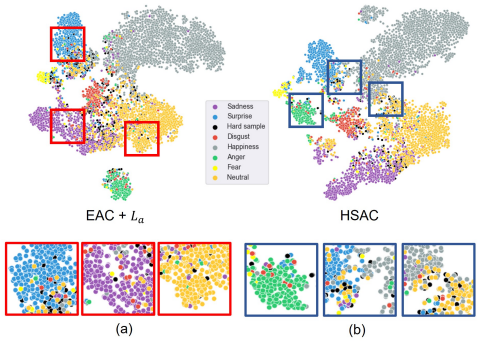

Figure 7. t-SNE visualization results of feature distributions about EAC $+L_a$, and HSAC on RAF-DB.

EAC $+L_a$ and HSAC. Hard samples are denoted by black dots. The red box corresponds to the scenario where hard samples are positioned amid the distribution of class samples (a). The blue box signifies the situation where hard samples are placed at the periphery of the class sample distribution (b). The feature distribution of EAC $+L_a$ encompasses cases (a) and (b), whereas HSAC exclusively contains case (b). In addition, it is clearly observed that the boundaries of the feature distributions generated from HSAC between different classes are more obvious. However, the feature distributions generated from comparison methods seem relatively vague. The sole distinction between EAC $+L_a$ and HSAC lies in the consideration of hard samples. The results suggest that controlling hard samples is important for creating a discriminative feature distribution and a generalized model for multi-scale LR FER.

4.7. Performance, s_A and τ Value According to Epoch

Figure 6 illustrates the evaluation results across down-sampling factors ($\times 2$, $\times 4$, and $\times 8$) over epochs, alongside the changes in the average values of s_A and τ. In the left figure, the black line represents the average accuracy of HSAC across all down-sampling factors. The network's learning stability tends to increase as the down-sampling factor decreases. The middle figure demonstrates the gradual decrease and convergence of the similarity between A_L and A_H, represented by s_A, during the training process. This phenomenon can be interpreted as an effort to balance the gap between A_{LF} and A_L, effectively maintaining a certain distance between HR samples and multi-resolution LR samples. This indicates that perfect alignment between hard samples and HR samples is not attainable. The right figure displays the average value of τ, representing the count of non-hard samples in each training epoch. Over the course of training, the average value of τ progressively increases and eventually stabilizes. This trend arises from L_{ca} aligning attention maps for LR samples with those of HR samples. However, HSAC is observed to designate 1% of the down-sampled training images as hard samples after 20 epochs.

5. Conclusion

We identify hard samples that lead to overfitting issues when applying consistency constraints across various resolutions of images. To address this challenge, we introduce HSAC—a framework comprising combined attention consistency, label distribution learning, and a classification difficulty indicator. The proposed method allows for linear behavior on multi-resolution samples and prevents the network from overfitting to hard samples. Our evaluation results demonstrated the superiority of our approach in multi-scale LR FER, surpassing the performance of state-of-the-art methods on original datasets. Our observations indicate that hard samples are placed on the outskirts of the distribution of the class sample, and it implies that HSAC creates a generalized model.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2023R1A2C2005916).

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2209–2218, 2019. 3

[2] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020. 3

[3] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4288–4298, 2022. 2, 3, 4

[4] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2106–2112. IEEE, 2011. 5

[5] 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. 7

[6] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 729–739, 2019. 3

[7] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 5

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[9] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[11] Nhat Le, Khanh Nguyen, Quang Tran, Erman Tjiputra, Bac Le, and Anh Nguyen. Uncertainty-aware label distribution learning for facial expression recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6088–6097, 2023. 3, 5

[12] Hanting Li, Mingzhe Sui, Feng Zhao, Zhengjun Zha, and Feng Wu. Mvt: mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520*, 2021. 7

[13] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 5

[14] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 5

[15] Fang Nan, Wei Jing, Feng Tian, Jizhong Zhang, Kuo-Ming Chao, Zhenxin Hong, and Qinghua Zheng. Feature super-resolution based facial expression recognition for multi-scale low-resolution images. *Knowledge-Based Systems*, 236:107678, 2022. 1, 3, 5, 6

[16] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6248–6257, 2021. 1, 3, 7

[17] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[18] Jiahe Wang, Heyan Ding, and Shangfei Wang. Occluded facial expression recognition using self-supervised learning. In *Proceedings of the Asian Conference on Computer Vision*, pages 1077–1092, 2022. 3

[19] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020. 1, 3, 7

[20] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 2, 3

[21] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017. 2, 3

[22] Bin Xia and Shangfei Wang. Occluded facial expression recognition with step-wise assistance from unpaired non-occluded images. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2927–2935, 2020. 3

[23] Yan Yan, Zizhao Zhang, Si Chen, and Hanzi Wang. Low-resolution facial expression recognition: A filter learning perspective. *Signal Processing*, 169:107370, 2020. 1, 3

[24] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 4

[25] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of*

the European conference on computer vision (ECCV), pages 286–301, 2018. 6

[26] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021. 1, 4, 7

[27] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 418–434. Springer, 2022. 2, 3, 6

[28] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3510–3519, 2021. 7

[29] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2