MIDAS: Mixing Ambiguous Data with Soft Labels for Dynamic Facial Expression Recognition

Ryosuke Kawamura[1], Hideaki Hayashi[2], Noriko Takemura[3], Hajime Nagahara[2]
[1]Fujitsu Research of America, Inc.
[2]Osaka University
[3]Kyushu Institute of Technology
rkawamura@fujitsu.com, {hayashi, nagahara}@ids.osaka-u.ac.jp, takemura@ai.kyutech.ac.jp

Abstract

Dynamic facial expression recognition (DFER) is an important task in the field of computer vision. To apply automatic DFER in practice, it is necessary to accurately recognize ambiguous facial expressions, which often appear in data in the wild. In this paper, we propose MIDAS, a data augmentation method for DFER, which augments ambiguous facial expression data with soft labels consisting of probabilities for multiple emotion classes. In MIDAS, the training data are augmented by convexly combining pairs of video frames and their corresponding emotion class labels, which can also be regarded as an extension of mixup to soft-labeled video data. This simple extension is remarkably effective in DFER with ambiguous facial expression data. To evaluate MIDAS, we conducted experiments on the DFEW dataset. The results demonstrate that the model trained on the data augmented by MIDAS outperforms the existing state-of-the-art method trained on the original dataset.

1. Introduction

Facial expressions play an important role in human communication, and facial expression recognition (FER) has broad applications in areas such as human-computer interaction, driver monitoring, and intelligent tutoring systems for education. To correctly understand emotions from facial expressions, the temporal cues of facial expressions are important for FER because facial expressions are based on facial muscle movements, as demonstrated in previous research [23, 44, 45]. Accordingly, our study focuses on dynamic FER (DFER), which is the task of recognizing an emotion class from a video clip.

Although deep learning-based techniques have shown remarkable performance in DFER on lab-controlled data, DFER on in-the-wild data is still a difficult problem because such data include ambiguous facial expressions, which cannot simply be categorized into a single emotion class. There are several factors that contribute to the ambiguity in facial expressions, with the coexistence of multiple emotions being one of the significant factors. Since emotions are not mutually exclusive but collective, multiple emotions can coexist at different intensities in ambiguous facial expressions captured under natural conditions. This is a significant difference from lab-controlled data, where the researcher usually instructs the subject to make facial expressions. In addition, facial expression varies over time, and multiple emotions can be contained even within a single video clip. For example, in Fig. 1, the annotators' evaluations for this video clip were split between "disgust," "neutral," and "fear" with different probabilities. These features of emotions and facial expressions are considered to be the main factor of ambiguity.

Attaching soft labels to training data, instead of hard labels, is an effective way to address the ambiguity in DFER. Hard labels, that is, one-hot encoded class labels, are mostly used in general recognition tasks such as object recognition, where the input sample is clearly categorized into a single class; however, they cannot appropriately represent an objective variable composed of a combination of multiple emotions with different intensities in ambiguous facial expressions. To correctly learn the ambiguity in DFER, soft labels consisting of probabilities for multiple emotions are helpful to maximize the use of information provided by annotators. One possible method of assigning soft labels to training data is to have multiple annotators evaluate the training data and use the ratio of their votes.

The disadvantage of soft labels is that they are more flexible than hard labels, making it difficult to collect a variety of labels in a uniform manner. There is an enormous amount of possible combinations of emotion classes and the corresponding probabilities, and therefore it is difficult to prepare training data that include all of these patterns. Furthermore, the size of the dataset itself also tends to be limited in DFER

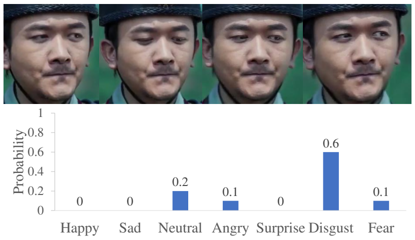

Figure 1. Example of an ambiguous facial expression. The images were taken from the DFEW dataset [14]. The bar chart in the bottom row shows the soft-labeled annotation constructed based on the proportions of votes by ten annotators. The annotations are split into four emotion classes.

due to the difficulty of manual annotation and data collection. To address this problem, it is necessary to augment data effectively and properly to learn from limited data.

In this paper, we propose a data augmentation method for DFER with ambiguous facial expressions called MIDAS (Mixing Ambiguous Data with Soft labels). In MIDAS, the mixing strategy is expanded to handle soft labels and dynamic facial expressions. The method convexly combines pairs of video frames of facial expressions and the corresponding soft labels that represent the probabilities of emotion classes after aligning the facial position. It then trains a model on the generated data.

Our contributions are summarized as follows:

- We proposed MIDAS, a data augmentation method for DFER with ambiguous facial expressions. MIDAS convexly combines pairs of video frames of facial expressions in a similar way to mixup. One significant difference from mixup is that MIDAS is applicable when the true hard labels are unknown and only soft labels consisting of multiple classes of probabilities are given.

- We showed that MIDAS corresponds to minimizing the vicinal risk in a situation where the true hard label is unknown with a vicinity distribution using a random ratio and virtual labels that are different from the original mixup.

- Through DFER experiments on the DFEW dataset, we showed that the proposed method outperforms existing state-of-the-art methods. Through an ablation study, we also showed that the combination of soft labels and mixing strategy has a synergistic effect although the use of each individually is also effective. Additionally, the effectiveness of MIDAS with hard labels was demonstrated on both the DFEW and FER39k datasets.

2. Related Work

2.1. DFER

While many in-the-wild datasets for static FER utilize images collected from the internet, most datasets for DFER, including CK+ [25] and Oulu-CASIA [50], are created under lab-controlled environments. In these datasets, the changes in facial expression are prompted by researchers' instructions. Although still relatively few, there has been a growing trend toward the development of large-scale in-the-wild datasets for DFER. Notably, AFEW, introduced by Dhall *et al.* [8] stands out as the pioneering in-the-wild DFER dataset, comprising short clips from movies annotated by a pair of annotators. Similarly, Jiang *et al.* [14] collected movie clips for the DFEW datasets. They assigned ten out of twelve annotators to one video clip. This dataset uniquely provides both single-labeled and seven-dimensional emotion class annotations. Furthermore, FERV39k [40] presents a large in-the-wild dataset tailored for DFER. This dataset contains video clips in 22 fine-grained contexts such as business, daily life, and school. The data are annotated by 20 crowd-sourcing annotators and 10 professional researchers.

Regarding FER in the video, methods based on selecting peak frames or aggregating features from each frame have been proposed by [18, 19, 28, 43, 46, 51]. Two- or three-dimensional convolutional neural networks (2D-CNN or 3D-CNN) combined with a sequential neural network, such as long short-term memory (LSTM) and gated recurrent unit (GRU), are commonly used in [1,4,15,17,22,37,42,49].

Several studies have proposed the utilization of a Transformer-based module for DFER [26,52]. For example, Zhao *et al.* [52] proposed Former-DFER, which is based on the Transformer module, and Ma *et al.* [26] used features processed by a 2D-CNN as input to a spatio-temporal Transformer (STT). Wang *et al.* [41] proposed the dual path multi-excitation collaborative network (DPCNet). DPCNet consists of two modules, a spatial-frame excitation module to extract spatial features and a channel-temporal aggregation module to aggregate channel and temporal aware features.

2.2. Ambiguity in FER

Facial expressions are known to contain multiple emotion classes [6, 53]. Ambiguous data are often regarded as noisy or inconsistent data. There are several kinds of approaches to deal with ambiguous data such as uncertainty estimation. She *et al.* [32] proposed an architecture with a latent label distribution module and uncertainty estimation module to address the ambiguity. Wang *et al.* [39] proposed an extra module to suppress harmful instances and find latent truths by ordering with an estimated confidence level. Li *et al.* [20] proposed a global convolution attention block and intensity aware loss (GCA+IAL). IAL is designed to have the network pay extra attention to the most confusing category. While their approach focuses on the uncertainty of facial expressions through attention blocks and loss functions, our approach handles ambiguity in DFER using data augmentation with soft labels. Soft labels are a simple approach against ambiguous data. Barsoum *et al.* [53] investigated whether soft labels annotated by multiple crowd workers improved the static FER performance of a deep learning architecture and showed a model trained with soft labels outperformed that trained with hard labels. In addition, Gan *et al.* [10] proposed a framework to generate pseudo soft labels for static FER. However, to the best of our knowledge, no prior research has focused on using a mixing strategy with soft labels for DFER.

2.3. Mixing strategy

The application of mixing strategies in data augmentation has widely been investigated. Zhang *et al.* [48] proposed a data augmentation method called *mixup*, where additional training samples are synthesized by convexly combining random pairs of images and their labels. Mixup is based on the vicinal risk minimization [2] principle, where the vicinity of each training sample is used to approximate the true distribution, thereby improving the generalization capability. Thulasidasan *et al.* [33] showed that mixup is also a better training strategy from the viewpoint of confidence calibration.

Inspired by mixup, other methods using mixing strategies have also been proposed. CutMix [47] and CutOut [7] use regional crop-and-paste techniques. Saliency information was employed in SaliencyMix [34] and PuzzleMix [16]. Some methods, such as Attentive-CutMix [38] and TransMix [3], employ activation or attention maps to achieve mixing. The mixing strategy is also applied to feature space in certain methods such as Manifold-mix [36] and PatchUp [9]. MixGen [11] enhances vision-language representation learning, employing multi-modal data augmentation based on image interpolation and text concatenation. Unlike our focus on label-based classification, MixGen targets tasks such as visual grounding and reasoning.

Most studies focus on image mixing, while few apply the mixing strategy to video data. Sahoo *et al.* [31] proposed background mixing for contrastive learning in action recognition; however, their method was not used for generating data belonging to a class different from that of the source data. In addition, existing studies use hard labels because class information is clearly different from others in other computer vision tasks such as object and action recognition. However, at the time of writing, there has been no prior research dedicated to a mixing strategy for DFER and soft labels.

3. MIDAS: Mixing Ambiguous Data with Soft Labels

MIDAS generates data in a similar way to mixup, that is, it convexly combines given training data and labels using a randomly generated mixing coefficient. It differs from mixup in that (i) the input data are video clips and (ii) soft labels representing class probabilities are given instead of single ground-truth class labels encoded in one-hot format, i.e., hard labels. The soft labels are assumed to be given based on the average of the votes by multiple annotators. This is due to the fact that recognizing ambiguous facial expressions is difficult even for the human eye, and each annotator's judgment is not necessarily correct.

The important point is that the true hard label is unknown, and the proposed method is designed to minimize the vicinal risk under this condition. The data mixing procedure and how it minimizes vicinal risk are described below.

3.1. Data mixing

Let $X_i = \left(x_i^{(1)}, \dots, x_i^{(T)}\right)$ be the i-th video clip with a length of T in the training dataset, where $x_i^{(t)} \in \mathbb{R}^{H \times W \times 3}$ is an image of height H and width W from the t-th frame in the video. In addition, let $y_i \in \mathbb{R}^C$ denote the soft-labeled ground truth for the i-th video clip whose elements are the probabilities for C emotion classes. MIDAS combines each frame of two video clips randomly selected from the training dataset and generates virtual samples, as illustrated in Fig. 2. The generated video clip $\tilde{X}$ and label $\tilde{y}$ are formulated as

$$\tilde{X} = \left(\tilde{x}^{(1)}, \dots, \tilde{x}^{(T)}\right), \tag{1}$$

$$\tilde{x}^{(t)} = \lambda x_i^{(t)} + (1 - \lambda) x_j^{(t)}, \tag{2}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j, \tag{3}$$

where $\lambda \in [0, 1] \sim \beta(\alpha, \alpha)$ is a random ratio that follows the beta distribution with α. It should be noted that MIDAS does not allow the same video to be selected. After that, to take annotation noise such as misjudgment into account [35], a combined soft label $\tilde{y}$ is normalized by using a softmax operation. By applying our method to the facial expression data, data that have multiple emotion classes with different intensities and temporal changes can be generated.

3.2. Vicinal risk minimization

The data augmentation that MIDAS performs is justified from the viewpoint of vicinal risk minimization [2], as is done in mixup [48]. The difference with mixup is that the true hard label of each sample in the training data is unknown, and a soft label that includes the variation of the annotators' evaluation is used instead. Accordingly, it can

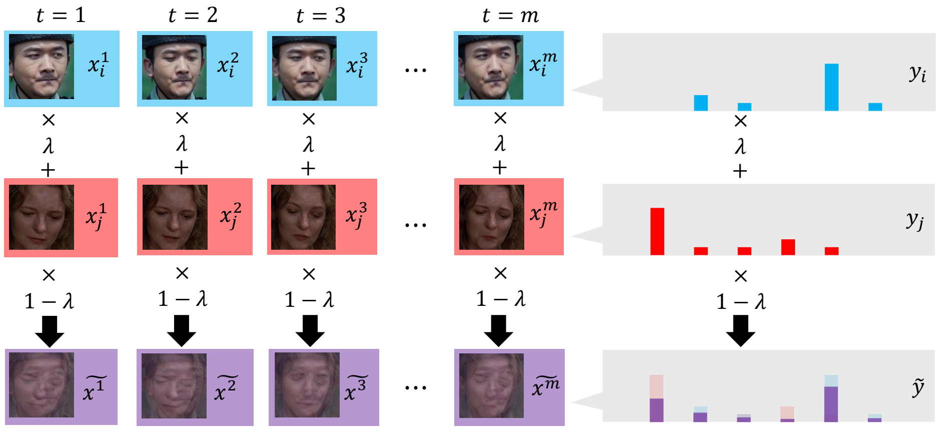

Figure 2. Outline of the data mixing procedure in MIDAS. In MIDAS, the training data are augmented by convexly combining pairs of video frames and their corresponding emotion class labels. The mixing coefficient λ is randomly generated from a beta distribution. The key point is that soft labels representing class probabilities are used instead of hard labels.

be explained that MIDAS calculates an empirical risk using a distribution different from that of mixup.

In supervised learning, we assume a joint probability distribution $P(x, y)$ over input and output variables and minimize the expectation of a given loss function ℓ.

$$R(f) = \int \ell(f(x), y) \mathrm{d}P(x, y), \tag{4}$$

where f is a classifier to be trained. However, eq. (4) cannot directly be computed because the joint distribution $P(x, y)$ is unknown. In general, we minimize instead the empirical risk given a training dataset $\{(x_i, y_i)\}_{i=1}^{M} \sim P(x, y)$.

$$R_{\text{emp}}(f) = \frac{1}{M} \sum_{i=1}^{M} \ell(f(x_i), y_i) \tag{5}$$

The empirical risk is derived by taking an expectation of the loss function over an empirical distribution $P_{\text{emp}}(x, y) = \frac{1}{M} \sum_{i=1}^{M} \delta(x = x_i, y = y_i)$, where δ is a Dirac measure.

There are other possible choices to approximate the true distribution, and a different choice of distribution results in different risk minimization. The empirical vicinal risk based on the vicinity distribution [2] is one of them. In [48], it was shown that mixup training minimizes the empirical vicinal risk:

$$R_{\text{mixup}}(f) = \frac{1}{M} \sum_{i=1}^{M} \ell(f(\tilde{x}_i), \tilde{y}_i), \tag{6}$$

where $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{M}$ is a set of virtual feature-target pairs generated from the vicinity distribution defined as

$$\begin{aligned} &P_{\text{mixup}}(\tilde{x}_i, \tilde{y}_i \mid x_i, y_i) \\ &= \frac{1}{n} \sum_{j}^{n} \mathop{\mathbb{E}}_{\lambda} [\delta(\tilde{x}_i = \lambda x_i + (1 - \lambda) x_j, \tilde{y}_i = \lambda y_i + (1 - \lambda) y_j)]. \end{aligned} \tag{7}$$

In the problem setting of this study, the hard label y_i[1] corresponding to the underlying true emotion is unknown, and instead a soft label q_i based on the voting average of multiple annotators is given. In MIDAS, the training data are sampled from the following distribution:

$$\begin{aligned} &P_{\text{MIDAS}}(\tilde{x}_i, \tilde{y}_i \mid x_i, q_i) \\ &= \frac{1}{n} \sum_{j}^{n} \mathop{\mathbb{E}}_{\lambda} [\delta(\tilde{x}_i = \lambda x_i + (1 - \lambda) x_j, \tilde{y}_i = \lambda q_i + (1 - \lambda) q_j)]. \end{aligned} \tag{8}$$

Equation (8) can be regarded as a variation of vicinity distribution. Assuming that S annotators give one-hot labels $v_i^{(s)}$ $(s = 1, \ldots, S)$ to each training sample x_i, the soft label q_i is given by the average of the annotators' votes as $q_i = \frac{1}{S} \sum_{s=1}^{S} v_i^{(s)}$. If l out of S votes are correct, q_i is expressed as $q_i = \frac{l}{S} y_i + \frac{1}{S} \sum_{s \in \mathcal{W}} v_i^{(s)}$, where $\mathcal{W}$ is a set of indices for wrong annotations. Using this expression, Eq. (8) can be written as

$$\begin{aligned} &P_{\text{MIDAS}}(\tilde{x}_i, \tilde{y}_i \mid x_i, q_i) \\ &= \frac{1}{n} \sum_{j}^{n} \mathop{\mathbb{E}}_{\lambda, \lambda'} [\delta(\tilde{x}_i = \lambda x_i + (1 - \lambda) x_j, \tilde{y}_i = \lambda' y_i + (1 - \lambda') y'_j)], \end{aligned} \tag{9}$$

where we defined $\lambda' = \frac{\lambda l}{S}$ and $y'_j = \frac{\lambda}{S - \lambda l} \sum_{s \in \mathcal{W}} v_i^{(s)} + \frac{S(1-\lambda)}{S - \lambda l} q_j$. MIDAS corresponds to minimizing the vicinal risk in a situation where the true hard label y_i is unknown, by defining the vicinity distribution using a random ratio and virtual labels that are different from the original mixup.

4. Experiments

The purpose of this experiment is to verify the validity of MIDAS for DFER using a deep learning-based automatic

[1] The superscript for the frame number is omitted for simplicity in this subsection.

DFER model. We evaluated the performance of MIDAS using a publicly available DFER dataset, and compared the results with those of existing methods for DFER including the state-of-the-art one.

4.1. Evaluation dataset

We used the dynamic facial expression in-the-wild (DFEW) dataset, a collection of 11,967 video clips sourced from movies. The DFEW stands as the singular dataset providing soft labels for each individual video clip. These video clips contain various challenging interferences in practical scenarios such as extreme illumination, occlusions, and capricious pose changes. Twelve expert annotators were hired for this dataset, and ten out of twelve annotators were assigned to each video clip. Each annotator was asked to select one out of seven emotion classes ("happy," "sad," "neutral," "angry," "surprise," "disgust," and "fear"). The voting results by the annotators are provided as seven-dimensional emotion distribution labels, which are used as soft labels in this experiment. This dataset also stores the class with the highest number of votes by the annotators for each sample, and we used them as a hard label in comparative experiments.

Fig. 3 shows examples of facial expression images and the corresponding emotion labels in the DFEW dataset. In the figure, the left and right panels show examples of a clear expression that all annotators judged as "happy" and an ambiguous facial expression, respectively. In the example of an ambiguous expression on the right panel, the votes by the annotators are split into five classes although more than half of the annotators judged this sample as "sad."

Fig. 4 shows the distribution of emotion classes. The DFEW dataset is a class-imbalanced dataset that contains relatively more "natural" and "happy" and less "disgust" and "fear."

4.2. Preprocessing

First, we detected a facial region using Face++ [27]. Face++ is a face recognition-related software that can be used for face detection, face comparison, and face retrieval, and we used its face detection function in this experiment. We then extracted facial landmarks from the detected face area and applied an affine transformation to landmarks to align the position of facial landmarks using Seeta [24] by referring to the method in [14].

4.3. Experimental conditions

We used the temporal shifted module (TSM) with the ResNet-18 backbone [21], which has been used for facial emotion recognition from video clips in previous studies. The ResNet-18 was pre-trained on ImageNet [30]. Since the length of the videos varied, we divided each video into eight segments and sampled one frame from each segment. The hyper-parameter settings for training were fixed. Regarding optimizers, we used SGD [29] with an initial learning rate of 0.02 and a momentum value of 0.9. Cosine learning rate decay was applied in training, and the number of training epochs was set to 450. We applied random scaling before inputting the video clip into the model at training time. We used a batch size of eight and a dropout with a rate of 0.5. The input image of each frame was resized to 224 × 224. For the loss function, we used cross-entropy loss. In addition, α for the beta distribution for MIDAS was set to 0.4.

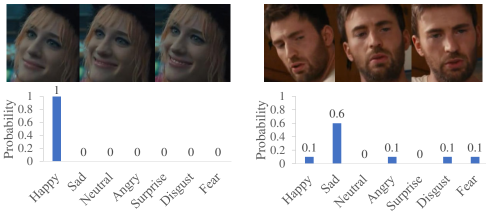

Figure 3. Examples of a clear facial expression (left) and ambiguous facial expression (right) with their soft label annotations in the DFEW dataset [14]

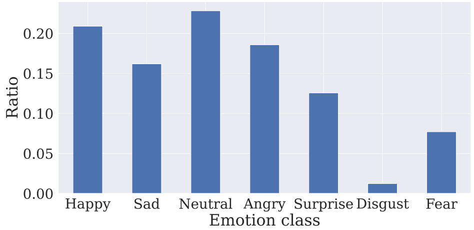

Figure 4. Emotion class distribution of the DFEW dataset

To evaluate the generalization capability, we performed five-fold cross-validation using the data split provided in the DFEW dataset. For the evaluation indexes, we employed the unweighted average recall (UAR) and weighted average recall (WAR), which are officially used in [14]. UAR represents the average prediction accuracy of each class and WAR represents accuracy. We calculated the averages of UAR and WAR over five groups of cross-validation. We compared the results with those of some existing methods for DFER proposed in [5, 12, 13, 26, 52] including the state-of-the-art method. We also employed the ResNet-18 with TSM trained simply on soft and hard labels for comparison methods to evaluate the effectiveness of MIDAS.

4.4. Result

Table 1 summarizes the accuracy for each emotion class, the WAR, and UAR, with the scores of comparative methods and our model trained on the original video (i.e., without data augmentation) with soft and hard labels. MIDAS achieved the best scores in WAR and UAR, thereby showing the effectiveness of the combination of soft labels and mixing strategy for DFER.

Compared to other methods, MIDAS showed the best

Table 1. Comparison of the accuracy for each emotion class, UAR, and WAR. Bold and underlined scores denote the best and second best, respectively.

Method	Label	Accuracy for each emotion class (%)							Metrics	
		happy	sad	neutral	angry	surprise	disgust	fear	UAR	WAR
3D Resnet18 [12, 52]	Hard	79.18	49.05	57.85	60.98	46.15	2.76	21.51	46.52	58.27
Resnet18+GRU [5, 13, 52]	Hard	82.87	63.83	65.06	68.51	52.00	0.86	30.14	51.68	64.02
Former-DFER [52]	Hard	84.05	62.57	67.52	70.03	56.43	3.45	31.78	53.69	65.70
STT [26]	Hard	87.36	**67.96**	64.97	71.24	53.10	3.49	34.04	54.58	66.65
DPCNet [41]	Hard	**89.59**	64.82	66.98	63.14	53.81	14.48	32.34	57.11	66.32
GCA+IAL [20]	Hard	87.95	67.21	**70.10**	**76.06**	**62.22**	0.00	26.44	55.71	**69.24**
Resnet18+TSM (Ours)	MIDAS	87.40	67.34	58.64	68.06	59.65	**28.69**	**44.50**	**57.45**	69.16

score in UAR with a 0.34% gap from the second-best method (DPCNet) and achieved the second-best score in WAR with only a 0.08% gap from GCA+IAL, which is the state-of-the-art method for the DFEW dataset. These results demonstrated the effectiveness of MIDAS in improving the performance of DFER.

Regarding the accuracy for each class, It should be emphasized that MIDAS scored higher than the other methods in "disgust" and "fear," whose number of training samples is much less than that of the other classes (see Fig. 4). In contrast, GCA+IAL [20], which represents the state-of-the-art in WAR and is another approach addressing facial expression ambiguity based on the intensity-aware loss function, showed remarkably low accuracy in these categories. These results indicate that our method improves the accuracy for emotion classes with smaller data sizes.

5. Analysis and Ablation Studies

5.1. Comparison with the model trained on soft and hard labels

While our approach has yielded noteworthy outcomes, it is essential to clarify MIDAS's effectiveness compared to models trained solely on hard and soft labels. To evaluate the effectiveness of MIDAS, we employed the ResNet-18 with TSM trained simply on soft and hard labels. The results are shown in Table 2. Regarding WAR, MIDAS achieved 69.16%, which outperforms the score of the models trained solely on hard labels (64.31%) and soft labels (67.27%). For UAR, MIDAS (57.45%) marked a better score than the models with soft labels (54.61%) and hard labels (54.03%). These results demonstrate the effectiveness of MIDAS in improving the performance of DFER. In addition, compared with the existing methods in Table 1, our model with soft labels achieved the third-best in both WAR and UAR, suggesting the effectiveness of simply using soft labels even without data augmentation.

Table 2. Comparison of the results of our method with the model trained on soft and hard label

Label	UAR	WAR
Hard	54.03	64.31
Soft	54.61	67.27
MIDAS	**57.45**	**69.16**

Table 3. Comparison of the results of our method with and without hard label

Label	UAR	WAR
Hard	54.03	64.31
Soft	54.61	67.27
MIDAS w/ hard label	54.93	65.66
MIDAS	**57.45**	**69.16**

5.2. Our method with hard labels

We investigated whether a mixing strategy for dynamic facial expressions with hard labels could improve performance. Creating soft labels is costly although our method can improve performance. It would be helpful if training on the single-labeled annotation that can be obtained more easily than soft labels also improves performance.

The procedure of this experiment is simple; we applied our method to video clips with hard labels instead of soft ones. The model was trained on this generated dataset after normalizing the combined labels in the same settings described above.

The results are shown in Table 3. The model trained on MIDAS with hard labels improved the WAR of the model trained on the original dataset with hard labels by 1.35%. For UAR, the score is 54.93%, which is higher than a model trained on the original dataset with hard and soft labels. This result indicates that our strategy can enhance the performance of FER when using hard labels.

Furthermore, we extended our investigation to another large-scale dataset, FERV39k [40]. Unlike the DFEW dataset, only hard-labeled annotations are given in FERV39k. The experimental setup remains consistent with Section 4.3, except for the initial learning rate, set at 0.01. Table 4 demonstrates that MIDAS with hard labels achieved a superior UAR score (39.2%) compared to current state-of-

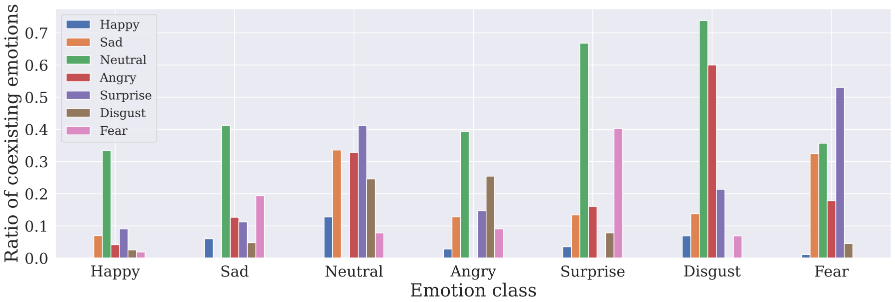

Figure 5. Ratio of coexisting emotions for each emotion class. The values in the figure were calculated by averaging the soft label values of the samples that belong to the corresponding emotion class. The higher the value, the more likely the emotion class is to be voted for by the annotators simultaneously, that is, the more likely it is to coexist.

Table 4. Comparison of the results of our method with hard label on FER39k [40]

Method	Label	UAR	WAR
Two VGG13-LSTM	Hard	31.28	43.2
Former-DFER [52]	Hard	37.20	46.85
GCA+ICL [20]	Hard	35.82	**48.54**
Resnet18+TSM (Ours)	MIDAS w/ hard label	**39.2**	47.37

the-art methods on the FERV39k dataset, and secured the second-best score (47.37%) in WAR. These results suggest that MIDAS possesses the potential for generalization even in scenarios involving only hard labels.

5.3. Cross-dataset evaluation

Generalization ability is crucial in facial expression recognition due to the need to accommodate variations such as lighting conditions, facial orientations, and the diversity in facial shapes. These factors can significantly influence the accuracy of recognition models in real-world situations.

We conducted a cross-dataset evaluation to examine whether our approach improves the model's generalization ability to unseen conditions. In this experiment, we trained the model on the DFEW dataset and tested it on the acted facial expressions in the wild (AFEW) dataset [8] to assess the performance on first-time encountered conditions. As in Section 4, we trained the TSM with ResNet-18 using MIDAS and compared its performance with those trained simply on soft and hard labels. These models were not further fine-tuned using the AFEW dataset, but simply evaluated on the AFEW test set. Due to the absence of a method providing results for models trained on DFEW and tested on AFEW datasets, our comparison involves simply contrasting the outcomes of MIDAS using models trained with soft and hard labels.

Table 5 presents the UAR and WAR on the AFEW test set for each method. MIDAS demonstrated superior performance in both UAR and WAR compared to the models trained on hard and soft labels. These results suggested that MIDAS potentially improves the model's generalization ability to unseen conditions.

Table 5. Comparison of our method's results on the AFEW test set

Label	UAR	WAR
Hard	38.20	40.72
Soft	36.67	39.61
MIDAS	**39.56**	**43.77**

5.4. Analysis of the impact of coexisting emotion

we analyzed the effect of coexisting emotions in DFER. The soft labels in the DFEW dataset were constructed based on the votes of ten annotators. During this voting process, the votes of all annotators do not necessarily coincide; some minority annotators vote for emotion classes other than the class that received the most votes. For example, "sad" received the most votes with eight votes, but "angry" also received two votes. We analyzed how such coexisting emotion classes that are often voted together affect the model capability.

Fig. 5 shows the average ratio of coexisting emotions for each emotion class. For example, in the case of the leftmost emotion class "happy," "neutral" is also voted for a lot for the instances where "happy" received the most votes. The values in the figure were calculated by averaging the soft label values of the samples that belong to the corresponding emotion class. From this figure, the following are observed.

- "Neutral" tends to coexist with all other emotion classes.
- "Happy" does not coexist much with other emotion classes.
- "Angry" and "disgust" coexist frequently.
- "Sad," "surprise," and "fear' often coexist.

Fig. 6 shows the confusion matrix of the Resnet-18 with TSM model trained with MIDAS. The impact of emotion

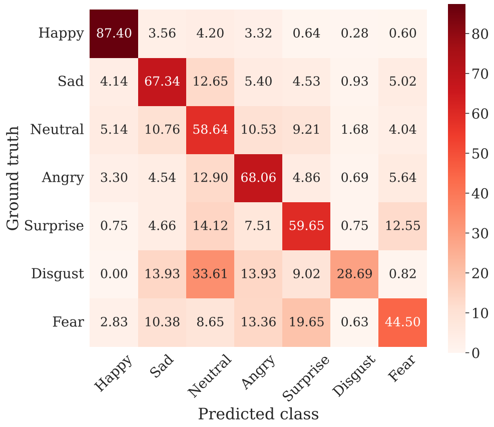

Figure 6. Confusion matrix of the model trained with MIDAS

class coexistence revealed in Fig. 5 on the classification results can be observed. "Neutral" coexists with other classes more than other emotion classes, the model wrongly predicted many samples of other classes as neutral. In particular, instances classified as "disgust" are most often recognized as "neutral." There are few coexisting emotions for "happy," and the instances classified as "happy" are almost always predicted correctly (87.40%). These results indicate that coexisting emotions affected model performance in the DFER.

5.5. The effect of ambiguous data on model performance

To confirm whether ambiguous data in the DFEW dataset affect the model performance, we investigated the effect of ambiguous data by comparing models trained on datasets with and without ambiguous data. In this comparison, we divided the original DFEW dataset into two groups: clear expression and mixed expression groups. The clear expression group consists of data with maximum soft label values of more than 0.9, e.g., the left example in Fig. 3. The mixed expression group contains data regardless of the soft labels' values and includes ambiguous facial expressions such as the right example in Fig. 3. To address the difference in data distribution, the distribution of each emotion class was matched to the original dataset by oversampling and down-sampling. Finally, the sizes of both data groups were set to an equal number (4275). We trained two models on these datasets with soft labels and evaluated them using a validation split of the original dataset.

The results are shown in Table 6. The model with the mixed expression group obtained higher UAR and WAR scores than the model trained with the clear expression group. These results demonstrated that the existence of ambiguous data can improve the performance of DFER.

Table 6. Comparison of the results of models trained with and without ambiguous data

Data	UAR	WAR
Clear expression group	45.42	58.46
Mixed expression group	**48.54**	**60.91**

Table 7. Results using the 2DCNN-GRU architecture, which is different from the architecture used in Table 1

Label	UAR	WAR
Hard	51.62	64.00
Soft	51.82	65.00
MIDAS	**53.70**	**67.01**

5.6. Comparison of different architectures

In our experiments, we used TSM [21], an architecture for dynamic data, and the results show our method with TSM improves the performance of DFER. However, we did not investigate whether our method is effective for different deep learning architectures.

We, hence, conducted an experiment with different architectures for DFER. In this experiment, we used a 2DCNN-GRU as a different deep learning architecture. The 2DCNN-GRU was trained using MIDAS and compared with a model trained on the original dataset with soft and hard labels. The settings of training were the same as those in the other experiments.

Table 7 summarizes the results using the 2DCNN-GRU architecture. The scores of MIDAS with 2DCNN-GRU (UAR: 53.70% and WAR: 67.01%) are higher than those of the soft- and hard-label supervised models. In addition, its WAR is higher than those of the existing works except for GCA+IAL [20] (see Table 1). These results indicate that MIDAS has the potential to improve the performance of DFER regardless of the architecture.

6. Conclusion

In this paper, to handle various ambiguous facial expressions, we proposed a data augmentation method called MIDAS, which is based on data mixing with soft labels for DFER. In our method, we combine two video clips of facial expressions and their soft labels to generate various combinations of emotions and intensities. We conducted experiments to evaluate our method with a dataset for DFER. The results showed that our method can enhance the performance of DFER and outperform the state-of-the-art method.

In future work, we plan to evaluate MIDAS using other domains of datasets, as it was evaluated using only the DFEW dataset. Although the MIDAS is developed for DFER with ambiguous facial expressions, it would be effective for other tasks that involve ambiguous class categorization soft-labeled annotations.

References

[1] Masih Aminbeidokhti, Marco Pedersoli, Patrick Cardinal, and Eric Granger. Emotion recognition with spatial attention and temporal softmax pooling. In *Proceedings of the International Conference on Image Analysis and Recognition (ICIAR)*, pages 323–331. Springer, 2019. 2

[2] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 13, 2000. 3, 4

[3] Jie-Neng Chen, Shuyang Sun, Ju He, Philip HS Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12135–12144, 2022. 3

[4] Weicong Chen, Dong Zhang, Ming Li, and Dah-Jye Lee. Stcam: Spatial-temporal and channel attention module for dynamic facial expression recognition. *IEEE Transactions on Affective Computing*, 2020. 2

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 5, 6

[6] Antitza Dantcheva, Piotr Bilinski, Hung Thanh Nguyen, Jean-Claude Broutart, and Francois Bremond. Expression recognition for severely demented patients in music reminiscence-therapy. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 783–787. IEEE, 2017. 2

[7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

[8] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, July 2012. 2, 7

[9] Mojtaba Faramarzi, Mohammad Amini, Akilesh Badrinaaraayanan, Vikas Verma, and Sarath Chandar. Patchup: A regularization technique for convolutional neural networks. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, volume 36, pages 589–597, 2022. 3

[10] Yanling Gan, Jingying Chen, and Luhui Xu. Facial expression recognition boosted by soft label with a diverse ensemble. *Pattern Recognition Letters*, 125:105–112, 2019. 3

[11] Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. Mixgen: A new multimodal data augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 379–389, 2023. 3

[12] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 5, 6

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5, 6

[14] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the International Conference on Multimedia*, pages 2881–2889, 2020. 2, 5

[15] Dae Hoe Kim, Wissam J Baddar, Jinhyeok Jang, and Yong Man Ro. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*, 10(2):223–236, 2017. 2

[16] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5275–5285. PMLR, 2020. 3

[17] Youngsung Kim, ByungIn Yoo, Youngjun Kwak, Changkyu Choi, and Junmo Kim. Deep generative-contrastive networks for facial expression recognition. *arXiv preprint arXiv:1703.07140*, 2017. 2

[18] Boris Knyazev, Roman Shvetsov, Natalia Efremova, and Artem Kuharenko. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. *arXiv preprint arXiv:1711.04598*, 2017. 2

[19] Vikas Kumar, Shivansh Rao, and Li Yu. Noisy student training using body language dataset improves facial expression recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 756–773. Springer, 2020. 2

[20] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 67–75, 2023. 2, 6, 7, 8

[21] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the International Conference on Computer Vision (CVPR)*, pages 7083–7093, 2019. 5, 8

[22] Daizong Liu, Xi Ouyang, Shuangjie Xu, Pan Zhou, Kun He, and Shiping Wen. Saanet: Siamese action-units attention network for improving dynamic facial expression recognition. *Neurocomputing*, 413:145–157, 2020. 2

[23] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1749–1756, 2014. 1

[24] Xin Liu, Meina Kan, Wanglong Wu, Shiguang Shan, and Xilin Chen. VIPLFaceNet: an open source deep face recognition SDK. *Frontiers of Computer Science*, 11(2):208–218, 2017. 5

[25] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, June 2010. 2

[26] Fuyan Ma, Bin Sun, and Shutao Li. Spatio-temporal transformer for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2205.04749*, 2022. 2, 5, 6

[27] MEGVII. Face++ face detection API. https://www.faceplusplus.com/. 5

[28] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression recognition in videos. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages 3866–3870. IEEE, 2019. 2

[29] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 5

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 5

[31] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 34:23386–23400, 2021. 3

[32] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6244–6253, Nashville, TN, USA, June 2021. IEEE. 2

[33] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 3

[34] A F M Shahab Uddin, Mst. Sirazam Monira, Wheemyung Shin, TaeChoong Chung, and Sung-Ho Bae. Saliencymix: A saliency guided data augmentation strategy for better regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 3

[35] Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. A case for soft loss functions. In *Proceedings of the Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177, 2020. 3

[36] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 6438–6447. PMLR, 2019. 3

[37] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the International Conference on Multimodal Interaction*, pages 569–576, 2017. 2

[38] Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3642–3646. IEEE, 2020. 3

[39] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 6897–6906, 2020. 2

[40] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. FERV39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20922–20931, 2022. 2, 6, 7

[41] Yan Wang, Yixuan Sun, Wei Song, Shuyong Gao, Yiwen Huang, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos. In *Proceedings of the International Conference on Multimedia*, pages 101–110, 2022. 2, 6

[42] Jingwei Yan, Wenming Zheng, Zhen Cui, Chuangao Tang, Tong Zhang, and Yuan Zong. Multi-cue fusion for emotion recognition in the wild. *Neurocomputing*, 309:27–35, 2018. 2

[43] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2168–2177, 2018. 2

[44] Peng Yang, Qingshan Liu, Xinyi Cui, and Dimitris N Metaxas. Facial expression recognition using encoded dynamic features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 1

[45] Peng Yang, Qingshan Liu, and Dimitris N Metaxas. Boosting encoded dynamic features for facial expression recognition. *Pattern Recognition Letters*, 30(2):132–139, 2009. 1

[46] Zhenbo Yu, Qinshan Liu, and Guangcan Liu. Deeper cascaded peak-piloted network for weak expression recognition. *The Visual Computer*, 34(12):1691–1699, 2018. 2

[47] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019. 3

[48] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 3, 4

[49] Hepeng Zhang, Bin Huang, and Guohui Tian. Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. *Pattern Recognition Letters*, 131:128–134, 2020. 2

[50] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 2

[51] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 425–442. Springer, 2016. 2

[52] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the International Conference on Multimedia*, pages 1553–1561, 2021. 2, 5, 6, 7

[53] Ying Zhou, Hui Xue, and Xin Geng. Emotion distribution recognition from facial expressions. In *Proceedings of the International Conference on Multimedia*, MM '15, pages 1247–1250, New York, NY, USA, 2015. Association for Computing Machinery. 2, 3