Deep Insights of Learning-Based Micro Expression Recognition: A Perspective on Promises, Challenges, and Research Needs

Monu Verma, Member, IEEE, Santosh Kumar Vipparthi, Senior Member, IEEE, and Girdhari Singh

Abstract—Micro expression recognition (MER) is a very challenging area of research due to its intrinsic nature and fine-grained changes. In the literature, the problem of MER has been solved through handcrafted/descriptor-based techniques. However, in recent times, deep learning (DL)-based techniques have been adopted to gain higher performance for MER. Also, rich survey articles on MER are available by summarizing the data sets, experimental settings, conventional, and DL methods. In contrast, these studies lack the ability to convey the impact of network design paradigms and experimental setting strategies for DL-based MER. Therefore, this article aims to provide a deep insight into the DL-based MER frameworks with a perspective on promises in network model designing, experimental strategies, challenges, and research needs. Also, the detailed categorization of available MER frameworks is presented in various aspects of model design and technical characteristics. Moreover, an empirical analysis of the experimental and validation protocols adopted by MER methods is presented. The challenges mentioned earlier and network design strategies may assist the affective computing research community in forge ahead in MER research. Finally, we point out the future directions, research needs, and draw our conclusions.

Index Terms—CNN models, deep learning (DL), facial expression recognition, micro-expression recognition (MER).

I. INTRODUCTION

MICRO-EXPRESSIONS (MEs) exhibit the insight into true feelings of a person even if he/she is trying to hide the genuine emotions within the manifested emotion (macro expression). Ekman and Friesen [1] introduced various deceptive expressions (known as MEs) after investigating a depressed patient’s interview video who attempted to commit suicide. Ekman observed that, the patient surpasses his intensive sadness in happiness within 1/12 s. However, these expressions were spotted in a few frames of the video, recorded through a standard 25-fps device but serves enough clues to sense the true sentiments of the patient. Thus, MEs can be decisive in the fields of spotting genuine psychological activities [2], e.g., lie detection, psychoanalysis, criminal interrogation, medical diagnosis, pain detection, autism disorder, and business negotiation.

Micro expression recognition (MER) can broadly be divided into three steps: 1) preprocessing; 2) feature extraction; and 3) emotion classification.

Preprocessing Techniques: The first step of MER is to spot the MEs and then detect the RoIs from them. ME spotting is a vital step for automatic ME analysis as it locates the segments of micro-movements in a MEs video. Thus, precise ME spotting can decrease the redundant information and improve the performance of MER further. Recently, few studies [3], [4], [5] focused on ME spotting using deep learning (DL) methods. Li et al. [6] introduced an ME spotting method for spontaneous ME data sets. Furthermore, Zhang et al. [3] designed a DL-based ME spotting method by extracting features from video clips. Tran et al. [5] proposed a deep sequence model for ME spotting. Moreover, Liong et al. [4] proposed an automatic apex frame spotting model. (A more detailed categorization and analysis of the ME spotting can be found in existing MER surveys [6], [7], [8], [9]). Further, face alignment and noise filtration are employed to systematize the input data samples for better feature extraction and learning [8]. Some of the MER frameworks also utilized the motion magnification (MM) [10], [11] and temporal normalization [12], [13] techniques to enhance the visibility of the minute temporal variations and normalize the frames. Recently, DL-based approaches require a huge data set for training. However, all available MEs data sets are far from the enough data samples. Therefore, data augmentation techniques such as: random crop and rotation in terms of the spatial domain, shifting, magnification [13], and synthetic data generation using generative adversarial networks (GANs) [14], [15] are also gaining the attention to enhance the data samples. The more details can be seen in [8] and [16].

Based on feature extraction and classification methods, MER approaches can be categorized into traditional and DL-based approaches.
**Traditional Handcrafted MER Methods:** The traditional MER methods rely on the predefined feature descriptors to encode the spatial and temporal changes from the MEs video sequences. In literature many robust spatio-temporal feature descriptors: local binary pattern (LBP) and its variants: three orthogonal planes (LBP-TOPs) [17], LBP with six intersection points (LBP-SIPs) [18], spatiotemporal LBP with integral projection (STLBP-IP) [19], revisited integral projection (DiSTLBP-RIP) [20], etc., were proposed for MER. Furthermore, some optical flow-based descriptors: main directional mean optical flow (MDMO) [21], sparseMDMO [22], FHOFO [23], facial dynamic map (FDM), and color-based descriptors: TICS [24] were introduced to encode the features of MEs. After that, the encoded features are forwarded to the classifiers, such as support vector machine (SVM), neural networks (NNs), etc., which learn the distinctive properties of the emotion classes. Sensitivity and specificity of the traditional descriptors have gained good performance as compared to professionally trained specialists. However, it is still difficult to manually design a robust descriptor for capturing quick subtle changes in MEs. The detailed summary of the traditional MER approaches are listed in the supplementary draft (supplementary: Table III). A more detailed categorization of traditional methods and classifiers can be found in [6] and [9].

**DL-Based MER Methods:** The supervised techniques of DL adaptively learn the features from the raw data and classify the emotion classes accordingly. This article aims to describe the details of network design strategies followed in the literature in-terms of downsampling, multistream, multiscale, deep, or shallow networks, kernels depth, sizes, etc., for MER. Moreover, there is no standard evaluation protocol, class settings, and metrics for the fair comparison of the models in the literature. Therefore, it is difficult to come up with a common conclusion of the performances for the existing state-of-the-art approaches. Thus, these factors motivated us to present a detailed survey by addressing the effects of selecting the input formats, evaluation strategies, implementation settings, evaluation metrics on MER performance. More details of input formats, such as apex frame, onset-apex-offset frames, compressed single instance image, and image sequences are studied and discussed their effect on the model’s overall performance. Similarly, the evaluation strategies, such as person dependent, a person independent, composite, cross-domain, and class settings like 3-, 4-, 5-, and 7-emotion classes, and implementation setting like learning rates, data augmentation, and evaluation metrics: recognition accuracy, F1-score, unweighted F1-score, unweighted average recall (UAR), mean diagonal value of the confusion matrix, and its impact on overall models performances is discussed in detailed.  

A. Comparison With Previous Reviews

In past years, many notable surveys related to MER approaches have been published and the details are summarized in Table I. Many articles focus to summarise the details of data sets, preprocessing techniques, traditional MEs spotting along with the feature extraction algorithms, classifiers, and experimental settings. Also, the survey of the MER framework is available in [8], [9], [16], [25], and [26]. First, Merghani et al. [25] briefly described CNN frameworks and the technical differences with the traditional approaches in MER. Further, a detailed study of MEs databases and the comparative analysis of the data with challenges in data accumulation and labeling, evaluation metrics, and strategies are discussed. Similarly, Goh et al. [16] presented a survey by including data sets, preprocessing techniques, MEs spotting, and feature extraction algorithms. The main focus of the study is to highlight the ME features by dividing them into three categories: 1) low; 2) mid; and 3) high-level features. Also, Zhou et al. [26] presented a brief survey of the available traditional and DL techniques with preprocessing techniques and data sets. Apart from the similar data collection, evaluation matrix, and categorization of the conventional and CNN methods, Xie et al. [8] presented the details of the macro-to-micro feature adoption and synthetic data generation to balance MEs data samples. Guerdelli et al. [28] presented a detailed survey of facial expression data sets by including both macro and micro expressions. In recent times, Ben et al. [9] presented a micro-and-macro expression warehouse (MMEW) data set by incorporating both micro and macro expressions. Also, a detailed study of the challenges while creating the data sets is presented. The apparent technical differences between the conventional approaches in the literature are presented for ME. In addition, a brief study of the CNN techniques is given in this article. From the above details, it is clear that the available survey articles focus on presenting a categorization of the available models and data sets. Similarly, Li et al. [27] presented a detailed study of DL methods for MER, challenging data sets, and comparative analysis between most influential MER methods. Moreover, the study also detailed the remaining challenges and future scope of the MER. However, they fail to comprehensively analyze the essential aspects, such as model design, evaluation strategies, challenges, and research needs for DL models. Therefore, this article focuses on presenting a detailed survey on.

1) The detailed survey of DL-based MER approaches has been presented by categorizing into three major categories: a) multistage; b) end-to-end; and c) transfer learning-based MER. Further, subcategorization of these categories is offered by considering the CNN frameworks, such as 2D-CNN, multi-stream/scale, capsule, 3D-CNN, CNN-LSTM, Graph-based, NAS-based, etc. Also, the detailed technical characteristics of these frameworks are discussed.

2) Promises toward the different modules of DL-based model designing, such as two stage and end-to-end learning, downsampling, kernel sizes, shallow, and deep networks, to design and develop an effective and efficient DL framework for MER are discussed and observations are concluded.

3) A detailed discussion on validation strategies [PDE: LOVO and 80/20split, PIE: LOSO, composite LOSO, and cross-domain evaluation (CDE)], evaluation metrics (accuracy, F1-score, Recall, U1, UAR, and confusion matrix) and their significance on MER frameworks are presented.
TABLE I
SUMMARIZATION OF EXISTING SURVEYS IN LAST DECADE

| No. | Pub-Year | Title                                                      | Highlights                                                                 |
|-----|----------|------------------------------------------------------------|-----------------------------------------------------------------------------|
| 1.  | Fr. Psy. 2018 [7] (Frontiers) | A Survey of Automatic Facial Micro-Expression Analysis: Databases, Methods, and Challenges | A survey of databases and traditional feature based methods for MEs spotting and recognition. The survey also presents the evaluation metrics (accuracy and F1 score), validation strategies and challenges. |
| 2.  | T-AFF-2018 [6] (IEEE) | Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-Expression Spotting and Recognition Methods | A survey of various preprocessing methods and feature descriptors for spontaneous ME spotting and recognition. The survey also conducted the competitive analysis of conventional MER approaches. |
| 3.  | Arxiv-2018 [25] | A Review on Facial Micro-Expressions Analysis: Datasets, Features and Metrics | A survey of various feature extraction algorithms, classifiers and evaluation metrics for MER. The survey also presents the datasets, challenges in data accumulation, comparative analysis between available datasets, evaluation metrics, validation strategies and analysis of learning approaches as compared to traditional approaches. |
| 4.  | Vis.Comp-2020 [16] (Springer) | Micro-expression recognition: an updated review of current trends, challenges and solutions | A survey of traditional approaches for MEs detection and recognition by categorizing solutions into low-level, mid-level, and high-level solutions. |
| 5.  | IVC-2020 [26] (Elsevier) | A survey of micro-expression recognition                     | A survey of datasets, preprocessing techniques and existing MER algorithms based on the issues: overfitting, data unbalancing and robustness. |
| 6.  | Arxiv-2020 [8] | An Overview of Facial Micro-Expression Analysis: Data, Methodology and Challenge | A survey of datasets, preprocessing techniques, handcrafted features, ME spotting and learning based MER. The learning based MER discussed by dividing into three aspects: macro- to micro-adaptation, recognition based on key apex frames, and recognition based on facial action units. |
| 7.  | T-PAMI-2021 [9] (IEEE) | Video-based Facial Micro-Expression Analysis: A Survey of Datasets, Features and Algorithms | A review that highlighted the key differences between macro- and micro-expressions, video-based micro-expression analysis, neuropsychological basis for MEs and datasets. The survey focused on the comprehensive study of existing MEs dataset and introduce a new MMIE dataset. Also, the survey included a brief summary of features spotting algorithms, recognition algorithms, applications and evaluation metrics for MER approaches. |
| 8.  | Arxiv-2021 [27] | Deep Learning for Micro-expression Analysis: A Survey       | A review that highlighted deep learning based MER methods, challenging datasets, and comparative analysis between most influential DL based MER methods. The survey also included remaining challenges, and future direction of MER. |
| 9.  | Sensors-2022 [28] | Macro-and Micro-Expressions Facial Datasets: A Survey      | A survey of facial expression datasets, including both macro and micro expressions. The comparative review of the existing deep learning frameworks with research needs and the deep study of the technical characteristics like, end-to-end vs two-stage architecture, downsampling, multi-stream/scale structure, deeper vs shallow network, etc. has done. Moreover, the MER approaches used different metric calculation or platforms or number of emotion classes or input strategies; therefore it is difficult to compare the performance of MER frameworks. We reviews these factors in detail and highlighted the effects on the performance of MER. |

4) A detailed study of different validation setups: PDE, PIE, and CDE, and other experimental settings: data augmentation, input selection, number of emotion classes, etc., is conducted to analyze their effects over the performance of the MER approaches. Thus, new researchers will get awareness about the selection of experimental settings in DL-based MER.

To the best of our knowledge this is the first attempt to comparatively analyze the role of various designing modules, evaluation strategies, and experimental settings for learning-based MER frameworks. The main aim of this study is to help researchers or affective computing community to concentrate on the effective designing modules and experimental settings to design a robust MER framework. The detailed comparison between recently published survey [9] and the proposed survey is presented in the supplementary file. In addition, the supplementary file also included the visual presentation of the different DL-based MER approaches. The detailed information about the handcrafted MER approaches is tabulated in Table III of the supplementary file.

II. DEEP LEARNING-BASED MICRO EXPRESSION RECOGNITION

The MER frameworks demand spatio-temporal feature learning with momentary changes to capture the subtle variations of MEs. These factors make the design and development of DL models for MER an incredibly challenging task. In this section, we present an empirical review of DL methods highlighted in Fig. 1. Based on available framework architectures
in the literature, we categorize these frameworks into three broad categories: 1) multistage (Section II-A); 2) an end-to-end (Section II-B); and 3) transfer learning (Section II-C)-based MER frameworks as shown in Table II. We further divide the categories into subcategories based on different network characteristics: 2D-CNN, multi-scale/stream, capsule, RCN, and 3D-CNN. More details of the sub categorization is discussed as follows.

1) **2D-CNN Models:** Almost all DL-based MER approaches adopt the 2D-CNN networks [11], [29], [30], [31], [32]. Learning the spatio-temporal features from 2D-CNN is an insignificant problem. Therefore, to maintain the compatibility with 2D-CNN, researchers have designed two-stage frameworks [11], [29], [30], [31], [32], [33], [34]. Where, in the first stage optical flow or single instance image is computed by applying handcrafted approaches. While in the second stage the 2D-CNN model is designed to learn the MEs specific features. Specifically, Verma et al. [29] introduced a 2D-CNN MER framework. The framework first computed the single image instances by applying the dynamic imaging. Furthermore, the 2D-CNN model LEARNet is designed to learn the MEs specific features. Gupta [35] introduced the 2D-CNN MERASTC approach for MER by encoding the subtle deformations through action units (AUs), landmarks, gaze, and appearance features of MER.

**Discussion:** For MER, 2D-CNN models require an auxiliary first stage to process the spatio-temporal information into 2-D formats (Supplementary Fig. 2). However, the 2D-CNN are easy to design as well as require very less computational cost as compared to spatio-temporal networks, such as CNN-LSTM, RCN, 3D-CNN, etc. In addition 2D-CNN approaches [33], [34], [35] have shown impressive performance as shown in Tables IV and V. Therefore, 2D-CNN networks gain much attention as compared to other techniques in MER as reported in Table II.

2) **Multistream Networks:** Multistream networks capture a diverse range of features from different streams to learn
TABLE II

| Pub-Yr   | Input       | Prepr. | Handcrafted Support | N/W Type | E-to-E | Macro Support | P.T. Weights | Protocol |
|----------|-------------|--------|---------------------|----------|--------|---------------|--------------|----------|
| ACM-2016 | Video       | No     | No                  | CNN-LSTM | Yes    | No            | No           | LOSO     |
| ICPR-16  | Video (20F) | ASM    | No                  | 2D-CNN   | Yes    | No            | No           | 4Fold-LOS0 |
| FG-18    | Apex        | CS, R, S | No                  | Transfer-Learning | Yes | Yes | No | SD-LOS0, CDE |
| ICIP-18  | Apex        | EVM    | No                  | Transfer-Learning | Yes | No | Yes | SD-LOS0 |
| IPIF-19  | Video (30F) | ASM, EVM | No | RCN | Yes | No | No | 5Fold-LOS0, SD-LOS0 |
| FG-18 [36] | Video (10F) | DLib, TIM | Yes | O.F | CNN-LSTM, Multi-Stream | No | No | Yes | CD-LOS0, CDE |
| FG-19 [12] | Onset, Apex | TV-L1 | Yes | O.F | 2D-CNN, Multi-Stream, Multi-scale | No | No | No | SD-LOS0, CD-LOS0 |
| FG-19 [39] | Onset, Apex | - | Yes | O.F | 3D-CNN | No | No | No | SD-LOS0, CD-LOS0 |
| FG-19 [45] | Onset-Apex | - | Yes | O.F | Multi-Stream | No | No | No | SD-LOS0, CD-LOS0 |
| MTA-19 [38] | Video (10F) | TIM | Yes | O.F | CNN-LSTM, Multi-Stream | No | No | Yes | SD-LOS0 |
| PAA-19 [46] | Video (10-15F) | TIM | Yes | O.F | 3D-CNN | No | No | No | SD-LOS0 |
| ICIP-19 [32] | Onset-Apex | TV-L1, | Yes | O.F | 2D-CNN, Multi-Stream | No | No | No | SD-LOS0 |
| IICNN-19 [47] | Onset-Apex | Video | - | No | 3D-CNN, Multi-Stream | Yes | No | No | 80/20 Split |
| FG-19 [48] | Apex | FacialToolkit | No | Capsule | Yes | No | Yes | SD-LOS0, CD-LOS0 |
| IEEE-Acc-19 [33] | Video | - | Yes | UPL, BP, O.F | 2D-CNN, Multi-Stream, Multi-Scale | No | Yes | Yes | SD-LOS0 |
| ICBEA-19 [49] | Onset-Apex | EVM | Yes | O.F | RCN | No | No | No | SD-LOS0, CD-LOS0 |
| Neu.Co-19 [50] | Apex | AAM | No | 2D-CNN, Multi-Scale | Yes | Yes | Yes | CD-LOS0, CDE |
| TIP-19 [29] | Video | VoilaJones | Yes | DI | 2D-CNN, Multi-Scale | No | No | No | 80/20 Split |
| MLSP-19 [51] | - | DLib | No | CNN-LSTM | Yes | No | No | SD-LOS0, CD-LOS0 |
| IICNN-20 [31] | Video | VoilaJones | Yes | AI | 2D-CNN, Multi-Scale | No | No | No | SD-LOS0 |
| ToM-20 [13] | Video | ASM, TIM, EVM | Yes | O.F | RCN | No | No | No | SD-LOS0, LOVO |
| TIP-20 [10] | Onset-Apex | EVM | Yes | O.F | RCN | No | No | No | SD-LOS0, CD-LOS0 |
| TAFI-20 [52] | Video, Apex, On-A-off | OpenFace | No | Teacher-Student | Yes | Yes | Yes | SD-LOS0 |
| ACMMM-20 [53] | Video | - | No | GAN, 3D-CNN | Yes | Yes | No | SD-LOS0, CDE, LOVO |
| ACMMM-20 [54] | Apex | - | No | 2D-CNN, Encoder-Decoder, Graph-TCN | Yes | Yes | Yes | SD-LOS0, CD-LOS0 |
| ACMMM-20 [55] | Onset-Apex | - | No | 2D-CNN, Multi-Scale | Yes | No | No | SD-LOS0, CD-LOS0 |
| Arxiv-20 [15] | On-A-Off | Landmark Detection | No | GAN, Capsule | Yes | No | Yes | SD-LOS0, CD-LOS0 |
| IEEE-Acc-20 [56] | Video | - | Yes | LMF | CNN-LSTM | No | Yes | Yes | CD-LOS0, CDE |
| TIP-21 [11] | Apex | EVM | Yes | 3D-FPT | 2D-CNN | No | No | Yes | SD-LOS0, 90/10 Split |
| IEEE-Acc-21 [57] | Image | AAM, CLAHE | Yes | 2D-CNN | No | No | No | SD-LOS0, CD-LOS0 |
| IEEE-MM-21 [30] | Video | Viola Jones | Yes | DI | 2D-CNN, Multi-Scale | No | No | No | SD-LOS0, CDE |
| TAFFI-21 [35] | Video | Viola Jones | Yes | Landmark & Gaze feature | 2D-CNN | No | No | No | SD-LOS0, CD-LOS0 |
| IEEEENL-S-21 [58] | Video | Viola Jones | No | NAS-3DCNN | Yes | No | Yes | SD-LOS0, CD-LOS0 |
| Neu.Co-21 [59] | Video | - | Yes | O.F | 3D-CNN | No | No | No | SD-LOS0, CD-LOS0 |
| Sig. Proc. [60] | Onset, Apex | - | Yes | O.F | 3D-CNN+GA | No | No | No | SD-LOS0, CD-LOS0 |
| PR-22 [61] | Apex | LibFace, TV-L1 | Yes | O.F | Transfer-Learning | No | No | No | SD-LOS0, CD-LOS0 |

Here, Prepr. E-to-E, P.T. weights, N/W On-A-Off, GA, and EVM represents preprocessing, end-to-end, pre-trained weights, network Onset-Apex Offset, Genetic Algorithm, and eulerian motion video motion.
proposed a three streams-based CNN network consisting of a static-spatial stream, local-spatial stream, and dynamic- temporal stream to capture three different clues in three different frames. Furthermore, spatial features are concatenated and fed to the single LSTM to learn the temporal features. Khor et al. [36] introduced a CNN model with two streams, channel-wise stacking for spatial enrichment and feature wise stacking for temporal enrichment. Liong et al. [39] designed a shallow three-stream network to learn the optical flows guided features. Similarly, to use the optical flow-guided features, Liu et al. [37] introduced two-stream networks for MER. Yang et al. [38] exploited the feature discriminative capability of the VGGNet-16 to capture the spatial feature of MEs in three streams: 1) ME sequences; 2) optical flow; and 3) optical strain. Liu et al. [37] introduced a five-stream network with capsuleNet to improve the performance of MER. Also, a detailed comparison of the performances of multistream networks is tabulated in Tables IV and V.

Discussion: In literature multistream networks achieve high performance in MER over sequential 2D-CNN networks. The multistream networks can capture the diverse range of features from different streams and boost the efficiency of the network. Moreover, the multistream networks benefited MER frameworks to learn enough features with shallow networks and small data samples. From Tables IV and V it is evident that multistream networks outperform the sequential networks in terms of performance.

3) Multiscale Networks: Multiscale feature representations have been successfully used in two-stage MER [12], [30]. Zhou et al. [12] proposed a dual-inception network that operates in three scales (1 × 1, 3 × 3, and 5 × 5) for feature encoding of MEs. Verma et al. [29] introduced a multiscale-based lateral assertive hybrid network to capture the micro-level features of an expression in the facial regions. Zhai et al. [62] extended the LearNet approach and introduced the displacement-generating module-based MER (DGMER) framework. Song et al. [33] encoded spatio-temporal features by employing 5 × 5 and 3 × 3 sized filters in a consecutive manner. Verma et al. [31] improved the robustness of MER with hybrid (fusion of 3 × 3 and 5 × 5) local receptive feature blocks. Furthermore, Verma et al. [30] designed the AffectiveNet by incorporating MICRoFeat block to conserve the scale-invariant features with 3 × 3, 5 × 5, 7 × 7, and 11 × 11 sized convolution (conv) filters.

Discussion: Similar to multistream MER frameworks, multiscale MER frameworks also achieve impressive performance. The multiscale convolution layers guide the network toward both minute and abstract level features, which are significant to describe the distinctive features of different micro expressions. The more detailed analysis based on literature results have been included in Section II-D and Tables IV and V.

4) Capsule Networks: CNNs have shown impressive performance in literature. However, CNN models are computationally expensive and need a lot of data to train a model for specific-domain tasks. Moreover, the CNN model can pay attention to the translation in variance but failed to learn the rotation in variance. To resolve these issues, Sabour et al. [63] introduced the concept of Capsule. Capsule is a group of neurons to maintain the part-whole relationship and handle the viewpoint in variance. Some two-stage MER approaches [37], [48] have exploited the capability of Capsule networks. Very first, Van Quang et al. [48] used the Capsule networks along with ResNet 18 and secured 4th position in the MEGC-2019 challenge). Liu et al. [37] utilized the Capsule module with a multistream CNN for MER.

Discussion: The capsule-based networks can handle both translation and rotation variations and design a more robust MER network as compared to CNNs. Moreover, capsule networks facilitate the more concrete features represented, which can be interpreted to understand the behavior of the network (how the network is learning the MEs’ features). However, the capsule-based MER approaches are not widely notable due to complex nature and computational cost, though they have shown great promise. The capsule-based networks are still evolving and there is lots of scope for the researchers to create better and faster architectures so that it will be the baseline for solving any expression (MaEs or MEs) classification problem.

5) Recurrent Convolutional Networks: Recurrent convolutional networks (RCNs) enable every unit to incorporate context information in an arbitrarily large region in the current conv layer and allows learning microlevel edge variations. Xia et al. [10], [13], [49] exploited the RCN by following two-stage architecture to learn the representation of subtle facial movements from image sequences. In these studies, the recurrent connection within the feed-forwarded conv layers are employed to learn the temporal variations of image sequences extracted by multiple-scale receptive fields.

Discussion: CNN’s were inspired by early findings in the study of biological vision and share properties of the visual system of the brain. One notable distinction is that CNN is often a feed-forward design, whereas the visual system of the brain is abundant with recurrent connections. Thus, RCN-based architectures are benefited with more microbiologically realistic than their feed-forward counterparts. In addition, the activities of RCN units evolve over time as the activity of each unit is modulated by the activities of its neighboring units, which allows the network to learn distinctive edge variations with temporal information in MEs under challenging conditions over CNN as reported in Tables III–VI.

6) 3D-CNN Models: Most of the existing MER approaches [11], [45], [48] rely only on the apex frame/single instance for the analysis of MEs through 2D-CNNs. However, some studies emphasize the importance of dynamic aspects for detecting the subtle changes [64] and its effect on the performance of MER. In MEs video, each frame has its own viewpoint in variance. Whereas some other CNN models [36], [38], [50], exploit the capability of 2-D CNN and LSTM/RNN to elicit the spatial and temporal features, respectively. However, these models are not capable of extracting joint features of spatial and temporal variations, simultaneously [47]. Therefore to overcome the above issues, recently, some of MER approaches [39], [46], [59], [60] have taken advantage of the 3D-CNN network to capture, both spatial and temporal features simultaneously by adopting two-stage architecture.
Discussion: The 3-D convolutional layers are reasonable to learn the spatio-temporal information simultaneously. However, 3D-CNNs do not gain much attention in MER as it holds huge parameters and requires more computation power as compared to others, such as 2D-CNN, CNN-LSTM, RCN, etc. Also, deciding the number of hyper parameters, such as layers, 3-D down-sampling, and number of filters, in the network is a challenging task.

B. End-to-End Frameworks
An end-to-end model means that the CNN model takes the raw data as input and gives the final response without

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
any aid of external modules or blocks [see supplementary Fig. 5(a)]. Based on characteristics of the DL models, end-to-end frameworks are also categorized as: multiscale, CNN-LSTM, 3D-CNN, Graph-based, and NAS-based MER.

1) Multiscale Networks: As discussed in Section II-A3, multiscale networks have shown great performance in two-stage networks. Similarly, to exploit the capability of multiscale conv layers, Wang et al. [50] proposed an end-to-end micro-attention for MER by utilizing two scales $1 \times 1$ and $3 \times 3$ to encode the micro expressive features.

Discussion: The multiscale convolution layers also achieve impressive performance in case of end-to-end MER models. MEs are holding very sensitive and subtle information, thereby extracting the minute muscle changes within coarse facial features. It is important to learn minute as well as abstract features of the facial appearance and multiscale convolution layers learn the distinctive features of true emotions. Thus, multiscale convolutional layers can be embedded in any type of DL MER models to extract the disparities between different micro expressions.

2) CNN-LSTM Models: In CNN-LSTM networks, CNN is used to extract the spatial features and LSTM is included to learn the time-scale dependent information that resides along with the frame sequences. First, Kim et al. [40] used 2D-CNN followed by LSTM to encode the spatial and temporal features in MEs videos. Similarly, other work [36], [38], [65] also utilized the combination of CNN and LSTM-based conv networks to design end-to-end MER frameworks. CNN is employed to encode the MEs frames into spatial feature vectors and then MEs classes are predicted by passing the resultant features through the LSTM module. Choi and Song [56] introduced an integrated framework of CNN and LSTM for landmark feature map-based MER.

Discussion: The CNN-LSTM-based MER framework allows to capture and classify the spatio-temporal features of MEs in an end-to-end manner. However, CNN-LSTM networks first learn the spatial features and then temporal features. Therefore, sometimes these networks fail to correlate the spatial and temporal features simultaneously and fail to achieve good performance. Thus, 3D-CNN-based models are introduced to resolve the issue of CNN-LSTM and learn the spatio-temporal features simultaneously. The detailed insights of models are discussed in Section II-D.

3) 3D-CNN Models: As discussed in Section II-A6, cascaded MER architectures, such as CNN with LSTM or RNN are not capable of extracting joint features of spatial and temporal variations, simultaneously [47]. The capability of 3D-CNN to describe the spatio-temporal features for MEs in an end-to-end manner was first presented in [47]. Xie et al. [53] adopted the 3-D ConvNet-based Pseudo-3D to design a light weighted end-to-end architecture. Furthermore, AU node features are extracted and processed through the AU graph relation learning module to describe the emotion classes of MEs.

4) Graph-Based CNN Models: Recently, the graph-based end-to-end CNN approaches have achieved attention and attracted the researchers [53], [55], [66], [67] in the field of MER. Lei et al. [55] exploited the capability of landmarks and proposed a graph temporal CNN (Graph-TCN) to capture the local muscle movements of the MEs. The Graph-TCN method consists of two streams: 1) node and 2) edge feature extraction streams. Finally, both node and edge features are merged to classify the emotion label of MEs. Xie et al. [53] used the AUs relation graph to learn the subtle facial muscle movements for MER.

Each expression originated due to facial muscle movements and divided the face into small regions named AUs to represent the affective expression regions, defined in facial action coding system (FACS) [68] and ME training tool (METT) [69]. Thus, AUs are frequently used to describe how emotions are physically expressed. Most of the graph-based DL models incorporate the AUs relationship. Lo et al. [70] introduced MER-GCN, an AU-oriented MER architecture based on graph convolutional network (GCN) [71], where GCN layers are able to explore the dependency laying between AU nodes for MER. Similarly, Xie et al. [53] and Lei et al. [67] utilized the GCN to discover the AUs relationship. Lei et al. [55] utilized the graph structure for node and edge feature extraction.

Discussion: The graph-based models are growing gradually in MER. The new affective computing researchers have a lot of scope in the graph-based FER/MER approaches. The graph-based models are complex to understand but need very less computation cost with great efficiency, which is the future demand to work with memory limited or hand-held devices.

5) GANs-Based MER: With the advancement in learning-based methods, the performance of the MER is improved but still limited due to the lack of large-scale training data, computation, and design expertise. Some recent works [15], [53], [54] have been focused on these issues and provide solutions by exploiting the power of GANs. Yu et al. [15] introduced a capsule-enhanced GAN to generate the synthetic MEs with identity-aware faces to increase the data samples for
better training. Xie et al. [53] came up with the AU intensity controller GAN to produce synthetic data for resolving the problem of limited and biased data samples.

6) NAS-Based MER: The robust DL models are designed manually based on trial-and-error engineering and need expert knowledge, further this is very time consuming and requires high-level domain expertise in designing CNN networks. Very first, Verma et al. [58] focused on these factors and came up with a conclusion that instead of spending time and effort designing the best possible CNN, in the hope of improved performance, it is prudent to design algorithms to search for the best CNN model for MER.

Discussion: NAS algorithms are able to search and automatically design the best optimal CNN model with minimum human intervention. Initially, NAS-based algorithms [72], [73] required huge computation and took many days to train a task-specific CNN model. However, recent NAS-based approaches [74] focus on the faster searching and training architectures. Specifically, in MEs, NAS-based algorithms need extra efforts to design MEs feature adaptive inner as well as outer architecture search. The MER field is still far from automatic model designing and has a lot of scope to develop better and faster NAS-based MER algorithms. Also, selecting robust operation and number of cells for MER application is one of the prominent steps for the performance improvements in MER.

C. Transfer Learning-Based Networks

All available data sets for MER are relatively smaller as compared to other computer vision tasks. However, it is a well-known fact that the direct training of deep networks from scratch over smaller data sets is prone to overfitting. To mitigate the effect of overfitting, many studies have taken advantage of pretrained weights (fine tuning) of well-known models: AlexNet [75], VGG [76], ResNet [77], etc., which are trained over large-scale data sets, such as ImageNet, FaceNet, etc. However, Patel et al. [41] explored that all pretrained weights are not fitting well to discriminate against the MEs due to the low intensity of facial movements. Although, features of MaEs and MEs share some feature similarities in facial texture and muscle movements. Therefore, to exploit the capability of pretrained features by considering domain adaption, Patel et al. [41] retrained ImageNet weights for macro data sets: CK+ and SPOS to train the model for expressive features. Next, resultant features are fine-tuned over micro expression data sets. On the same hypothesis, in MEGC-18 [78] and MEGC-19 challenge [79], researchers [42], [45], [50] utilized the pretrained weights of ResNet and its variants [77], trained over ImageNet and further to learn the expression specific features model retrained over MaE data sets: CK+, OULU, JAFFE, and MUG. Further, by adopting the domain features of the expressions from MaEs the CNN models are fine-tuned for ME tasks. The pertained weights of ResNet-18 have been widely used in [42], [45], and [54] due to their feature learning capability. Further, Wang et al. [65] designed a CNN network: transferring long-term convolutional NNs (TLCNNs) by combining 2D-CNN and LSTM. First, TLCNN is trained over MaE large-scale data sets: KDEF, MMI, and TFID to acquire the knowledge of emotion-specific features. Afterward, TLCNN weights are used to retrain the model for MER. Khor et al. [36] fine-tuned the VGG-16 with VGG-face weights to capture the enriched spatial feature of MEs. Moreover, Yang et al. [38] introduced an MERTA network by combining VGG-16 and LSTM to exploit the spatio-temporal features of ME sequences and their respective optical flow and strain. Li et al. [11] have benefited from pretrained weights of VGG Face to guide the deep CNN network for smaller MEs data sets. Besides fine-tuning and domain adaption, knowledge distillation is another effective transfer learning approach. To provide a compact solution for training data and computation, Sun et al. [33] proposed knowledge distillation to transfer knowledge of AUs to MER through teacher–student CNN learning. The main aim of the framework is to guide the shallow student network by transferring the knowledge of features from a pretrained deep teacher network.

Discussion: The pretrained weights or transfer learning is a sure-fire concept to solve the problem of overfitting and speed up the learning process with smaller-sized data sets. The well-trained models: VGG-16, ResNet, GoogleNet, etc., benefited the MER approaches to reduce the problem of overfitting up-to some extent. However, most of these models are trained over ImageNet data set, which has contrast data samples related to MEs with low muscle intensity, subtle and rapid changes. Therefore, pretrained weights of ImageNet are not suitable for MER. Whereas pretrained weights over face images or macro data sets are more advisable as these data sets hold analogous features in terms of facial structure and shape. Moreover, pretrained weights can be utilized in both multistage and end-to-end MER networks. More detailed technical characteristics are discussed in Section II-D.

1) End-to-End Versus Multistage Frameworks: The multistage framework, especially two stage network, a handcrafted feature descriptor, such as dynamic imaging [29], affective motion imaging [30], optical flows [39], [46], [60], etc., is used in the first stage to capture the primary features. While, in the second stage, the CNN network is used to learn stage-1 features. For example, Verma et al. [29], [30] utilized the dynamic and affective motion imaging to capture the spatiotemporal features into a single instance and then CNN network is designed to learn the MEs features [see supplementary Figs. 2(a) and 3(b)]. Therefore, the performance of the CNN models is also dependent on the handcrafted feature descriptors. Moreover, the two-stage networks require auxiliary computation to compute the handcrafted features and inconvenience in real-time applications. Whereas, end-to-end MER architectures process the data into a single shot and advisable for real-time applications. The model categorization in terms of end-to-end and two-stage is represented in Table II. Moreover, efficacy of end-to-end and two-stage networks over different experimental settings: number of classes, input formats, validation protocols, learning rates, data augmentation, etc. is reported in Tables III–VI.

D. Discussion on Technical Characteristics of DL-Based MER

This section represents the deep insights of technical characteristics in various decisive aspects, such as down sampling,
multi-scale/stream CNNs, shallow versus deeper CNNs, and effects of kernel sizes in convolutional layers of the DL-based MER approaches.

1) Impact of Downsampling With Convolution and Pooling: Down sampling plays a significant role to reduce the number of parameters and ensures higher computational speed in CNN framework. In the case of MER, majorly, max pooling, and convolution with stride operations were used for dimensionality reduction. The max-pooling layer captures the high-level edge information by applying the max operation. However, max pooling operations tend to focus on high-level stack information by ignoring minors, which play a key role in the recognition of micro level features. However, the convolution adds to the interfeature dependencies and reduces the dimension by parameter learning among channels rather than fixing it. The impact of convolution with strides over max pooling operations on MER data sets are shown in Fig. 3. Fig. 3(a) represents that the convolution with stride gains significant improvement in accuracy as compared to pooling over CASME-I, CASME-II, SMIC, CAS(ME)², and SAMM data sets. Moreover, Fig. 3(b) depicts the visual effects of the max pooling over convolution with stride. From Fig. 3(b), it is clearly visible that the response map of pooling loses more information as compared to convolution. The more details about the technical differences with qualitative and quantitative measures are discussed in [29], [30], and [80]. From Fig. 3 and [29], [30], [80] it is clear that convolution with stride outperforms the max pooling operation in MER.

2) Impact of Multi-Scale/Stream and Sequential CNN Frameworks on MER: Multi-scale/stream feature representations have been successfully used in MER and achieve good performance in [12], [33], and [38]. From [12], [33], and [38], it is evident that linearly coupled conv layers with identical filter size (VGG-16, VGG-19, and ResNet) have failed to capture heterogeneous scaled receptive fields and avoid fine-tuned edge variations. However, multi-scale/stream CNN models could capture detailed features from small to extensive regions, by applying multi-conv layers with different scale filters. The qualitative comparison between single-scale/stream and multi-scale/stream conv layers are depicted in Fig. 4. From Fig. 4(a), it is quite clear that CNN-based models built on single branch linearly connected conv layers, lack in gathering adequate features of facial appearance due to repetitive cross-correlation operation. Whereas, Fig. 4(b) represents the capability of multi-scale/stream conv layers in learning of significant discriminable features from the expressive regions of the MEs. Moreover, the quantitative results for models [11], [29], [31], [33] on different data sets: CASME-I, CASME-II, CAS(ME)², SAMM, and SMIC, are represented in Fig. 5. The quantitative results have also proven a higher generalization capabilities of the multi-scale/stream over single-scale/linear CNN frameworks. Based on both qualitative and quantitative results analysis, we can conclude that multi-scale/stream CNNs acquired more MEs features and outperformed the single stream/linear CNN frameworks. The existing multi-scale/stream models are detailed in Table II and corresponding results for different data sets over single-domain leave one subject out, composite-domain leave one subject out and cross-domain/data set validation protocols are tabulated in Tables III–VI.

3) Does Deeper Network Affect the Performance of the MER?: Yes, in general, the CNN model requires a large amount of data samples for efficient training. However, publicly available data sets for MER consist of limited data samples and tend to cause over-fitting. Moreover, deep/dense networks, such as AlexNet, VGG-11, VGG-16, SqueezeNet, GoogleNet, ResNet-18, etc., also failed to capture minute features but were liable in emotion classification. Deep/dense
networks [39], [84] may vanish the micro-level features of the expressive regions due to progressive convolution and pooling operation. Therefore, most state-of-the-art MER approaches [34], [35], [39], [48] adopt shallow and light weighted CNN models for MER. The quantitative results of the literature study are analyzed in Fig. 6. In Fig. 6, we can observe that, deep/dense networks: AlexNet, VGG-11, VGG-16, SqueezeNet, GoogleNet, and ResNet-18 have failed to achieve superior performance as compared to shallower networks: OffApex, STSTNet, CLF, and MTC. Based on results, we can conclude that shallow networks are preferable and achieve high performance in terms of accuracy as well as F1-score as compared to deeper networks. Moreover, deeper networks are computationally expensive as compared to shallow networks.

4) Does Kernel Sizes Have Any Impact on CNN Layers?: Yes, the kernel sizes also come under important paradigms of the CNN model designing. Kernel sizes directly affect the performance of the model along with computation cost. The comparative analysis between different sized kernels on data sets: CASME-I and CASME-II for MER frameworks [30], [31] are demonstrated in Fig. 7. Moreover, the qualitative effect of various kernel sizes is depicted in Fig. 4. Based on literature study [31], it is clear that kernel sizes 3 × 3 and 5 × 5 are more capable to define the MEs features and achieve higher performance in MER. From the observations the smaller kernel sizes are preferable for MER applications. Kernels with large scales (7 × 7 and 11 × 11) have a larger receptive field per layer and allow the extraction of generic features spread across the image. Therefore, these filters focus on abstract transitional information and skip the minute information, which is quite important in MER.

III. TRAINING AND EVALUATION STRATEGIES

The performance of any framework is affected by two factors: 1) the model uncertainty in architecture design and 2) the evaluation strategy, i.e., the strictness in the data division to validate the generalization strength of the framework. The technical aspects of designing paradigms are already discussed in the previous section (learning-based MER models and studying technical characteristics). In this section, first we focus on the data sets and evaluation metrics (Section III-A). Further, we discussed the validation strategies available in the literature to validate the robustness of the MER frameworks (Section III-C). Moreover, we have studied the different paradigms of experimental strategies in detail and observations are provided in Section III-D.

A. Data Sets and Evaluation Metrics

The current research needs across all the computer vision applications are motivated by the success of DL algorithms.
The success of the DL algorithms highly depends on the availability of sufficient training data with variations of the populations and environments as much as possible. The higher the diversity in the present training data, the more robustly one can estimate the model parameters. In this section, we primarily discuss the publicly available MEs data sets (highlighted by purple color in Fig. 1) that have been used for evaluating the MER methods. In the literature, ME data sets can be broadly classified into a data set in lab environment and a data set in-wild.

1) Traditional Data Sets (Lab Environment): The six traditional data sets: 1) CASME-I; 2) CASME-II; 3) CAS(ME)²; 4) SMIC; 5) SAMM; and 6) MMEW have been widely used in the literature for MER evaluation. Among them, the CASME-II data set is most prominently used in literature. The CASME-II data set holds two sets: Part 1-247 and Part 2-255 with five and eight emotion classes, respectively. The CASME-I elicited 195 image sequences of 35 participants (22 males and 13 females) with eight emotions. Moreover, CAS(ME)² data set is prepared to spot and recognize MEs in long videos. The CAS(ME)² data set contains two parts: part A and part B. Part A includes 87 long video sequences of macro and micro expressions which are used for MEs spotting tasks. While part B included 300 macro and 57 micro expressions, which are used to evaluate MER tasks. The CAS(ME)² data set samples were collected from 22 participants (9 males and 13 females). All the samples of CAS(ME)² data set are annotated by using AUs and four emotion labels: Positive (8), Negative (21), Surprise (9), and Others (19). Moreover, to provide more challenging scenarios, SMIC has been introduced with three different illumination conditions: high speed (HS), visual (VIS), and near infrared (NIR), including 164, 71, and 71 samples of 16, 8, and 8 subjects, respectively. The SMIC data set is the second most utilized data set to validate the performance of MER models. All the above data sets involve either one or three ethnic participants. Therefore, to include the challenge of diverse ethnicity SAMM data set has been furnished with 32 participants having 13 different ethnicities. The SAMM data set comprises 159 video samples annotated with seven emotion classes: Happiness (24), Surprise (13), Anger (20), Disgust (8), Sadness (3), Fear (7), and Others (84). Recently, an MMEW [9] data set with the largest pool of MEs was introduced. The MMEW data set contains 300 macro and micro image sequences of 36 participants. The MMEW data set is annotated with FACS and seven emotion classes: Happiness (36), Anger (8), Surprise (89), Disgust (72), Fear (16), Sadness (13), and Others (102).

2) Wild Data Set: All traditional data sets samples are elicited in a lab-controlled environment and lack the detail and divinity of real-life challenges. The MEVIEW data set [85] is the first MEs data set that incorporated wild environment challenges: occlusion, illumination variations, candid faces, etc. The video samples are collected from the websites. Specifically, all the samples are downloaded from YouTube videos of poker games. The database contains 31 samples from 16 subjects, having both macro and micro expressions. More detailed analysis of data sets can be found in [85].

B. Evaluation Metrics and Significance

The widely used evaluation metrics for MER are recognition accuracy (Acc.), weighted f1-score (F1), weighted average recall (WAR), Un-weighted f1-score (UF1), UAR, and mean diagonal value of the confusion matrix. The accuracy is calculated by computing the average hit rate across the all emotion class samples. Let TP, TN, FP, and FN be the true positive, true negative, false positive, and false negative, respectively. The recognition accuracy is computed by using

\[
\text{Recognition accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{1}
\]

The accuracy can be largely contributed by a large number of true negatives and not focus on false negative and false positive. Thus, Acc. is liable to bias data and reflect partial effectiveness of the MER frameworks. Whereas, F1 score is a better measure to balance between TP, TN, FP, represents the total number and FN. The F1-score is calculated by using

\[
F_1 \text{ Score} = \frac{\sum_{i=1}^{C} T_i}{T} \times \frac{2 \times \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i}}{\sum_{i=1}^{C} \frac{2 \times \text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i}} \tag{2}
\]

where \(T\) and \(C\) represent the total number of samples and emotion classes, respectively. F1 Score is considered as a better measure because of its balancing nature with uneven class distribution (large number of actual negatives).

Though MER data sets have heavy imbalanced annotations, both Acc and F1 score failed to justify the efficacy of the MER models. Recently, UAR and UF1 have drawn much attention due to their unbiased nature of evaluation. Both UAR and UF1 computed the performance of a model with respect to a number of classes without consideration of samples per class. The UAR and UF1 are calculated by using (3) and (4), respectively

\[
\text{UAR} = \frac{1}{C} \sum_{i=1}^{C} \frac{\text{TP}_i}{T_i} \tag{3}
\]

\[
\text{UF1} = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \times \text{TP}_i}{(2 \times \text{TP}_i) + \text{FP}_i + \text{FN}_i}. \tag{4}
\]

Moreover, some of the methods [42] utilized the WAR for evaluation. The WAR is computed by using

\[
\text{WAR} = \sum_{i=1}^{C} \frac{\text{TP}_i}{T}. \tag{5}
\]

Some of the MER frameworks [11], [43], [48], [56] also adopt the mean diagonal value of the confusion matrix to show the detailed generalization of the model.

C. Validation Strategies/Protocols

In literature, many state-of-the-art techniques adopted diverse data-division strategies to prove the robustness of the algorithm. Therefore, the supervised techniques we broadly categorize into person-dependent evaluation (PDE), person-independent evaluation (PIE), and cross domain evaluation (CDE). In “Person-Dependent setup” the training and testing set contains frames from the same category, whereas in Person-independent setup’ completely unseen data is used for testing. Similarly, in CDE train the model with one data set and test
Fig. 8. Differences between person dependent, person independent, and cross-domain validation setups.

Fig. 9. Evolution in validation strategies in past years.

the model accuracy with a completely different data set. The sample technical differences between PDE, PIE, and CDE are depicted in Fig. 8. Also, the impact of data-division strategies over the model’s performances are tabulated in Tables II–VI. More details of the data-division strategies are discussed as follows.

1) Person-Dependent Evaluation: Based on the literature study, PDE setup can be further divided into two categories 1) \(k\)-fold cross validation and 2) leave one video out (LOVO). In \(k\)-fold cross validation data is randomly divided into a ratio of 100-P:P. Where, in each iteration P\% of data is used for inference to validate the performance of the model and remaining data samples are reserved for the training. Some studies [29], [47] follow the PDE setup with 80:20 ratio. Recently, Thuseethan et al. [57] used a 90:10 ratio with tenfold cross-validation to validate the effectiveness of the framework. In LOVO, one expression video of a person is used for inference, and remaining all data samples are used for training. Thus, there are immense chances to use the same person’s expression in both training and testing data. The articles [10], [21], [23] in literature adopted LOVO setup to prove the robustness of their models.

2) Person-Independent Evaluation: The PIE setup follows a strict division for training and testing sets. It ensures the evaluation over videos with unseen subject’s identities. Like PDE, the PIE setup can also be further categorized in single-domain leave-one-subject-out (SD-LOSO) and composite-domain leave-one-subject-out (CD-LOSO). In SD-LOSO, a single data set is used, and samples are partitioned in such a way that all expressions of a particular subject at a particular iteration act as a testing set, and the remaining data is considered as the training set. The SD-LOSO is the most widely used validation strategy in [10], [13], [48], [52], and [53]. The list of models with SD-LOSO is tabulated in Table II. The CD-LOSO is one of the recently emerging validation setups to validate the robustness of the model with domain shifts. First MEGC 2018 [78] introduced the CD-LOSO setup for the MEs challenge. The CD-LOSO data division was introduced by combining CASME-II, SAMM, and SMIC with three emotion classes: 1) positive; 2) negative; and 3) surprise. Further, the composite data set is used by adopting an LOSO set up to evaluate the robustness and generalization of the model for MER with domain shifts. Based on literature study [10], [12], [36], [42], [45], [48], [54] CD-LOSO is a highly recommended validation strategy as it ensures evaluation on unseen faces with ethnicity variations. It is observed from the literature, recent works on MER adopted the CD-LOSO along with SD-LOSO to validate the performance of the MER frameworks as shown in Fig. 9.

3) Cross Domain Evaluation: CDE is another setup that ensures PIE by training a model over a particular data set and testing on a different data set. The MEGC-2018 [78] used the CDE setup to evaluate the efficacy of the submitted MER frameworks. Peng et al. [42] and Khor et al. [36] have successfully proven the performance of the models on the CDE setup of the MEGC-18 challenge. Wang et al. [50] performed two cross-domain experiments: CASME-II data set is used for training and testing results are evaluated on the SAMM data set and vice versa. Choi and Song [56] have utilized the pair of CASME-II and SMIC data set for CDE. First, CASME-II is considered for training, while SMIC is used for inference. Further, SMIC is used for training, and CASME-II is reserved for testing. Moreover, Verma et al. [30] have conducted nine experiments over four data sets: 1) CASME-I; 2) CASME-II; 3) CAS(ME)\(^2\); and 4) SAMM. Three experiments were conducted by using CASME-I as a training data set and CASME-II, CAS(ME)\(^2\), SAMM as testing data sets, individually. Another three experiments were evaluated by using CASME-II as a training data set and CASME-I, CAS(ME)\(^2\), SAMM as testing data sets, individually. Similarly, the other three experiments were performed by using SAMM as a training data set and the remaining three as testing data sets. Based on the literature, CDE setups are very less popular but have the most recommendable data validation strategies to gain much generalization capabilities of the FER models.
D. Discussion on Experimental Setups and Validation Strategies

In the literature, we noticed that the standard evaluation protocols for MER techniques are not available as many authors follow contrast experimental setups to prove the robustness of the MER techniques. The term contrast is defined in terms of a number of samples used for training and testing, input selection strategies, the number of expression classes adopted in training or dropped some of the emotion classes due to a smaller number of images [10], [30], type of validation strategy adopted to prove the robustness, etc. Based on the above contrast settings, it is harder to compare the performance of these techniques directly. Therefore, this section aims to provide a detailed study on experimental setups adopted in the literature. Also, analyzed the effect of adopting different evaluation setups, such as PDE, PIE, and CDE, data augmentation, input selection, number of emotion classes, etc., over the performance of the MER approaches.

1) Impact of PDE, PIE, and CDE on MER Performance: In PDE, the training and testing set contains frames from the same category. Thus, there is a possibility of similar samples of the same subject present on both training and testing data sets, which leads to inflated performance during testing. However, it may fail in real-world scenarios. Thus, there is a need to evaluate the model performance over unseen or person-independent scenarios. This also makes the process of model design much more challenging to ensure robust performance even in real-world scenarios. Therefore, the PIE ensures stable performance over PDE on unseen data.

In literature many existing works [10], [12], [36], [42], [45], [48], [54], [58] have opted the PIE setup’s evaluation (More detailed categorization is indexed in Table II). Further, to study the effect of PDE as compared to PIE setup, we accumulated some results for PDE (including both 10-folds and LOVO) and PIE (SD-LOSO) setups as shown in Fig. 10. More specifically, Fig. 10(a) illustrated the results analysis for the models [57], evaluated over tenfold and SD-LOSO validation strategies. Whereas, Fig. 10(b) shows the comparative results for model [44] evaluated over LOVO and SD-LOSO validation strategies. Based on Fig. 10, it is evident that the models that adopted PDE setups outperform in accuracy over the PDE setup. Thereby, the PDE results are unreliable to validate the actual robustness of the DL models. Similarly, the model’s performance depends on a person’s identity, the cross culture, and ethnicity variations.

Therefore, recent works [15], [35], [39], [60] also focused on the CD-LOSO validation strategy as it ensures person independence with sparse diversity in domains. However, it is hard to achieve impressive performance with the strictness of CD-LOSO. As we can see in Fig. 9, the best accuracy results for CD-LOSO are 76.9%, which has enough margin as compared to the results of PDE and SD-LOSO. Furthermore, some work [30], [36], [42], [50] also adopted the CDE validation strategy to evaluate the models with more challenging scenarios such as cross-ethnicities, out-group, illumination, and resolution variations. Therefore, the CDE validation protocol evaluates the model’s robustness more toward real-world scenarios. From Fig. 9, the best accuracy over CDE setup is 62.8%, which is very less as compared to PDE as well as PIE validation setups. Therefore, benchmarking the performances of different CNN, GAN models in a standard evaluation setup (PDE/PIE/CDE) is an important scope in micro expression research.

2) Impact of Data Augmentation on MER Performance: Since MEs data sets consist of a limited number of samples and imbalanced classes, which leads to model overfit. Many existing MER approaches [11], [13], [30], [31], [40], [42], [48], [55], [57], [65], [86] adopt the data augmentation techniques to create a sufficient pool of data samples for training. 2D-CNN-based approaches [30], [31], [40], [42], [48], [55], [57], [65], [86] performed the basic operations, such as flipping, rotating, color shift, and smoothing to increase the data samples. Xia et al. [13] introduced the two new augmentation techniques with temporal connectivity. While the recent work [15], [53] utilized the GAN-based model to generate the synthetic MEs data for data augmentation. In [15] and [53], the patchGAN network is used to generate synthetic images/samples using the apex frame. The quality of the synthetic (fake/augmented) images are validated by using, i.e., adversarial loss, consistency loss, attention loss, AU intensity loss, SSIM loss, ME loss, and sequence authenticity loss. Similarly, Xie et al. [53] utilized a 3DConvNet architecture to distinguish the synthetic sequences from the real ones. Li et al. [11] collected the five nearest frames to the apex frame to enhance the size of data sets. From [11], [13], [30], [31], [40], [42], [48], [53], [55], [57], [65], and [86], we observed that data-augmentation resolves the issue of overfitting up to some extent and improves the performance of the model as shown in Fig. 11(a) and (b). Fig. 11(a) depicted the results of STRCN-A and STRCN-G models over CASME-II, SMIC, and SAMM data sets with and without data augmentation (based on temporal connectivity) in terms of recognition accuracy. Fig. 11(b) represents the results of four other MER models: STCN, CapsuleNet, MER-GCN, and AU-GACN models over CASME-II and SAMM data sets with and without data augmentation (based on GANs) in terms of recognition accuracy. From the analysis of the results, it is clear that the models trained over augmented data sets achieved higher accuracy as compared to models without augmented data sets. However, some of the MER approaches [32], [34], [36], [39], [46] significantly improves the models performance.
without augmentation by introducing optical flow and shallow network designing. Therefore, it is observed that there is scope to design robust models without increasing the sample size.

3) Impact of Input Selection Strategies on MER Performance: Most of the existing MER approaches follow three types of input formats: 1) apex frame [39], [48], [55]; 2) onset-apex-offset frames [15], [52]; and 3) whole videos [46], [47] as shown in Fig. 2. Majority of the existing MER approaches [11], [39], [45] rely only on the apex frame for the analysis. However, some studies emphasize the importance of dynamic aspects for detecting the subtle changes [64] and its effect on the performance of MER. In an MEs video, each frame has its own significance toward the identification of the emotion class. Therefore, apex frame-based approaches are lacking to analyze the motion information, which has its own potential to describe the MEs classes. Therefore, whole video input is more effective and reliable in MER. Nevertheless, the whole video input is incredibly challenging to handle and achieves less results as compared to apex frames input as shown in Fig. 12. In Fig. 12, we included top three accuracy results achieved by different models over CASME-II and SAMM data sets with apex and whole video inputs, respectively. From Fig. 12, it is quite clear that models with whole video input formats acquire less accuracy results as compared to apex frames but have more reliability to the capability of delivering appearance information along with time variants. Therefore, utilizing complete frames in a video is more effective and reliable in MER than using a single apex frame.

4) Do the Number of Emotion Classes Affect the MER Performance?: The number of emotion classes plays a key role in the estimation of the MER’s performance. In literature, there is no standard for the emotion class settings. Therefore, it is difficult to compare the performance of the MER models directly. In literature, work has been done using 03 emotion classes (P, N, S), 04 emotion classes (P, N, S, O), 05 emotion classes (H, D, S, R, T), and 7 or 8 emotion classes for training the MER model. The list of several emotion classes adopted to test the MER model performance is tabulated in Tables II and III.

To analyze the impact of emotion classes over the performance of a model, we analyzed the existing MER approaches and results are highlighted in Fig. 13. From Fig. 13, it is clear that the performance of the models is improved by reducing the number of emotion classes. More specifically, models for 3-emotion classes gained the highest recognition accuracy, and models for 4-emotion classes attained second highest and so on. Based on the observations and results of existing MER frameworks, we can conclude that a greater number of classes (7/8 emotion classes) create more confusion for models to generalize the emotion classes in true positives. However, in real life scenarios humans can exhibit a wide range of facial expressions. Thus, using a reduced set of emotions to carry on the experiments is a bit misleading. The fact that existing data sets lack enough labels for some emotions is a challenge that the research on the topic owes to face. Thereby, there is enough scope to provide a persuasive solution to handle such emotion classes instead of simply merging or dropping.

IV. RESEARCH NEEDS AND FUTURE DIRECTIONS

This section describes the research needs and future directions of the MER approaches. More details of the critical issues in the literature which are unresolved or get the least attention in the MER approaches are discussed. Therefore,
FR [61].

IM-CNN, CNN-LSTM, STRCN-A, STRCN-G, and FR are grab from the direct taken from the AU-GACNN [53] published results. While, results for STRCNN, ELRCN, CapsuleNet, MER-GCN, and AU-GACNN are Xie [87], the study on single cultured samples lead to in-lab environments with single-cultured subjects. According to B. Cross-Cultured ME Data Set

imbancing emotion class problems for MER. From the above discussion, it is evident that there is a need to generate the synthetic data samples to alleviate the problem set. Furthermore, Xie et al. [53] proposed a GAN-based model and proposed temporal augmentation to create a balanced data set. Meanwhile, results of IM-CNN, CNN-LSTM, STRCN-A, STRCN-G, and FR are grab from the FR [61].

there is scope for the upcoming researchers to design robust solutions for MER applications.

A. Unbiased Learning

The publicly available ME’s data sets have imbalanced sample sizes in the emotion classes. The imbalancing nature of the data set leads the CNN models bias toward a dominating class. Thus, there is an immense need to develop an unbiased learning algorithm to enhance the generalization ability of the model. To handle the imbalancing issue in the emotion classes, some of the existing MER approaches [46], [65] dropped the emotion classes, which contain very few samples. While some other approaches [44], [53] have created new emotion classes by merging the existing emotions as positive, negative, surprise, and other. However, combined emotion classes are still imbalanced. Therefore, Xia et al. [13] took a step forward to resolve the imbalanced data samples problem and proposed temporal augmentation to create a balanced data set. Furthermore, Xie et al. [53] proposed a GAN-based model to generate the synthetic data samples to alleviate the problem of unbalancing and trained a model with unbiased learning. From the above discussion, it is evident that there is a need to develop a balanced data set and robust DL technique to handle imbalancing emotion class problems for MER.

B. Cross-Cultured ME Data Set

Most of the available ME data sets are developed in-lab environments with single-cultured subjects. According to Xie [87], the study on single cultured samples lead to in-group downside problems. Thus, the model trained over single targeted ethnic participants may be biased and lead to underperform on cross-cultured MEs. Therefore, there is a need to collect cross-cultured data samples of MEs. The cross-cultured ME data sets will ensure the fair analysis of MER frameworks and it is a great addition to the affective computing research community.

C. Group Emotion Recognition

Identifying the common emotion, shared among a group of people is known as group-level emotion recognition (GER). The GER plays a significant role in a wide variety of applications, such as security, surveillance, early event prediction, image retrieval, and social era. Much study on GER using macro expression recognition is available in [89] and [90]. However, due to the challenges of MEs there is no group emotion data set available for MER. Thus, there is scope to develop balanced and cross-cultured group emotion micro expression data set.

D. Motion Magnification in MER

MEs are involuntary which cannot be captured in normal sight as they usually occur only for the minute interval. Therefore, it is hard to train a system to detect these variations and identify the relevant emotion class in MEs video sequences. In literature, Eulerian video magnification (EVM) [91] and learning-based MM [92] are adopted for magnifying the MEs. However, both approaches are not specifically designed for the MEs and sometimes destroy the emotional features. Therefore, there is a need to magnify the micro variations of emotion, developing application-dependent MM algorithms could improve the performance of the model.

E. Multimodal in MER

Ongoing MER research is not achieving enough performance to use in real-time applications due to the low intensity and subtle nature. Moreover, the available data set samples are not enough to train the MER frameworks. Therefore, to enhance the ability to recognize micro variations, in recent times multimodal algorithms [93], [94] gain the attention of researchers by supplying sufficient information to enhance the model performance. Thus, there is a need to develop an application-specific (MER) data set with multimodalities, such as body gestures, eye gaze, electrocardiogram (ECG), electroencephalogram (EEG), etc. Also, there is an immense need to design a robust algorithm to handle these multi modalities.

F. MER In-Wild

In real-world applications, to analyze the emotional state of a person, models designing in the lab environment may fail as the movement of the subject is dynamic, viewing angle of the camera is not static, illumination, and lighting conditions are dynamic in nature. Due to the challenging nature of micro expressions, in-lab data set in-wild for MER [85] is available in the literature. Also, extremely limited articles on MER in the wild are available in the literature. Therefore, there is a lot of scope to develop a robust algorithm to handle the challenges of the in-wild and scope to develop cross-cultured data set in the wild for MER.

G. MER in Psychological Disorders

Psychological disorders, such as autism spectrum, bipolar, anxiety, and stress-related disorders affect a person’s thoughts, behavior, feelings, and sense of well-being. In these disorders, people have a state of low mood and aversion to activity. In such a low mood and lack of interest, the facial expression appears different from the ones in normal states. Some
research [95, 96] have been made for psychological disorders through facial expression recognition in past years. However, incredible challenges in data set accumulation due to privacy of neurotic subjects, limited labeled data sets, and lack of DL-based solutions make it an appealing research area to be explored. The available data sets are not focusing on the privacy of neurotic subjects which is a very crucial aspect in social lives. The future work requires more attention toward both data collection and algorithm development for psychological disorders analysis through FER/MER.

H. MER in Entertainment
Online games have gained a lot of popularity due to the features of collaboration, communication, and interaction. However, the virtual world of gaming is still primitive and far from real-world communication. For example, players still communicate through text chats, avatars have no activities related to natural body gestures, facial expressions, and so forth. Therefore, there is a need for a robust automatic expression system that can be integrated into the gaming system to control the facial expressions of avatars and enhance the interest of players by providing a virtual interface near the real world. Moreover, facial expression analysis can be integrated into video-controlled devices for entertainment (music, movies, games, YouTube, etc.) and the dynamic balancing system will automatically adjust the entertainment level based on the user’s facial expressions.

I. MER in Education System
In recent COVID pandemic years, the standard offline education systems transferred to the online education system. Effective online education is the primary need to maintain the education gap of each age group of students. The online education system allows students from anywhere to access the classes as well as experimental work from a distance via the Internet. However, this is way far away from physical education. In such critical situations students as well as tutors face many challenges, such as lack of attention, high chances of distraction, etc. Resultant performance of the students degrades and it increases the chances of many mental problems such as stress, anxiety, and depression. Thus, there is a lot of scope to develop a robust online education system by integration of facial expression analysis to scan the expressive features of the students. This will allow tutors to survey the attentiveness of the students in class and help them accordingly.

V. CONCLUSION
This article presents deep insights of learning-based MER frameworks with a perspective on promises in model designing, experiment strategies, challenges, and research needs. Particularly, the existing learning-based MER frameworks are analyzed in terms of model design and evaluation frameworks. The variety of existing DL architectures is examined and their effect on MER performances are discussed. The important paradigms in model designing and evaluation for MER are presented. Also, the impact of data division strategies, such as PDE, PIE, and CDE on the model performance and the limitations of these strategies are discussed. The challenges in designing robust MER models and the current research needs are discussed. Further, the available data sets, challenges and the evaluation metrics utilized to test the efficacy of MER frameworks are discussed. In addition, presented useful insights of future needs and research guidance to carry forward the research in MER.

REFERENCES
[1] P. Ekman and W. V. Friesen, “Nonverbal leakage and clues to deception,” Psychiatry, vol. 32, no. 1, pp. 88–106, 1969.
[2] M. Takalkar, M. Xu, Q. Wu, and Z. Chaczk, “A survey: Facial micro-expression recognition,” Multimedia Tools Appl., vol. 77, no. 15, pp. 19301–19325, 2018.
[3] Z. Zhang, T. Chen, H. Meng, G. Liu, and X. Fu, “SMEConvNet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos,” IEEE Access, vol. 6, pp. 71143–71151, 2018.
[4] S.-T. Liong, J. See, K. Wong, A. C. Le Ngo, Y.-H. Oh, and R. Phan, “Automatic apex frame spotting in micro-expression database,” in Proc. 3rd IAPR Asian Conf. pattern Recognit. (ACPR), IEEE, 2015, pp. 665–669.
[5] T.-K. Tran, Q.-N. Vo, X. Hong, X. Li, and G. Zhao, “Micro-expression spotting: A new benchmark,” Neurocomputing, vol. 443, pp. 356–368, 2021.
[6] X. Li et al., “Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods,” IEEE Trans. Affect. Comput., vol. 9, no. 4, pp. 563–577, 2017.
[7] Y.-H. Oh, J. See, A. C. Le Ngo, R. C.-W. Phan, and V. M. Baskaran, “A survey of automatic facial micro-expression analysis: Databases, methods, and challenges,” Front. Psychol., vol. 9, p. 1128, 2018.
[8] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, “An overview of facial micro-expression analysis: Data, methodology and challenge,” arXiv preprint arXiv:2012.11307, 2020.
[9] X. Ben et al., “Video-based facial micro-expression analysis: A survey of datasets, features and algorithms,” IEEE Trans. Pattern Anal. Mach. Intell., 2021.
[10] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, and G. Zhao, “Revealing the invisible with model and data shrinking for composite-database micro-expression recognition,” IEEE Trans. Image Process., vol. 29, pp. 8590–8605, 2020.
[11] Y. Li, X. Huang, and G. Zhao, “Joint local and global information learning with single apex frame detection for micro-expression recognition,” IEEE Trans. Image Process., vol. 30, pp. 249–263, 2020.
[12] L. Zhou, Q. Mao, and L. Xue, “Dual-inception network for cross-database micro-expression recognition,” in Proc. 14th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG 2019). IEEE, 2019, pp. 1–5.
[13] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, “Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions,” IEEE Trans. Multimedia, vol. 22, no. 3, pp. 626–640, 2019.
[14] F. Zhang, T. Zhang, Q. Mao, and C. Xu, “Joint pose and expression modeling for facial expression recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3359–3368.
[15] J. Yu, C. Zhang, Y. Song, and W. Cai, “ICE-GAN: Identity-aware and capsule-enhanced GAN for micro-expression recognition and synthesis,” arXiv preprint arXiv:2005.04370, 2020.
[16] K. M. Goh, C. H. Ng, L. L. Lim, and U. U. Sheikh, “Micro-expression recognition: An updated review of current trends, challenges and solutions,” Vis. Comput., vol. 36, no. 3, pp. 445–468, 2020.
[17] G. Zhao and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 915–928, 2007.
[18] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, “Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition,” in Proc. Asian Conf. Comput. Vis. Springer, 2014, pp. 525–537.
[19] X. Huang, S.-J. Wang, G. Zhao, and M. Pietikainen, “Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection,” in Proc. IEEE Int. Conf. Comput. Vis. workshops, 2015, pp. 1–9.
[20] X. Huang, S.-I. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, “Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition,” IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 32–47, 2017.

[21] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, “A main directional mean optical flow feature for spontaneous micro-expression recognition,” IEEE Trans. Affect. Comput., vol. 7, no. 4, pp. 299–310, 2015.

[22] Y.-J. Liu, B.-J. Li, and Y.-K. Lai, “Sparse MDMO: Learning a discriminative feature for spontaneous micro-expression recognition,” IEEE Trans. Affect. Comput., 2018.

[23] S. Happy and A. Routray, “Fuzzy histogram of optical flow orientations for micro-expression recognition,” IEEE Trans. Affect. Comput., vol. 10, no. 3, pp. 394–406, 2017.

[24] S.-J. Wang et al., “Micro-expression recognition using color spaces,” Proc. 25th IEEE Int. Conf. Multimedia Syst. (ICMR), 2019.

[25] W. Merghani, A. K. Davison, and M. H. Yap, “A review on facial micro-expression recognition,” Pattern Anal. Appl., vol. 22, no. 4, 2019, pp. 1313–1339.

[26] S. Happy and A. Routray, “Fuzzy histogram of optical flow orientations for micro-expression recognition,” IEEE Trans. Affect. Comput., vol. 10, no. 3, pp. 394–406, 2017.

[27] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between recurrent neural networks,” in Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit., 2017, pp. 2258–2263.

[28] Z. Xia, Z. Wang, X. Hong, and G. Zhao, “Selective deep features for micro-expression recognition,” IEEE Trans. Affect. Comput., vol. 22, no. 4, pp. 1524–1527, 2022.

[29] D. H. Kim, W. J. Baddar, and Y. M. Ro, “Micro-expression recognition with expression-state constrained spatio-temporal feature representations,” in Proc. 24th ACM Int. Conf. Multimedia, 2016, pp. 382–386.

[30] Y. Liu, X. Hu, and G. Zhao, “Selective deep features for micro-expression recognition,” in Proc. 23rd Int. Conf. Pattern Recognit. (ICPR), IEEE, 2016, pp. 2258–2262.

[31] M. Peng, Z. Wu, Z. Zhang, and T. Chen, “From macro to micro expression recognition: Deep learning on small datasets using transfer learning,” in Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG 2019), IEEE, 2019, pp. 657–661.

[32] Y. Li, X. Hong, and G. Zhao, “Can micro-expression be recognized based on single apex frame?” in Proc. 25th IEEE Int. Conf. Image Process. (ICIP), IEEE, 2018, pp. 3094–3098.

[33] Z. Xia, X. Feng, X. Hong, and G. Zhao, “Spontaneous facial micro-expression recognition via deep convolutional network,” in 2018 8th Int. Conf. Image Process. Theory, Tools Appl. (IPTA), IEEE, 2018, pp. 1–6.

[34] Y. Liu, H. Du, L. Zheng, and T. Gedeon, “A neural micro-expression recognizer,” in Proc. 14th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG 2019), IEEE, 2019, pp. 1–4.

[35] J. Li, Y. Wang, J. See, and W. Liu, “Micro-expression recognition based on 3D flow convolutional neural network,” Pattern Anal. Appl., vol. 22, no. 4, 2019, pp. 1313–1339.

[36] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, “Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks,” in Proc. 2019 Int. Joint Conf. Neural Netw. (IJCNN), IEEE, 2019, pp. 1–8.

[37] N. Van Quang, J. Chun, and T. Tokuyama, “CapsuleNet for micro-expression recognition,” in Proc. 14th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG 2019), IEEE, 2019, pp. 1–7.

[38] Z. Xia, H. Liang, X. Hong, and X. Feng, “Cross-database micro-expression recognition with deep convolutional networks,” in Proc. 3rd Int. Conf. Biometric Eng. Appl., 2019, pp. 56–60.

[39] C. Wang, M. Peng, T. Bi, and T. Chen, “Micro-attention for micro-expression recognition,” Neurocomputing, vol. 410, pp. 354–362, 2020.

[40] M. Aouayeb, W. Hamidouche, K. Kpalma, and A. Benazza-Benyahia, “A spatiotemporal deep learning solution for automatic micro-expressions recognition from local facial regions,” in Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process. (MSP), IEEE, 2019, pp. 1–6.

[41] B. Sun, S. Cao, D. Li, J. He, and L. Yu, “Dynamic micro-expression recognition using knowledge distillation,” IEEE Trans. Affect. Comput., 2020.

[42] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, “AU-assisted graph attention convolutional network for micro-expression recognition,” in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 2871–2880.

[43] B. Xia, W. Wang, S. Wang, and E. Chen, “Learning from macro-expression: A micro-expression recognition framework,” in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 2936–2944.

[44] L. Lei, J. Li, T. Chen, and S. Li, “A novel graph-TCN with a graph structured representation for micro-expression recognition,” in Proc. 28th ACM Int. Conf. Multimedia, 2020, pp. 2237–2245.

[45] D. Y. Choi and B. C. Song, “Facial micro-expression recognition using two-dimensional landmark feature maps,” IEEE Access, vol. 8, pp. 121549–121563, 2020.

[46] S. Thuseethan, S. Rajasegarar, and J. Yearwood, “Complex emotion profiling: An incremental learning based approach with sparse annotations,” IEEE Access, vol. 8, pp. 147711–147727, 2020.

[47] M. Verma, M. S. K. Reddy, Y. R. Meedimalle, M. Mandal, and S. K. Vipparthi, “AutoMER: Spatiotemporal neural architecture search for Microexpression recognition,” IEEE Trans. Neural Netw. Learn. Syst., 2021.

[48] S. Zhao et al., “A two-stage 3D CNN based learning method for spontaneous micro-expression recognition,” Neurocomputing, vol. 448, pp. 276–289, 2021.

[49] K.-H. Liu, Q.-S. Jin, H.-C. Xu, Y.-S. Gan, and S.-T. Liong, “Micro-expression recognition using advanced genetic algorithm,” Signal Process.: Image Communication, vol. 93, pp. 116153, 2021.

[50] L. Zhou, Q. Mao, X. Huang, F. Zhang, and Z. Zhang, “Feature refinement: An expression-specific feature learning and fusion method for microexpression recognition,” Pattern Recognit., vol. 122, p. 108275, 2022.

[51] Z. Zhai, H. Sun, J. Zhao, Z. Dong, S. He, and H. Zhao, “Displacement on landmark relations with graph attention convolutional network,” in Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC), IEEE, 2021, pp. 2994–2999.

[52] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[53] Z. Ambadar, J. W. Schooler, and J. F. Cohn, “Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions,” Psychological Sci., vol. 16, no. 5, pp. 403–410, 2005.

[54] S. J. Wang et al., “Micro-expression recognition with small sample size by transferring long-term convolutional neural network,” Neurocomputing, vol. 312, pp. 251–262, 2018.

[55] A. J. R. Kumar and B. Bhanu, “Micro-expression classification based on landmark relations with graph attention convolutional network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 1511–1520.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
VERMA et al.: DEEP INSIGHTS OF LEARNING-BASED MICRO EXPRESSION RECOGNITION 1069

[67] L. Lei, T. Chen, S. Li, and J. Li, “Micro-expression recognition based on facial graph representation learning and facial action unit fusion,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 1571–1580.

[68] P. Ekman and W. V. Friesen, “Facial action coding system,” Environmental Psychol. Nonverbal Behav., 1978.

[69] P. Ekman, “METT. micro expression training tool,” CD-ROM. Oakland, 2003.

[70] L. Lo, H.-X. Xie, H.-H. Shuai, and W.-H. Cheng, “MER-GCN: Micro-expression recognition based on relation modeling with graph convolutional networks,” in Proc. IEEE Conf. Multimedia Inf. Process. Retrieval (MIPR). IEEE, 2020, pp. 79–84.

[71] L. Lei, T. Chen, S. Li, and J. Li, “Micro-expression recognition based on facial graph representation learning and facial action unit fusion,” in Proc. IEEE Conf. Multimedia Inf. Process. Retrieval (MIPR). IEEE, 2020, pp. 79–84.

[72] H. Liu, K. Simonyan, and Y. Yang, “DARTS: Differentiable architecture search,” arXiv preprint arXiv:1806.09055, 2018.

[73] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Int. Conf. Learn. Representations (ICLR), 2012.

[74] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 25, pp. 1097–1106, 2015.

[75] X. Zheng et al., “MIGO-NAS: Towards fast and generalizable neural architecture search,” J. Int. Conf. Learn. Representations (ICLR), 2017.

[76] H. Liu, K. Simonyan, and Y. Yang, “DARTS: Differentiable architecture search,” arXiv preprint arXiv:1806.09055, 2018.

[77] C. Liu et al., “Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 82–92.

[78] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Adv. Neural Inf. Process. Syst., vol. 25, pp. 1097–1106, 2012.

[79] S. S. Bengio, “Deep learning,” in Proc. 13th Int. Conf. Digit. Image Comput.: Techn. Appl. (DICTA). IEEE, 2017, pp. 1–7.

[80] T.-H. Oh et al., “Learning-based video motion magnification,” in Proc. Eur. Conf. Comput. Vis. (ECCV). Springer, 2018, pp. 633–648.

[81] A. G. Howard et al., “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” arXiv preprint arXiv:1704.04861, 2017.

[82] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size,” arXiv preprint arXiv:1602.07360, 2016.

[83] M. Verma, S. K. Vipparthi, and G. Singh, “HiNet: Hybrid inherited feature learning network for facial expression recognition,” IEEE Lett. of Comput. Soc., vol. 2, pp. 36–39, 2019.

[84] M. Verma, S. K. Vipparthi, and G. Singh, “HiNet: Hybrid inherited feature learning network for facial expression recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 9, pp. 2936–2952, 2021.

[85] M. F. Hashmi et al., “LARNet: Real-time detection of facial micro-expressions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 21, no. 4, p. 1098, 2021.

[86] A.-X. Xie, “Downside Risk Measure: Theory and Applications in Portfolio Choice and Risk Management.” Fordham University, 2002.

[87] M. A. Takalkar and M. Xu, “Image based facial micro-expression recognition using high accuracy on small datasets,” in Proc. 2017 Int. Conf. Digit. Image Comput.: Techn. Appl. (DICTA). IEEE, 2017, pp. 1–7.

[88] M. Verma, S. K. Vipparthi, and G. Singh, “HiNet: Hybrid inherited feature learning network for facial expression recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 9, pp. 2936–2952, 2021.

[89] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, “M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues,” in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 02, 2020, pp. 1359–1367.

[90] M. Verma, J. K. Bhui, S. K. Vipparthi, and G. Singh, “EXPERTNet: Exigent features preservative network for facial expression recognition,” in Proc. 11th Indian Conf. Comput. Vis., Graph. Image Process., 2018, pp. 1–8.

[91] E. A. Veltmeijer, C. Gerritsen, and K. Hindriks, “Automatic emotion recognition for groups: A review,” IEEE Trans. Affect. Comput., 2021.

[92] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, “Eulerian video magnification for revealing subtle changes in the world,” ACM Trans. Graph. (TOG), vol. 31, no. 4, pp. 1–8, 2012.

[93] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, “Emoticon: Context-aware multimodal emotion recognition system using frege’s principle,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 14234–14243.

[94] M. Verma, J. K. Bhui, S. K. Vipparthi, and G. Singh, “HiNet: Hybrid inherited feature learning network for facial expression recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 9, pp. 2936–2952, 2021.

[95] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, “Emoticon: Context-aware multimodal emotion recognition system using frege’s principle,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 14234–14243.

[96] M. Koizumi and H. Takagishi, “The relationship between child maltreatment and emotion recognition,” PLoS One, vol. 9, no. 1, p. e86093, 2014.

[97] M. Vaškoci and H. Takagishi, “The relationship between child maltreatment and emotion recognition,” PLoS One, vol. 9, no. 1, p. e86093, 2014.