Micro-Expression Recognition: A Comprehensive Survey of Methods, Challenges, and Future Directions in Facial Analysis

Priyansh Sharma, Sanju Choudhary
Department of Information Technology, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, 302017 (India)
Email: m210004@skit.ac.in,sanju@skit.ac.in
Received 11.05.2025 **received in revised form** 03.12.2025, **accepted** 04.12.2025
DOI: 10.47904/IJSKIT.15.2.2025.28-33

Abstract-Facial expression recognition (FER) has become a prominent field of research in computer vision, human-computer interaction, and artificial intelligence due to its potential applications in various domains such as healthcare, education, marketing, and entertainment. This survey presents a comprehensive review of the methods, techniques, and advancements in FER, focusing on the identification and analysis of human emotions through facial expressions. The paper explores traditional techniques, including geometric feature-based methods, and compares them with more recent advancements that utilize "deep learning, particularly convolutional neural networks (CNNs) and other neural network architectures". These deep learning approaches have revolutionized FER by improving accuracy, robustness, and real-time performance, especially in complex and unconstrained environments. The paper also discusses the challenges in FER, such as variations in lighting, pose, and occlusions, and emphasizes the importance of large, diverse datasets for training effective models. Furthermore, the survey reviews the most widely used FER datasets and evaluation metrics, highlighting their impact on the development of reliable systems. Finally, it provides insights into future research directions, including cross-cultural considerations, multimodal emotion recognition, and the integration of FER with other artificial intelligence technologies for more sophisticated, human-like interactions.

Keywords– FER, Machine Learning, CNN, Multi-model

1. INTRODUCTION

Face recognition is "a task that humans perform routinely and effortlessly in their daily lives. Robert Axelrod has also shown the ability to recognize that they have met before and distinguish them from strangers is one of the bases for humans to form cooperation [1]. The last decade has witnessed a trend towards an increasingly ubiquitous computing environment, where powerful and low-cost computing systems are being integrated into mobile phones, cars, medical instruments and almost every aspect of our lives. This has created an enormous interest in automatic processing of digital images and videos in a number of applications, including biometric authentication, surveillance, human-computer interaction, and multimedia management. Research and development in automatic face recognition follows naturally. Face recognition is a visual pattern recognition problem where a three-dimensional object is to be identified based on its two-dimensional image. In recent years, significant progress has been made in the area; owing to better face models and more powerful computers, face recognition system can achieve good results under constrained situations".

However, "because face images are influenced by several factors: illumination, head pose, expression and soon, in general conditions, face recognition is still challenging. From a computer vision point of view, among all these “noises” facial expression maybe the toughest one in the sense that expressions actually change the three-dimensional object while other factors, such as illumination and position, only affect imaging parameters. To get rid of expression “noise”, one first needs to estimate the expression of an image; this is called “Facial Expression Recognition”. Another, maybe more important motivation of facial expression recognition is that expression itself is an efficient way of communication: it’s natural, non-intrusive, and [2] has shown that, surprisingly, expression conveys more information than spoken words and voice tone. To build a friendlier Human Computer Interface, expression recognition is essential".

In this chapter, we talk about and examine a number of works that have to do with human emotions, especially those that have to do "with recognising emotions based on facial expressions". It will give an overview of the different kinds of computer technology that "can be used for face expression recognition". Existing works will be shown so that you can see what kinds of research are being done now and what kinds of methods are being used to in measure how people feel. In the 21st century, everyone uses computers, and they are an important part of daily life. Computers are getting better at recognising human feelings, and the not-too-distant future, they might even be able to "have emotions" of their own. People think that the Stoics of Ancient Greece, Plato, and Aristotle were some of the first people to study emotions. The famous Aristotelian theory of emotions, which was made by Aristotle, looks at how his ideas about feelings changed over time by defining, explaining, comparing, and

contrasting the following: "Emotions are the things on account of which the ones altered differ with respect to their judgments, and are accompanied by pleasure and pain: such are anger, pity, fear, and all similar emotions and their contraries" [3].

1.1 Typical FER System

To make a good FER system, you need to know how a normal system is built and what steps you need to take to get from the image to the expression. Figure 1 shows that the designs of these common systems are made up of three main parts that are used in many different kinds of applications, such as FER and Depression analysis. In the next line, we'll talk more about what these building blocks are and how they work. Most systems that deal with images of faces or items use three main building blocks as the main units of analysis. The first block takes care of the data, which can be raw photos, audio, or video sequences. In this step, some pre-processing steps are done to the data to get it ready for the next step, which is called "feature extraction." There are many different ways to do these pre-processing steps, such as face recognition, alignment, normalization, and augmentation for facial images, in addition to the many other methods that are already available and are being made. After all of the steps that the system needs to be done are finished, the next building block pulls characteristics from the samples that have already been handled.

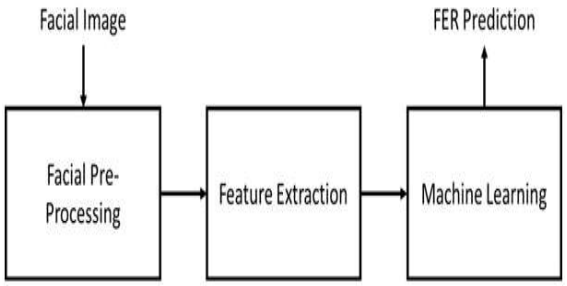

Figure 1: Building blocks for a typical FER system

"There are many different techniques and approaches that exist to extract features from the sample images, some of which are detailed later. These approaches can extract features in the form of appearance information, geometric information, temporal and spatiotemporal information. Once these features are obtained, the final building block will try to learn them using machine learning techniques; also mentioned later in this chapter. To get a better understanding of these systems, the following sections will go deeper into the existing feature extraction techniques used on different modalities of data. This will also include the popular machine learning techniques widely used in divers".

1.2 Prototypic Emotional Expressions

Instead of describing "the detailed facial features, most FER systems attempt to recognize a small set of prototypic emotional expressions. The most widely-used set is perhaps human universal facial expressions of emotion which consists of six basic expression categories that have been shown to be recognizable across cultures (Fig.2). These expressions, or facial configurations have been recognized in people from widely divergent cultural and social backgrounds and they have been observed even in the faces of individuals born deaf and blind" [4].

"These 6 basic emotions, i.e., disgust, fear, joy, surprise, sadness and anger plus neutral which means no facial expression are considered in this work. Given a facial image, our system either works as a conventional classifier to determine the most likely emotion or estimates the weights (or possibility) of each emotion as a fuzzy classifier does".

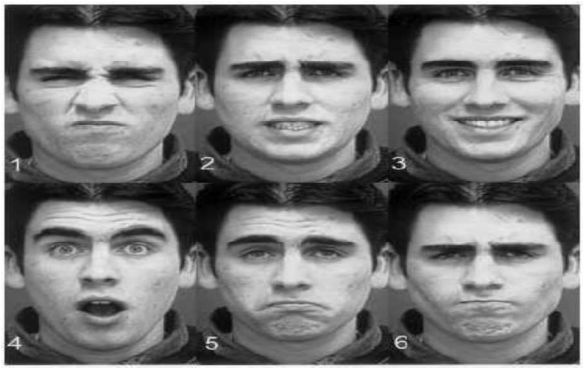

Figure 2: Basic facial expression phenotypes. 1, disgust; 2, fear; 3, joy; 4, surprise; 5,Sadness; 6, anger

2. LITERATURE REVIEW

FER came up with the first way "to automatically analyse facial emotions by tracking the movement of 20 chosen spots in a series of images". This was the first method of this kind. Since then, many different systems have been made that can analyse face expressions automatically from both still images and moving image sequences. In the fields of "human-computer interaction, affective computing, intelligent control, psychological analysis, pattern recognition, security monitoring, social cognition, and machine vision", this has been a very popular subject. Entertainment in social settings and other areas.From a person's facial movements, you can get a sense of how they are feeling, what they are thinking, what they want to do, and who they are as a person. Communication plays a part in how people get along with each other. It's not easy to make a computer system that can recognize faces and figure out how people are feeling by looking at them. There are a few related problems that need to be answered, such as how to tell if a part of an image is a face, how to get information about facial expressions, and how to put those expressions into different emotional states. To create a more human-like interface between people and computers, it would be a big step forward to create a system that can do these things correctly and in real time.

Alzahrani et al. [5] presented a dual-attention frame -work that combines global and local adversarial learning to obtain domain-invariant facial representations for cross-database FER. The model achieves strong performance on RAF-DB and AffectNet. However, the network is relatively heavy and its adversarial optimization remains sensitive to domain drift, which may hinder deployment in resource-constrained or highly dynamic environments.

Yu et al. [6] designed a multi-modal expression recognition pipeline that fuses audio–visual cues with a global channel–spatial attention mechanism, complemented by decision-level fusion and key-frame alignment. The method reports competitive accuracy on the ABAW 2025 expression validation set, highlighting the benefit of exploiting complementary modalities for subtle affect cues. Its dependence on multiple synchronized modalities and intensive training, however, increases system complexity and limits applicability in unimodal or low-resource scenarios.

Jayaraman et al. [7] introduce a GAN-based system that leverages depth information and adversarial transfer learning to improve recognition under multi-view facial poses. The approach reports notable gains on FER2013, CK+, and JAFFE, demonstrating that depth-guided synthesis can mitigate viewpoint variability. Nonetheless, the architecture is complex and requires substantial data to stabilize GAN training, which can restrict reproducibility and practical adoption.

Wang et al. [8] combine quantum-inspired feature extraction with an enhanced ResNet backbone to enrich the representational capacity for FER. Experiments on CK+ and JAFFE show accuracy improvements over conventional ResNet variants, suggesting that quantum-based descriptors can complement deep convolutional features. Yet, limited tooling and hardware support for quantum techniques, along with potential reproducibility issues, make the approach challenging to replicate in standard research settings.

Junhuan Wang [9] proposes a GAN architecture augmented with residual learning and dimensionality reduction to generate more discriminative facial representations while controlling computational cost. The method achieves high recognition accuracy on JAFFE, CK+, and FER2013, indicating that residual blocks and feature compression can enhance GAN-based FER. On the downside, the framework relies on careful hyperparameter tuning and sizable labeled datasets, which may limit its robustness across new domains.

Yan et al. [10] construct a dedicated neonatal facial expression dataset (FENP) and benchmark several CNN variants for pain assessment in infants. DenseNet achieves strong accuracy, showing that deep convolutional models can reliably capture neonatal pain cues from limited visual signals. However, the dataset's narrow age range raises concerns about generalizability to older children or adults, restricting the broader applicability of the findings.

Luo et al. [11]develop an enhanced CNN-based FER pipeline that integrates dedicated pre- processing, feature extraction, and classification stages. The approach attains competitive accuracy on CK+, suggesting that careful input normalization and feature handling can significantly bolster CNN performance. Nevertheless, its effectiveness drops on non-standard expression datasets, indicating that the pipeline may be over-optimized for controlled conditions.

Kumar et al. [12] present a real-time FER frame-work that uses Haar cascades for face detection followed by a CNN classifier for expression recognition from webcam streams. The system achieves reasonable accuracy while maintaining interactive frame rates, making it suitable for basic real-time applications. Its reliance on traditional detection and a relatively simple CNN, however, limits robustness under occlusion and complex in-the-wild scenarios.

Otberdout et al. [13] propose a manifold-based GAN that synthesizes dynamic facial motions on hyperspherical spaces to augment data for expression analysis. The generated sequences enhance recognition performance on Oulu-CASIA, indicating the value of motion-aware generative models for video-based FER. Yet, the method incurs high computational cost and may overfit to synthetic motion patterns, calling for careful regularization and validation.

Yu et al. [14] explore unsupervised neutral face generation using CycleGAN combined with cross-entropy-based training to learn neutral counterparts of expressive images without manual labels. The model reaches high accuracy on JAFFE, showing that neutral-expression synthesis can be automated for data balancing and normalization. However, its focus on neutral face creation and reliance on a single dataset restrict the generality and diversity of the produced samples.

Xia et al. [15] design LGP-GAN, a generative model that cascades local and global perception stages to improve expression realism and fine-grained facial detail. Evaluations on RAF-DB and AffectNet confirm that the method produces more detailed and expressive faces than previous GAN-based approaches, benefiting downstream recognition. This accuracy comes at the expense of high computational demand and architectural complexity, which complicate training and deployment.

Ju et al. [16] introduce MAP-Net, which exploits facial landmarks and spatial mask attention to achieve robust FER under partial occlusions. The network sets strong results on RAF-DB, AffectNet, and FEDRO under occluded conditions, under-scoring the importance of structured attention for visibility-aware recognition. Nonetheless, perfor-mance degrades for extreme poses and profile views, suggesting limitations in handling severe viewpoint variation.

Khine et al. [17] employ EfficientNet-B0 with transfer learning to build a compact yet accurate FER system on the CK+ dataset. The approach offers a good balance between performance and model size, achieving strong accuracy with relatively simple training. Its dependence on pretraining quality and evaluation on a single

controlled dataset, however, raises questions about robustness in more challenging, unconstrained settings.

Teng et al. [18] propose TFEN, a deep spatio-temporal network that couples 3D CNNs with a feature decoupler and adversarial training to separate identity from expression information. Results on CK+ indicate that decoupling spatial and temporal cues enhances dynamic FER performance, particularly for subtle changes over time. The model's deep and dynamic architecture, however, incurs substantial runtime cost, which may limit scalability to long videos or real-time use.

Cai et al. [19] develop IF-GAN, an identity-free conditional GAN intended to reduce subject-specific variation by mapping faces to a canonical identity space before expression classification. The method improves generalization across CK+, MMI, Oulu-CASIA, and SFEW, demonstrating that suppressing identity cues can aid cross-subject FER. Still, GAN training instability and incomplete subject invariance remain challenges, indicating that identity disentanglement is not fully resolved.

3. Conclusions

The reviewed works collectively show that recent FER research has moved strongly toward attention mechanisms, adversarial learning, and generative models to tackle domain shift, pose variation, occlusion, and identity bias while targeting higher accuracy on in-the-wild datasets. Hybrid designs that blend global–local attention, multi-modal fusion, and identity- or occlusion-aware modules consistently outperform conventional CNN baselines, but they introduce higher architectural and computational complexity. GAN-based and quantum-inspired approaches further enhance representation power and data diversity, yet they frequently demand large labeled datasets, careful hyperparameter tuning, and specialized hardware or tool support, which can impede reproducibility and real-time deployment. Overall, the table highlights an accuracy–efficiency trade-off and indicates that future FER work must prioritize lightweight, robust, and more easily generalizable models that preserve the strengths of advanced attention and generative strategies while reducing their practical overheads.

4. Future Scope

The future scope of facial expression recognition using CNN classifier and DRLBP is promising. Here are some potential areas of development and advancement:

Improved Accuracy and Robustness:There is "ongoing research to improve the accuracy and robustness of facial expression recognition systems".

Real-time and Low-power Implementations:Efforts are being made to optimize facial expression recognition algorithms for real-time performance and low-power consumption. This involves designing lightweight CNN architectures, developing efficient feature extraction methods, and exploring hardware acceleration techniques such as specialized processors or dedicated hardware accelerators.

Large-scale and Diverse Datasets: Availability of large-scale and diverse datasets plays a crucial role in improving facial expression recognition systems. Future research will focus on collecting and annotating more extensive datasets, encompassing diverse demographics, age groups, cultural backgrounds, and expressions.

Multimodal Approaches: "Combining facial expression recognition with other modalities like voice, body language, or physiological signals can provide a more comprehensive understanding of human emotions. The future scope involves integrating multiple modalities to enhance the accuracy and richness of emotion recognition systems".

Personalized and Adaptive Systems: Future systems may focus on personalization and adaptation, tailoring the recognition models to individual users' unique facial expressions and emotional patterns. This could involve developing user-specific models or employing "transfer learning techniques to adapt pre-trained models to new individuals".

Cross-cultural and Cross-ethnic Recognition: Advancements in facial expression recognition will address the challenge of cross-cultural and cross-ethnic variations. Research will focus on developing "models that can generalize well across different ethnicities" and cultural backgrounds, reducing biases and improving recognition accuracy for diverse populations.

Ethical Considerations and Privacy Protection: As facial expression recognition technology continues to evolve, it is crucial to address ethical considerations and privacy protection. Future developments will involve incorporating fairness and transparency into the algorithms, ensuring data privacy, and addressing potential biases and societal implications of facial recognition systems.

S.No	Title (Year)	Author(s)	Methodology	Result	Drawback
1	Dynamic Cross-Domain Dual Attention Network for FER (2025) [1]	AO Alzahrani et al.	DCD-DAN using global/local adversarial learning for domain-invariant FER	High accuracy and robust cross-domain performance (RAF-DB 93.18%, Affect Net 82.13%)	Performance sensitive to adversarial domain drift; heavy model size
2	Expression Recognition Solution via Global Channel-Spatial Attention (2025) [2]	Jun Yu et al.	Multi-modal fusion with channel-spatial attention; decision fusion and key frame alignment	8th ABAW, CVPR 2025 Expression Validation Set: 63.20%	Needs multiple modalities, intensive training
3	GAN-Based Multi-Angle FER System (2025) [3]	S Jayaraman et al.	GAN method using depth data and adversarial transfer learning	Improved accuracy for multi-view datasets FER 2013: 82.89%, CK+: 96.78%, JAFFE: 95.87%	Complex GAN architecture; needs extensive data for stability
4	Deep Quantum & Advanced ResNet for FER (2024) [4]	Wang et al.	Quantum-based features and modified ResNet extraction	Better accuracy than standard ResNet on tested benchmarks. CK+: 98.19%, JAFFE: 99.68%	Quantum techniques not widely supported; reproducibility issue.
5	Improved GAN-Based FER via Residual Networks (2024) [5]	Junhuan Wang	GAN with residual learning and dimensionality reduction	High accuracy on JAFFE: 96.6%, CK+: 95.6%, FER2013: 72.8% with lower computation	Challenging hyperparameter tuning; needs large labeled datasets
6	FENP: Neonatal Facial Expression Dataset with CNNs (2023) [6]	Jingjie Yan et al.	Created neonatal facial expression dataset; used CNN variants	CNNs efficient for pain detection in neonates. DenseNet: 93.1%	Limited age range; generalization to adults unclear
7	Improved CNN for FER with Preprocessing (2023) [7]	Yue Luo et al.	Enhanced CNN pipeline for FER; includes preprocessing, extraction, classification	Achieved 88% accuracy in CK+ emotion recognition	Lower consistency on non-standard expression datasets
8	Haar Cascade & CNN Real-Time FER (2023) [8]	Rajesh Kumar et al.	Haar cascade for face detection, CNN for classification	Real-time recognition from webcam with good accuracy. FER: 67%	Less adaptive to occlusion; basic CNN restricts expressiveness
9	Manifold GAN for Dynamic Motion Expression (2022) [9]	Naima Otberdout et al.	Motion synthesis on hyperspheres by GAN, data augmentation	Generated realistic dynamic facial video. Recognized 94.75% on Oulu-CASIA	Expensive computation; may overfit synthetic motions
10	CycleGAN Unsupervised Neutral Face Generation (2022) [10]	Yating Yu et al.	Cycle GAN +cross entropy for unsupervised neutral face generation	98.81% accuracy on JAFFE, no manual labeling needed	Focused only on neutral face creation, limited to JAFFE
11	LGP-GAN: Local and Global Perception for FER (2022) [11]	Yifan Xia et al.	GAN with cascaded local-global feature stages	Superior expression realism/detail over prior models. RAF-DB: 89.42%, AffectNet: 64.08%	Computational cost and multi-stage complexity
12	MAP-Net: Mask-Based Attention Parallel Net (2022) [12]	Lingzhao Ju et al.	Landmarks and spatial mask attention for occlusion-robust FER	Best accuracy under occlusion (RAFDB:94.44%, AffectNet 60.13%, FEDRO: 77.45%)	Accuracy drops on profiles/extreme pose images
13	EfficientNet-BO Transfer Learning for FER (2022) [13]	Win Shwe Sin Khine et al.	EfficientNet BO and transfer learning on CK+	CK+:92.14% accuracy, faster and simpler network	Sensitive to pretraining quality; limited to CK+ dataset
14	TFEN: Deep Spatio-Temporal Facial Feature Decoupling (2021) [14]	Jianing Teng et al.	3D CNN, facial feature decoupler, adversarial training	CK+: 95.1%	High runtime cost with deep dynamic networks
15	IF-GAN: Identity-Free Conditional GAN (2021) [15]	Jie Cai et al.	Identity-free GAN architecture to reduce subject variation	CK+: 88.1%, MMI: 69.6%, Oulu-CASIA: 66.4%, SFEW: 47.9%	GAN training instability; subject transferability not universal

Table 1: Comparisons of Different Techniques

REFERENCES

[1]. Majumder, L. Behera, and V. K. Subramanian, "Emotion recognition from geometric facial features using self-organizing map," Pattern Recognition, (2014), vol. 47, no. 3, pp. 1282–1293.

[2]. Y. Tong, R. Chen, J. Yang, and M. Wu, "Robust facial expression recognition based on local tri-directional coding pattern," in Proceedings of the 12th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS), (2019), pp. 606–614.

[3]. H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), (2015), pp. 2983–2991.

[4]. H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2018), pp. 2168–2177.

[5]. A. O. Alzahrani et al., "A novel facial expression recognition framework using dynamic cross-domain dual attention network (DCD-DAN)," PubMed, (2025), PMID: 40567646.

[6]. J. Yu, Y. Zheng, L. Wang, Y. Wang, and S. Xu, "Design of an expression recognition solution employing the global channel–spatial attention mechanism," in Proceedings of the CVPR 2025 Workshop on Affective Behavior Analysis in the Wild (ABAW), (2025), arXiv:2503.11935.

[7]. S. Jayaraman et al., "Generative adversarial network based multi-angle facial expression recognition using depth data and adversarial learning," Scientific Reports, (2025).

[8]. S. Alsubai, A. Alqahtani, A. Alanazi, M. Sha, and A. Gumaei, "Facial emotion recognition using deep quantum and advanced transfer learning mechanism," Frontiers in Computational Neuroscience, (2024).

[9]. J. Wang, "Improved facial expression recognition method based on GAN," Journal of Computer Applications, (2021).

[10]. J. Yan et al., "FENP: A database of neonatal facial expression for pain analysis," IEEE Transactions on Affective Computing, (2020).

[11]. Y. Luo, J. Wu, Z. Zhang, H. Zhao, and Z. Shu, "Design of facial expression recognition algorithm based on CNN model," in Proceedings of the IEEE 3rd International Conference on Power Electronics and Computer Applications (ICPECA), (2023).

[12]. R. Kumar, "A deep learning approach to recognizing emotions through facial expressions," in Proceedings of the Global Conference on Wireless and Optical Technologies (GCWOT), (2023).

[13]. N. Otberdout, M. Daoudi, A. Kacem, L. Ballihi, and S. Berretti, "Dynamic facial expression generation on Hilbert hypersphere with conditional Wasserstein generative adversarial nets," IEEE Transactions on Pattern Analysis and Machine Intelligence, (2020).

[14]. Y. Yu, Y. Sun, and Z. Yang, "An unsupervised facial expression recognition method based on CycleGAN," in Proceedings of the International Conference on Big Data, Information and Computer Network (BDICN), (2022).

[15]. Y. Xia, W. Zheng, Y. Wang, H. Yu, J. Dong, and F.-Y. Wang, "Local and global perception generative adversarial network for facial expression synthesis," IEEE Transactions on Circuits and Systems for Video Technology, (2022).

[16]. L. Ju and X. Zhao, "Mask-based attention parallel network for in-the-wild facial expression recognition," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2022).

[17]. W. S. S. Khine, P. Siritanawan, and K. Kotani, "Facial expression features analysis with transfer learning," in Proceedings of the 14th International Conference on Knowledge and Systems Engineering (KSE), (2022).

[18]. J. Teng, D. Zhang, W. Zou, M. Li, and D.-J. Lee, "Typical facial expression network using a facial feature decoupler and spatio-temporal learning," IEEE Transactions on Affective Computing, (2023)..

[19]. J. Cai, Z. Meng, A. S. Khan, J. O'Reilly, Z. Li, S. Han, and Y. Tong, "Identity-free facial expression recognition using conditional generative adversarial network," in Proceedings of the IEEE International Conference on Image Processing (ICIP), (2021).