International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 11 | Nov 2026 www.irjet.net p-ISSN: 2395-0072

Dataset-Aware Comparison of Lightweight, Attention-Based and Hybrid Convolutional Neural Networks for Facial Expression Recognition: A Review

Ashraf Osman Sayed[1], Khalid Hamid Bilal[2], Yasir Mohamed[3]

[1]Karary University, Sudan
[2] Omdurman Islamic University, Sudan
[3]A'Sharqiyah University, Oman

Abstract - *Convolutional neural networks (CNNs) remain the basis of most facial expression recognition (FER) systems, and almost every new architecture is presented as state of the art. The reported figures are hard to compare, because they come from datasets that differ in size, label quality, class setting and collection conditions, from small posed laboratory collections to large web-scale corpora with noisy labels. We review 22 studies published between 2020 and 2026 and group them into six families: lightweight and real-time CNNs, attention-based networks, transformer and state-space models, hybrid CNN-recurrent or graph models, multi-branch fusion, and transfer-learning or in-the-wild studies. We then regroup the headline results reported by the original authors by dataset rather than by method. The regrouped view shows that the choice of dataset explains most of the variation. Reported accuracy lies between 96% and 99% on the posed CK+ and JAFFE data, between 88% and 96% on RAF-DB, and between 60% and 68% on AffectNet, so laboratory and in-the-wild results are separated by almost 30 percentage points. Moving a single model between the 7-class and 8-class AffectNet settings changes its reported accuracy by 3.6 to 4.9 points, which is more than many of the margins used to claim superiority, and one pair of models swaps order between RAF-DB and 8-class AffectNet. Among three models with reported size and cost, the step from FERMam to POSTER++ added 62% more parameters for a gain of 0.08 points on RAF-DB. We close with a twelve-item reporting checklist covering dataset version, split protocol, evaluation set, averaging, repeated runs, model size, latency and hardware, pretraining and same-split baselines, which is intended to make future FER results comparable.*

Keywords: deep learning; benchmarking; reproducibility; affective computing; label noise

1. INTRODUCTION

Facial expressions carry much of what people communicate without words, and recognizing them automatically matters for human–computer interaction, driver monitoring, teaching and care settings [1]. Most current benchmarks still label images with the six basic emotions described by Ekman and Friesen, plus a neutral class [2]. Handcrafted descriptors were largely replaced by deep learning in this area after 2012, and CNNs form the backbone of most recent systems [1].

The literature has grown quickly, and so have the claims. Lightweight networks are said to run in real time on embedded hardware [10]. Attention modules are credited with separating look-alike expressions [14], [18]. Transformers and state-space models hold the highest accuracies on in-the-wild data [12], [22], [23], while recurrent hybrids approach saturation on laboratory sets [25], [26]. A reader who wants to choose among these options meets an immediate difficulty: the numbers cannot be set side by side. A model scoring 99% on CK+ [4] has not been shown to beat one scoring 67% on AffectNet [7]. The two figures answer different questions about different data. Even on one dataset, the class setting (7-class or 8-class AffectNet), the evaluation subset (validation or test), the pretraining source and the number of runs behind a figure are often left unstated.

Surveys of deep FER exist and are valuable [1], but they predate much of the 2020–2026 work and do not treat comparability as a subject in itself. This paper does. It makes three contributions:

- a taxonomy of recent CNN-based FER architectures in six families, built from 22 studies (Section 4);
- a dataset-aware comparison in which published headline results are grouped by dataset and class setting, and accuracy–cost statements are made only from same-dataset pairs (Section 5);
- a twelve-item reporting checklist that authors, reviewers and practitioners can use to judge whether two FER results are comparable (Section 6).

Section 2 describes how studies were selected. Section 3 summarizes the datasets, Section 4 reviews the six families, and Section 5 gives the comparison. The checklist is in Section 6, followed by a discussion, the limitations of the review and the conclusion.

2. REVIEW METHOD

This is a narrative review with a structured selection procedure. It is not a systematic review in the PRISMA

sense, and we do not claim exhaustive coverage. We searched the web and the public pages of arXiv and of IEEE, Elsevier, Springer Nature, MDPI and the Computer Vision Foundation, combining terms such as facial expression recognition, lightweight, attention, CNN, transformer, BiLSTM, transfer learning and in the wild. The period covered is January 2020 to 2026.

A study was kept when its core model was convolutional or a CNN-based hybrid, when it reported quantitative results on at least one public FER dataset, and when it was a journal article, a conference paper or an arXiv preprint. Work based only on non-visual signals such as speech was left out. For each study we recorded the dataset, the class setting where it was stated, the headline metric, and the parameter count and floating-point operations (FLOPs) where given. Every number in this paper is as stated by the original authors, taken from the abstract or the main results of the primary source. We re-ran no model. When a source did not state a value, the cell is empty rather than estimated.

Twenty-two studies met the criteria [10]–[31]. They are not a random sample of the field, and the set leans toward open-access and English-language venues, a point we return to in Section 8.

3. BENCHMARK DATASETS

Our central argument is that the dataset largely determines the reported accuracy, so we start with the datasets that recur in the reviewed studies (Table 1). CK+ [4] and JAFFE [5] are laboratory collections with posed or elicited, mostly frontal expressions, small subject pools and controlled lighting. FER2013 [3], RAF-DB [6], AffectNet [7] and AFEW [9] come from the web or from films. They vary widely in pose, illumination, occlusion and age, and their labels are noisier.

Table 1. Benchmark datasets recurring in the reviewed studies.

Dataset	Ref.	Size	Setting	Labels
FER2013	[3]	35,887 images, 48×48 grayscale	In the wild, web search	7 classes; noisy (human agreement reported at about 65–68%)
FER+	[8]	FER2013 images, relabelled	In the wild	10 crowd taggers per image; label distributions
CK+	[4]	593 sequences, 123 subjects	Laboratory, posed or spontaneous	7–8 classes (FACS-coded)
JAFFE	[5]	213 images, 10 subjects	Laboratory, posed, frontal	7 classes
RAF-DB	[6]	About 29,672 images	In the wild, crowd-annotated	7 basic classes (plus compound classes)
AffectNet	[7]	About 1 million web images, roughly 450,000 manually annotated	In the wild, web	8 categories (a 7-class subset is also used); valence and arousal
AFEW	[9]	Video clips from films	In the wild, temporal	7 classes

Sizes and settings are as described in the cited dataset papers.

Two points matter for what follows. The FER2013 labels came from a partly automated pipeline, and human agreement with them has been reported at roughly 65–68% [3]. That figure describes annotation noise, not a ceiling on classifier accuracy, but it makes absolute accuracy on FER2013 hard to read. FER+ was built to reduce the problem by having ten crowd taggers relabel each image [8]. The second point is that AffectNet is used in both 7-class and 8-class settings, which are not interchangeable [7]; Section 5 shows how much this alone can change a reported result.

4. TAXONOMY OF RECENT CNN-BASED FER ARCHITECTURES

Figure 1 shows the six families used to organize the studies. They overlap. FCCA [11] and FERMam [12], for example, pair lightweight design with attention or fusion. We place each study where its main contribution lies and note cross-references. Table 2 lists all 22 studies.

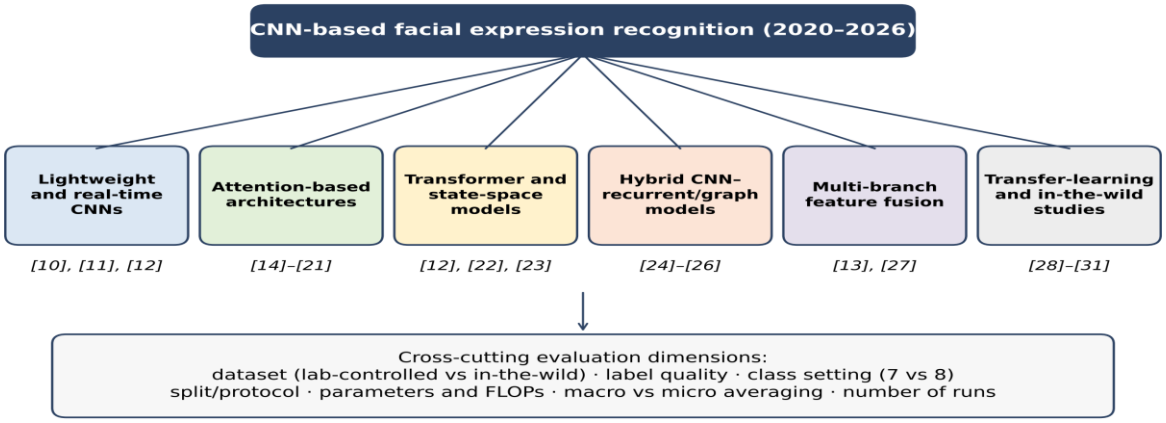

Figure 1. Taxonomy of recent CNN-based FER architectures and cross-cutting evaluation dimensions. Numbers in brackets are reference numbers.

4.1 Lightweight and real-time CNNs

Another theme in recent work has been to reduce computation while maintaining accuracy, mainly to allow models to run on embedded or resource-constrained hardware. Zhou et al. [10] developed a shallow CNN for real time expression detection on embedded platforms and sacrificed some accuracy against deeper networks. FCCA [11] employs a FasterNet-S backbone with partial convolution to reduce memory access, a center-weighted attention module, and a flip-consistency loss against label uncertainty, achieving 91.30% on RAF-DB and 65.51% on 8-class AffectNet. FERMam [12] integrates a CNN, a Mamba-style state-space model, and a landmark branch. It reports 92.13% on RAF-DB with 26.99M parameters and 5.97GFLOPs, which the authors claim to be close to POSTER++ at lower cost. The trend in these studies is that the accuracy of recent lightweight models is attributed to attention or state-space components, not to extra depth.

4.2 Attention-based architectures

Attention directs a network to discriminative regions or channels rather than to each position equally. Li et al. [14] combined local binary pattern texture with a learned attention module to localise informative regions, at some complexity cost. Liu et al. [15] replaced the first convolutional layer by a multi-scale module and inserted wavelet channel attention after the first pooling layer, and achieved gains over a plain CNN, which demonstrates that frequency-domain attention enhances spatial features. Yu [16] combined a global multi-scale module and a local multi-granularity module that segments face and applies light attention to fine-grained cues.

There exist a few well-known in-the-wild methods that leverage attention to handle occlusion, pose or label noise. The Region Attention Network [17] learns the importance of each face region, pools variable number of regions into a fixed-length vector and introduces a region-biased loss. It was evaluated on FERPlus, AffectNet, RAF-DB and SFEW datasets focusing on occluded and non-frontal subsets. Deep attentive center loss [18] weights the feature dimensions such that the center loss is mainly applied to informative dimensions and achieves 87.78% and 65.20% on the RAF-DB test set and the AffectNet validation set, respectively. The Self-Cure Network [19] down-weights uncertain samples through importance weighting, ranking regularization and relabelling, and reports 88.14% on RAF-DB, 60.23% on AffectNet and 89.35% on FERPlus. In the Distract Your Attention Network [20], several cross-attention heads attend to different regions and are pushed to be diverse; the four-head model reaches 89.70% on RAF-DB and 65.69% and 62.09% on 7-class and 8-class AffectNet, with 19.72 M parameters and 2.23 GFLOPs. Zhao et al. [21] combine global multi-scale features with local attention and report 88.40% on RAF-DB and 64.53% and 60.29% on 7-class and 8-class AffectNet. Across the family, attention is reported to help separate similar expressions, at the price of extra modules and sometimes heavier training.

4.3 Transformer and state-space models

Transformers, and more recently state-space models, have been attached to or combined with CNN backbones. Ma et al. [22] use visual transformers with attentional selective fusion and report 88.14% on RAF-DB, 88.81% on FERPlus and 61.85% on AffectNet. POSTER++ [23] fuses image and landmark streams in a simplified two-stream design, with 92.21% on RAF-DB and 67.49% and 63.77% on 7-class and 8-class AffectNet, using 43.7 M parameters and 8.4 GFLOPs. FERMam [12] follows the same two-source idea with a state-space refinement module aimed at lowering parameters and FLOPs, and reports 66.38% and 61.45% on 7-class and 8-class AffectNet and 91.68% on FERPlus.

Accuracy on in-the-wild benchmarks keeps rising in this family, but the compute behind it differs several-fold between designs.

4.4 Hybrid CNN–recurrent and graph models

Hybrid models pair a CNN with a recurrent or graph module to capture sequence structure. Liu et al. [24] combined graph convolutional networks with CNN–RNN models for video, modelling temporal dynamics and regional spatial features; dynamic expressions were recognized better than with static-image CNNs, at higher cost. Febrian et al. [25] merged convolutional features with bidirectional LSTM layers and report 99.43% on CK+ with augmentation, above the authors' own CNN-only and CNN–LSTM baselines. Jayaraman and Mahendran [26] added dual spatial and channel attention to a CNN–BiLSTM and report 82.89% on FER2013, 96.78% on CK+, 95.78% on RAF-DB and 95.87% on JAFFE. Figures above 96% on CK+ call for caution, since CK+ is small, posed and collected in a laboratory, and such figures do not transfer to large naturalistic corpora.

4.5 Multi-branch feature fusion

Multi-branch designs extract complementary information along parallel paths and fuse it before classification. Mahmoudi et al. [13] studied kernelized dense layers and report better recognition with less computational overhead than standard dense layers. Liu et al. [27] used a two-channel CNN whose parallel pathways extract appearance and geometric features; this adds feature diversity and tolerance to the loss of one pathway, but roughly doubles the extraction cost of a single-channel network of similar depth. FERMam [12] and POSTER++ [23] can also be read as fusion architectures.

4.6 Transfer-learning and in-the-wild studies

Transfer learning addresses the shortage of large labelled FER sets. Podder et al. [28] adapted a VGG-Face backbone with a lightweight dense CNN and joint softmax and center-loss supervision, and tested on FER2013, JAFFE, CK+ and KDEF; the work shows that a face-recognition backbone can lower the data and compute needed compared with training from scratchVerschae and Bugueño-Córdova [29] generalised the idea to event-based FER using neuromorphic cameras, therefore the method is general and covers different imaging modalities and datasets. In less controlled conditions, Begaj et al. [30] compared CNN configurations and pre-processing on the iCV MEFED dataset and reported performance to remain moderate on hard samples. Anas et al. [31] submitted a deep CNN trained on FER2013 to ABAW 2020 Track 2 and achieved 50.77% accuracy and 29.16% F1-score on the validation set, which is a clear example of the drop when evaluated unconstrained.

Table 2. Summary of the 22 reviewed studies (results as reported by the authors).

Study	Year	Family	Core idea	Datasets and headline result
Zhou et al. [10]	2021	Lightweight	Shallow CNN for embedded real-time detection	Fast inference; slight accuracy loss against deeper CNNs
Sun et al. [11]	2025	Lightweight, attention	FasterNet-S, center-weighted attention, flip-consistency loss	RAF-DB 91.30%; AffectNet (8-class) 65.51%
Gao et al. [12]	2026	Lightweight, state-space	CNN, Mamba-style refinement, landmark branch	RAF-DB 92.13%; AffectNet-7 66.38%; AffectNet-8 61.45%; FERPlus 91.68%; 26.99 M params, 5.97 GFLOPs
Mahmoudi et al. [13]	2020	Fusion	Kernelized dense layers	Better accuracy with lower overhead
Li et al. [14]	2020	Attention	LBP texture plus learned attention	Better localization; added complexity
Liu et al. [15]	2025	Attention	Multi-scale wavelet channel attention	Gain over a plain CNN baseline
Yu [16]	2025	Attention	Global multi-scale and local multi-granularity fusion	Better sensitivity to local cues

Study	Year	Family	Core idea	Datasets and headline result
Wang et al. [17]	2020	Attention	Region attention, region-biased loss	FERPlus, AffectNet, RAF-DB, SFEW; robust to occlusion and pose
Farzaneh and Qi [18]	2021	Attention	Deep attentive center loss	RAF-DB 87.78% (test); AffectNet 65.20% (validation)
Wang et al. [19]	2020	Attention, label noise	Self-Cure Network	RAF-DB 88.14%; AffectNet 60.23%; FERPlus 89.35%
Wen et al. [20]	2023	Attention	Multi-head cross attention	RAF-DB 89.70%; AffectNet-7 65.69%; AffectNet-8 62.09%; 19.72 M params, 2.23 GFLOPs
Zhao et al. [21]	2021	Attention	Global multi-scale and local attention	RAF-DB 88.40%; AffectNet-7 64.53%; AffectNet-8 60.29%; SFEW 59.40%
Ma et al. [22]	2021	Transformer	Visual transformers, attentional selective fusion	RAF-DB 88.14%; FERPlus 88.81%; AffectNet 61.85%
Mao et al. [23]	2023	Transformer	Two-stream image and landmark cross-fusion	RAF-DB 92.21%; AffectNet-7 67.49%; AffectNet-8 63.77%; 43.7 M params, 8.4 GFLOPs
Liu et al. [24]	2020	Hybrid	GCN with CNN–RNN for video	Better dynamic expressions; higher cost
Febrian et al. [25]	2023	Hybrid	BiLSTM-CNN	CK+ 99.43% (with augmentation)
Jayaraman and Mahendran [26]	2025	Hybrid, attention	DCNN-BiLSTM with dual attention	FER2013 82.89%; CK+ 96.78%; RAF-DB 95.78%; JAFFE 95.87%
Liu et al. [27]	2020	Fusion	Two-channel CNN	More diverse features; about twice the extraction cost
Podder et al. [28]	2026	Transfer	VGG-Face adaptation, dense CNN, center loss	FER2013, JAFFE, CK+, KDEF; lower data and compute needs
Verschae and Bugueño-Córdova [29]	2026	Transfer	Event-camera FER	Transfer across imaging modalities
Begaj et al. [30]	2020	In the wild	CNN and pre-processing comparison	iCV MEFED; moderate performance
Anas et al. [31]	2020	In the wild	Deep CNN for ABAW 2020 Track 2	Validation accuracy 50.77%; F1 29.16%

Where no figure is given, the source reported a qualitative result or no number we could confirm.

5. DATASET-AWARE COMPARISON

Table 3 and Figure 2 regroup the headline results by dataset and, for AffectNet, by class setting. They use only numbers that the original authors stated, and they come from no new experiment. Protocols still differ between studies, because splits, evaluation subsets and pretraining are not identical. The figures are therefore ranges within a dataset and should not be read as a league table.

Table 3. Reported headline accuracy grouped by dataset and class setting.

Dataset	Study	Accuracy (%)	Note
CK+ (laboratory)	Febrian et al. [25]	99.43	BiLSTM-CNN, with augmentation
	Jayaraman and Mahendran [26]	96.78	DCNN-BiLSTM, dual attention
JAFFE (laboratory)	Jayaraman and Mahendran [26]	95.87	
RAF-DB	Jayaraman and Mahendran [26]	95.78	Hybrid, high complexity
	Mao et al. [23]	92.21	43.7 M params, 8.4 GFLOPs
	Gao et al. [12]	92.13	26.99 M params, 5.97 GFLOPs
	Sun et al. [11]	91.30	Lightweight attention
	Wen et al. [20]	89.70	19.72 M params, 2.23 GFLOPs
	Zhao et al. [21]	88.40	
	Wang et al. [19]	88.14	
	Ma et al. [22]	88.14	
	Farzaneh and Qi [18]	87.78	Test set
FERPlus	Gao et al. [12]	91.68	
	Wang et al. [19]	89.35	
	Ma et al. [22]	88.81	
FER2013	Jayaraman and Mahendran [26]	82.89	
AffectNet, 7-class	Mao et al. [23]	67.49	
	Gao et al. [12]	66.38	
	Wen et al. [20]	65.69	
	Zhao et al. [21]	64.53	
AffectNet, 8-class	Sun et al. [11]	65.51	
	Mao et al. [23]	63.77	
	Wen et al. [20]	62.09	
	Gao et al. [12]	61.45	
	Zhao et al. [21]	60.29	
AffectNet, class setting not confirmed	Farzaneh and Qi [18]	65.20	Validation set
	Ma et al. [22]	61.85	
	Wang et al. [19]	60.23	
ABAW 2020 Track 2	Anas et al. [31]	50.77	Validation set; F1 29.16%

Values are as stated by the original authors. Protocols differ between studies.

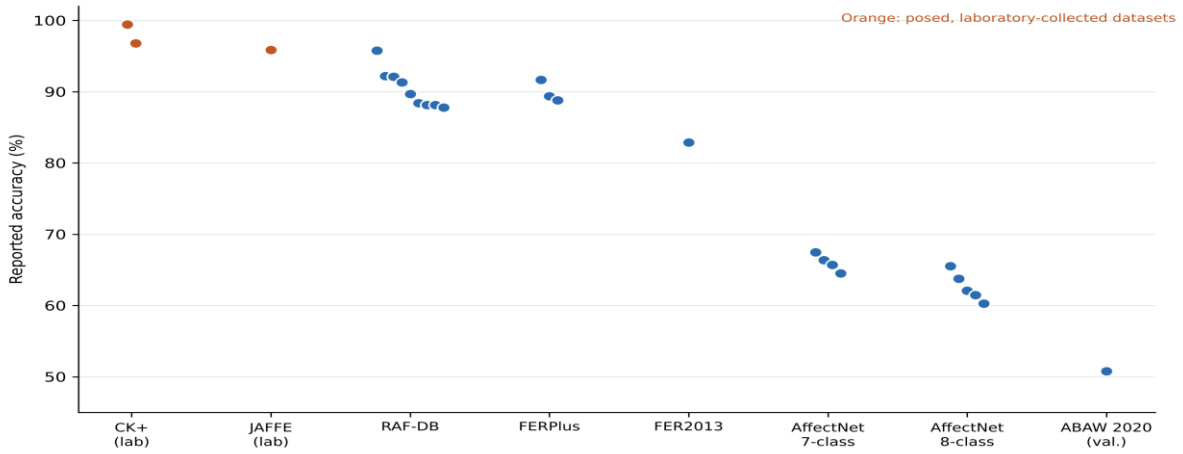

Figure 2. Reported accuracy by dataset and class setting. Each point is one study's headline result; protocols differ between studies.

5.1 Dataset and class setting

The method explains more of the spread than the data set does. The minimum CK+ score in Table 3 (96.78%) is approximately 29 points higher than the maximum 7-class AffectNet score (67.49%) and the maximum CK+ score is almost 32 points higher. The nine studies on RAF-DB span 8.0 points (87.78–95.78%) or 4.4 points when excluding the CNN–BiLSTM hybrid [26]. These within-dataset differences are minuscule in comparison to the difference between datasets, but they represent the margins on which superiority is usually claimed.

Class setting matters as well. Four models report both 7-class and 8-class AffectNet results, and each drops when the eighth class is added: DAN by 3.60 points (65.69 to 62.09), POSTER++ by 3.72 (67.49 to 63.77), Zhao et al. by 4.24 (64.53 to 60.29) and FERMam by 4.93 (66.38 to 61.45). A change of 3.6 to 4.9 points from the label set alone is larger than most of the differences between published methods on either setting.

5.2 Rank stability

Rankings do not always carry from one benchmark to another. FCCA [11] trails POSTER++ [23] on RAF-DB (91.30% against 92.21%) but leads it on 8-class AffectNet (65.51% against 63.77%). The two papers may not use identical AffectNet protocols, so we draw no conclusion about which model is better. The point is that a ranking built on one benchmark would have been reversed on the other. FERMam shows a similar pattern against DAN: it is ahead on RAF-DB (92.13% against 89.70%) and on 7-class AffectNet, but behind on 8-class AffectNet (61.45% against 62.09%).

5.3 Accuracy against cost

Three models in Table 3 come with parameter and FLOP counts, and they allow a same-dataset reading (Table 4). Going from DAN [20] to FERMam [12] raises parameters by 37% and FLOPs by a factor of 2.7, for 2.4 points on RAF-DB and 0.7 points on 7-class AffectNet. The next step, from FERMam to POSTER++ [23], adds a further 62% in parameters and 41% in FLOPs but only 0.08 points on RAF-DB and 1.1 points on 7-class AffectNet. Pretraining, augmentation and evaluation details are not identical across the three papers, so this is an indication rather than a controlled result. It does show the kind of accuracy–cost statement that only same-dataset reporting makes possible.

Table 4. Reported accuracy and cost for the three models with published parameter and FLOP counts.

Model	Params (M)	GFLOPs	RAF-DB (%)	AffectNet-7 (%)	AffectNet-8 (%)
DAN [20]	19.72	2.23	89.70	65.69	62.09
FERMam [12]	26.99	5.97	92.13	66.38	61.45
POSTER++ [23]	43.70	8.40	92.21	67.49	63.77

As stated by the original authors; protocols are not identical.

5.4 Saturation on laboratory data

Results on CK+ and JAFFE lie between 95.87% and 99.43%. At that level, differences of a few tenths of a point on a few hundred images are well within what the choice of split could produce, and the datasets separate modern architectures poorly.

6. A REPORTING CHECKLIST FOR COMPARABLE FER RESULTS

The observations above show that a headline accuracy cannot be interpreted without context that papers often omit. We therefore propose the checklist in Table 5. It is deliberately short. Each item can be answered in a sentence in the experimental-setup section, and each corresponds to a way in which published accuracies differ in the studies we reviewed.

Table 5. Twelve-item reporting checklist for FER studies.

#	Item	What to report
1	Dataset and version	Name, version and subset, for example FER2013 or FER+, and AffectNet 7-class or 8-class.
2	Label source	Original, relabelled or consensus labels, and the known level of label noise.
3	Split protocol	Official or custom split, proportions, random seed, and whether subjects overlap across splits.
4	Evaluation set	Whether results are on a validation or a held-out test set, and that no tuning used the test set.
5	Metrics and averaging	Overall accuracy and per-class precision, recall and F1, with macro or micro averaging stated.
6	Repeated runs	Mean and standard deviation over several seeds, or a statement that a single run was made.
7	Model size	Parameter count and FLOPs at the stated input resolution.
8	Latency and hardware	Inference time per image with batch size, device (CPU or GPU) and software versions.
9	Pretraining	Pretraining data and backbone, for example ImageNet or VGG-Face, or training from scratch.
10	Augmentation and test-time methods	Augmentation, class balancing, and any test-time augmentation or ensembling.
11	Same-split baselines	At least one simpler baseline trained under identical conditions.
12	Cross-dataset check	Evaluation on a dataset not used for training, or a statement that none was done.

Items 3 to 6 and 11 guard against inflated or irreproducible accuracy. Items 7 and 8 make efficiency claims checkable, which matters because a large share of the reviewed work aims at lightweight design [10]–[12]. Item 12 addresses the generalization question that single-dataset accuracy cannot answer.

7. DISCUSSION

From the dataset-aware view, we derive three practical implications. Researchers should be evaluating new architectures by comparing to baselines trained on the same split and reported with some measure of variability, not by numbers copied from previous papers with different protocols. The most striking numbers in Table 3 are from small laboratory datasets, where evidence for any method's superiority is weakest. A checklist for reviewers and editors can act as a quick screen: if a manuscript cannot answer questions 1 to 6, it does not yet support a ranking claim. For practitioners selecting a model to deploy, the accuracy-cost pairs in Table 4 are more informative than leaderboard position, as compute varies several-fold between designs while accuracy varies only a few points.

The review gives rise to five directions. The first one consists of shared evaluation protocols with fixed splits and released seeds for the most used datasets in the wild. The second is to regularly report parameters, FLOPs, and measured latency on stated hardware, especially for lightweight models. The third is more in-depth analysis of label noise and class setting differences which currently obscure comparisons on FER2013 and AffectNet. Fourth, the cross-dataset evaluation as a standard step instead of optional extra. The fifth are low-resource attention and state-space modules, as initiated in [11] and [12].

8. LIMITATIONS

This review has several limitations. It is a narrative review with a structured but not an exhaustive search, thus relevant studies may be missing. The 22 included studies tend to open access and English language sources. We took numbers as given by original authors, mostly from abstract and main results, and did not reproduce them. In some cases the class setting or evaluation subset underlying a figure could not be confirmed. Therefore, Table 3 has a separate group for results of unconfirmed class setting. A corresponding headline number is not tabulated for many of the studies [10], [13]–[17], [24], [27]–[30]. Thus the dataset-aware comparison is limited to a small portion of the reviewed work. Pretraining, augmentation, and compute budgets vary across studies, making same-dataset comparisons only indicative. Also the field is moving fast and work published after the search closed is not covered.

9. CONCLUSION

Recent FER research based on CNN provides lightweight, attention-based, transformer, hybrid, fusion and transfer-learning designs with steadily rising reported accuracy. This review shows that the dataset and even the class setting within a dataset contribute more to the variation in reported accuracy than does architecture. ~30 points between laboratory and in-the-wild benchmarks. Moving from 7 to 8 classes in AffectNet moves a model's score by 3.6 to 4.9 points. Rankings can change between datasets and the last few bits of accuracy can come at a large increase in parameters and compute. We recommend comparing FER results within a dataset and class situation, stating the protocol and efficiency, and we propose a twelve-item checklist to aid that practice. If used consistently, such reporting would allow the field to separate real architectural progress from variations in data and protocol.

REFERENCES

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Trans. Affective Computing, vol. 13, no. 3, pp. 1195–1215, 2022, doi: 10.1109/TAFFC.2020.2981446.

[2] P. Ekman and W. V. Friesen, Unmasking the Face. Englewood Cliffs, NJ, USA: Prentice-Hall, 1975.

[3] I. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59–63, 2015, doi: 10.1016/j.neunet.2014.09.005.

[4] P. Lucey et al., "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in Proc. IEEE CVPR Workshops, 2010, pp. 94–101.

[5] M. J. Lyons et al., "Japanese Female Facial Expression (JAFFE) Database," Zenodo, 1998, doi: 10.5281/zenodo.3451524.

[6] S. Li et al., "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in Proc. IEEE CVPR, 2017, pp. 2584–2593, doi: 10.1109/CVPR.2017.277.

[7] A. Mollahosseini et al., "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Trans. Affective Computing, vol. 10, no. 1, pp. 18–31, 2019, doi: 10.1109/TAFFC.2017.2740923.

[8] E. Barsoum et al., "Training deep networks for facial expression recognition with crowd-sourced label distribution," in Proc. ACM Int. Conf. Multimodal Interaction (ICMI), 2016, pp. 279–283, doi: 10.1145/2993148.2993165.

[9] A. Dhall et al., "Collecting large, richly annotated facial-expression databases from movies," IEEE MultiMedia, vol. 19, no. 3, pp. 34–41, 2012, doi: 10.1109/MMUL.2012.26.

[10] N. Zhou et al., "A lightweight convolutional neural network for real-time facial expression detection," IEEE Access, vol. 9, pp. 5573–5584, 2021, doi: 10.1109/ACCESS.2020.3046715.

[11] R. Sun et al., "FCCA: Fast Center Consistency Attention for facial expression recognition," Electronics, vol. 14, no. 6, art. 1057, 2025, doi: 10.3390/electronics14061057.

[12] C. Gao et al., "FERMam: A lightweight dual-source and multi-scale fusion framework for facial expression recognition," Scientific Reports,

vol. 16, art. 13826, 2026, doi: 10.1038/s41598-026-44396-6.

[13] M. Mahmoudi et al., "Kernelized dense layers for facial expression recognition," in Proc. IEEE Int. Conf. Image Processing (ICIP), 2020, doi: 10.1109/ICIP40778.2020.9190694.

[14] J. Li et al., "Attention mechanism-based CNN for facial expression recognition," Neurocomputing, vol. 411, pp. 340–350, 2020, doi: 10.1016/j.neucom.2020.06.014.

[15] J.-W. Liu et al., "Multiscale wavelet attention convolutional network for facial expression recognition," Scientific Reports, vol. 15, art. 22219, 2025, doi: 10.1038/s41598-025-07416-5.

[16] G. Yu, "A multi-granularity feature fusion approach with attention for facial expression recognition," Scientific Reports, vol. 15, 2025, doi: 10.1038/s41598-025-26533-9.

[17] K. Wang et al., "Region attention networks for pose and occlusion robust facial expression recognition," IEEE Trans. Image Processing, vol. 29, pp. 4057–4069, 2020, doi: 10.1109/TIP.2019.2956143.

[18] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), 2021, pp. 2402–2411.

[19] K. Wang et al., "Suppressing uncertainties for large-scale facial expression recognition," in Proc. IEEE/CVF CVPR, 2020, pp. 6897–6906.

[20] Z. Wen et al., "Distract your attention: Multi-head cross attention network for facial expression recognition," Biomimetics, vol. 8, no. 2, art. 199, 2023, doi: 10.3390/biomimetics8020199.

[21] Z. Zhao et al., "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," IEEE Trans. Image Processing, vol. 30, 2021, doi: 10.1109/TIP.2021.3093397.

[22] F. Ma et al., "Facial expression recognition with visual transformers and attentional selective fusion," arXiv:2103.16854, doi: 10.1109/TAFFC.2021.3122146.

[23] J. Mao et al., "POSTER++: A simpler and stronger facial expression recognition network," arXiv:2301.12149, 2023.

[24] D. Liu et al., "Video-based facial expression recognition using graph convolutional networks," in Proc. Int. Conf. Pattern Recognition (ICPR), 2020, arXiv:2010.13386.

[25] R. Febrian et al., "Facial expression recognition using bidirectional LSTM-CNN," Procedia Computer Science, vol. 216, pp. 39–47, 2023, doi: 10.1016/j.procs.2022.12.109.

[26] S. Jayaraman and A. Mahendran, "An interactive information based DCNN-BiLSTM model with dual attention mechanism for facial expression recognition," Scientific Reports, vol. 15, art. 26287, 2025, doi: 10.1038/s41598-025-09709-1.

[27] C. Liu et al., "Two-channel feature extraction convolutional neural network for facial expression recognition," J. Advanced Computational Intelligence and Intelligent Informatics, vol. 24, no. 6, pp. 792–801, 2020, doi: 10.20965/jaciii.2020.p0792.

[28] T. Podder et al., "Facial expression recognition in real time with domain adaptive transfer learning and feature integrated CNN," Neural Processing Letters, vol. 58, art. 19, 2026, doi: 10.1007/s11063-025-11824-z.

[29] R. Verschae and I. Bugueño-Córdova, "evTransFER: A transfer learning framework for event-based facial expression recognition," Neurocomputing, vol. 671, art. 132641, 2026, doi: 10.1016/j.neucom.2026.132641.

[30] S. Begaj et al., "Emotion recognition based on facial expressions using convolutional neural network (CNN)," in Proc. Int. Conf. Computing, Networking, Telecommunications and Engineering Sciences Applications (CoNTESA), 2020, pp. 58–63, doi: 10.1109/CoNTESA50436.2020.9302866.

[31] H. Anas et al., "Deep convolutional neural network based facial expression recognition in the wild," arXiv:2010.01301, 2020.