Technical University of Cluj - Napoca
Computer Science Department

Interfete om-calculator

Recunoasterea expresiei faciale (FER)

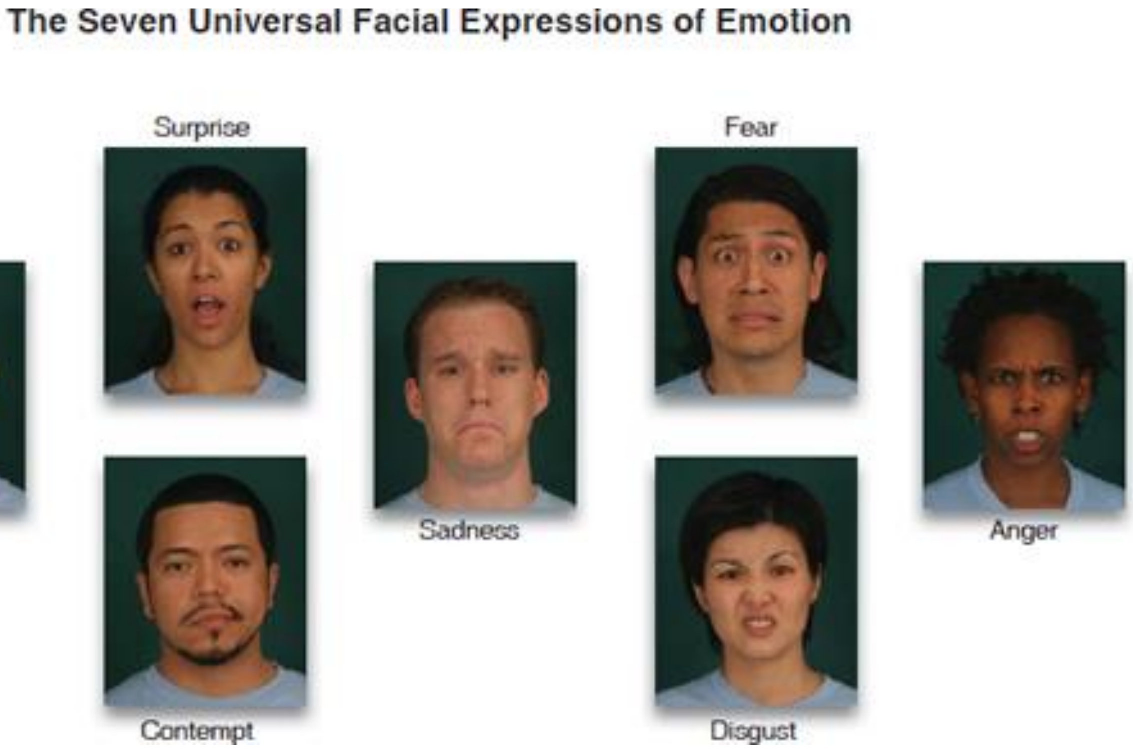
References

[1] LI, Shan; DENG, Weihong. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

https://arxiv.org/abs/1804.08348

Ekman and Friesen - defined six /seven basic emotions: anger, disgust, fear, happiness, sadness, and surprise (+ contempt (dispret))

⇒ culture specific



[2] https://leb.fbi.gov/image-repository/truth_8.jpg/view

Overview

Facial expression - one of the most powerful, natural and universal signals for human beings to convey their emotional states and intentions

Usage: sociable robotics, medical treatment, driver fatigue surveillance, affective computing and many other human-computer interaction systems

FER methods:

- *static-based methods* – feature representation is encoded with only spatial information from a single image
- *dynamic-based methods* - temporal relation among contiguous frames in the input facial expression sequence
- *multimodal systems* – additional sensorial channels: audio , physiological (HR, BP, EEG probes etc.).

Overview

Traditional methods: handcrafted features (shallow learning)

local binary patterns (LBP)

LBP on three orthogonal planes (LBP-TOP)

non-negative matrix factorization (NMF)

Emotion recognition competitions: FER2013, Emotion Recognition in the Wild (EmotiW) ⇒ collected relatively sufficient training data on challenging real-world scenarios

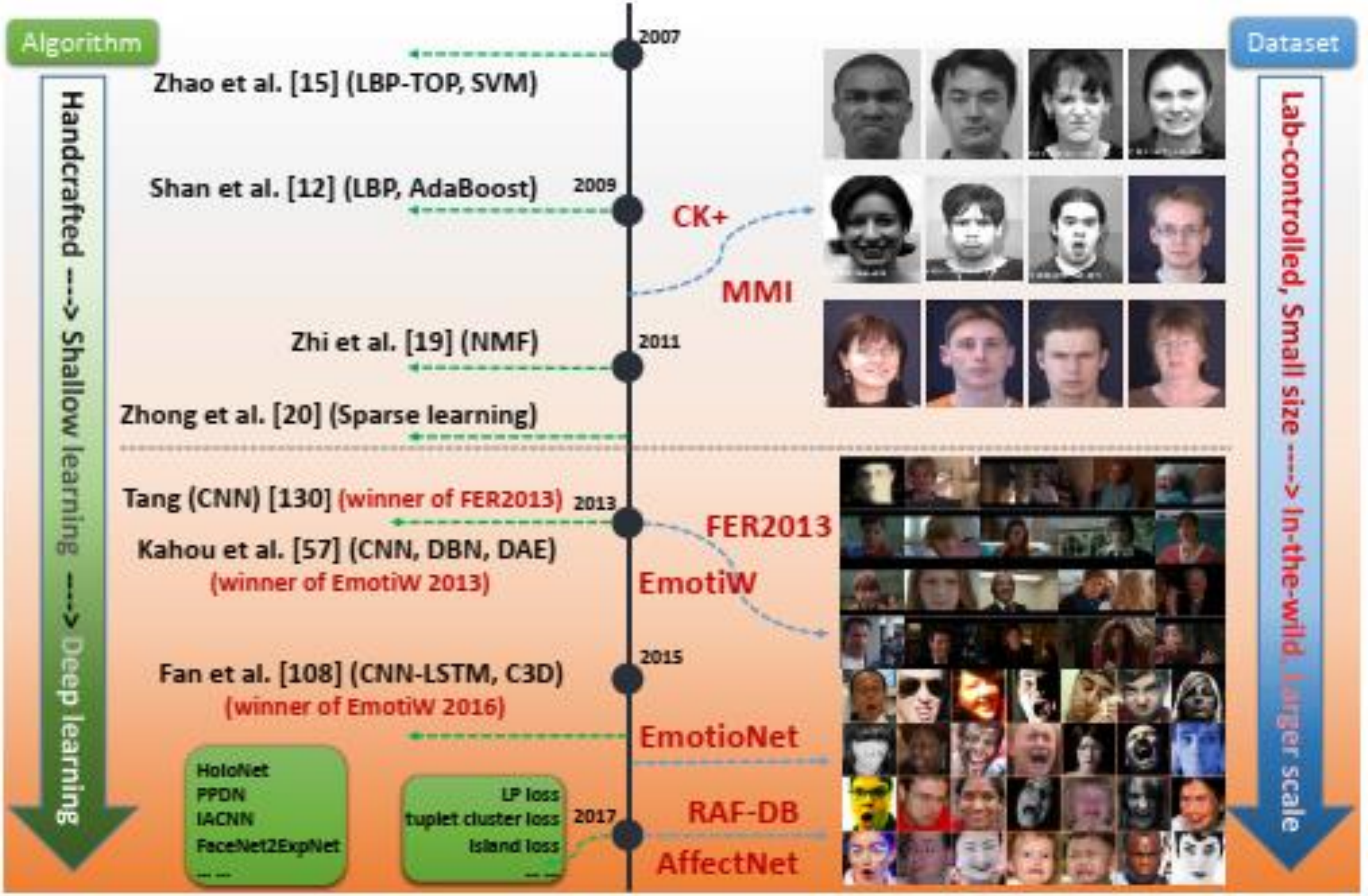
⇒ transition of FER from lab-controlled to in-the-wild settings

⇒ transition from traditional to deep learning methods

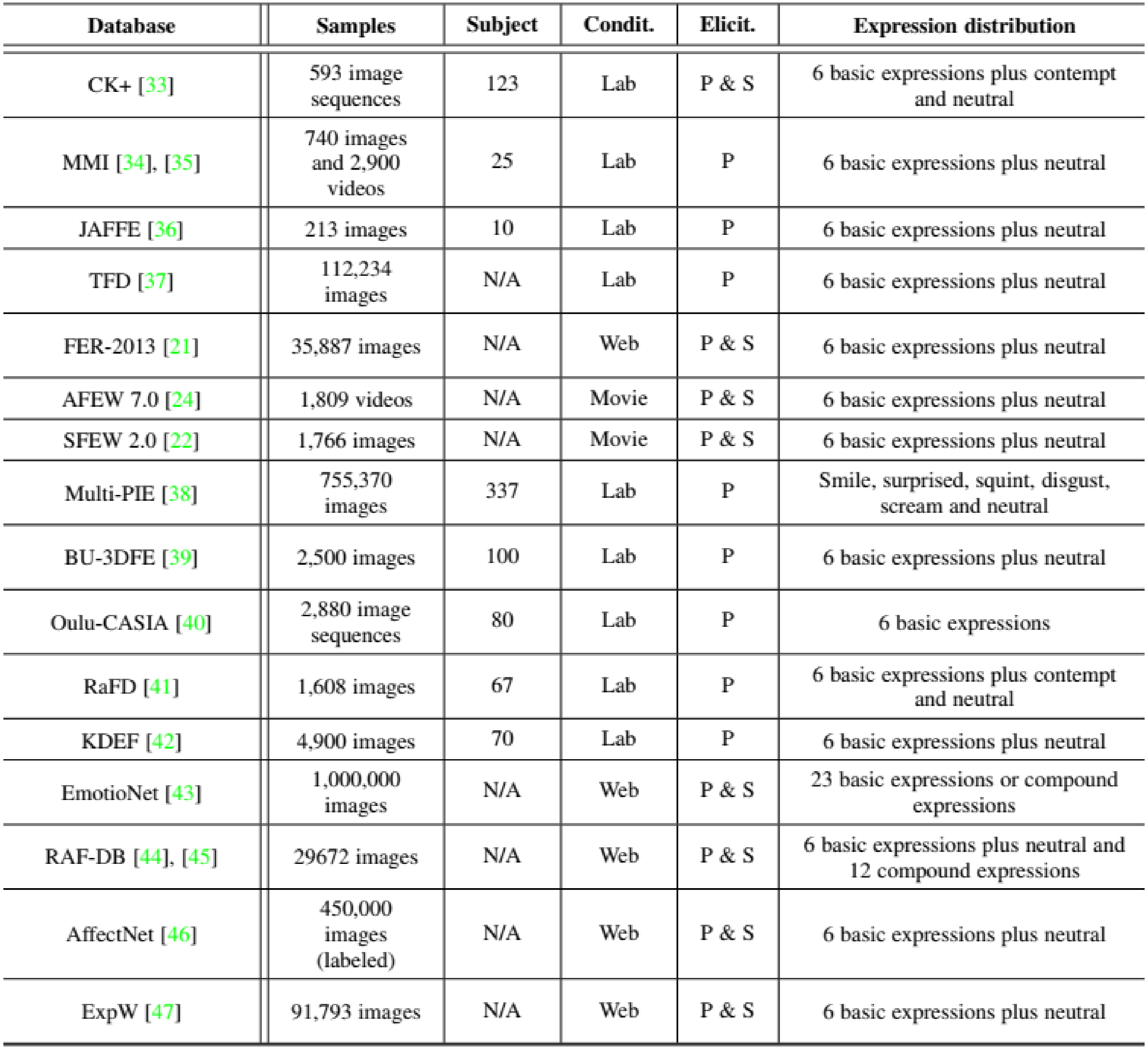
Challenges:

- high inter-subject variations due to different personal attributes (age, gender, race / ethnic backgrounds, level of expressiveness)
- subject identity bias, variations in pose, illumination and occlusions (unconstrained facial expression scenarios)

FER evolution (datasets + methods)


Algorithm
Dataset
Handcrafted ---> Shallow learning ---> Deep learning
Lab-controlled, Small size ---> In-the-wild, Larger scale
2007
Zhao et al. [15] (LBP-TOP, SVM)
Shan et al. [12] (LBP, AdaBoost)
2009
CK+
MMI
Zhi et al. [19] (NMF)
2011
Zhong et al. [20] (Sparse learning)
Tang (CNN) [130] (winner of FER2013)
2013
FER2013
Kahou et al. [57] (CNN, DBN, DAE)
(winner of EmotiW 2013)
EmotiW
2015
Fan et al. [108] (CNN-LSTM, C3D)
(winner of EmotiW 2016)
EmotioNet
HoloNet
PPDN
IACNN
FaceNet2ExpNet
--- ---
LP loss
tuplet cluster loss
Island loss
--- ---
2017
RAF-DB
AffectNet

FER datasets



Database	Samples	Subject	Condit.	Elicit.	Expression distribution
CK+ [33]	593 image sequences	123	Lab	P & S	6 basic expressions plus contempt and neutral
MMI [34], [35]	740 images and 2,900 videos	25	Lab	P	6 basic expressions plus neutral
JAFFE [36]	213 images	10	Lab	P	6 basic expressions plus neutral
TFD [37]	112,234 images	N/A	Lab	P	6 basic expressions plus neutral
FER-2013 [21]	35,887 images	N/A	Web	P & S	6 basic expressions plus neutral
AFEW 7.0 [24]	1,809 videos	N/A	Movie	P & S	6 basic expressions plus neutral
SFEW 2.0 [22]	1,766 images	N/A	Movie	P & S	6 basic expressions plus neutral
Multi-PIE [38]	755,370 images	337	Lab	P	Smile, surprised, squint, disgust, scream and neutral
BU-3DFE [39]	2,500 images	100	Lab	P	6 basic expressions plus neutral
Oulu-CASIA [40]	2,880 image sequences	80	Lab	P	6 basic expressions
RaFD [41]	1,608 images	67	Lab	P	6 basic expressions plus contempt and neutral
KDEF [42]	4,900 images	70	Lab	P	6 basic expressions plus neutral
EmotioNet [43]	1,000,000 images	N/A	Web	P & S	23 basic expressions or compound expressions
RAF-DB [44], [45]	29672 images	N/A	Web	P & S	6 basic expressions plus neutral and 12 compound expressions
AffectNet [46]	450,000 images (labeled)	N/A	Web	P & S	6 basic expressions plus neutral
ExpW [47]	91,793 images	N/A	Web	P & S	6 basic expressions plus neutral

Elicit.(Elicitation method)

- P = posed;
- S = spontaneous;
- Condit. = Collection condition

FER datasets

CK+ The Extended CohnKanade (CK+) database

- most extensively used laboratory-controlled database for evaluating FER
- 593 video sequences from 123 subjects (10 … 60 frames) shift from a neutral facial expression to the peak expression.
- 327 sequences from 118 subjects are labeled with seven basic expression labels (anger, contempt, disgust, fear, happiness, sadness, and surprise) based on the Facial Action Coding System (FACS).
- CK+ does not provide specified training, validation and test sets, ⇒ algorithms evaluated on this database are not uniform.

MMI

- laboratory-controlled 326 sequences from 32 subjects.
- 213 sequences are labeled with six basic expressions (no contempt”),
- 205 sequences are captured in frontal view.
- onset-apex-offset labeled sequences (begin with a neutral expression and reaches peak near the middle before returning to the neutral)
- MMI has more challenging conditions: large inter-personal variations (subjects perform the same expression non-uniformly and many of them wear accessories (e.g., glasses, mustache).

FER datasets

JAFFE (Japanese Female Facial Expression)

- laboratory-controlled image database (213 samples of posed expressions from 10 Japanese females).
- Each person has 3..4 images with each of six basic facial expressions (anger, disgust, fear, happiness, sadness, and surprise) + one image with a neutral expression.
- challenging due to few examples per subject/expression

TFD (Toronto Face Database)

- amalgamation of several facial expression datasets: 112,234 images, 4,178 of which are annotated with one of seven expression labels: anger, disgust, fear, happiness, sadness, surprise and neutral.
- The faces are localized and normalized to a size of 48*48 such that all the subjects eyes are the same distance apart and have the same vertical coordinates.
- Five official folds are provided; each fold contains a training, validation, and test set consisting of 70%, 10%, and 20% of the images, respectively.

FER datasets

FER2013

ICML 2013 Challenges in Representation Learning

- large-scale and unconstrained database collected automatically by the Google image search API.
- Images are registered and resized to 48*48 pixels after rejecting wrongfully labeled frames and adjusting the cropped region.
- 28,709 training images, 3,589 validation images and 3,589 test images with seven expression labels (anger, disgust, fear, happiness, sadness, surprise and neutral)

AFEW (Acted Facial Expressions in the Wild)

- evaluation platform for the annual Emotion Recognition In The Wild Challenge (EmotiW) since 2013.
- video clips from movies with spontaneous expressions, various head poses, occlusions and illuminations.
- temporal and multimodal database (audio and video).
- seven expressions labeled: anger, disgust, fear, happiness, sadness, surprise and neutral.
- **AFEW 7.0** in EmotiW 2017: Train (773 samples), Val (383 samples) and Test (653 samples)

FER datasets

EmotioNet [43]:

- one million facial expression images collected from the Internet.
- 950,000 images were annotated by the automatic action unit (AU)
- the remaining 25,000 images were manually annotated with 11 AUs. the EmotioNet Challenge provides six basic expressions and ten compound expressions, and 2,478 images with expression labels are available.

RAF-DB (The Real-world Affective Face Database)

- real-world database that contains 29,672 highly diverse facial images downloaded from the Internet.
- manually crowd-sourced annotation and reliable estimation
- Seven basic and eleven compound emotion labels are provided
- 15,339 images from the basic emotion set are divided into two groups (12,271 training samples and 3,068 testing samples) for evaluation

FER datasets

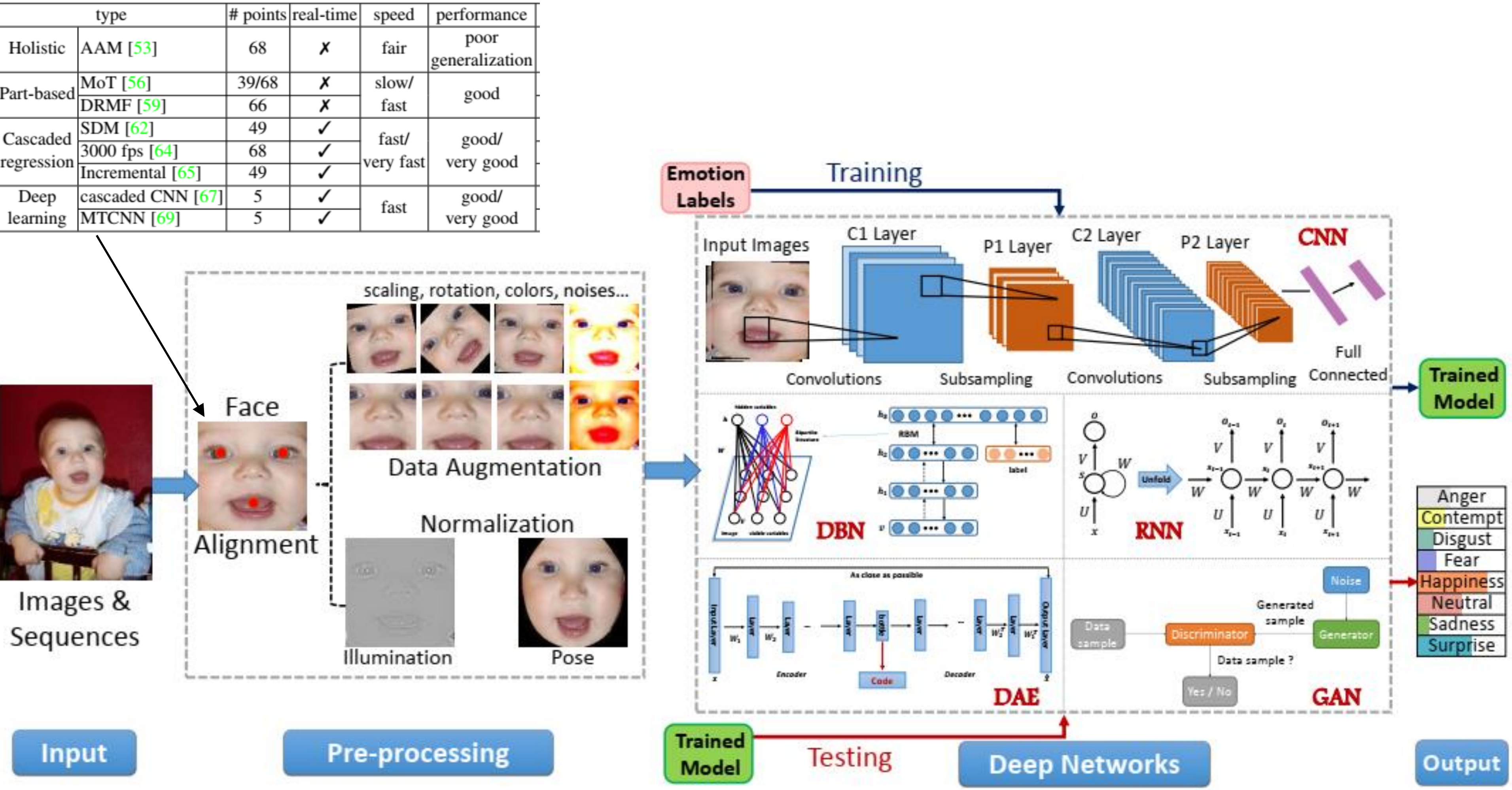
AffectNet

- one million images from the Internet that were obtained by querying different search engines using emotion-related tags.
- the largest database that provides facial expressions in two different emotion models (categorical model and dimensional model), of which 450,000 images have manually annotated labels for eight basic expressions.

ExpW (The Expression in-the-Wild Database)

- 91,793 faces downloaded using Google image search.
  Each of the face images was manually annotated as one of the seven basic expression categories.
- Non-face images were removed in the annotation process

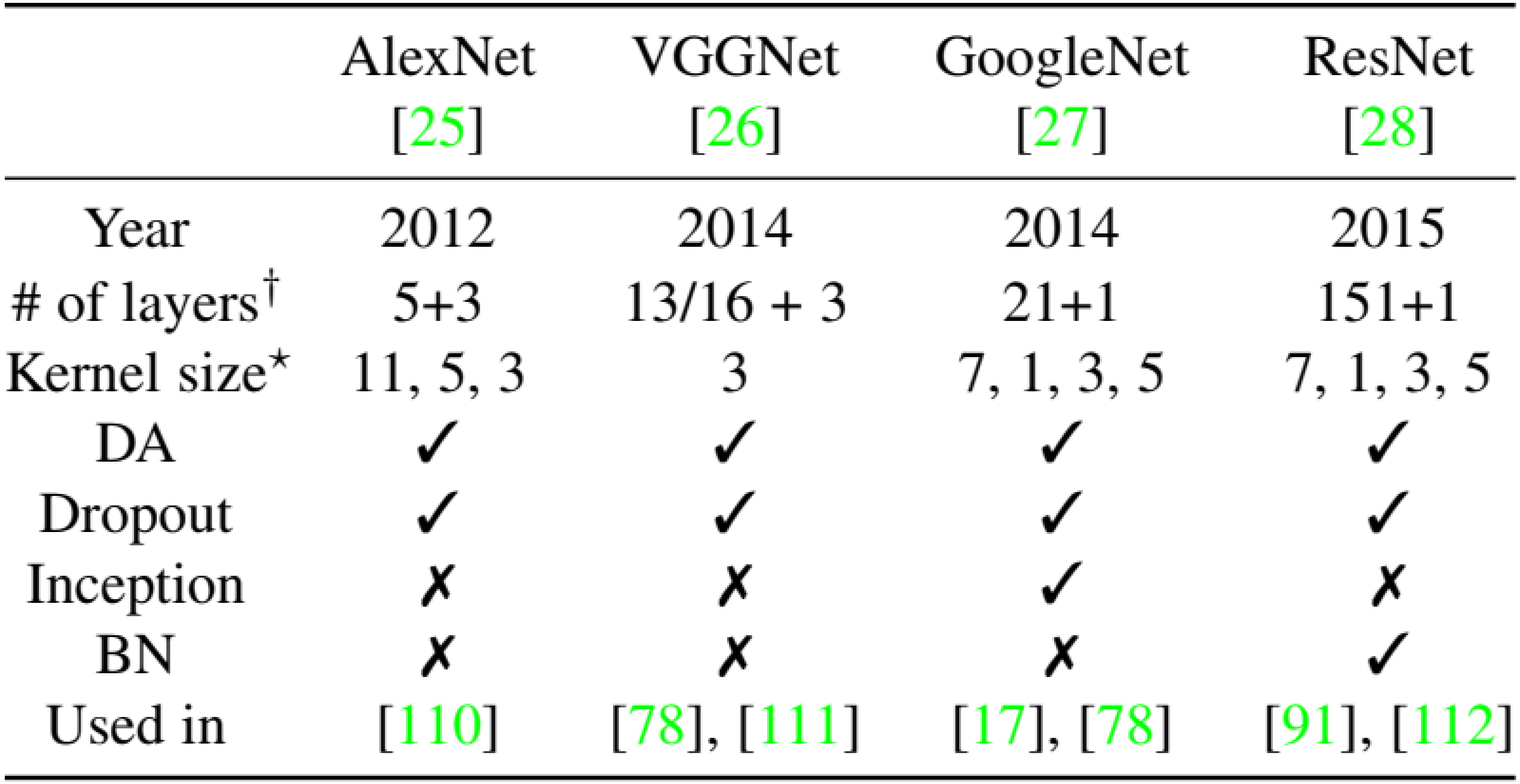
Deep facial expression recognition



type		# points	real-time	speed	performance
Holistic	AAM [53]	68	✗	fair	poor generalization
Part-based	MoT [56]	39/68	✗	slow/ fast	good
	DRMF [59]	66	✗		
Cascaded regression	SDM [62]	49	✓	fast/ very fast	good/ very good
	3000 fps [64]	68	✓		
	Incremental [65]	49	✓		
Deep learning	cascaded CNN [67]	5	✓	fast	good/ very good
	MTCNN [69]	5	✓		

Feature map extraction | Feature extraxction + Classification [1]

CNN



	AlexNet [25]	VGGNet [26]	GoogleNet [27]	ResNet [28]
Year	2012	2014	2014	2015
# of layers[†]	5+3	13/16 + 3	21+1	151+1
Kernel size[*]	11, 5, 3	3	7, 1, 3, 5	7, 1, 3, 5
DA	✓	✓	✓	✓
Dropout	✓	✓	✓	✓
Inception	✗	✗	✓	✗
BN	✗	✗	✗	✓
Used in	[110]	[78], [111]	[17], [78]	[91], [112]

CNN based approaches [1]

Deep facial expression recognition

Facial expression classification

- end-2-end way: a loss layer is added to the end of the network (to regulate the back-propagation error)

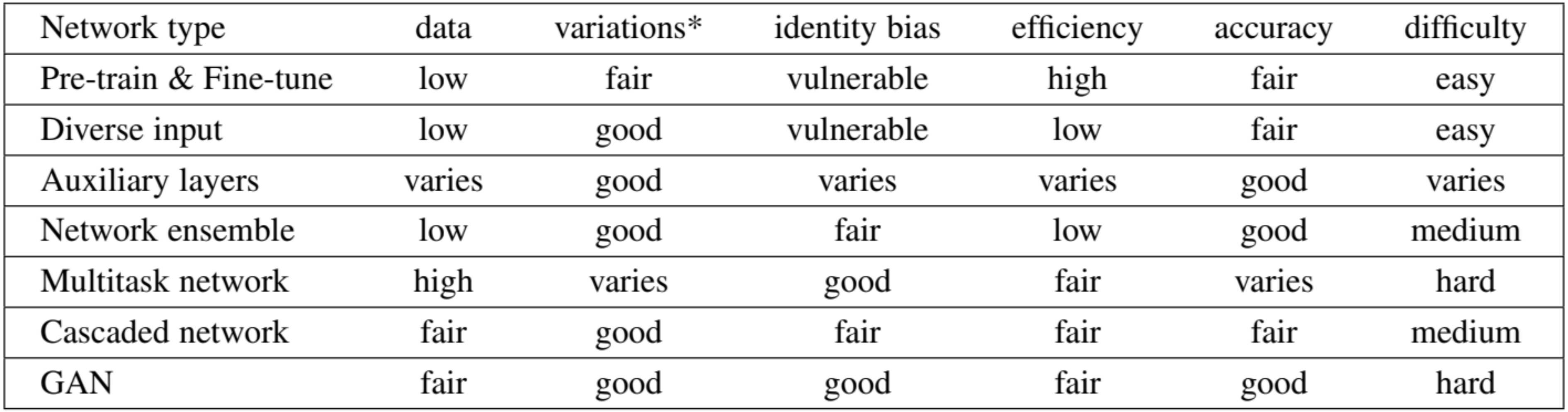
CNN: softmax loss
SVM loss
neural forests (NFs)

- additional independent classifiers

support vector machine
random forest

State of the art

Deep FER networks for static images

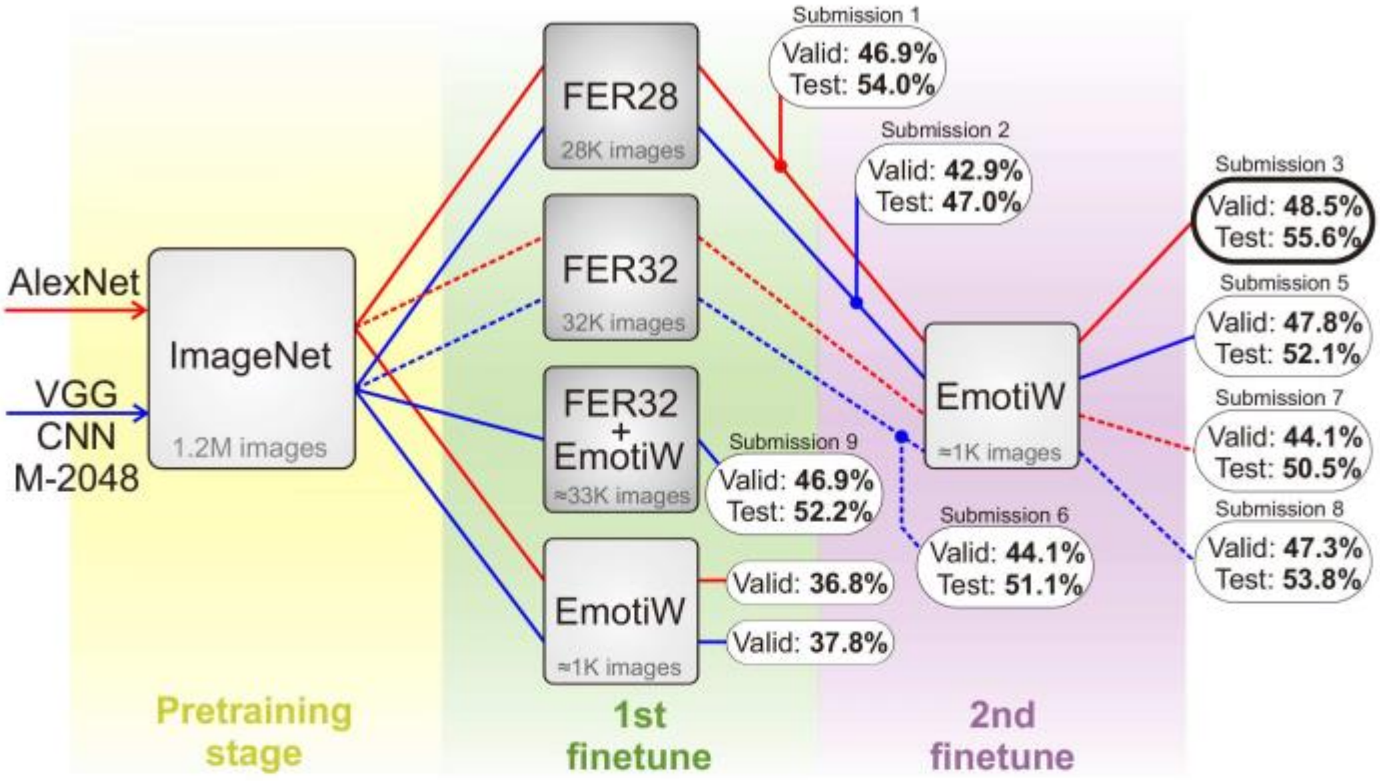


Network type	data	variations*	identity bias	efficiency	accuracy	difficulty
Pre-train & Fine-tune	low	fair	vulnerable	high	fair	easy
Diverse input	low	good	vulnerable	low	fair	easy
Auxiliary layers	varies	good	varies	varies	good	varies
Network ensemble	low	good	fair	low	good	medium
Multitask network	high	varies	good	fair	varies	hard
Cascaded network	fair	good	fair	fair	fair	medium
GAN	fair	good	good	fair	good	hard

Comparison of different types of methods for static images in terms of data size requirement, variations* (head pose, illumination, occlusion and other environment factors), identity bias, computational efficiency, accuracy, and difficulty on network training [1]

State of the art

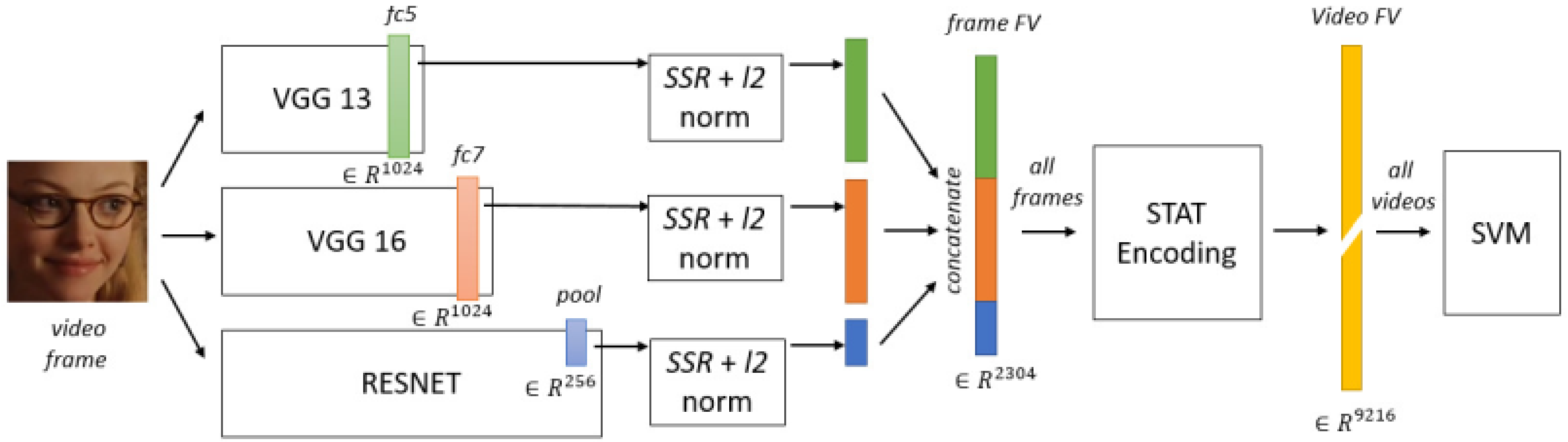
Deep FER networks for static images



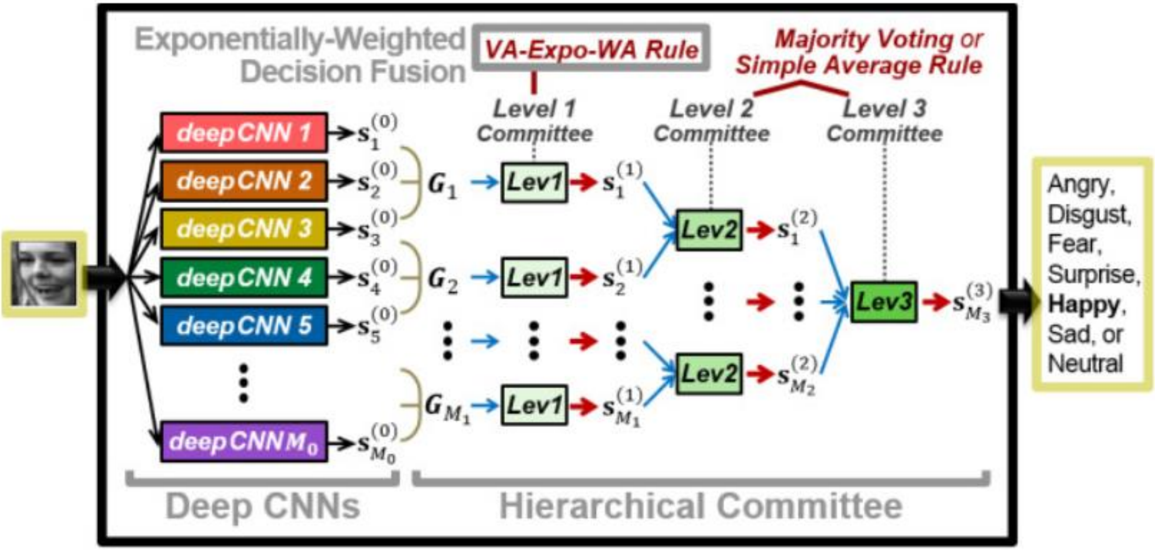
Pre-training and fine-tuning [1]

State of the art

Deep FER networks for static images - *Network ensemble [1]*



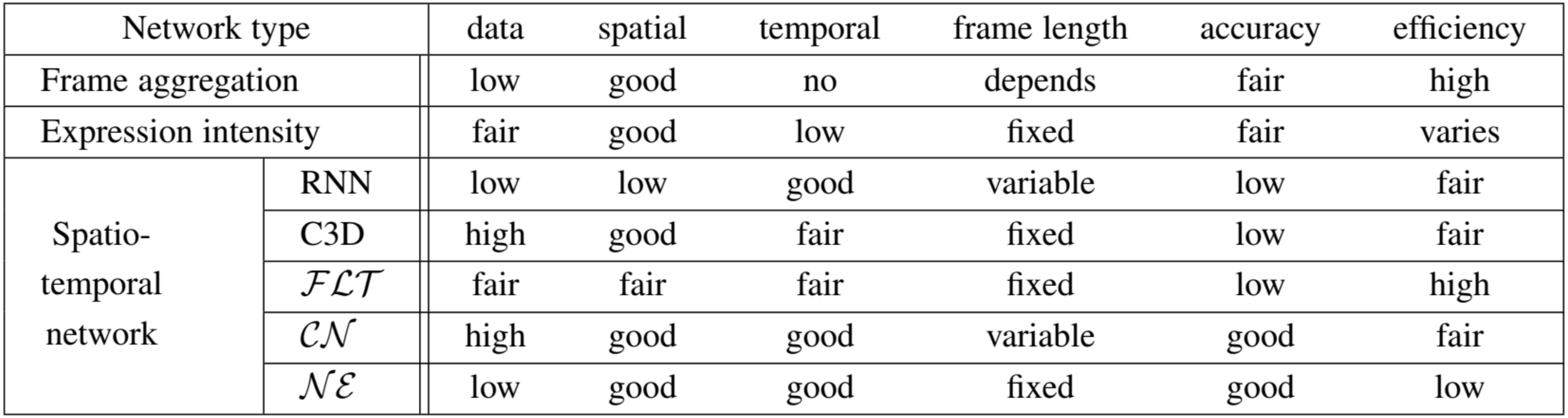
Feature-level ensemble [1]



Decision-level ensemble [1]

State of the art

Deep FER networks for dynamic image sequences



Network type		data	spatial	temporal	frame length	accuracy	efficiency
Frame aggregation		low	good	no	depends	fair	high
Expression intensity		fair	good	low	fixed	fair	varies
Spatio-temporal network	RNN	low	low	good	variable	low	fair
	C3D	high	good	fair	fixed	low	fair
	$\mathcal{FLT}$	fair	fair	fair	fixed	low	high
	$\mathcal{CN}$	high	good	good	variable	good	fair
	$\mathcal{NE}$	low	good	good	fixed	good	low

Comparison of different types of methods for dynamic image sequences in terms of data size requirement, representability of spatial and temporal information, requirement on frame length, performance, and computational efficiency. *FLT* = Facial Landmark Trajectory; *CN* = Cascaded Network; *N E* = Network Ensemble [1]