BAB 2
LANDASAN TEORI

2.1 Facial Expression Recognition

Facial Expression Recognition (FER) merupakan cabang penting dalam bidang Computer Vision yang bertujuan mengidentifikasi emosi manusia melalui analisis ekspresi wajah secara otomatis [3, 5]. Teknologi ini memiliki peran fundamental dalam memahami komunikasi non-verbal manusia, di mana ekspresi wajah menyampaikan informasi emosional yang tidak dapat diungkapkan hanya melalui kata-kata [6]. Dalam beberapa dekade terakhir, FER telah berkembang dari metode berbasis *handcrafted features* seperti Local Binary Patterns (LBP) dan Histogram of Oriented Gradients (HOG) menjadi pendekatan berbasis *deep learning* yang mampu mengekstraksi fitur secara otomatis dengan akurasi yang lebih tinggi [16, 17, 18].

Menurut teori Ekman, terdapat tujuh emosi dasar universal yang dapat dikenali melalui ekspresi wajah di seluruh budaya manusia, yaitu *happy* (bahagia), *sad* (sedih), *angry* (marah), *fear* (takut), *surprise* (terkejut), *disgust* (jijik), dan *neutral* (netral) [6]. Setiap emosi memiliki karakteristik visual yang distingtif pada region wajah tertentu, seperti sudut mulut yang terangkat untuk ekspresi bahagia atau alis yang mengerut untuk ekspresi marah. Pemahaman terhadap karakteristik visual ini menjadi dasar pengembangan sistem FER yang akurat dan robust.

Aplikasi FER sangat luas dan mencakup berbagai domain [5, 16]. Dalam bidang pendidikan, FER digunakan untuk memantau keterlibatan dan kondisi emosional siswa selama pembelajaran daring, memungkinkan instruktur memberikan intervensi yang tepat waktu [7, 11, 19, 2]. Dalam layanan konsumen, sistem FER dapat menganalisis kepuasan pelanggan secara *real-time* untuk meningkatkan kualitas layanan. Di bidang kesehatan mental, teknologi ini mendukung deteksi dini kondisi psikologis seperti depresi atau kecemasan [20, 21, 10, 22]. Aplikasi lainnya mencakup pengembangan sistem Human-Computer Interaction yang responsif, analisis perilaku dalam penelitian psikologi, dan sistem keamanan untuk deteksi emosi mencurigakan.

Meskipun telah mengalami kemajuan signifikan, FER masih menghadapi berbagai tantangan fundamental [3, 5]. Tantangan utama meliputi *class imbalance* pada dataset di mana beberapa emosi memiliki jumlah sampel jauh lebih banyak

dibanding emosi lain, *noise label* akibat subjektivitas anotasi manusia, variasi pencahayaan dan pose wajah, oklusi parsial seperti wajah tertutup tangan atau rambut, serta perbedaan demografis yang dapat mempengaruhi ekspresi wajah [8, 6, 23, 17]. Selain itu, mayoritas sistem FER menggunakan model *black box* yang tidak menyediakan penjelasan tentang dasar keputusan klasifikasi, sehingga sulit untuk memvalidasi apakah model benar-benar belajar dari fitur wajah yang relevan atau hanya mengandalkan artefak pada data [16].

2.2 MobileNetV2

MobileNetV2 adalah arsitektur *deep learning* yang dikembangkan oleh Google pada tahun 2018 untuk aplikasi *mobile* dan *embedded devices* yang memerlukan efisiensi komputasi tinggi tanpa mengorbankan akurasi secara signifikan. Arsitektur ini telah menjadi salah satu model *lightweight* yang paling populer dan banyak digunakan dalam penelitian Facial Expression Recognition berkat keseimbangan optimal antara akurasi dan efisiensi komputasi [3, 15, 24]. MobileNetV2 merupakan evolusi dari MobileNetV1 dengan penambahan *inverted residual blocks* dan *linear bottleneck layers* yang meningkatkan representasi fitur sekaligus mengurangi kehilangan informasi selama proses *forward propagation*.

Inovasi utama MobileNetV2 terletak pada penggunaan *inverted residual blocks* dengan struktur *expand–depthwise–project*. Berbeda dengan *residual blocks* pada arsitektur ResNet yang mempersempit dimensi *channel* terlebih dahulu, MobileNetV2 justru memperlebar *channel* dengan faktor ekspansi t (umumnya $t = 6$) sebelum melakukan *depthwise convolution*, lalu memproyeksikan kembali ke dimensi rendah [3, 15]. Operasi pada setiap blok tanpa *skip connection* dapat diformulasikan sebagai [15]:

$$\mathbf{y} = \mathrm{PW}_{\mathrm{proj}}\big(\mathrm{DW}\big(\mathrm{PW}_{\mathrm{exp}}(\mathbf{x})\big)\big) \tag{2.1}$$

di mana $\mathbf{x}$ adalah *input feature map*, $\mathrm{PW}_{\mathrm{exp}}$ adalah *pointwise convolution* untuk ekspansi *channel*, DW adalah *depthwise convolution* 3×3, dan $\mathrm{PW}_{\mathrm{proj}}$ adalah *pointwise convolution* untuk proyeksi kembali ke dimensi rendah. Apabila dimensi input dan output sama, ditambahkan *skip connection* sehingga:

$$\mathbf{y} = \mathbf{x} + \mathrm{PW}_{\mathrm{proj}}\big(\mathrm{DW}\big(\mathrm{PW}_{\mathrm{exp}}(\mathbf{x})\big)\big) \tag{2.2}$$

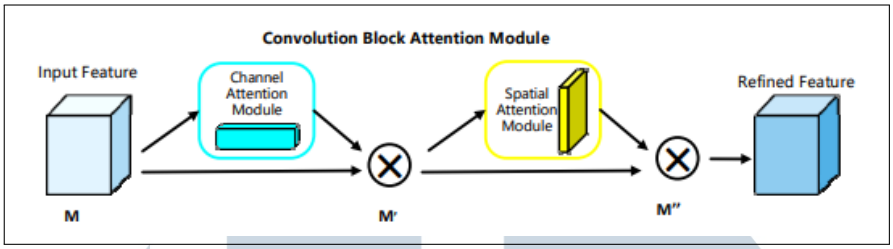
Komponen kunci dari setiap blok adalah *depthwise separable convolution*

yang memisahkan operasi spasial dan penggabungan fitur menjadi dua tahap. *Depthwise convolution* memproses setiap *channel* input secara independen dengan satu *kernel* berukuran 3×3, sedangkan *pointwise convolution* (1×1) menggabungkan hasilnya lintas *channel*. Pemisahan ini mengurangi kompleksitas komputasi sekitar 8–9 kali lipat dibandingkan konvolusi standar untuk *kernel* 3×3 [3, 15].

MobileNetV2 menggunakan aktivasi ReLU6 pada expansion layer untuk meningkatkan robustness pada representasi fixed-point, sedangkan pada *projection layer* digunakan aktivasi linear untuk mencegah kehilangan informasi akibat non-linearitas pada ruang dimensi rendah. Dalam konteks penelitian ini, MobileNetV2 digunakan sebagai *backbone* untuk mengekstraksi *feature map* dari gambar wajah dengan bobot *pretrained* ImageNet, menghasilkan representasi fitur berukuran $[B, 1280, 7, 7]$ yang kemudian diproses lebih lanjut oleh CBAM [13, 11]. Arsitektur *lightweight* ini memungkinkan *training* yang lebih cepat dan penggunaan memori yang lebih efisien, dengan total parameter sekitar 2,2 juta dibandingkan arsitektur *heavy* seperti ResNet50 yang memiliki 25 juta parameter [3, 24, 25, 26].

2.3 Convolutional Block Attention Module (CBAM)

Attention mechanism telah menjadi komponen penting dalam arsitektur *deep learning* modern, memungkinkan model untuk fokus pada informasi yang paling relevan dan mengabaikan *noise* atau informasi yang kurang penting. Convolutional Block Attention Module (CBAM), yang diperkenalkan oleh Woo et al. pada tahun 2018, adalah modul *attention* yang *lightweight* dan efektif yang dapat diintegrasikan ke dalam arsitektur apapun untuk meningkatkan representasi fitur [5, 17, 18]. CBAM menerapkan *attention mechanism* pada dua dimensi secara berurutan: *channel attention* dan *spatial attention*, sehingga memungkinkan model mempelajari *what* (apa yang penting) dan *where* (di mana lokasinya) secara eksplisit.



Gambar 2.1. Arsitektur *Convolutional Block Attention Module* (CBAM) yang memproses *input feature* secara berurutan melalui *Channel Attention Module* dan *Spatial Attention Module* untuk menghasilkan *refined feature*. Sumber: Liao et al. [1].

Channel attention module berfokus pada hubungan antar-*channel* pada *feature map*. Modul ini menggunakan *global average pooling* dan *global max pooling* untuk mengagregasi informasi spasial, menghasilkan dua deskriptor yang kemudian diproses melalui *shared* MLP dengan *reduction ratio r* [16]. Formulasi *channel attention map* $\mathbf{M}_c \in \mathbb{R}^{C\times 1\times 1}$ adalah [1]:

$$\mathbf{M}_c(\mathbf{F}) = \sigma(\text{MLP}(\text{AvgPool}(\mathbf{F})) + \text{MLP}(\text{MaxPool}(\mathbf{F}))) \tag{2.3}$$

di mana $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ adalah *input feature map*, σ adalah fungsi sigmoid, dan MLP adalah *multilayer perceptron* yang digunakan bersama untuk kedua jalur. Output *channel attention* diperoleh melalui perkalian elemen (*element-wise multiplication*):

$$\mathbf{F}' = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F} \tag{2.4}$$

Spatial attention module kemudian menerima $\mathbf{F}'$ sebagai input dan berfokus pada lokasi spasial yang paling informatif. Modul ini melakukan *average pooling* dan *max pooling* sepanjang dimensi *channel*, lalu men-*concatenate* hasilnya dan memprosesnya melalui konvolusi 7×7 [16]. Formulasi *spatial attention map* $\mathbf{M}_s \in \mathbb{R}^{1\times H\times W}$ adalah [1]:

$$\mathbf{M}_s(\mathbf{F}') = \sigma\left(f^{7\times 7}\left(\left[\text{AvgPool}(\mathbf{F}'); \text{MaxPool}(\mathbf{F}')\right]\right)\right) \tag{2.5}$$

di mana $f^{7\times 7}$ adalah operasi konvolusi dengan *kernel* berukuran 7×7, dan $[\cdot;\cdot]$ adalah operasi *concatenation* sepanjang dimensi *channel*. Output akhir CBAM adalah:

$$\mathbf{F}'' = \mathbf{M}_s(\mathbf{F}') \otimes \mathbf{F}' \tag{2.6}$$

Dalam aplikasi Facial Expression Recognition, CBAM memberikan kontribusi signifikan dalam mengatasi tantangan seperti oklusi parsial dan variasi pose wajah. *Channel attention* membantu model untuk fokus pada *channel* yang mengandung informasi emosi yang distingtif, sementara *spatial attention* membantu model fokus pada region wajah yang paling informatif seperti mata, alis, dan mulut [1, 21, 23]. Keunggulan CBAM adalah *overhead* komputasi yang minimal karena parameter tambahan yang dibutuhkan sangat sedikit, menjadikannya cocok untuk diintegrasikan pada arsitektur *lightweight* seperti MobileNetV2 tanpa menambah kompleksitas secara signifikan.

2.4 Focal Loss

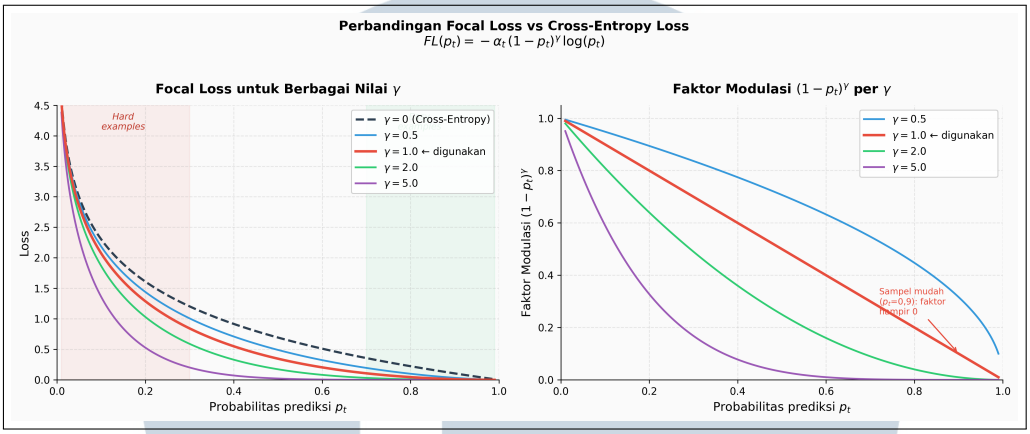
Class imbalance merupakan salah satu tantangan fundamental dalam machine learning, di mana distribusi sampel antar kelas sangat tidak merata. Pada dataset dengan class imbalance ekstrem, model cenderung bias terhadap kelas mayoritas dan mengabaikan kelas minoritas karena Cross-Entropy Loss standar memberikan kontribusi yang sama untuk setiap sampel terlepas dari kesulitan klasifikasinya. Focal Loss, yang diperkenalkan oleh Lin et al. pada tahun 2017, dikembangkan untuk mengatasi masalah ini dengan cara memodifikasi Cross-Entropy Loss agar model lebih fokus pada sampel yang sulit diklasifikasikan (*hard examples*) dan mengurangi kontribusi dari sampel yang mudah (*easy examples*) [14, 27, 28]. Pendekatan ini terbukti sangat efektif dalam task dense object detection dengan class imbalance ekstrem, dan telah diadaptasi untuk berbagai aplikasi termasuk Facial Expression Recognition.

Sebagai acuan, *Cross-Entropy Loss* standar didefinisikan sebagai $\mathrm{CE}(p_t) = -\log(p_t)$ di mana p_t adalah probabilitas model untuk kelas yang benar. *Focal Loss* memodifikasi rumus ini dengan menambahkan *modulating factor* $(1-p_t)^\gamma$ dan *weighting factor* α_t [14]:

$$\mathrm{FL}(p_t) = -\alpha_t (1-p_t)^\gamma \log(p_t) \tag{2.7}$$

di mana $\gamma \geq 0$ adalah *focusing parameter* yang mengontrol tingkat *down-weighting* pada *easy examples*, dan $\alpha_t \in [0,1]$ adalah *weighting factor* per kelas untuk menangani *class imbalance*. Ketika sampel mudah diklasifikasikan (probabilitas tinggi untuk kelas yang benar), nilai $(1-p_t)$ mendekati 0 sehingga kontribusi *loss*-nya menjadi sangat kecil. Sebaliknya, ketika sampel sulit diklasifikasikan (probabilitas rendah), nilai $(1-p_t)$ mendekati 1 sehingga *loss* mendekati *standard*

Cross-Entropy. Parameter $\gamma = 0$ menghasilkan *standard Cross-Entropy* dan $\gamma = 2$ umumnya memberikan hasil terbaik pada berbagai *task*.



Gambar 2.2. Perbandingan Focal Loss dan Cross-Entropy Loss untuk berbagai nilai *focusing parameter* γ. Grafik kiri menunjukkan nilai *loss* terhadap probabilitas prediksi p_t, di mana $\gamma = 0$ ekuivalen dengan Cross-Entropy standar. Grafik kanan menunjukkan *modulating factor* $(1-p_t)^\gamma$ yang mendekati nol untuk sampel mudah (p_t tinggi), sehingga model lebih fokus pada *hard examples*.

Dalam penelitian ini, α_c untuk setiap kelas c dihitung menggunakan akar kuadrat dari *inverse frequency*:

$$\alpha_c = \sqrt{\frac{N_{\text{total}}}{K \cdot N_c}} \tag{2.8}$$

di mana N_{total} adalah total jumlah sampel *training*, $K = 7$ adalah jumlah kelas, dan N_c adalah jumlah sampel pada kelas c. Penggunaan akar kuadrat bertujuan untuk memoderasi perbedaan bobot agar kelas minoritas seperti *disgust* ($N_c = 175$) tidak mendominasi proses pelatihan secara berlebihan.

Perbandingan antara Focal Loss dan Cross-Entropy Loss menunjukkan perbedaan fundamental dalam cara kedua loss function memperlakukan sampel yang berbeda tingkat kesulitannya. Cross-Entropy Loss memberikan kontribusi yang signifikan bahkan untuk sampel yang sudah mudah diklasifikasikan dengan confidence tinggi, yang dapat mendominasi total loss dan menghalangi pembelajaran pada sampel yang sulit. Focal Loss mengurangi kontribusi dari easy examples secara dramatis, memungkinkan model untuk mengalokasikan lebih banyak learning capacity pada hard examples yang lebih informatif. Dalam konteks class imbalance, hard examples seringkali berasal dari kelas minoritas yang memiliki representasi terbatas dalam training data. Dengan memberi fokus lebih pada sampel-sampel ini, Focal Loss efektif meningkatkan performa pada

kelas minoritas tanpa memerlukan teknik sampling atau data augmentation yang kompleks [27, 28].

Dalam aplikasi Facial Expression Recognition, Focal Loss sangat relevan untuk menangani class imbalance yang ekstrem pada dataset seperti FERPlus, di mana distribusi kelas berkisar dari 175 sampel untuk disgust hingga 9.365 sampel untuk neutral pada *training set*, sebagaimana dijabarkan pada Bab 3. Penelitian sebelumnya menunjukkan bahwa model yang dilatih dengan Cross-Entropy Loss cenderung memiliki performa sangat buruk pada kelas minoritas meskipun overall accuracy tinggi [5]. Implementasi Focal Loss, dikombinasikan dengan class weighting (α_t) yang disesuaikan dengan inverse frequency setiap kelas, terbukti meningkatkan F1-score pada kelas minoritas secara signifikan [14]. Penggunaan Focal Loss juga mengurangi kebutuhan akan oversampling agresif yang dapat menyebabkan overfitting, karena loss function sendiri sudah secara implisit memberikan bobot lebih pada sampel yang kurang terwakili dalam data [14, 28].

2.5 Grad-CAM (*Gradient-weighted Class Activation Mapping*)

Interpretabilitas merupakan aspek krusial dalam pengembangan sistem *deep learning*, terutama untuk aplikasi yang memerlukan transparansi keputusan. Model *deep learning* seringkali dipandang sebagai *black box* karena kompleksitas internalnya yang sulit dipahami. Grad-CAM (*Gradient-weighted Class Activation Mapping*), yang diperkenalkan oleh Selvaraju et al. pada tahun 2017, adalah teknik visualisasi yang menghasilkan *heatmap* untuk menunjukkan region pada gambar yang paling berpengaruh terhadap prediksi model untuk kelas tertentu [5, 16, 2, 29]. Keunggulan Grad-CAM adalah tidak memerlukan modifikasi arsitektur maupun *retraining* model, sehingga dapat diterapkan secara *post-hoc* pada model yang sudah terlatih.

Secara matematis, Grad-CAM menghitung bobot kepentingan α_k^c untuk setiap *feature map* k pada *target layer* terhadap kelas c menggunakan *global average pooling* dari gradien [2]:

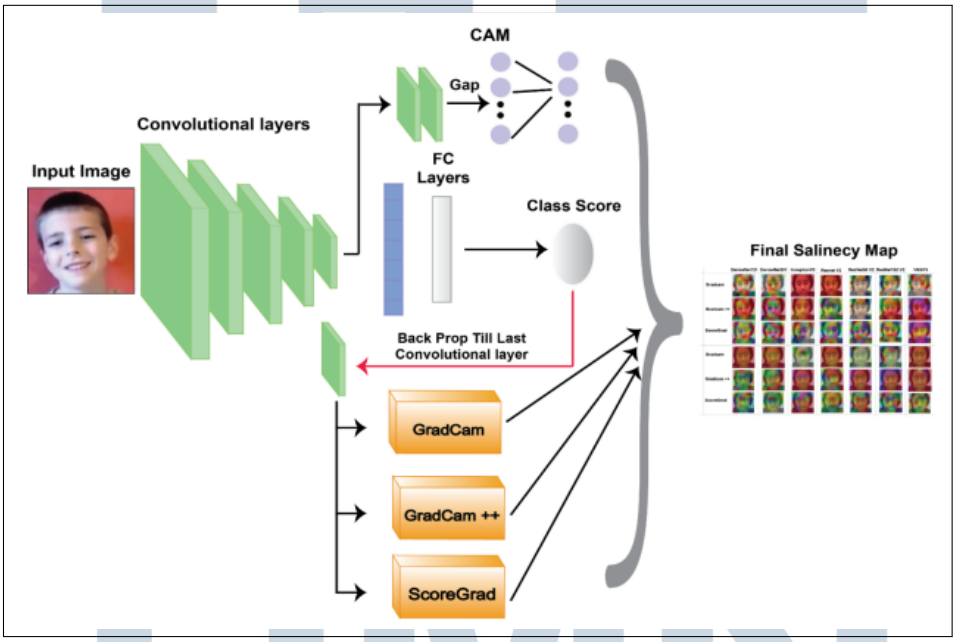
$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2.9}$$

di mana y^c adalah skor logit untuk kelas c sebelum *softmax*, A_{ij}^k adalah aktivasi pada posisi (i, j) dari *feature map* ke-k, dan $Z = H \times W$ adalah jumlah piksel

pada *feature map*. *Heatmap* Grad-CAM untuk kelas c kemudian diperoleh dengan kombinasi linear berbobot dari *feature maps*, diikuti operasi ReLU untuk hanya mempertahankan fitur yang berkontribusi positif terhadap kelas yang dimaksud:

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha^{c}_{k} A^{k}\right) \tag{2.10}$$

di mana $L^{c}_{\text{Grad-CAM}} \in \mathbb{R}^{H \times W}$ adalah *heatmap* akhir yang kemudian di-*upsample* ke ukuran gambar asli menggunakan interpolasi bilinear. Nilai tinggi pada *heatmap* menandakan region yang paling berpengaruh terhadap prediksi kelas c.



Gambar 2.3. Alur implementasi Grad-CAM: gambar input diproses melalui lapisan konvolusi, gradien kelas target di-*backpropagate* hingga lapisan konvolusi terakhir, lalu dikombinasikan untuk menghasilkan *heatmap* akhir. Sumber: Rathod et al. [2].

Dalam aplikasi Facial Expression Recognition, Grad-CAM digunakan untuk memvalidasi apakah model benar-benar fokus pada region wajah yang relevan secara semantik, seperti mata untuk ekspresi *surprise* dan *fear*, mulut untuk *happy* dan *sad*, serta alis untuk *angry* [7, 2, 29]. Visualisasi ini memberikan bukti interpretabilitas model sekaligus membantu analisis kesalahan klasifikasi untuk mendukung perbaikan dan pengembangan model lebih lanjut.

2.6 Dataset FERPlus

Dataset FER2013 (Facial Expression Recognition 2013) merupakan salah satu benchmark dataset paling populer dan banyak digunakan dalam penelitian Facial Expression Recognition. Dataset ini pertama kali diperkenalkan pada kompetisi Kaggle ”Challenges in Representation Learning” yang diselenggarakan pada International Conference on Machine Learning (ICML) 2013. FER2013 berisi 35.887 gambar grayscale berukuran 48×48 piksel yang dikumpulkan secara otomatis dari internet menggunakan Google Image Search API dengan query terkait ekspresi emosi [8, 30, 23]. Proses anotasi awal dilakukan melalui crowdsourcing menggunakan Amazon Mechanical Turk, di mana setiap gambar hanya diberikan label oleh satu annotator, sehingga rentan terhadap subjektivitas individual dan noise label yang signifikan. Untuk mengatasi keterbatasan tersebut, penelitian ini menggunakan FERPlus, yaitu versi re-anotasi dari FER2013 yang dikembangkan oleh Microsoft, di mana setiap gambar diberi label ulang melalui *majority voting* dari 10 annotator, sehingga meningkatkan kredibilitas label secara signifikan [31].

Dataset FERPlus menggunakan gambar yang sama dengan FER2013, namun dengan label yang telah diperbarui melalui proses *majority voting*. Pada proses tersebut, gambar dengan hasil voting seri (*tie*) serta gambar yang mayoritas terlabel sebagai *contempt*, *unknown*, atau *NF* (*Not a Face*) dibuang, menyisakan 33.500 gambar yang kemudian terbagi menjadi tiga subset: training set (26.778 gambar), validation set (3.372 gambar), dan test set (3.350 gambar). Setiap gambar diberi label salah satu dari tujuh kategori emosi: angry (marah), disgust (jijik), fear (takut), happy (bahagia), sad (sedih), surprise (terkejut), dan neutral (netral). Karakteristik utama dataset ini adalah resolusi gambar yang sangat rendah (48×48 piksel) dibandingkan dengan dataset FER lainnya, yang membuat task menjadi lebih challenging karena hilangnya detail halus pada wajah. Selain itu, gambar-gambar dalam dataset mengandung variasi yang sangat besar dalam hal pencahayaan, pose wajah, oklusi parsial, etnis, usia, dan jenis kelamin, yang merefleksikan kondisi ”in-the-wild” yang lebih realistis namun juga lebih sulit untuk diproses [8, 30].

Salah satu tantangan terbesar dalam FERPlus adalah class imbalance yang ekstrem, bahkan lebih parah dibandingkan FER2013 asli. Tabel 2.1 menunjukkan distribusi kelas pada dataset ini. Kelas neutral memiliki jumlah sampel terbanyak dengan 11.773 gambar (35,14% dari total dataset), sementara kelas disgust hanya memiliki 228 gambar (0,68% dari total dataset), menghasilkan rasio imbalance

hingga 53,5 kali lipat. Class imbalance ini menjadi tantangan signifikan karena model cenderung bias terhadap kelas mayoritas dan memiliki performa yang sangat buruk pada kelas minoritas jika tidak ditangani dengan teknik khusus seperti oversampling, class weighting, atau loss function yang adaptif [30, 17]. Distribusi yang tidak merata ini mencerminkan frekuensi natural ekspresi emosi dalam kehidupan sehari-hari, di mana ekspresi seperti neutral dan happy lebih sering muncul dibandingkan disgust.

Tabel 2.1. Distribusi Kelas pada Dataset FERPlus

Kelas	**Training**	**Validation**	**Test**	**Total**	**Persentase**
Angry	2.399	310	316	3.025	9,03%
Disgust	175	31	22	228	0,68%
Fear	648	73	92	813	2,43%
Happy	7.410	880	909	9.199	27,46%
Sad	3.403	395	424	4.222	12,60%
Surprise	3.378	441	421	4.240	12,66%
Neutral	9.365	1.242	1.166	11.773	35,14%
Total	**26.778**	**3.372**	**3.350**	**33.500**	**100,0%**

Tantangan lain yang signifikan pada FER2013 asli adalah noise pada label yang tinggi akibat proses crowdsourcing oleh satu annotator dan ambiguitas inherent dalam pengenalan emosi. Studi menunjukkan bahwa akurasi pengenalan emosi oleh manusia pada FER2013 hanya sekitar 65%, yang mengindikasikan tingkat kesulitan dataset ini dan subjektivitas dalam interpretasi ekspresi wajah [8, 12]. Beberapa gambar memiliki ekspresi yang ambigu atau berada di antara dua kategori emosi, sementara gambar lainnya memiliki label yang jelas-jelas salah. Penggunaan FERPlus pada penelitian ini secara signifikan mengurangi permasalahan tersebut melalui mekanisme *majority voting*, meskipun ambiguitas inherent pada ekspresi wajah tetap menjadi tantangan yang melekat pada task FER secara umum. Meskipun demikian, karena FERPlus dibangun di atas gambar FER2013 yang sama, dataset ini tetap mempertahankan posisinya sebagai benchmark standard dalam penelitian FER karena ukurannya yang substantial, variasi yang tinggi, dan tingkat kesulitan yang realistis yang mendorong pengembangan metode yang robust dan generalizable untuk aplikasi *real-world*.

2.7 Perbandingan Penelitian Terdahulu

Tabel 2.2 menyajikan perbandingan komprehensif antara penelitian-penelitian terdahulu dengan penelitian ini, mencakup metode yang digunakan, dataset evaluasi, serta hasil berupa akurasi, kelebihan, dan gap yang menjadi motivasi penelitian ini.

Tabel 2.2. Perbandingan Penelitian Terdahulu dengan Penelitian Ini

No	Judul Penelitian	Dataset	Metode	Hasil & Analisis
1	Facial Expression Recognition Based on Improved Depthwise Separable Convolutional Network [13] *(Huo et al., 2023)*	FER2013, CK+	Xception + MobileNetV2	**Akurasi: 70,76% (FER2013)** *Kelebihan:* Arsitektur lightweight (2,3M parameter), efisien secara komputasi, melampaui human-level recognition pada FER2013. *Gap:* Tidak menangani class imbalance secara eksplisit, tidak ada mekanisme explainability, lemah pada oklusi berat.

Tabel 2.2. Perbandingan Penelitian Terdahulu dengan Penelitian Ini (lanjutan)

No	Judul Penelitian	Dataset	Metode	Hasil & Analisis
2	Expression-Guided Deep Joint Learning for Facial Expression Recognition [14] *(Fang et al., 2023)*	FER2013, RAF-DB, CK+	ACNN + SE-Net + Joint Learning	**Akurasi: 75,42% (FER2013)** *Kelebihan:* Memanfaatkan unlabeled data, parameter sangat efisien (1,56M), SE-Net attention menguatkan fitur relevan. *Gap:* Augmentasi tidak efektif untuk kelas minoritas, tidak ada explainability, bias demografis belum dieksplorasi.
3	Human Emotion Recognition with an Advanced Vision Transformer Model [31] *(Huynh et al., 2025)*	FER2013+, AffectNet, RAF-DB	EfficientViT-M5	**Akurasi: 94,28% (FER2013+)** *Kelebihan:* State-of-the-art dengan Vision Transformer, transfer learning dari ImageNet, lebih efisien dari ViT standar. *Gap:* Kompleksitas tinggi untuk deployment, membutuhkan pre-training dataset besar, tidak dievaluasi pada FER2013 asli.

Tabel 2.2. Perbandingan Penelitian Terdahulu dengan Penelitian Ini (lanjutan)

No	Judul Penelitian	Dataset	Metode	Hasil & Analisis
4	Facial Expression Recognition Methods in the Wild Based on Fusion Feature of Attention Mechanism and LBP [1] *(Liao et al., 2023)*	FER2013, FERPlus, RAF-DB	ResNet + CBAM + LBP	**Akurasi: 73,71% (FER2013)** *Kelebihan:* Kombinasi CBAM (channel + spatial attention) dengan LBP texture features, robust terhadap oklusi dan variasi pose. *Gap:* Tidak menangani class imbalance secara khusus, dual-branch meningkatkan waktu inference.
5	Facial Expression Recognition of Students in Classroom Using Hybrid MobileNetV3-Vision Transformer [11] *(Isya et al., 2024)*	FER2013	MobileNetV3 + Vision Transformer	**Akurasi: 71,24% (FER2013)** *Kelebihan:* Hybrid CNN-Transformer menangkap fitur lokal dan global, token downsampling mengurangi kompleksitas komputasi. *Gap:* Performa belum optimal, akurasi masih di bawah baseline hybrid lainnya, tidak menangani class imbalance.

Tabel 2.2. Perbandingan Penelitian Terdahulu dengan Penelitian Ini (lanjutan)

No	Judul Penelitian	Dataset	Metode	Hasil & Analisis
6	Revolutionizing Online Education: Advanced FER for Real-time Student Progress Tracking [7] *(Aly, 2024)*	FER2013, RAF-DB, CK+	ResNet50 + CBAM + TCN	**Akurasi: 91,71% (FER2013)** *Kelebihan:* Multi-dataset evaluation, kombinasi CBAM + TCN untuk temporal dynamics, real-time monitoring. *Gap:* Fokus pada konteks pendidikan online, performa FER2013 kemungkinan dihasilkan dengan preprocessing khusus.
7	Towards a Better Performance in FER: A Data-Centric Approach [12] *(Mejia-Escobar et al., 2023)*	FER2013 reclassified	CNN + Dataset Refinement	**Akurasi: 73,11% (FER2013 original baseline)** *Kelebihan:* Fokus pada data quality, progressive refinement mengurangi noise label secara signifikan. *Gap:* Membutuhkan multiple training cycles untuk reclassification, tidak applicable untuk dataset baru tanpa preprocessing.

Tabel 2.2. Perbandingan Penelitian Terdahulu dengan Penelitian Ini (lanjutan)

No	Judul Penelitian	Dataset	Metode	Hasil & Analisis
8	CNN-LSTM Facial Expression Recognition Method Fused with Two-Layer Attention Mechanism [32] *(Ming et al., 2022)*	FER2013, CK+	CNN-LSTM + Two-Layer Attention	**Akurasi: 72,84% (FER2013)** *Kelebihan:* Attention pada dua layer (spatial + temporal), mampu menangkap contextual information dengan LSTM. *Gap:* Kompleksitas tinggi, waktu training lama, membutuhkan sequence data agar LSTM optimal.
9	A Student FER Model Based on Multi-Scale and Deep Fine-Grained Feature Attention Enhancement [20] *(Wang et al., 2024)*	FER2013, FERPlus, RAF-DB	Multi-Scale + Fine-Grained Attention	**Akurasi: 76,18% (FER2013)** *Kelebihan:* Multi-scale feature aggregation, fine-grained attention untuk detail ekspresi, evaluasi pada 5 dataset. *Gap:* Kompleksitas arsitektur tinggi, membutuhkan backbone pre-trained pada dataset besar (MS-Celeb-1M).

Tabel 2.2. Perbandingan Penelitian Terdahulu dengan Penelitian Ini (lanjutan)

No	Judul Penelitian	Dataset	Metode	Hasil & Analisis
10	Advancing FER in Online Learning Education Using a Homogeneous Ensemble CNN Approach [8] *(Sitti et al., 2024)*	FER2013	Ensemble CNN (VGG16, ResNet50, DenseNet121)	**Akurasi: 74,5% (FER2013)** *Kelebihan:* Ensemble meningkatkan robustness, membahas karakteristik FER2013 secara detail termasuk class imbalance. *Gap:* Kompleksitas tinggi (3 model bersamaan), tidak menangani class imbalance secara eksplisit, tidak ada explainability.

Berdasarkan Tabel 2.2, terdapat tiga gap utama yang belum diatasi oleh penelitian-penelitian terdahulu. Pertama, tidak satupun dari 10 penelitian yang menangani *class imbalance* secara eksplisit menggunakan *loss function* adaptif, bahkan mayoritas mengabaikannya [13, 1, 11, 7, 32, 20, 8], sementara Fang et al. [14] hanya mengandalkan augmentasi sederhana yang tidak efektif untuk kelas minoritas ekstrem seperti *disgust* (547 sampel) pada FER2013. Kedua, penelitian yang menggunakan CBAM [1, 7] mengintegrasikannya dengan arsitektur *heavy* seperti ResNet50 (25 juta parameter), sehingga tidak efisien untuk *deployment* pada perangkat terbatas. Ketiga, seluruh penelitian terdahulu tidak menyediakan mekanisme *explainability* sehingga tidak dapat divalidasi apakah model benar-benar belajar dari fitur wajah yang relevan [13, 14, 1, 11, 7, 12, 32, 20, 8]. Penelitian ini mengusulkan MobileNetV2-CBAM (2,4 juta parameter) dengan Focal Loss berbasis *sqrt inverse frequency* untuk mengatasi *class imbalance*, serta Grad-CAM untuk *explainability* model.