BAB II
LANDASAN TEORI

2.1 Dasar Teori

2.1.1 Emosi

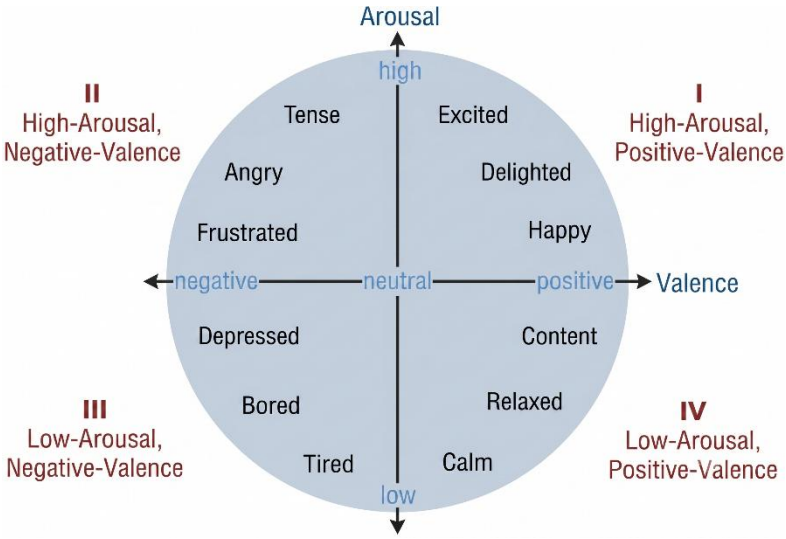
Emosi didefinisikan sebagai respons kompleks yang dipicu oleh peristiwa stimulus baik yang bersifat eksternal maupun internal setelah melalui proses evaluasi terhadap signifikansinya bagi individu (Barradas dkk., 2025). Sebagai kondisi afektif yang muncul secara spontan emosi tidak hanya merefleksikan reaksi terhadap pengalaman internal dan eksternal tetapi juga berperan signifikan dalam membentuk persepsi serta pola interaksi sosial individu. Khusus pada anak-anak, ekspresi emosional sering kali termanifestasi melalui media nonverbal, salah satunya adalah coretan visual yang mampu menyediakan jendela untuk melihat dunia representasi mereka (Fabris dkk., 2023). Hal ini menjadikan aktivitas menggambar sebagai instrumen krusial dalam mengidentifikasi dan memahami kondisi psikologis anak secara lebih mendalam karena gambar diyakini dapat mengungkap perasaan mereka sebagaimana tercermin dalam representasi pengalaman yang mereka miliki (Manomenidis dkk., 2025).

Dalam ranah *affective computing* representasi emosi secara komputasional terbagi ke dalam dua pendekatan utama. Model kategoris (diskret) yang berakar pada teori Ekman (1992) mengklasifikasikan emosi ke dalam kategori spesifik seperti *anger, fear, sadness, enjoyment, disgust,* dan *surprise*, dimana unggul dalam simplisitas anotasi namun terbatas dalam menangkap gradasi emosional yang halus. Model dimensional (kontinu) khususnya *Russell's Circumplex Model of Affect*, memetakan emosi dalam ruang dua dimensi *Valence* dan *Arousal* sehingga mampu merepresentasikan variasi afektif secara lebih bernuansa (Boateng dkk., 2022).

Penelitian ini menggunakan kedua pendekatan tersebut secara saling melengkapi yaitu klasifikasi utama menggunakan label diskret yang diperkaya dengan fitur dimensional VA untuk representasi emosi yang lebih beragam dan informatif.

2.1.2 *Russell's Circumplex Model of Affect*

Boateng dkk. (2022) mengimplementasikan konsep dari *Russell's Circumplex Model of Affect* yang memetakan emosi manusia ke dalam dimensi *Valence* dan *Arousal*. Pendekatan dimensional ini memandang emosi sebagai titik koordinat dalam ruang afektif di mana *valence* merepresentasikan kualitas perasaan dan *arousal* merepresentasikan tingkat aktivasi psikologis. Hal ini memungkinkan representasi emosi yang melampaui pelabelan kategorikal diskret, memberikan fleksibilitas lebih dalam mengidentifikasi kondisi mental subjek.



Gambar 2.1 *Russell's Circumplex Model of Affect* (Boateng dkk., 2022)

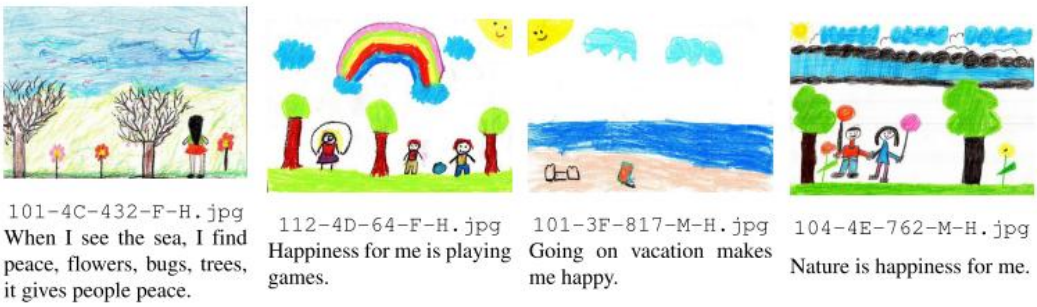
Gambar 2.1 menunjukkan distribusi emosi dalam empat kuadran Kartesius berdasarkan nilai ambang dimensi tersebut. Sebagai contoh, emosi *angry* dan *depressed* dikelompokkan dalam domain valensi negatif dengan tingkat *arousal* yang kontras sedangkan *relaxed* dan *excited* berada dalam domain valensi positif (Boateng dkk., 2022). Penggunaan koordinat dua dimensi ini memfasilitasi

pendeteksian nuansa emosional yang bersifat kontinu dan dinamis. Dengan demikian sistem dapat mengenali perubahan emosi yang bersifat halus, kemampuan yang sulit dicapai jika hanya mengandalkan model klasifikasi sederhana.

Nilai *Valence* dan *Arousal* dapat diimplementasikan sebagai fitur pendukung untuk memperkuat keputusan klasifikasi utama dengan mengeksploitasi informasi bersama antar representasi untuk generalisasi yang lebih baik (Bautista & Shin, 2025).

2.1.3 Gambar Coretan Anak

Sebagai media ekspresi non-invasif, gambar atau coretan anak secara konsisten berfungsi sebagai jendela yang sensitif untuk memahami aspek afeksi, kognisi, serta lingkungan yang dialami oleh subjek (Al Saadi, 2025). Secara teknis, emosi yang terkandung dalam gambar dapat diidentifikasi melalui sejumlah karakteristik visual yang menonjol seperti ukuran dan warna yang ditemukan secara sistematis berdasarkan valensi karakterisasi afektif (positif atau negatif) pada objek yang digambar (Burkitt dkk., 2009).



Gambar 2.2 Sampel Gambar Coretan dan Teks Refleksi (Çiftçi dkk., 2025)

Integrasi antara media visual dan refleksi tekstual merupakan pendekatan krusial dalam memahami emosi anak, mengingat penggunaan warna dalam gambar tidak secara konsisten merefleksikan muatan afektif karena sangat dipengaruhi oleh preferensi warna pribadi anak terlepas dari konten emosional yang digambarkan

(Crawford dkk., 2012). Gambar 2.2 merupakan contoh gambar gambar hasil coretan anak beserta teks refleksi yang berisi pemikiran orisinal anak mengenai kondisi emosional mereka seperti kebahagiaan dan kesedihan (Çiftçi dkk., 2025).

Dual-Coding Theory menjelaskan bahwa otak memproses informasi visual dan verbal melalui dua saluran berbeda yang saling melengkapi. Gambar memicu pemrosesan afektif secara langsung, sementara teks memberi pelabelan kognitif yang memperjelas makna emosi tersebut. Hal ini didukung oleh temuan Alvarado (2025) yang menegaskan bahwa properti visual, khususnya kombinasi dimensi fisik warna seperti *chroma* dan *lightness*, memiliki pola sistematis yang kuat dalam merepresentasikan dimensi emosi dari *valence, arousal,* dan *dominance*. Di sisi lain, Al Mascaty dkk. (2026) menunjukkan bahwa efektivitas pemrosesan kognitif dan retensi memori terhadap stimulus visual sangat bergantung pada kekayaan semantik dan konseptual dari stimulus tersebut, di mana dimensi emosi dan karakteristik inheren visual saling berinteraksi secara kontekstual. Kedua temuan ini menjadi dasar teoretis mengapa gambar coretan dan teks refleksi harus diproses secara berdampingan.

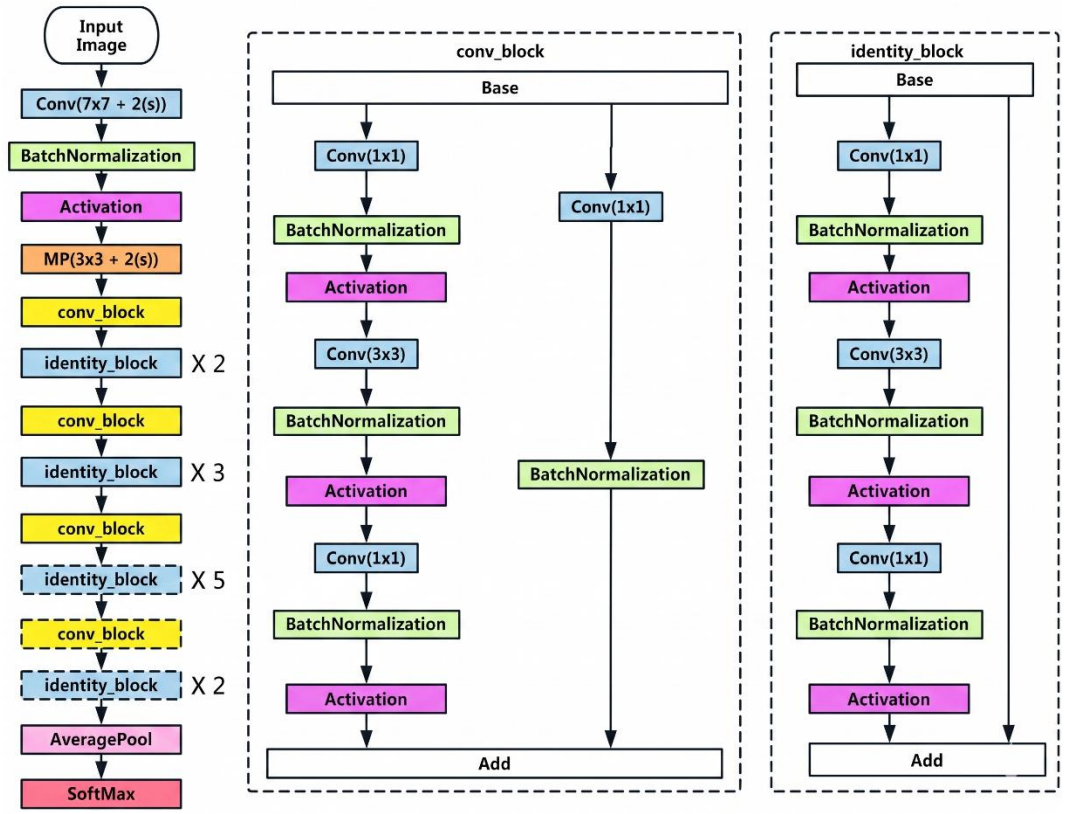
Tingkat abstraksi pada gambar coretan anak mempersulit sebagian arsitektur CNN standar sebagai alat digital meskipun mampu melakukan kuantifikasi fitur dalam skala besar, sering kali kehilangan nuansa atau konteks penting dalam interpretasinya (Al Saadi, 2025). Bagaimanapun, terdapat ketergantungan makna yang kuat terhadap konteks naratif atau perspektif verbal anak untuk mengidentifikasi emosi dan indikator kesejahteraan anak secara akurat, karena suara anak merupakan sumber informasi terbaik bagi penilaian kesehatan dan kesejahteraan mereka sendiri (Moula dkk., 2021).

2.1.4 ResNet (*Residual Network*)

ResNet (He dkk., 2016) memperkenalkan inovasi berupa *residual learning* yang memungkinkan pelatihan jaringan saraf dengan kedalaman yang sangat ekstrim. Konsep inti dari arsitektur ini adalah *skip connection* yang secara matematis didefinisikan pada Persamaan (1) (He dkk., 2016).

$$y = F(x) + x \qquad \textbf{(1)}$$

Persamaan (1) menunjukkan bahwa *output* dari sebuah blok residual merupakan hasil penjumlahan antara transformasi yang dipelajari *F(x)* dengan input asli x. Mekanisme ini menjamin aliran gradien secara langsung melalui jaringan tanpa mengalami degradasi yang signifikan. Maka dari itu, ResNet mampu mengatasi kendala *vanishing gradient* yang umumnya menjadi hambatan utama pada arsitektur CNN konvensional dengan lapisan yang sangat dalam.

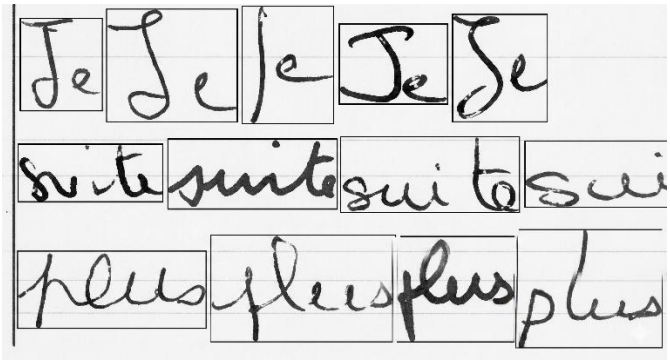


Gambar 2.3 Arsitektur ResNet 50 (Karim dkk., 2022)

Gambar 2.3 menunjukkan arsitektur ResNet50 yang digunakan sebagai *backbone* klasifikasi citra. Karim dkk. (2022) mengonfigurasi ResNet50 dengan tujuh lapisan awal yang *non-trainable* dan lapisan berikutnya *trainable*, membuktikan efektivitasnya untuk *transfer learning* pada domain klasifikasi citra.

2.1.5 Deteksi Region Berbasis Tinta (*Ink-Based Region Detection*)

Deteksi *region* berbasis tinta merupakan teknik prapemrosesan yang berfokus pada pemisahan goresan tinta dari piksel latar belakang untuk mengidentifikasi wilayah penarik (*region of interest*) secara akurat. Teknik ini sangat efektif dalam analisis gambar coretan anak yang sering kali didominasi oleh ruang kosong (*white space*), abstrak, dan pola goresan yang tidak beraturan (Lee & Yoo, 2025). Dengan mengidentifikasi area spesifik tempat tinta menempel seperti pada Gambar 2.4, model dapat menghindari pemrosesan informasi yang tidak relevan sehingga fitur wilayah yang dihasilkan untuk mekanisme *cross-modal attention* menjadi lebih padat makna dan objektif dalam merepresentasikan emosi subjek.



Gambar 2.4 Sampel implementasi *ink-based region* (Mahadevkar dkk., 2024)

Mahadevkar dkk. (2024) menyebutkan tahapan utama dalam pendekatan deteksi berbasis tinta ini meliputi:

a. Grayscale dan Binarisasi

Transformasi citra ke skala abu-abu yang dilanjutkan dengan binarisasi menggunakan metode *thresholding* seperti *Otsu* yang dinyatakan pada Persamaan (2) dan Persamaan (3) untuk menonjolkan piksel tinta terhadap latar belakang secara kontras.

$$\sigma^2_{between}(T) = \omega_0(T) \cdot \omega_1(T) \cdot [\mu_0(T) - \mu_1(T)]^2 \quad \textbf{(2)}$$

$$T^* = \arg\max_T \sigma^2_{between}(T) \quad \textbf{(3)}$$

Pada persamaan tersebut, ω_0, ω_1 adalah proporsi piksel di bawah dan di atas threshold T, serta μ_0, μ_1 adalah rata-rata intensitas masing-masing kelompok. Persamaan (2) berfungsi mengukur seberapa besar perbedaan antara dua kelompok piksel pada nilai *threshold T* tertentu, semakin besar variansi antar-kelas maka semakin tegas pemisahan antara tinta dan latar belakang. Persamaan (3) berfungsi menentukan nilai *threshold* optimal T^* dengan mencari nilai *T* yang menghasilkan variansi antar-kelas terbesar dari Persamaan (2), sehingga hasil binarisasi gambar coretan menghasilkan pemisahan tinta dan latar yang paling akurat secara otomatis tanpa memerlukan konfigurasi manual.

b. Operasi Morfologi

Penggunaan elemen matematis melalui proses dilasi (*dilation*) untuk menyambungkan garis tinta yang terputus serta erosi (*erosion*) untuk menghilangkan derau (*noise*) supaya membentuk kelompok goresan utama sebagai satu entitas wilayah.

c. Ekstraksi Kontur

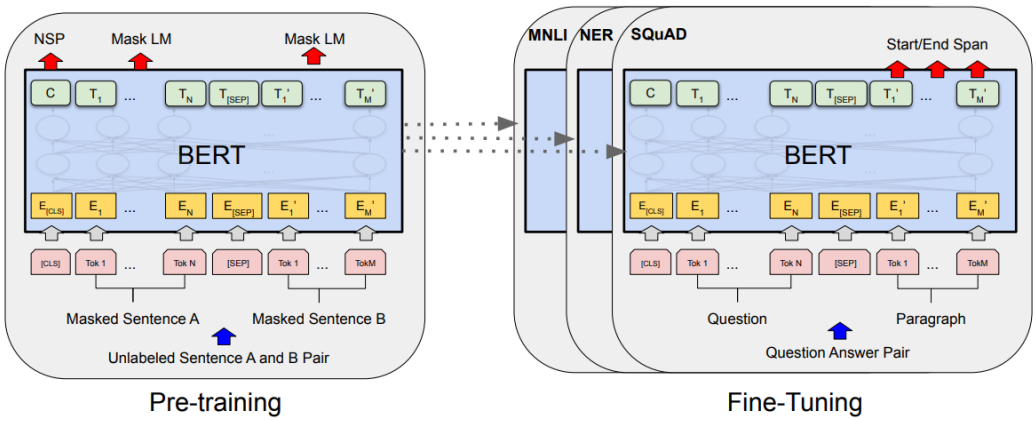
Penerapan algoritma topologis untuk melacak batas antara piksel tinta dan latar belakang dalam menemukan kurva tertutup berkelanjutan yang dapat merepresentasikan goresan signifikan.

d. Pembentukan *Bounding Box*

Pembuatan kotak pembatas terkecil yang membungkus setiap kontur untuk dijadikan masukan spasial bersifat *fine-grained* dalam arsitektur CNN. Gambar 2.4 memperlihatkan bagaimana proses tersebut dapat menjamin model supaya hanya berfokus pada representasi visual dari tarikan tinta saja.

2.1.6 BERT (*Bidirectional Encoder Representations from Transformers*)

BERT (*Bidirectional Encoder Representations from Transformers*) yang diperkenalkan oleh Devlin dkk. (2019) merupakan model bahasa berbasis arsitektur *Transformer* yang memanfaatkan konteks dua arah (*bidirectional*) dalam pemrosesan teks. Siklus pengembangan BERT terdiri dari dua fase krusial yaitu *pre-training* dan *fine-tuning* seperti yang terlihat di Gambar 2.5.



Gambar 2.5 Cara Kerja BERT (Delvin dkk., 2019)

Selama fase *pre-training* pada korpus data berskala besar, model ini menjalankan tugas *Masked Language Model* (MLM) untuk memprediksi token tersembunyi serta *Next Sentence Prediction* (NSP) untuk memahami koherensi antar-kalimat. Setelah melalui tahap tersebut, BERT dapat diadaptasi (*fine-tune*) untuk berbagai tugas spesifik seperti klasifikasi teks. Sun dkk. (2019) menegaskan bahwa BERT merupakan model representasi berbasis *fine-tuning* yang mencapai

hasil *state-of-the-art* pada berbagai tugas NLP, serta menunjukkan potensi besar dari pendekatan *fine-tuning* untuk klasifikasi teks.

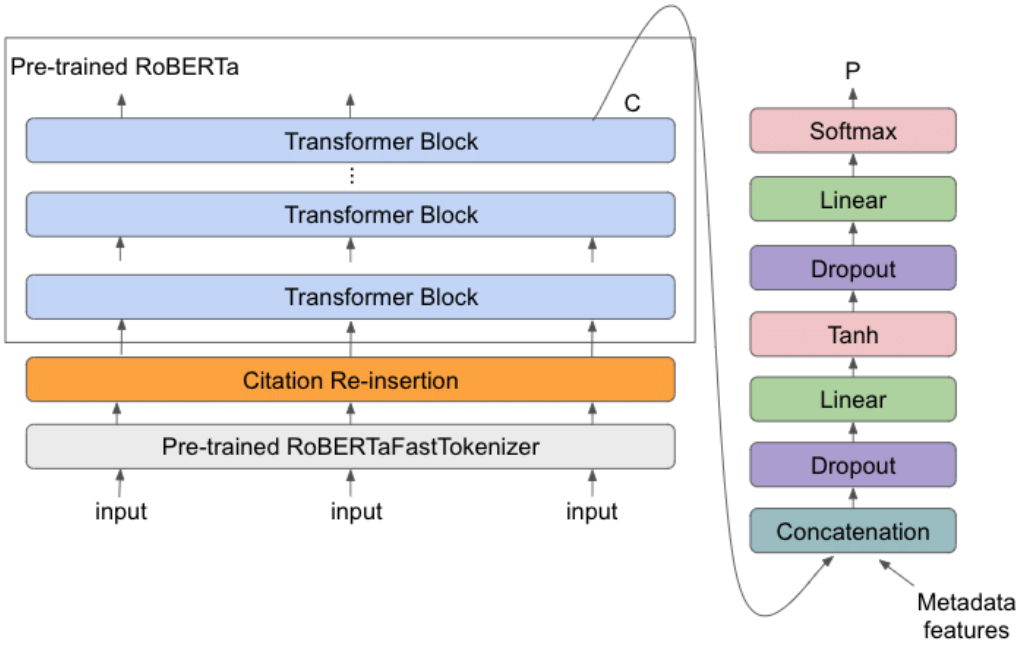
Arsitektur dasar pada kedua fase tersebut tetap sama dan perbedaannya terutama terletak pada lapisan output, sementara seluruh parameter turut disesuaikan pada tahap *fine-tuning*. Dalam skema ini, token [CLS] digunakan sebagai simbol khusus di awal input, sedangkan [SEP] berfungsi sebagai token pemisah antarbagian teks.

2.1.6.1 BERT-base Uncased

BERT-base Uncased merupakan konfigurasi standar dari arsitektur BERT yang terdiri dari 12 lapisan enkoder (*Transformer blocks*), 768 unit tersembunyi (*hidden units*), dan 12 *self-attention heads* dengan total parameter mencapai 110 juta (Devlin dkk., 2019). Istilah *"uncased"* merujuk pada tahap pra-pemrosesan teks di mana seluruh karakter diubah menjadi huruf kecil atau *lowercase* dan tanda aksen (diakritik) dihapus sebelum proses tokenisasi dilakukan. Pendekatan ini sangat efektif untuk menangani korpus teks yang memiliki variasi penggunaan huruf kapital yang tidak beraturan, seperti pada teks refleksi atau deskripsi yang ditulis oleh anak-anak.

2.1.6.2 RoBERTa (Robustly Optimized BERT Approach)

RoBERTa (*Robustly Optimized BERT Approach*) yang dikembangkan oleh Liu dkk. (2019) merupakan varian BERT yang telah dioptimasi secara menyeluruh untuk meningkatkan kualitas representasi bahasa. Berbeda dengan arsitektur aslinya, RoBERTa dilatih dengan durasi yang lebih lama serta ukuran *batch* (*batch size*) yang lebih besar dalam menjamin stabilitas proses optimasi.



Gambar 2.6 Arsitektur RoBERTa (Huang dkk., 2021)

Gambar 2.6 menunjukkan arsitektur RoBERTa yang digunakan sebagai model klasifikasi berbasis *Transformer*. Pada arsitektur ini, teks masukan terlebih dahulu diproses menggunakan *pre-trained* RoBERTaFastTokenizer kemudian hasil tokenisasi yang telah melalui tahap *citation re-insertion* dimasukkan ke dalam beberapa *Transformer Block* untuk membentuk representasi kontekstual. Selanjutnya, keluaran dari lapisan akhir diteruskan ke rangkaian *dropout*, linear, dan *tanh activation* untuk menghasilkan vektor akhir lalu dipetakan melalui softmax menjadi probabilitas keluaran.

Inovasi utamanya terletak pada penghapusan tugas *Next Sentence Prediction* (NSP) serta penerapan strategi *dynamic masking* yang memvariasikan pola *masking* pada setiap *epoch* pelatihan (Liu dkk., 2019). Dengan memanfaatkan korpus data yang lebih besar serta optimasi proses pre-training, RoBERTa terbukti mampu mengungguli BERT pada berbagai metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score* dalam tugas NLP (Fahmi dkk., 2025).

2.1.7 *Gated Recurrent Unit* (GRU)

Gated Recurrent Unit (GRU) merupakan varian dari *Recurrent Neural Network* (RNN) yang diperkenalkan oleh Cho dkk. (2014). Arsitektur ini memanfaatkan dua mekanisme gerbang yaitu *reset gate* (r_t) dan *update gate* (z_t), yang secara adaptif mengatur aliran informasi pada unit tersembunyi sehingga model dapat menentukan informasi mana yang perlu dipertahankan atau diabaikan saat memproses data sekuensial. Alur perhitungan pada sel GRU didefinisikan melalui rangkaian Persamaan (4), (5), (6), dan (7) (Cho dkk., 2014).

$$z_t^j = \sigma(W_z x_t + U_z h_{t-1})^j \quad \textbf{(4)}$$

$$r_t^j = \sigma(W_r x_t + U_r h_{t-1})^j \quad \textbf{(5)}$$

$$\tilde{h}_t^j = tanh\big(W\, x_t + U(r_t \odot h_{t-1})\big)^j \quad \textbf{(6)}$$

$$h_t^j = \big(1 - z_t^j\big) h_{t-1}^j + z_t^j \tilde{h}_t^j \quad \textbf{(7)}$$

Pada Persamaan (4) hingga (7), x_t adalah input pada waktu ke-t, h_{t-1}^j adalah *hidden state* sebelumnya, σ dan tanh adalah fungsi aktivasi sigmoid dan hiperbolik tangen, W dan U adalah matriks bobot yang dipelajari, serta $\odot$ adalah perkalian elemen per elemen. *Update gate* (z_t) mengontrol informasi *hidden state* yang dipertahankan, sementara *reset gate* (r_t) mengatur penggunaan informasi masa lalu. GRU mampu menangkap dependensi jangka panjang secara efisien dengan jumlah parameter lebih sedikit dibandingkan LSTM namun dengan performa yang sebanding (Chung dkk., 2014).

2.1.7.1 Bidirectional Gated Recurrent Unit (BiGRU)

Berbeda dengan GRU standar yang hanya memproses informasi secara kronologis, BiGRU mengadopsi konsep pemrosesan dua arah seperti yang telah diimplementasikan oleh Yu dkk. (2019) dimana BiGRU bertujuan untuk

menangkap konteks masa lalu dan masa depan secara simultan melalui dua lapisan tersembunyi (*forward* dan *backward*) yang bekerja terpisah. Proses penghitungan *hidden state* pada BiGRU dirumuskan dari Persamaan (8), (9), dan (10) (Yu dkk., 2019).

$$h_t^{fwd} = \mathrm{GRU}\left(x_t, h_{t-1}^{fwd}\right) \quad \textbf{(8)}$$

$$h_t^{bwd} = \mathrm{GRU}\left(x_t, h_{t+1}^{bwd}\right) \quad \textbf{(9)}$$

$$H_t = \sigma\left(h_t^{fwd} \oplus h_t^{bwd}\right) \quad \textbf{(10)}$$

Berdasarkan Persamaan (8) hingga (10), simbol $\oplus$ melambangkan konkatenasi *hidden state* maju h_t^{fwd} dan mundur h_t^{bwd} menghasilkan H_t yang merangkum konteks dari kedua arah. Kemampuan ini memungkinkan model menangkap informasi dari awal maupun akhir kalimat secara simultan, mencegah hilangnya konteks penting pada kalimat panjang (Bahdanau dkk., 2016).

2.1.8 Mekanisme *Attention*

Mekanisme *attention* telah menjadi komponen fundamental dalam pemrosesan data sekuensial karena mampu menghubungkan ketergantungan antar elemen tanpa terbatasi oleh jarak dalam urutan input maupun output (Vaswani dkk., 2023). Melalui mekanisme ini, model dapat secara dinamis memberikan bobot kepentingan yang berbeda pada setiap bagian dari data masukan sehingga fitur yang paling relevan dapat diberikan bobot lebih besar.

2.1.8.1 *Self-Attention*

Self-attention atau sering disebut sebagai *scaled dot-product attention* memungkinkan model untuk menghitung hubungan antar elemen dalam suatu urutan. Dalam mekanisme ini, setiap input diproyeksikan menjadi tiga representasi

vektor yaitu *Query* (Q), *Key* (K), dan *Value* (V). Secara matematis mekanisme *attention* dirumuskan menggunakan Persamaan (11) (Wang dkk., 2020).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad \textbf{(11)}$$

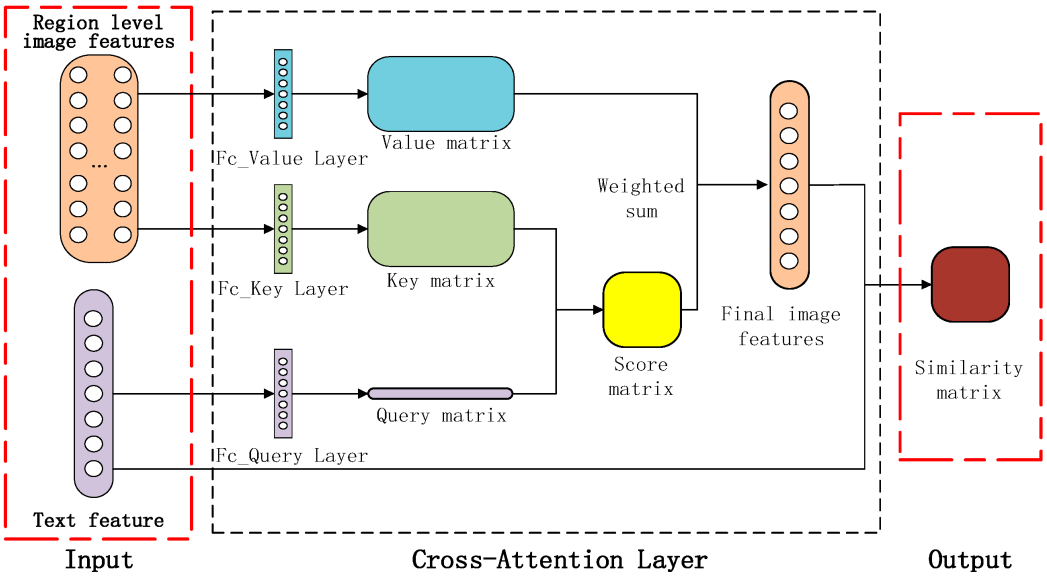
Berdasarkan Persamaan (11), skor atensi diperoleh dari hasil perkalian titik (*dot-product*) antara Q dan K yang kemudian dinormalisasi dengan akar kuadrat dari dimensi kunci ($\sqrt{d_k}$) untuk menjaga stabilitas gradien. Fungsi *softmax* digunakan untuk menghasilkan bobot probabilitas yang menentukan seberapa besar kontribusi setiap elemen terhadap representasi akhir.

Mekanisme ini berbeda dengan *additive attention* yang menghitung fungsi kompatibilitas menggunakan jaringan *feed-forward* dengan satu lapisan tersembunyi. Meskipun keduanya memiliki kompleksitas teoretis yang serupa, *dot-product attention* lebih cepat serta lebih efisien dalam penggunaan memori karena dapat diimplementasikan menggunakan operasi perkalian matriks yang telah dioptimalkan (Vaswani dkk., 2023). Hal ini memungkinkan model untuk memahami hubungan antara kata sifat dan kata benda dalam satu kalimat refleksi anak meskipun keduanya terpisah jarak yang jauh.

2.1.8.2 *Cross-Modal Attention*

Cross-modal attention merupakan pengembangan dari konsep atensi yang digunakan khusus untuk mengintegrasikan informasi dari dua modalitas yang berbeda, dalam hal ini adalah fitur visual dari gambar coretan dan fitur tekstual dari narasi anak. Mekanisme ini memiliki kemampuan untuk mempelajari relevansi yang mendalam di antara modalitas yang berbeda tersebut (Chen dkk., 2021). Berbeda dengan *self-attention*, pada mekanisme ini vektor *Query* biasanya berasal

dari satu modalitas (misalnya teks) sementara *Key* dan *Value* berasal dari modalitas lainnya, seperti fitur pada level regional gambar (Zheng dkk., 2022).

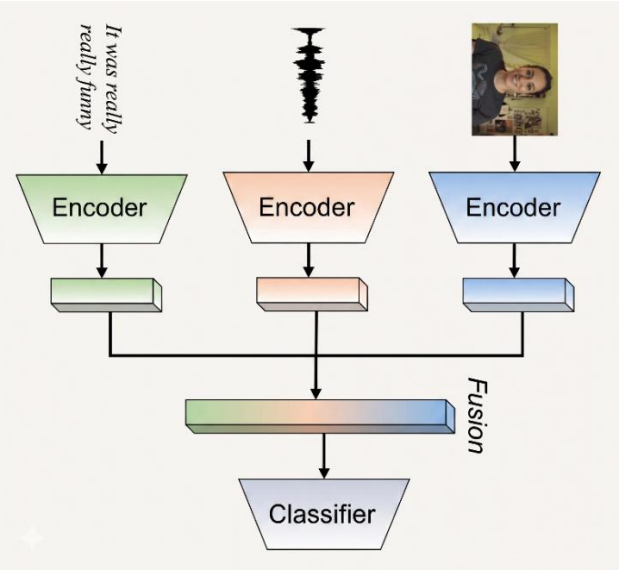


Gambar 2.7 Cara Kerja *Cross-Modal Attention* (Zheng dkk., 2022)

Penerapan *cross-modal attention* yang divisualisasikan pada Gambar 2.7 memungkinkan model untuk melakukan pencarian relasional antara dua domain data yang berbeda. Mekanisme ini mengevaluasi relevansi fitur visual dari gambar terhadap fitur semantik dari teks secara dinamis sehingga informasi yang saling mendukung dapat diberikan bobot lebih besar. Proses penyelarasan (*alignment*) ini sangat penting untuk mengatasi ambiguitas interpretasi misalnya sebuah coretan abstrak dapat dipahami maknanya secara lebih akurat berkat informasi kontekstual dari refleksi anak. Maka dari itu, *cross-modal attention* dapat berfungsi sebagai komponen inti yang mengoptimalkan kualitas fusi multimodal yang menghasilkan output fitur dengan kedalaman informasi jauh melampaui metode penggabungan fitur konvensional.

2.1.9 *Multimodal Learning*

Berdasarkan taksonomi yang dikemukakan oleh Baltrušaitis dkk. (2018), *multimodal learning* bertujuan untuk membangun model yang mampu memproses dan menghubungkan informasi dari berbagai modalitas. Pendekatan ini dapat memberikan peluang dalam menangkap korelasi antar-modalitas supaya memperoleh pemahaman yang lebih mendalam terhadap informasi dunia nyata yang kompleks.



Gambar 2.8 Kerangka Kerja *Multimodal Learning* (Wei dkk., 2024)

Gambar 2.8 memperlihatkan bahwa keberhasilan paradigma ini tidak hanya ditentukan oleh ketersediaan data tetapi pada strategi fusi yang mampu menyatukan karakter modalitas yang beragam. Kelebihannya, tahap fusi juga bisa dirancang secara fleksibel sehingga tidak bergantung pada satu strategi fusi tertentu melainkan mampu mengakomodasi berbagai metode penggabungan fitur dari yang sederhana hingga yang kompleks untuk menghasilkan representasi multimodal yang optimal (Wei dkk., 2024).

Tabel 2.1 Strategi Fusi Multimodal (Baltrušaitis dkk., 2018)

Strategi Fusi	Deskripsi	Kelebihan	Kelemahan
Early Fusion	Integrasi fitur setelah ekstraksi (konkatenasi)	Menangkap korelasi fitur tingkat rendah (low-level)	Sulit menangani data yang sangat heterogen
Late Fusion	Integrasi setelah keputusan unimodal dibuat	Fleksibilitas model dalam ketahan terhadap data hilang	Mengabaikan interaksi fitur tingkat rendah
Hybrid Fusion	Gabungan early fusion dan prediktor unimodal	Mengeksploitasi keuntungan kedua metode sebelumnya	Arsitektur lebih kompleks dan biaya komputasi tinggi

Tabel 2.1 menunjukkan *trade-off* ketiga strategi dimana *early fusion* menangkap korelasi antar-modalitas lebih dalam namun komputasinya berat, *late fusion* lebih ringan namun melewatkan interaksi halus, sedangkan *hybrid fusion* menyeimbangkan keduanya melalui integrasi bertahap.

2.1.9.1 *Weighted Logit Fusion*

Dalam konteks klasifikasi multimodal berbasis dua cabang (*dual-branch*), salah satu pendekatan *late fusion* pada level keputusan yang umum digunakan adalah *weighted logit fusion*. Teknik ini menggabungkan output *logits* dari masing-masing cabang modalitas dengan bobot tertentu sebelum dikenakan fungsi aktivasi akhir (misalnya *softmax*). Secara matematis *weighted logit fusion* dirumuskan sebagai Persamaan Persamaan (12) (Ye, 2025).

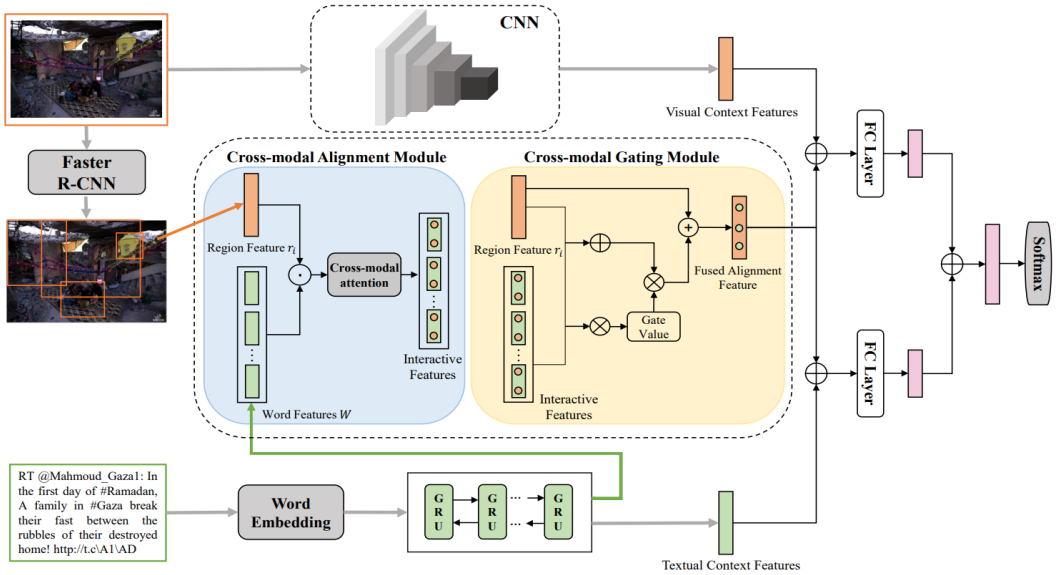
$$logits_{final} = w_A \cdot logits_A + w_B \cdot logits_B \quad \textbf{(12)}$$

Pada Persamaan (12), logits_A dan logits_B adalah vektor *logits* dari masing-masing cabang klasifikasi sedangkan w_A dan w_B adalah parameter bobot yang menentukan kontribusi relatif setiap cabang terhadap keputusan akhir dengan syarat $w_A + w_B = 1$. Pendekatan ini memberikan fleksibilitas untuk mengatur

prioritas modalitas yang dianggap lebih dominan dalam konteks tugas tertentu. Sebagai contoh, pada tugas klasifikasi emosi yang sangat bergantung pada informasi visual, bobot cabang visual dapat ditetapkan lebih besar dibandingkan cabang tekstual. Keunggulan *weighted logit fusion* dibandingkan dengan *concatenation* sederhana terletak pada kemampuannya untuk mempertahankan ruang keputusan masing-masing cabang secara independen sehingga noise dari satu modalitas tidak langsung mendominasi keputusan akhir (Baltrušaitis dkk., 2018).

2.1.10 *Image-Text Interaction Network* (ITIN)

Image-Text Interaction Network merupakan arsitektur multimodal yang dirancang untuk mengeksplorasi hubungan antara wilayah afektif pada gambar dan kata-kata relevan dalam teks pada tingkat halus (*fine-grained*), berbeda dengan metode konvensional yang hanya menggabungkan fitur global (Zhu dkk., 2022). Kerangka kerja ITIN terdiri dari empat komponen utama yaitu (a) Ekstraksi fitur regional dan tekstual menggunakan Faster R-CNN (pretrained Visual Genome) untuk region gambar dan BERT-Base dengan BiGRU untuk encoding teks, (b) *Cross-modal Alignment Module* berbasis *cross-modal attention* untuk mengintegrasikan region gambar dan kata dalam *embedding space*, (c) *Adaptive Cross-modal Gating Module* dengan *soft gate* yang mengontrol intensitas fusi secara adaptif dengan nilai tinggi untuk pasangan yang selaras, nilai rendah untuk menyaring yang tidak selaras, dan (d) Integrasi konteks menggunakan ResNet18 untuk konteks visual global dan rata-rata *word embedding* untuk konteks tekstual.



Gambar 2.9 Kerangka Kerja ITIN (Zhu dkk., 2022)

Gambar 2.9 mengilustrasikan alur ITIN dimana gambar diproses melalui Faster R-CNN dan teks melalui BiGRU, kemudian *Cross-modal Alignment* dan *Gating Module* bekerja berurutan untuk mengintegrasikan interaksi lokal antar fragmen kedua modalitas. Seluruh representasi yang terintegrasi kemudian dipetakan melalui lapisan *multilayer perceptron* dan *softmax* untuk menghasilkan prediksi label sentimen akhir.

2.1.11 *Valence-Arousal Regression*

Berbeda dengan pendekatan klasifikasi diskret yang terbatas pada label emosi kategorikal, regresi *Valence-Arousal* (VA) dirancang untuk memprediksi nilai dalam spektrum kontinu yang memungkinkan representasi emosi secara lebih bervariasi. Pendekatan dimensional ini menawarkan solusi yang lebih bernuansa untuk membedakan berbagai emosi yang direpresentasikan dalam ruang numerik kontinu, di mana urutan teks dipetakan ke dalam koordinat dua dimensi VA (Mendes & Martins, 2023).

Secara teoretis, model regresi VA berbasis *transformer* umumnya terdiri dari sebuah *language encoder* primer yang dipadukan dengan *regression head*. Mekanisme kerja teoretis dari sistem regresi ini mencakup beberapa tahapan sistematis sebagai berikut:

a. Proses Tokenisasi

   Merupakan tahap awal untuk mengubah masukan teks mentah menjadi representasi numerik dan *attention mask* agar dapat diproses oleh arsitektur saraf.

b. Ekstraksi Fitur Kontekstual

   Tahapan ini memanfaatkan lapisan *self-attention* untuk mengekstraksi informasi semantik dan hubungan antar kata dalam sebuah kalimat supaya menghasilkan ringkasan representasi yang lengkap.

c. Proyeksi Ruang Kontinu

   Fitur semantik yang telah diekstraksi kemudian dipetakan melalui lapisan linier yang berfungsi mentransformasikan dimensi vektor tersembunyi ke dalam ruang dua dimensi.

d. Kuantifikasi Skor Dimensi

   Tahap akhir menghasilkan keluaran berupa prediksi nilai numerik yang merepresentasikan intensitas emosi pada sumbu *Valence* dan *Arousal*, di mana standar skala yang umum digunakan merujuk pada konvensi dataset internasional seperti EmoBank.

Penggunaan regresi VA dalam pemodelan emosi memberikan keunggulan dalam menangkap hubungan rumit antara representasi teks dan intensitas perasaan

sehingga menghasilkan generalisasi yang lebih baik dibandingkan pendekatan kategorikal sederhana.

2.1.12 Metrik Evaluasi

Evaluasi kinerja pemodelan secara empiris dapat diukur menggunakan beberapa metrik standar dalam bidang *machine learning* (Sokolova & Lapalme, 2009). Pada riset multimodal, evaluasi disesuaikan dengan jenis objektif fungsional dari masing-masing lapisan keluaran seperti klasifikasi, regresi, dan keandalan probabilitas.

2.1.12.1 Metrik Klasifikasi

Tabel 2.2 Metrik Evaluasi Klasifikasi

Metrik	Persamaan	Deskripsi
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Persentase jumlah tebakan model yang tepat dari keseluruhan data uji
Precision	$\frac{TP}{TP + FP}$	Ketepatan prediksi positif
Recall	$\frac{TP}{TP + FN}$	Sensitivitas deteksi
F1-*Score*	$2 \cdot \frac{P \cdot R}{P + R}$	Rata-rata harmonik *precision* dan *recall*

Keterangan:

TP (*True Positive*) : Jumlah data yang sebenarnya positif dan diprediksi positif oleh model

TN (*True Negative*) : Jumlah data yang sebenarnya negatif dan diprediksi negatif oleh model

FP (*False Positive*) : Jumlah data yang sebenarnya negatif tetapi diprediksi positif oleh model

FN (*False Negative*) : Jumlah data yang sebenarnya positif tetapi diprediksi negatif oleh model

P (*Precision*) : Tingkat ketepatan model dalam memprediksi kelas positif

R (*Recall*) : Kemampuan model dalam menemukan seluruh data positif

Tabel 2.2 merupakan metrik klasifikasi yang dipilih untuk memberi gambaran yang berimbang dimana nilai dari *precision*, *recall*, dan F1-score akan diperlukan untuk membaca keseimbangan akurasi prediksi model pada masing-masing kelas emosi tanpa bergantung mutlak ke skala akurasi agregat (Sokolova & Lapalme, 2009). Secara keseluruhan, presisi gabungan mendeteksi kecenderungan apakah model memihak *false positive* maupun *false negative* dibandingkan tebakan aslinya.

Selain metrik evaluasi di atas, tugas klasifikasi emosi diskret juga memerlukan fungsi loss yang sesuai untuk mengoptimasi distribusi probabilitas kelas selama pelatihan. *Cross-Entropy Loss* merupakan fungsi loss standar untuk tugas klasifikasi yang mengukur ketidaksesuaian antara distribusi probabilitas prediksi model dan distribusi target sebenarnya. Secara sederhana *Cross-Entropy Loss* untuk klasifikasi multi-kelas atau *Categorical Cross-Entropy* didefinisikan sebagai Persamaan (13) (Zhu dkk., 2022).

$$L_{CE} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \qquad \textbf{(13)}$$

Di mana C adalah jumlah kelas, y_c adalah label *ground truth* biner (1 untuk kelas target, 0 untuk lainnya), dan $\hat{y}_c$ adalah probabilitas prediksi *softmax* untuk kelas c. Formula ini secara efektif menghitung $-\log(\hat{y}_{c^*})$, memberikan penalti lebih besar ketika model memprediksi probabilitas rendah untuk kelas yang benar. *Categorical Cross-Entropy* dipilih karena konvergensinya lebih cepat dan generalisasinya lebih kuat pada tugas klasifikasi gambar dan sentimen (Li, 2025).

2.1.12.2 Metrik Regresi

Untuk komponen pemodelan nilai afektif kontinu seperti regresor *Valence* dan *Arousal*, parameter model mengandalkan perbandingan prediksi menggunakan pendekatan fungsi jarak. Indikator metrik yang digunakan mencakup jarak galat numerik absah atau *Mean Squared Error* (MSE) dan koefisien keselarasan atau *Concordance Correlation Coefficient* (CCC). Perhitungan *Mean Squared Error* dapat didefinisikan sebagai Persamaan (14) (Al Qohar dkk., 2025).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \textbf{(14)}$$

Persamaan (14) mendefinisikan perhitungan *Mean Squared Error* (MSE) yang mengukur rata-rata kuadrat selisih antara nilai aktual (y_i) dan nilai prediksi ($\hat{y}_i$) untuk mengevaluasi besarnya galat pada model regresi. Selanjutnya, pengukuran keselarasan antara dua variabel tersebut dilakukan melalui Persamaan (15) (Shoer & Erzin, 2025).

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \qquad \textbf{(15)}$$

Keterangan:

ρ_c : *Concordance Correlation Coefficient* (rentang -1 hingga 1)

ρ : Koefisien korelasi Pearson antara nilai prediksi dan nilai aktual

σ_x : Standar deviasi dari distribusi nilai prediksi

σ_y : Standar deviasi dari distribusi nilai aktual

μ_x : Rata-rata (*mean*) dari nilai prediksi

μ_y : Rata-rata (*mean*) dari nilai aktual

CCC mengevaluasi derajat kesepakatan antara prediksi model dan *ground truth* dengan mengintegrasikan korelasi Pearson (ρ), perbedaan mean, dan

perbedaan variansi dalam satu metrik. Nilai ρ_c mendekati 1 menunjukkan prediksi yang selaras sempurna dengan distribusi target (Hutson & Yu, 2021).

Selain MSE dan CCC, *Root Mean Squared Error* (RMSE) digunakan untuk menyajikan galat dalam skala unit yang sama dengan data asli dengan menggunakan Persamaan (16) (Galvão dkk., 2021).

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \textbf{(16)}$$

Penggunaan RMSE sangat krusial dalam analisis afektif karena metrik ini memberikan penalti yang lebih besar terhadap *outlier* (galat yang besar) namun tetap menghasilkan nilai yang lebih intuitif untuk diinterpretasikan dibandingkan MSE.

2.1.12.3 Metrik Kalibrasi

Untuk evaluasi performa model, kemampuan probabilitas prediksi klasifikasi diukur juga menggunakan skor *Expected Calibration Error* (ECE) (Guo dkk., 2017).

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \qquad \textbf{(17)}$$

Pada Persamaan (17), ECE mengukur deviasi rata-rata antara skor kepercayaan (*confidence score*) prediksi model dan akurasi aktualnya. Sampel dikelompokkan ke dalam *M bins* berdasarkan kepercayaan, lalu selisih absolut antara $acc(B_m)$ dan $conf(B_m)$ pada setiap *bin* dihitung dan dijumlahkan secara berbobot. Metrik ini penting untuk memastikan kepercayaan model benar-benar merepresentasikan akurasi di dunia nyata (Guo dkk., 2017).

2.1.13 *Deployment* Model dan Antarmuka *Web*

Deployment model ke pengguna akhir dapat direalisasikan menggunakan FastAPI yaitu *web framework* Python berkinerja tinggi berbasis ASGI/Uvicorn yang mendukung inferensi *non-blocking* sehingga cocok untuk *model serving* multimodal. Visualisasi hasil prediksi dirender melalui Jinja2 yaitu *templating engine* Python yang menginjeksi nilai prediksi emosi dan *Valence-Arousal* langsung ke dokumen HTML dari sisi server (*Server-Side Rendering*).

2.2 *State of the Art*

Berbagai peneliti telah memanfaatkan pendekatan *deep learning* untuk pengenalan emosi dan analisis sentimen multimodal khususnya dengan mengintegrasikan modalitas gambar, teks, audio, maupun sinyal fisiologis. Penelitian-penelitian tersebut menggali berbagai aspek mulai dari pengembangan dataset emosi dimensi *Valence-Arousal*, desain arsitektur jaringan interaksi image-text seperti DINN dan ITIN, strategi fusi multimodal, hingga analisis emosi pada gambar coretan anak menggunakan dataset KIDO. Perbedaan fokus penelitian, modalitas yang digunakan, serta metode yang diterapkan menunjukkan keragaman pendekatan dalam menyelesaikan permasalahan pengenalan emosi secara komputasional. Berikut Tabel 2.3 adalah ringkasan penelitian-penelitian terkait yang menjadi rujukan dalam penelitian ini.

Tabel 2.3 *State of The Art*

No.	***State of The Art***	
1.	**Peneliti, Tahun**	Zhu dkk. (2022)
	Judul	*Multimodal Sentiment Analysis With Image-Text Interaction Network*
	Objek	Analisis sentimen multimodal pada pasangan gambar-teks dari media sosial Twitter, dengan klasifikasi tiga kelas sentimen (positive, neutral, negative), menggunakan dataset MVSA-Single (4.511 pasangan) dan MVSA-Multiple (17.024 pasangan).
	Algoritma	*Image-Text Interaction Network (ITIN)* dengan tiga komponen utama yaitu (1) *Cross-modal Alignment Module* yang menggunakan Faster R-CNN (ResNet-101, pra-latih pada Visual Genome) untuk deteksi region gambar dan BERT-Base + Bidirectional GRU untuk encoding kata, dengan region-word affinity matrix untuk alignment lintas modalitas, (2) *Cross-modal Gating Module* berbasis soft gate untuk menyeleksi pasangan region-word yang relevan secara adaptif, dan (3) integrasi konteks visual (ResNet18) dan konteks tekstual (rata-rata word embedding) dengan parameter penyeimbang lambda (lambda = 0,2). Optimisasi menggunakan Adam dengan cross-entropy loss.
	Hasil	**Hasil:** Pada MVSA-Single: akurasi 75,19% dan F1-score 74,97%. Pada MVSA-Multiple: akurasi 73,52% dan F1-score 73,49%. ITIN melampaui model baseline terbaik (MVAN) sebesar 2,21% akurasi pada MVSA-Single dan 1,16% pada MVSA-Multiple, membuktikan efektivitas alignment region-word tingkat halus disertai mekanisme gating adaptif.
2.	**Peneliti, Tahun**	Huang dkk. (2025)
	Judul	*An Enhanced Valence-Arousal Multimodal Emotion Dataset for Emotion Recognition*
	Objek	Pengembangan dataset emosi multimodal baru (EVA-MED) yang mencakup sinyal EEG, ECG, dan Pulse Interval (PI) dari 64 partisipan untuk meningkatkan presisi pemodelan dimensi

		Valence-Arousal dengan mempertimbangkan perbedaan individual.
	Algoritma	Dua paradigma induksi emosi yaitu *video stimuli* afektif untuk valence (positif, netral, negatif), dan Mannheim Multicomponent Stress Test (MMST) untuk induksi arousal tinggi. Validasi performa menggunakan SVM, Decision Tree, Random Forest, XGBoost, MLP, dan *1D Convolutional Neural Network (1dCNN).*
	Hasil	**Hasil :** Model 1dCNN menunjukkan performa terbaik dengan akurasi 90,46% dan 93,44% untuk klasifikasi valence dan arousal berbasis EEG, serta 82,00% dan 86,65% untuk ECG. Dataset EVA-MED berhasil menyediakan cakupan emosional yang mendalam, meliputi rendah hingga tinggi arousal dan spektrum penuh valence, yang diperkuat oleh data kuesioner psikologis (kepribadian, kecemasan, depresi).
3.	**Peneliti, Tahun**	Almulla (2024)
	Judul	*A Multimodal Emotion Recognition System Using Deep Convolution Neural Networks*
	Objek	Pengembangan sistem pengenalan emosi multimodal yang mampu mengklasifikasikan tujuh emosi dasar (happy, angry, sad, afraid, disgust, surprise, neutral) dari tiga modalitas: audio, video (ekspresi wajah), dan teks, menggunakan deep convolutional neural networks dengan strategi decision-level fusion.
	Algoritma	Deep Convolutional Neural Networks (DCNN) dengan tiga modul paralel yaitu (1) ekstraksi fitur visual wajah menggunakan CNN berbasis TensorFlow dengan 64 landmark wajah, (2) ekstraksi fitur audio menggunakan *Mel-Frequency Cepstral Coefficient (MFCC)* dengan Conv2D dan Adam optimizer, dan (3) klasifikasi teks menggunakan Logistic Regression, SVM, Random Forest, SGD, MLP, Naive Bayes, dan Decision Tree. Fusi akhir menggunakan *decision-level fusion* berbasis bobot Mehrabian (video 55%, audio 38%, teks 7%). Dataset: FER-2013 (video), RAVDESS (audio), dan ETD (teks).

	Hasil	**Hasil :** Sistem mencapai akurasi 100% untuk modalitas audio, 69% untuk video, dan 64% untuk teks. Setelah penerapan decision-level fusion, akurasi keseluruhan sistem meningkat menjadi 80%. Performa ini sebanding atau bahkan melampaui beberapa metode terkini yang menggunakan dataset yang sama.
4.	**Peneliti, Tahun**	Wang dkk. (2020)
	Judul	*Dynamic Interaction Networks for Image-Text Multimodal Learning*
	Objek	Pemodelan interaksi dinamis antara modalitas gambar dan teks pada empat tugas multimodal: Visual Question Answering (VQA), penilaian estetika gambar berbasis teks (text-IAA), klasifikasi burung (text-Bird), dan klasifikasi bunga (text-Flower), menggunakan dataset VQA, CUB, Oxford Flowers-102, dan AVA-review.
	Algoritma	***Dynamic Interaction Neural Network (DINN)*** terdiri dari dua modul yaitu *Meta Network* yang menggunakan *Multimodal Transformer* dengan multi-dimensional spatial embeddings untuk menyatukan representasi gambar (ResNet-152) dan teks (Bidirectional LSTM dan GloVe) ke dalam ruang fitur bersama, kemudian menghasilkan parameter dinamis W(z) melalui dekomposisi low-rank. Kemudian modul *Main Network* yang menerapkan fungsi interaksi (We-IF, Se-IF, DANs, Co-Att) dengan parameter kontekstual dinamis tersebut. Seluruh modul dilatih secara end-to-end.
	Hasil	**Hasil :** DINN dengan Co-Att mencapai akurasi 60,05% pada VQA (test-std), 97,9% pada text-Flower, 79,0% pada text-Bird, dan 90,4% pada text-IAA. Secara rata-rata, DINN meningkatkan performa lebih dari 2% dibandingkan model interaksi statis maupun model interaksi dinamis tanpa Multimodal Transformer, menunjukkan efektivitas pendekatan parameter learning yang mampu menangkap makna emosional secara lebih menyeluruh.
5.	**Peneliti, Tahun**	Çiftçi dkk. (2025)
	Judul	*Emotion Analysis of Children's Drawings*
	Objek	Analisis emosi komputasional pada gambar coretan anak menggunakan dataset KIDO yang baru diperkenalkan: 10.860

		gambar pindaian beserta refleksi tertulis dari 5.430 siswa sekolah dasar dan menengah (usia 6-14 tahun) di Sanliurfa, Turki, mencakup dua kelas emosi (happy/sad) beserta metadata gender dan tingkat pendidikan.
	Algoritma	**Klasifikasi gambar** menggunakan transfer learning berbasis CNN (VGG19, ResNet101, DenseNet121, EfficientNet, ConvNeXt) dan Vision Transformer (ViT, DeiT, CaiT, Swin, BEiT) dengan optimasi hyperparameter Optuna (10-fold cross-validation). **Klasifikasi teks** menggunakan model transformer pra-latih (BERT, RoBERTa, DistilBERT, ALBERT). **Generasi gambar** menggunakan Deep Convolutional Generative Adversarial Network (DCGAN) yang dimodifikasi untuk mensintesis gambar happiness dan sadness. Augmentasi data meliputi flipping, rotasi acak, color jittering, dan random resized cropping.
	Hasil	**Hasil :** DenseNet121 mencapai F1-score terbaik 79,03% untuk klasifikasi emosi dari gambar, kemudian ConvNeXt terbaik untuk klasifikasi tingkat pendidikan (F1 84,28%) dan gender (F1 77,45%). Untuk teks, ALBERT mencapai F1-score 97,47% pada klasifikasi emosi. Eksperimen subjektif dengan 70 siswa menunjukkan bahwa anak-anak sulit membedakan gambar buatan AI dari gambar teman sebaya, dengan 80% anak memilih gambar AI sebagai representasi happiness.
6.	**Peneliti, Tahun**	John & Kawanishi (2022)
	Judul	*Audio and Video-based Emotion Recognition using Multimodal Transformers*
	Objek	Pengenalan emosi audio-visual untuk interaksi manusia-robot menggunakan tiga dataset publik: RAVDESS (24 aktor, 8 emosi, 2 kalimat), CREMA-D (91 aktor, 6 emosi, 12 kalimat), dan SAVEE (4 aktor, 7 emosi, 15 kalimat) yang kemudian diuji melalui tiga skema eksperimen yaitu individual dataset, speaker-dependent, dan speaker-independent.
	Algoritma	***Multimodal Transformers*** arsitektur deep learning dengan tiga cabang transformer yaitu Log-Mel spectrogram diembed oleh 3 lapis Conv-1D dan learnable position embedding atau *audio*

		self-attention, (2) *video self-attention* yaitu MobileNet pre-trained untuk ekstraksi fitur frame kemudian diembed Conv-2D dan *Block Embedding* sebagai skema temporal embedding baru, dan (3) *audio-video cross-attention* yaitu video sebagai query dan audio sebagai key dan value. Output tiga cabang digabungkan oleh dua fully connected layers dan softmax. Optimisasi menggunakan Adam (lr=0,001), categorical cross-entropy loss, 50 epoch, batch size 6, TensorFlow 2.
	Hasil	**Hasil :** Pada eksperimen individual dataset, model mencapai akurasi 93,59% (RAVDESS), 72,45% (CREMA-D), dan 99,17% (SAVEE), melampaui semua baseline (CNN, RNN, LSTM) dengan selisih signifikan. Pada eksperimen gabungan, akurasi speaker-dependent mencapai 71,10% dan speaker-independent 51,01%. Studi ablasi mengonfirmasi bahwa kombinasi audio/video self-attention dan cross-attention memberikan performa terbaik, menegaskan efektivitas Block Embedding untuk mempertahankan informasi temporal video.
7.	**Peneliti, Tahun**	Yang dkk. (2021)
	Judul	*Image-Text Multimodal Emotion Classification via Multi-View Attentional Network*
	Objek	Penelitian ini berfokus pada klasifikasi emosi berbasis multimodal yang memanfaatkan pasangan data gambar dan teks dari media sosial. Studi ini juga memperkenalkan dataset baru bernama TumEmo yang dikumpulkan dari platform Tumblr dengan lebih dari 190.000 pasangan gambar dan teks yang diberi label emosi seperti angry, bored, calm, fear, happy, love, dan sad. Dataset ini dikembangkan untuk mengatasi keterbatasan dataset sebelumnya yang umumnya hanya memiliki label polaritas sentimen seperti positif, negatif, dan netral.
	Algoritma	Penelitian ini mengusulkan model deep learning bernama Multi-View Attentional Network (MVAN) untuk melakukan klasifikasi emosi multimodal. Model ini terdiri dari tiga tahap utama yaitu feature mapping, interactive learning, dan feature fusion. Pada tahap feature mapping, fitur teks diekstraksi menggunakan word embedding berbasis GloVe, sedangkan fitur

		gambar diekstraksi dari dua perspektif yaitu object view dan scene view untuk menangkap informasi visual yang lebih kaya. Pada tahap interactive learning digunakan mekanisme attention memory network untuk memodelkan interaksi antara modalitas teks dan gambar sehingga kedua modalitas dapat saling mempengaruhi dalam proses pembelajaran fitur. Selanjutnya pada tahap feature fusion, fitur teks dan gambar digabungkan menggunakan metode intermediate fusion melalui multilayer perceptron dan stacking-pooling module untuk menghasilkan representasi multimodal yang lebih mendalam sebelum dilakukan klasifikasi akhir menggunakan softmax.
	Hasil	**Hasil :** Hasil eksperimen menunjukkan bahwa model MVAN mampu mencapai performa yang lebih baik dibandingkan berbagai metode baseline pada dataset MVSA-Single, MVSA-Multiple, dan TumEmo. Model multimodal ini menunjukkan peningkatan akurasi dan F1-score karena mampu memanfaatkan informasi emosional yang saling melengkapi dari gambar dan teks secara simultan. Selain itu, performa model pada dataset TumEmo yang memiliki jumlah data lebih besar menunjukkan stabilitas yang lebih baik dibandingkan dataset yang lebih kecil.
8	**Peneliti, Tahun**	Ham dkk. (2022)
	Judul	*A Negative Emotion Recognition System with Internet of Things-Based Multimodal Biosignal Data*
	Objek	Penelitian ini berfokus pada pengembangan sistem pengenalan emosi negatif berbasis data biosinyal multimodal yang dikumpulkan melalui perangkat Internet of Things (IoT). Sistem ini dirancang untuk mendukung pemantauan kesehatan mental dengan mengumpulkan berbagai sinyal fisiologis seperti electroencephalogram (EEG), detak jantung, galvanic skin response (GSR), dan suhu kulit. Data tersebut diperoleh dari perangkat wearable seperti EEG headset dan smartband yang terhubung dengan aplikasi berbasis IoT.
	Algoritma	Sistem yang diusulkan terdiri dari beberapa komponen utama yaitu perangkat pengumpulan data biosinyal, aplikasi Android berbasis IoT, server IoT yang kompatibel dengan standar

		oneM2M, serta modul klasifikasi emosi. Data biosinyal yang dikumpulkan meliputi lima sinyal EEG dari headset serta sinyal fisiologis tambahan seperti heart rate, galvanic skin response, dan skin temperature dari smartband. Data tersebut kemudian diproses dan digunakan sebagai input untuk model klasifikasi yang bertujuan mengenali beberapa jenis emosi negatif. Pendekatan multimodal digunakan agar sistem dapat menangkap pola emosional secara lebih akurat dibandingkan pendekatan berbasis satu jenis sinyal saja.
	Hasil	**Hasil :** Hasil eksperimen menunjukkan bahwa integrasi berbagai biosinyal dalam kerangka multimodal mampu meningkatkan performa sistem dalam mengenali emosi negatif. Sistem ini juga menunjukkan potensi penerapan dalam pemantauan kondisi emosional secara real-time melalui perangkat wearable dan platform IoT. Dengan memanfaatkan beberapa jenis biosinyal secara simultan, sistem dapat menangkap respons fisiologis pengguna yang berkaitan dengan kondisi emosional secara lebih lengkap.
9	**Peneliti, Tahun**	Liang dkk. (2021)
	Judul	*Deep Metric Network via Heterogeneous Semantics for Image Sentiment Analysis*
	Objek	Penelitian ini berfokus pada analisis sentimen gambar dengan memanfaatkan informasi semantik yang berasal dari berbagai sumber untuk meningkatkan akurasi klasifikasi emosi pada gambar. Studi ini mencoba mengatasi keterbatasan metode sebelumnya yang hanya mengandalkan fitur visual objek atau scene tanpa mempertimbangkan makna semantik global dari gambar. Untuk itu, penelitian ini memanfaatkan kombinasi fitur visual dan informasi semantik dari caption gambar supaya menangkap hubungan yang lebih kaya antara konten visual dan makna emosional yang terkandung di dalamnya.
	Algoritma	Penelitian ini mengusulkan metode bernama Deep Metric Network via Heterogeneous Semantics (DMN-HS). Model ini memanfaatkan pendekatan deep metric learning untuk mempelajari representasi fitur yang mampu membedakan

		sentimen gambar secara lebih efektif. Dalam pendekatan ini, fitur visual gambar diekstraksi menggunakan jaringan convolutional neural network (CNN), sementara informasi semantik tambahan diperoleh melalui proses image captioning yang menghasilkan deskripsi tekstual dari gambar. Informasi visual dan semantik tersebut kemudian digabungkan dalam kerangka heterogeneous semantics untuk mempelajari hubungan antara representasi gambar dan label sentimen melalui metric learning. Dengan pendekatan ini, model dapat memetakan gambar dengan sentimen serupa ke dalam ruang fitur yang berdekatan sehingga meningkatkan kemampuan klasifikasi sentimen gambar.
	Hasil	**Hasil :** Hasil eksperimen menunjukkan bahwa metode DMN-HS mampu meningkatkan performa klasifikasi sentimen gambar dibandingkan metode konvensional yang hanya menggunakan fitur visual. Integrasi informasi semantik melalui caption gambar terbukti membantu model memahami konteks emosional secara lebih baik. Pendekatan deep metric learning juga memungkinkan pemisahan yang lebih jelas antara kategori sentimen dalam ruang fitur sehingga meningkatkan akurasi pengenalan sentimen gambar.
10	**Peneliti, Tahun**	(Pinto dkk., 2019)
	Judul	*Biosignal-Based Multimodal Emotion Recognition in a Valence-Arousal Affective Framework Applied to Immersive Video Visualization*
	Objek	Pengenalan emosi multimodal berbasis biosignal (ECG, EDA, BVP, Respirasi) dari 23 partisipan menggunakan framework *Valence-arousal* dengan visualisasi video VR immersive.
	Algoritma	SVM (LIBSVM), fitur fisiologis statistik per biosignal, preprocessing filter (FIR/Butterworth), segmentasi BioSPPy, nested 4-fold CV, kernel RBF γ=0.01. Fusi: $\text{Valence} = 0.26 * \text{ECG} + 0.26 * \text{EDA} + 0.25 * \text{BVP} + 0.23 * \text{Resp}$ $\text{Arousal} = 0.27 * \text{ECG} + 0.23 * \text{EDA} + 0.27 * \text{BVP} + 0.24 * \text{Resp}$

	Hasil	**Hasil:** Akurasi fusi multimodal mencapai 69,13% untuk arousal dan 67,75% untuk valence (video terfilter), dan 58,11%/57,12% untuk seluruh video.
11	**Peneliti**	(Zhang dkk., (2025)
	Judul	*TAG-fusion: Two-stage Attention Guided Multi-modal Fusion Network for Semantic Segmentation*
	Objek	Segmentasi semantik multimodal menggunakan empat dataset benchmark publik yang mencakup tiga kombinasi modalitas: RGB-D dengan NYU Depth V2 (1.448 sampel, 40 kategori) dan SUN-RGBD (10.335 sampel, 37 kategori), RGB-L dengan DELIVER (7.885 pasang, 25 kelas, mencakup kondisi adversarial berupa hujan, kabut, malam, serta kegagalan sensor seperti motion blur dan overexposure), dan RGB-T dengan MFNet (1.569 sampel, 8 kelas, distribusi merata siang dan malam).
	Algoritma	Arsitektur TAG-fusion berbasis Transformer dengan dual-stream menggunakan GroupMixFormer (backbone dengan Group-Mix Attention/GMA) untuk ekstraksi fitur dari modalitas RGB dan modalitas bantu secara paralel. Terdiri dari empat tahap utama yaitu (1) *Early Fusion* pada tahap input RGB melalui skip connection dengan operasi konvolusi dan Hardswish untuk mengintegrasikan informasi antarmodalitas sejak dini, (2) *Feature Guidance Module* yang menerapkan CBAM (channel attention dan spatial attention) secara lintas-modal untuk saling membimbing dan meredam noise antarmodalitas secara hierarkis di setiap lapisan ekstraksi, (3) *Feature Interaction Module* menggunakan mekanisme cross-attention yang ditingkatkan dengan operasi agregasi multi-scale convolutional (MSC) berstride berbeda (6, 12, 18) dan linear attention untuk pemodelan dependensi global antarmodalitas secara efisien, dan (4) *Channel Fusion* melalui concatenation, reduksi dimensi MLP, dan depthwise separable convolution dengan skip connection untuk menghasilkan fitur terfusi akhir. Pada tahap decoding, diterapkan supervisi terpisah (MLPDecoder ringan) untuk stream RGB, stream modalitas

		bantu, dan stream fusi, dengan total loss gabungan tiga komponen cross-entropy ($\lambda_1=\lambda_2=1$).
	Hasil	**Hasil:** Pada NYU Depth V2, model mencapai mIoU tertinggi 58,4% (multi-scale), PixAcc 81,0%, dan mAcc 72,2% dengan GroupMixFormer-B, melampaui seluruh baseline termasuk DFormer-L (57,2%) dan CMX MiT-B5 (56,9%). Pada SUN-RGBD, mIoU 53,3% (multi-scale) juga melampaui semua pembanding. Pada DELIVER (RGB-L), mIoU 60,4% melampaui CMNext (58,0%) dan CMX (56,4%). Pada MFNet (RGB-T), mIoU 60,6% (multi-scale) dengan keunggulan khusus pada kondisi malam hari (mIoU 61,2% vs CMX 59,4%). Studi ablasi mengkonfirmasi kontribusi positif setiap modul, dengan peningkatan bertahap dari baseline RGB murni (49,9%) hingga model penuh (56,5%) pada NYU Depth V2.

Berdasarkan Tabel 2.3 yang berisikan rangkuman berbagai penelitian terdahulu, menunjukkan bahwa pendekatan deep learning multimodal telah banyak dikembangkan untuk pengenalan dan klasifikasi emosi dari berbagai modalitas. Pinto dkk. (2019) dan Huang dkk. (2025) membuktikan efektivitas framework *Valence-Arousal* dua dimensi dalam merepresentasikan emosi secara kontinu, meskipun masing-masing masih terbatas pada modalitas biosignal dan sinyal fisiologis. Yang dkk. (2021) dan Zhu dkk. (2022) menunjukkan bahwa fusi gambar-teks melalui mekanisme attention mampu meningkatkan akurasi analisis sentimen dan klasifikasi emosi multimodal secara signifikan dibandingkan pendekatan unimodal. Wang dkk. (2020), John & Kawanishi (2022), Almulla (2024), serta Zhang dkk. (2025) mengonfirmasi pentingnya pemodelan interaksi lintas modalitas yang dalam untuk rekognisi emosi atau sentimen dari berbagai sumber data.

Namun demikian, sebagian besar penelitian tersebut masih menghadapi keterbatasan berupa ketergantungan pada modalitas audio dan video yang tidak

tersedia pada data coretan anak, kurangnya eksplorasi pada domain ekspresi visual anak usia dini yang bersifat abstrak dan tidak terstruktur, serta belum adanya integrasi representasi *Valence-Arousal* kontinu dengan mekanisme interaksi gambar-teks secara terpadu untuk analisis emosi berbasis gambar diam. Kesenjangan tersebut menunjukkan perlunya pendekatan baru yang secara khusus dirancang untuk memahami ekspresi emosi anak melalui modalitas coretan dan teks.

Mengacu pada temuan tersebut, penelitian ini mengusulkan pengembangan *Image-Text Interaction Network* (ITIN) yang mengintegrasikan representasi *Valence-Arousal* ke dalam mekanisme *cross-modal interaction* dengan menggunakan *two-stage attention* dan *adaptive gating* antara fitur visual coretan anak dan fitur semantik teks emosi. Pendekatan ini bertujuan untuk menghasilkan model klasifikasi emosi yang akurat, dapat diinterpretasikan secara psikologis melalui dimensi *Valence-Arousal*, dan mampu menangkap nuansa emosi yang unik pada ekspresi visual non-verbal anak usia dini.

2.3 Matrik Penelitian

Tabel 2.4 Matriks Penelitian

No	Penulis	Model				Dataset / Anotasi		Representasi Emosi		Modalitas Input				Evaluasi			Integrasi VA
		CNN	Trans-former	ITIN*	Lainnya	Manual	Otomatis	VA Space	Diskret	Gambar	Teks	Audio	Bio-signal	Akurasi	F1	CCC	
1	Zhu dkk. (2022)	✓	-	-	-	✓	-	-	✓	✓	✓	-	-	✓	✓	-	-
2	Huang dkk. (2025)	✓	-	-	-	✓	-	✓	-	-	-	-	✓	✓	-	-	✓
3	Almulla (2024)	✓	-	-	-	✓	-	-	✓	✓	✓	✓	-	✓	-	-	-
4	Wang dkk. (2020)	-	✓	✓	-	✓	-	-	✓	✓	✓	-	-	✓	-	-	-
5	Ciftci dkk. (2025)	✓	✓	-	-	✓	-	-	✓	✓	✓	-	-	✓	✓	-	-
6	John & Kawanishi (2022)	-	✓	-	-	✓	-	-	✓	✓	-	✓	-	✓	-	-	-

No	Penulis	Model				Dataset / Anotasi		Representasi Emosi		Modalitas Input				Evaluasi			Integrasi VA
		CNN	Trans-former	ITIN*	Lainnya	Manual	Otomatis	VA Space	Diskret	Gambar	Teks	Audio	Bio-signal	Akurasi	F1	CCC	
7	Yang dkk. (2021)	✓	-	-	-	✓	-	-	✓	✓	✓	-	-	✓	✓	-	-
8	Ham dkk. (2022)	-	-	-	✓	-	✓	-	✓	-	-	-	✓	✓	-	-	-
9	Liang dkk. (2021)	✓	-	-	-	✓	-	-	✓	✓	✓	-	-	✓	-	-	-
10	Pinto dkk. (2019)	-	-	-	✓	✓	-	✓	-	-	-	-	✓	✓	-	-	✓
11	Zhang dkk. (2025)	✓	✓	-	-	✓	-	-	-	✓	-	-	-	-	-	-	-
	Penelitian Ini	✓	✓	✓	-	-	✓	✓	✓	✓	✓	-	-	✓	✓	✓	✓

**ITIN = Image-Text Interaction Network*

Berdasarkan Tabel 2.4, mayoritas penelitian menggunakan CNN atau Transformer untuk ekstraksi fitur, dengan anotasi label emosi secara manual. Çiftçi dkk. (2025) mengombinasikan keduanya untuk analisis emosi gambar coretan anak, namun masih terbatas pada klasifikasi diskret tanpa integrasi Valence-Arousal. Hanya Huang dkk. (2025) dan Pinto dkk. (2019) yang menggunakan kerangka VA, sementara metrik CCC belum digunakan oleh satu pun penelitian terdahulu dan belum ada yang secara sekaligus mengevaluasi model menggunakan akurasi, F1, dan CCC.

Kesenjangan ini mendorong penelitian ini untuk mengusulkan arsitektur ITIN yang mengintegrasikan fitur dimensi kontinu Valence-Arousal dengan mekanisme *two-stage attention* dan *adaptive gating*, dievaluasi secara menyeluruh menggunakan metrik akurasi, F1, dan CCC dalam satu kerangka kerja terpadu.