JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1047-1058
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.531

Pengenalan Ekspresi Wajah Peserta Didik di Ruang Kelas Menggunakan Vision Transformer (ViT)

Muhammad Fakhri Fadhlurrahman[1*], Munir[2], Yaya Wihardi[3]

[1,2,3] Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam, Ilmu Komputer, Universitas Pendidikan Indonesia, Bandung, Indonesia
Email: [1,*]mfakhrif@upi.edu, [2]munir@upi.edu, [3]yayawihardi@upi.edu
(* Email Corresponding Author: mfakhrif@upi.edu)
Received: July 31, 2025 | Revision: August 17, 2025 | Accepted: August 28, 2025

Abstrak

Ekspresi wajah merupakan bentuk komunikasi non-verbal yang penting dalam memahami kondisi emosional peserta didik di ruang kelas. Pemahaman ini dapat membantu pendidik menyesuaikan metode pengajaran sesuai dengan keadaan emosional siswa, sehingga proses belajar mengajar menjadi lebih efektif. Penelitian ini bertujuan untuk mengembangkan dan menerapkan sistem pengenalan ekspresi wajah secara real-time di ruang kelas dengan memanfaatkan arsitektur *Vision Transformer* (ViT). Dua pendekatan sistem dikembangkan dalam penelitian ini: sistem *dual-stage* yang memanfaatkan kombinasi model deteksi wajah YOLOv11s dan model pengenalan ekspresi wajah HybridViT (ResNet-50), serta sistem *single-stage* yang menggunakan model YOLOv11s untuk langsung mendeteksi emosi dari citra wajah. *Dataset* yang digunakan meliputi *Real-world Affective Face Database* (RAF-DB), *Face Detection Dataset*, dan *Facial Expression in Classroom*, yang masing-masing digunakan untuk pelatihan awal dan *fine-tuning* model. Hasil pengujian menunjukkan bahwa sistem *dual-stage* memiliki performa klasifikasi yang lebih baik dengan nilai *mean Average Precision* (mAP) sebesar 0,2846, dibandingkan sistem *single-stage* dengan mAP sebesar 0,1603. Sebaliknya, dari segi efisiensi inferensi, sistem *single-stage* lebih unggul dengan latensi rata-rata per wajah sebesar 0,290 ms (6.539 FPS) di GPU dan 1,862 ms (545 FPS) di CPU, dibandingkan sistem *dual-stage* yang memiliki latensi lebih tinggi. Selain itu, evaluasi menunjukkan ketidakseimbangan performa antar kelas emosi akibat distribusi data yang tidak merata. Secara keseluruhan, kedua pendekatan menunjukkan potensi yang menjanjikan untuk implementasi sistem pengenalan ekspresi wajah di ruang kelas. Keduanya masih dapat ditingkatkan dari segi akurasi, generalisasi antar emosi, serta efisiensi waktu inferensi melalui peningkatan kualitas *dataset* dan eksplorasi teknik pelatihan lanjutan.

Kata Kunci: Pengenalan Ekspresi Wajah, *Vision Transformer*, YOLOv11s, *Real-Time*, Ruang Kelas, *Dual-Stage*, *Single-Stage*

Abstract

Facial expressions serve as an essential form of non-verbal communication in understanding students' emotional states in the classroom. This understanding enables educators to adjust their teaching methods according to students' emotions, thus improving the effectiveness of the learning process. This study aims to develop and implement a real-time facial expression recognition system in classroom settings by utilizing the Vision Transformer (ViT) architecture. Two system approaches were developed: a dual-stage system combining a YOLOv11s face detection model with a HybridViT (ResNet-50) facial expression recognition model, and a single-stage system using a YOLOv11s model to directly detect emotions from facial images. The datasets used include the Real-world Affective Faces Database (RAF-DB) and the Facial Expression in Classroom Dataset, which were employed for model training and fine-tuning, respectively. Evaluation results demonstrate that the dual-stage system achieves superior classification performance with a mean Average Precision (mAP) of 0.2846, compared to the single-stage system's mAP of 0.1603. However, in terms of inference efficiency, the single-stage system outperforms the dual-stage system, achieving a lower average latency per face of 0.290 ms (6.539 FPS) on GPU and 1.862 ms (545 FPS) on CPU. The evaluation also highlights an imbalance in classification performance across emotion classes, primarily due to the uneven distribution of training and fine-tuning data. Overall, both approaches exhibit promising potential for facial expression recognition applications in classroom environments. Further improvements in accuracy, emotional generalization, and computational efficiency can be achieved through enhanced dataset quality, balanced emotion representation, and exploration of advanced training techniques.

Keywords: *Facial Expression Recognition, Vision Transformer, YOLOv11s, Real-Time, Classroom, Dual-Stage, Single-Stage*

1. PENDAHULUAN

Ekspresi wajah merupakan salah satu bentuk komunikasi non-verbal yang penting dalam interaksi manusia [1]. Dalam konteks pembelajaran di kelas, ekspresi wajah dapat merefleksikan berbagai kondisi emosional seperti kemarahan, kebahagiaan, ketakutan, keterkejutan, rasa jijik, dan kesedihan, yang tidak selalu diungkapkan secara verbal [2]. Pemahaman terhadap ekspresi wajah ini sangat penting karena memungkinkan pendidik untuk mengidentifikasi kebutuhan emosional dan akademik peserta didik, sehingga mereka dapat menyesuaikan metode pengajaran secara lebih tepat dan efektif [3].

Dengan kemajuan teknologi, pengenalan ekspresi wajah atau *Facial Expression Recognition* (FER) telah berkembang menjadi sistem komputer yang mampu menganalisis dan mengenali perubahan gerakan wajah dari informasi visual [4]. Teknologi ini memiliki beragam manfaat di berbagai bidang, termasuk interaksi manusia-komputer, *virtual reality*, sistem bantuan pengemudi canggih, hiburan, serta pendidikan [5]. Dalam ranah pendidikan, FER memainkan peran penting untuk memahami emosi siswa melalui ekspresi wajah, memberikan wawasan yang bernilai bagi pendidik. Penelitian sebelumnya menunjukkan bahwa emosi seperti keterlibatan dan frustrasi dapat berdampak signifikan terhadap hasil pembelajaran, dan ekspresi tertentu seperti *mouth dimpling* dapat menjadi indikator positif dalam peningkatan proses

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1047-1058
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.531

belajar [6]. Dengan demikian, pemantauan ekspresi wajah siswa memberikan potensi untuk menyesuaikan gaya pengajaran secara adaptif [7].

Topik FER juga telah menjadi fokus utama dalam bidang *artificial intelligence* (AI) dan *computer vision*, karena kemampuannya untuk menginterpretasi emosi manusia secara otomatis dari citra visual [8]. Teknologi berbasis machine learning, khususnya *Convolutional Neural Networks* (CNN), telah banyak digunakan dalam berbagai aplikasi seperti klasifikasi gambar, pengenalan objek, segmentasi, pemrosesan video, *natural language processing*, hingga pengenalan suara [9]. Dalam konteks Indonesia, CNN telah digunakan di berbagai sektor, mulai dari keamanan [10], ketertiban lalu lintas [11], deteksi kendaraan [12], kesehatan [13], pertanian [14], peternakan [15], hingga pendidikan [16].

Dalam tugas pengenalan ekspresi wajah, CNN memiliki keunggulan dalam mengekstraksi fitur penting dari gambar secara otomatis, membuatnya sangat efektif dalam klasifikasi visual [17]. Meskipun demikian, CNN memiliki keterbatasan, seperti sensitivitas terhadap pencahayaan, sudut pengambilan gambar, serta kesulitan dalam mendeteksi ekspresi mikro atau micro-expressions [18]. Oleh karena itu, penelitian beralih ke pendekatan yang lebih canggih, salah satunya adalah *Vision Transformer* (ViT), yang diperkenalkan oleh [19]. Berbeda dengan CNN, ViT memanfaatkan *attention mechanism* yang memungkinkan model untuk menangkap hubungan spasial antar piksel secara lebih kontekstual melalui pemrosesan gambar dalam bentuk *patch*. Pendekatan ini telah terbukti mampu mengenali pola visual dengan tingkat akurasi yang tinggi [20].

Berbagai penelitian terkait telah dilakukan untuk meningkatkan performa sistem FER. Gunawan [21] menerapkan arsitektur CNN VGG16 pada *dataset* FER2013 yang terdiri dari 35.887 citra wajah dengan tujuh kategori emosi dasar. Dengan menerapkan teknik *pre-processing*, augmentasi data, serta pelatihan menggunakan 100 *epoch* dan *learning rate* 0,001, model Modified VGG16 berhasil mencapai akurasi uji sebesar 70,63%. Namun, peningkatan performa ini memerlukan sumber daya komputasi yang lebih besar dibandingkan model standar, menjadikan efisiensi sebagai salah satu tantangan.

Guntoro [22] merancang model CNN dengan tiga *hidden layer* dan melatihnya menggunakan *dataset* FFHQ. Hasilnya, akurasi pelatihan mencapai 71%, sementara akurasi validasi hanya 65%, menunjukkan adanya masalah *overfitting*. Model juga menunjukkan keterbatasan dalam mengenali ekspresi tertentu seperti *disgust*, yang mengindikasikan kurangnya kemampuan generalisasi terhadap variasi emosi.

Sang [23] mengusulkan arsitektur *deep learning* berbasis CNN menggunakan *dataset* FERC-2013 dan mencatatkan akurasi sebesar 71,9% pada *private test set*, mengungguli pendekatan sebelumnya. Keunggulan utama terletak pada efisiensi jumlah parameter, menjadikan model ini cocok untuk aplikasi *real-time*. Namun, tantangan tetap ada dalam hal risiko *overfitting* dan penerapan di lingkungan dengan sumber daya terbatas.

Minaee [24] mengintegrasikan *attention mechanism* ke dalam arsitektur CNN, menghasilkan akurasi tinggi di berbagai *dataset* seperti FER-2013 (70,02%), FERG (99,3%), JAFFE (92,8%), dan CK+ (98%). Meskipun memiliki keunggulan dalam mengekstraksi fitur yang relevan secara lebih selektif, pendekatan ini tetap menghadapi tantangan dalam hal kebutuhan komputasi yang tinggi dan sensitivitas terhadap ukuran *dataset* yang terbatas.

Faikar [25] berhasil meningkatkan akurasi pengenalan emosi pada data non-frontal melalui pendekatan Facial Region Segmentation (FRS) yang dikombinasikan dengan Log-Gabor Convolutional Networks (Log-GCNs). Metode ini berhasil mencapai akurasi 66.34%, melampaui baseline secara signifikan, meskipun memiliki kelemahan utama yaitu kegagalan total dalam mengenali emosi disgust (muak) dan kebutuhan komputasi yang jauh lebih besar.

Wahyono [26] berfokus pada evaluasi kepuasan pelanggan melalui model dua tahap yang efisien, menemukan bahwa kombinasi Real Time Detection Transformer (RT-DETR) dengan backbone ResNet-18 untuk deteksi wajah dan Real-Time CNN untuk klasifikasi 3 kelas ekspresi (positif, negatif, netral) memberikan performa optimal dengan F1-score 68.4% pada tahap klasifikasi. Namun, model integrasi akhirnya menunjukkan kinerja yang sangat rendah (4.7% AP) karena ketidaksesuaian pose wajah antara data latih dan data uji.

Pendekatan berbasis Vision Transformer juga semakin dikembangkan, seperti yang dilakukan oleh Chaudhari [27] melalui sistem ViTFER. Model ini di-*fine-tune* dari ImageNet untuk tugas pengenalan delapan kelas emosi, menggunakan *dataset* gabungan AVFER (FER-2013, AffectNet, CK+48). ViT-B/16/SAM mencatatkan akurasi tertinggi sebesar 53,10% dan *F1-score* sebesar 0,6220, melampaui *baseline* seperti ResNet-18 (50,05%). Keunggulan ViT terletak pada kemampuannya menangkap fitur global melalui perhatian *multi-head* dan peningkatan stabilitas prediksi melalui teknik augmentasi dan optimisasi. Namun, tantangan besar tetap ada, seperti sensitivitas terhadap kualitas anotasi (noisy labels) dan peningkatan waktu komputasi akibat penggunaan Sharpness-Aware Minimizer (SAM), yang hampir dua kali lebih berat dibanding metode standar—membuatnya kurang ideal untuk aplikasi real-time.

Terinspirasi dari studi-studi tersebut, penelitian ini merancang dan membandingkan dua pendekatan sistem pengenalan ekspresi wajah untuk implementasi di ruang kelas secara real-time. Pendekatan pertama adalah sistem dual-stage yang mengombinasikan model deteksi wajah YOLOv11s dan model klasifikasi ekspresi wajah HybridViT berbasis ResNet-50 sebagai *backbone*. Pendekatan kedua adalah sistem *single-stage*, menggunakan YOLOv11s sebagai model yang langsung mendeteksi emosi sekaligus wajah dalam satu proses *end-to-end*. Kedua pendekatan akan dievaluasi berdasarkan akurasi (*mean Average Precision* - mAP) dan efisiensi waktu inferensi pada CPU maupun GPU, untuk mengetahui kelayakannya dalam implementasi real-time di lingkungan pembelajaran. Dengan demikian, hasil dari penelitian ini diharapkan dapat memberikan kontribusi nyata terhadap pengembangan teknologi pendidikan berbasis AI, serta menjadi referensi bagi penelitian dan implementasi sistem FER di masa mendatang.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1047-1058
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.531

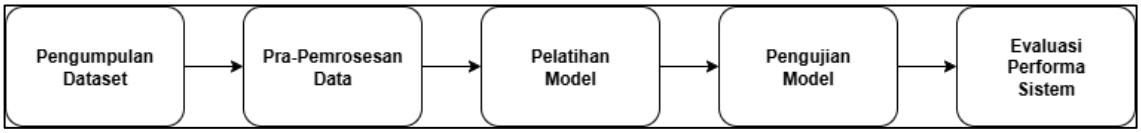
2. METODOLOGI PENELITIAN

2.1 Rancangan Penelitian

Penelitian ini menggunakan pendekatan eksperimen kuantitatif untuk mengembangkan dan membandingkan dua pendekatan sistem pengenalan ekspresi wajah secara *real-time* di ruang kelas:

a. *Dual-stage*: deteksi wajah oleh YOLOv11s dilanjutkan klasifikasi ekspresi oleh HybridViT-ResNet-50.
b. *Single-stage*: deteksi wajah dan klasifikasi emosi dilakukan sekaligus oleh YOLOv11s.

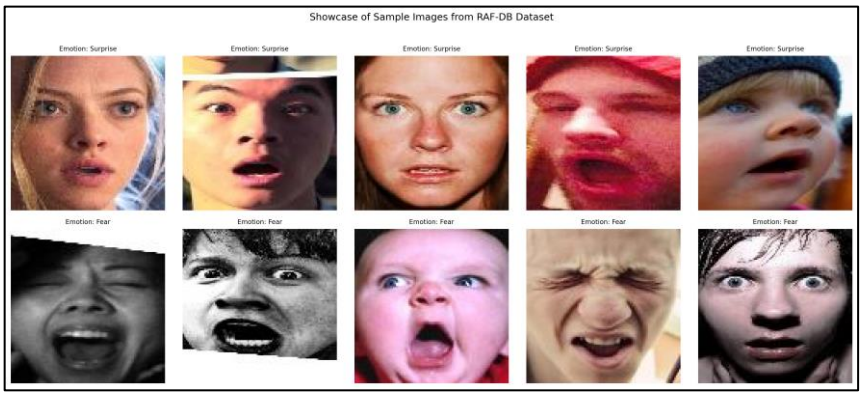
Tahapan penelitian terdiri atas: (1) pengumpulan *dataset*, (2) pra-pemrosesan data, (3) pelatihan model, (4) pengujian model, dan (5) evaluasi performa sistem. Alur umum ditampilkan pada Gambar 1.



Gambar 1. Diagram Alur Tahapan Penelitian

2.2 Dataset dan Pra-pemrosesan

Tiga *dataset* digunakan:



Gambar 2. Cuplikan *Dataset* RAF-DB

a. RAF-DB (Real-world Affective Face Database): terdiri dari 15.300 citra wajah resolusi 100×100 piksel dalam format RGB, diklasifikasikan ke dalam tujuh emosi dasar. *Dataset* ini digunakan untuk *pretraining* model FER dan dipilih karena ekspresinya lebih subtil, menyerupai kondisi alami peserta didik saat pembelajaran. Format dataset yang konsisten dan jumlah datanya yang cukup besar menjadikan RAF-DB sangat ideal sebagai dasar pelatihan awal.

Gambar 3. Cuplikan *Dataset* Face Detection Dataset

b. Face Detection Dataset: mencakup sekitar 16.700 gambar dengan satu atau lebih wajah disertai anotasi *bounding box* dalam format YOLO. Dataset ini digunakan untuk *fine-tuning* YOLOv11s sebagai model deteksi wajah. Setiap *bounding box* mencerminkan posisi wajah dengan akurasi tinggi, memungkinkan pelatihan detektor yang andal dan *robust* terhadap variasi posisi dan orientasi wajah.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1047-1058
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.531

Gambar 4. Cuplikan *Dataset* Facial Expression in Classroom

c. Facial Expression in Classroom: berisi 1.698 gambar dari tiga sesi berbeda. *Dataset* ini digunakan untuk *fine-tuning* model dalam konteks kelas nyata. Sesi 1 dan 2 (1.162 citra) digunakan untuk pelatihan, sedangkan Sesi 3 (536 citra) digunakan untuk pengujian akhir. Ekspresi dalam dataset ini sangat subtil dan natural, mencerminkan variasi emosi peserta didik dalam lingkungan belajar yang sebenarnya. *Dataset* ini juga menyertakan label emosi dan *bounding box* secara eksplisit, memudahkan proses pelatihan *pipeline dual-stage* dan *single-stage*.

Pra=pemrosesan pada dua model:

a. Pra-pemrosesan HybridViT (FER):
   1. Ukuran Input: 224×224 piksel
   2. Augmentasi:
      a. *RandomHorizontalFlip* (50%)
      b. *RandomRotation* (±10°)
      c. *RandomBrightnessContrast*, *HueSaturationValue*, *Affine*, *GaussNoise* (via Albumentations, peluang 30–50%) untuk meningkatkan variasi citra dan mengurangi risiko *overfitting*.
   3. Normalisasi:
      a. Mean: [0.485, 0.456, 0.406],
      b. Std: [0.229, 0.224, 0.225]
   4. *Bounding Box Crop*: wajah dipotong berdasarkan koordinat YOLO dengan *padding* 20% agar konteks wajah tetap terjaga, serta minimum resolusi dipertahankan di atas 64×64 piksel.

b. Pra-pemrosesan YOLOv11s:
   1. *Resize* Otomatis: 640×640 piksel
   2. Augmentasi Internal: mencakup rotasi acak, *flipping* horizontal, dan *scaling* untuk menambah keragaman data
   3. Format Label: YOLO (*class_id, x_center, y_center, width, height*)
   4. Konfigurasi YAML: mendefinisikan jalur file *dataset*, jumlah kelas (nc=7), dan daftar label emosi secara eksplisit.

2.3 Arsitektur Model

a. HybridViT-ResNet-50
   1. *Backbone*: ResNet-50 *pretrained* (tanpa FC layer)
   2. Adaptasi Saluran: Conv2d 1×1 dari 2048 → 3
   3. *Upsample*: *Bilinear* ke 224×224
   4. Vision Transformer:
      a. *Hidden size*: 768
      b. *Layers*: 12
      c. *Heads*: 12
      d. *Patch size*: 16
   5. *Classifier*: *Linear Layer* (768 → 7 kelas emosi)

b. YOLOv11s
   1. *Backbone*: CSPNet dengan C3k2 dan SPPF
   2. *Neck*: *Feature Pyramid Network* (FPN) + *Path Aggregation Network* (PAN)
   3. *Head*: prediksi *bounding box* dan kelas (1 kelas untuk deteksi wajah, 7 kelas untuk deteksi emosi).

2.4 Pelatihan Model

Seluruh pelatihan dilakukan di Google Colaboratory dengan GPU T4 (16 GB). Konfigurasi tiap program dirinci pada tabel berikut:

Tabel 1. Konfigurasi Pelatihan Tiap Program

Program	Model	Dataset	Epoch	Batch Size	Optimizer	Learning Rate	Scheduler

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1047-1058
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.531

1	HybridViT	RAF-DB	20	32	Adam	1e-4	StepLR (step_size=10, gamma=0.1)
2	YOLOv11s (Face)	Face Detection Dataset	16	32	Ultralytics (default)	default	Early Stopping (patience=2)
3	YOLOv11s (Face) + HybridViT	Facial Expression in Classroom (Sesi 1 & 2)	max 50	32	Adam	5e-5	CosineAnnealingLR + Early Stopping (patience=5)
4	YOLOv11s (Emotion)	Facial Expression in Classroom (Sesi 1 & 2)	50	32	AdamW	0.001	Early Stopping (patience=5)

Selain *early stopping*, monitoring dilakukan menggunakan validasi mAP antar *epoch* untuk mencegah *overfitting* dan mengidentifikasi *checkpoint* terbaik.

2.5 Evaluasi dan Pengujian

Pengujian dilakukan pada *dataset* Sesi 3 menggunakan berbagai metrik evaluasi:

a. *mean Average Precision* (mAP): mAP@0.5 dan mAP@0.5–0.95
b. *Average Precision* (AP) per kelas emosi
c. *Precision*, *Recall*, *F1-score*
d. *Confusion Matrix*
e. Waktu inferensi per wajah dan per citra
f. FPS (*Frame per Second*) untuk evaluasi efisiensi sistem

Selain penghitungan numerik, hasil prediksi divisualisasikan dan dibandingkan langsung dengan data *ground truth* untuk mendukung interpretasi kualitatif.

3. HASIL DAN PEMBAHASAN

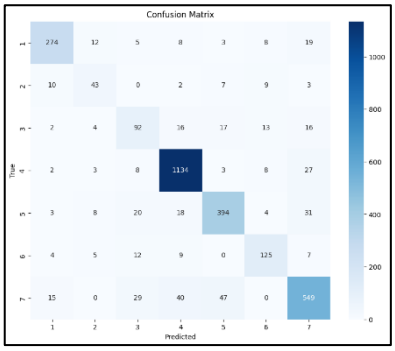
3.1 Lingkungan Penelitian

Penelitian ini dilaksanakan menggunakan dua perangkat utama untuk mendukung proses komputasi, yaitu platform komputasi cloud Google Colaboratory dan laptop pribadi. Google Colaboratory digunakan untuk keperluan pelatihan, *fine-tuning*, evaluasi, dan pengujian berbasis GPU, dengan spesifikasi perangkat keras sebagai berikut: CPU Intel Xeon @ 2.00 GHz, GPU NVIDIA Tesla T4 dengan memori 16 GB GDDR6, dan RAM sebesar 13 GB. Sementara itu, laptop pribadi digunakan untuk pengujian berbasis CPU dan memiliki spesifikasi perangkat keras berupa CPU AMD Ryzen 5 7535HS @ 3.30 GHz, RAM 16 GB, penyimpanan 1 TB SSD, serta sistem operasi Windows 11 Pro 24H2. Dari sisi perangkat lunak, penelitian ini memanfaatkan bahasa pemrograman Python versi 3.11 yang didukung oleh berbagai pustaka, yaitu Albumentations versi 2.0.8, Jupyter Notebook versi 7.4.4, Matplotlib versi 3.10.0, NumPy versi 2.0.2, OpenCV (cv2) versi 4.12.0, Pandas versi 2.2.2, PyYAML versi 6.0.2, Seaborn versi 0.13.2, Scikit-learn versi 1.6.1, Torch versi 2.6.0, Torchvision versi 0.21.0, Transformers versi 4.53.3, dan Ultralytics versi 8.3.170. Seluruh eksperimen dilakukan dalam lingkungan pemrograman Jupyter Notebook pada *platform* Google Colaboratory yang telah terintegrasi dengan GPU untuk mendukung komputasi intensif. Untuk pengujian lokal berbasis CPU, digunakan aplikasi Visual Studio Code pada laptop pribadi dengan konfigurasi perangkat keras dan perangkat lunak sebagaimana disebutkan di atas.

3.2 Evaluasi dan Pembahasan

3.2.1 Model Hybrid Vision Transformer (HybridViT) ResNet-50 (Pengenalan Ekspresi Wajah)

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1047-1058
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.531



Gambar 5. *Confusion Matrix* dari Evaluasi Model HybridViT

Berdasarkan *confusion matrix* pada Gambar 5 terlihat bahwa model secara umum menunjukkan performa yang baik dalam mengklasifikasikan ekspresi wajah. Namun, terdapat beberapa kelas emosi yang menunjukkan performa yang kurang optimal, khususnya *Fear*, *Disgust*, dan *Angry*. Hal ini kemungkinan disebabkan oleh ketidakseimbangan jumlah data pelatihan untuk masing-masing kelas emosi, yang rinciannya disajikan dalam bentuk tabel berikut:

Tabel 2. Jumlah Data Pelatihan untuk Setiap Emosi pada Dataset RAF-DB

ID	Nama	Jumlah Data Pelatihan
1	Surprise	1.290
2	Fear	281
3	Disgust	717
4	Happy	4.772
5	Sad	1.982
6	Angry	705
7	Neutral	2.524

Berdasarkan Tabel 2, dapat dilihat bahwa kelas *Happy* memiliki jumlah data terbanyak, yaitu sebanyak 4.772 gambar. Sementara itu, kelas *Fear* memiliki jumlah data paling sedikit, yaitu hanya 281 gambar. Ketidakseimbangan ini secara signifikan memengaruhi performa model, sebagaimana terlihat pada *confusion matrix* di Gambar 5 dan Gambar 6, yang akan dibahas lebih lanjut pada bagian berikutnya.

```
Classification Report:
              precision    recall  f1-score   support

    Surprise       0.88      0.83      0.86       329
        Fear       0.57      0.58      0.58        74
     Disgust       0.55      0.57      0.56       160
       Happy       0.92      0.96      0.94      1185
         Sad       0.84      0.82      0.83       478
       Angry       0.75      0.77      0.76       162
     Neutral       0.84      0.81      0.82       680

    accuracy                           0.85      3068
   macro avg       0.77      0.76      0.76      3068
weighted avg       0.85      0.85      0.85      3068
```

Gambar 6. *Classification Report* dari Model HybridViT yang Telah Dilatih

Classification Report pada Gambar 6 semakin menegaskan dampak dari ketidakseimbangan jumlah data antar kelas terhadap performa model, baik secara keseluruhan maupun per kategori emosi. Kelas *Happy*, yang merupakan kelas dengan jumlah data terbanyak, memperoleh nilai *f1-score* tertinggi sebesar 0.94. Sebaliknya, kelas-kelas dengan jumlah data di bawah 1.000 seperti *Fear*, *Disgust*, dan *Angry* menunjukkan performa yang relatif lebih rendah, dengan *f1-score* masing-masing sebesar 0.58, 0.56, dan 0.76. Meski demikian, kelas *Angry* dapat dikatakan memiliki performa yang cukup baik dibandingkan *Disgust*, walaupun jumlah datanya serupa. Adapun kelas-kelas dengan jumlah data di atas 1.000, seperti *Sad* dan *Neutral*, umumnya memperoleh *f1-score* di atas 0.80. Secara umum, model menunjukkan performa yang cukup baik dengan *accuracy*, *precision*, *recall*, dan *f1-score* rata-rata di 0.85. Dengan capaian ini, model dapat dikatakan layak untuk di-*fine-tune* lebih lanjut dalam konteks aplikasi pengenalan emosi di lingkungan ruang kelas.

3.2.2 Model YOLOv11s (Deteksi Wajah)

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1047-1058
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.531

```
Ultralytics 8.3.167 🚀 Python-3.11.13 torch-2.6.0+cu124 CUDA:0 (Tesla T4, 15095MiB)
YOLO11s summary (fused): 100 layers, 9,413,187 parameters, 0 gradients, 21.3 GFLOPs
                 Class     Images  Instances      Box(P          R      mAP50  mAP50-95)
                   all       3347      10299      0.892      0.823      0.891      0.597
Speed: 0.2ms preprocess, 3.4ms inference, 0.0ms loss, 2.7ms postprocess per image
```

Gambar 7. Hasil *Fine-Tuning* Model YOLOv11s (Deteksi Wajah)

Berdasarkan Gambar 7, hasil evaluasi metrik menunjukkan keberhasilan proses *fine-tuning*, dengan model YOLOv11s mencapai akurasi deteksi yang tinggi. Nilai mAP@50 sebesar 0,891 mengindikasikan kemampuan model dalam mengidentifikasi objek secara akurat. Selain itu, rata-rata waktu inferensi model tergolong rendah, yaitu sekitar 3,4 ms, menunjukkan efisiensi yang optimal untuk aplikasi *real-time*. Dalam *pipeline* FER, proses deteksi wajah berjalan secara efisien tanpa menjadi hambatan, mendukung performa keseluruhan sistem.

3.2.3 Pipeline Model Hybrid Vision Transformer (HybridViT) ResNet-50 (Pengenalan Ekspresi Wajah) dan YOLOv11s (Deteksi Wajah)

```
Epoch 39/50, Average Train Loss: 0.3014
Epoch 39/50, Validation Loss: 0.4697, Validation mAP: 0.9557
AP for Angry: 0.8851
AP for Disgust: 0.8788
AP for Fear: 1.0000
AP for Happy: 0.9792
AP for Neutral: 0.9471
AP for Sad: 1.0000
AP for Surprise: 1.0000
Saved best model with Validation mAP: 0.9557
```

Gambar 8. Epoch dengan Performa Model HybridViT Terbaik dalam Proses *Fine-Tuning*

```
mAP: 0.5692
AP for Angry: 0.6829
AP for Disgust: 0.4018
AP for Fear: 0.3371
AP for Happy: 0.8536
AP for Neutral: 0.7822
AP for Sad: 0.4027
AP for Surprise: 0.5238
Inference Time per Face (Face Detection + FER): 18.35 ms
Inference Time per Image: 215.10 ms
FPS (Faces): 54.49
FPS (Images): 4.65
```

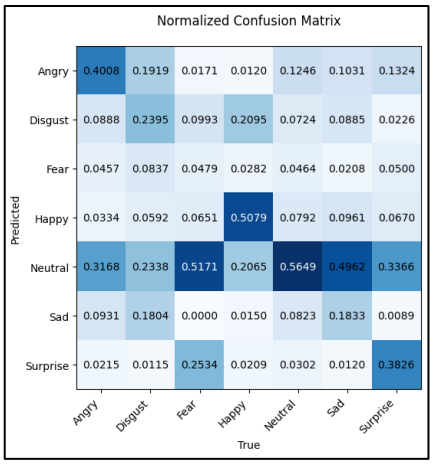
Gambar 9. Hasil Evaluasi Performa Model HybridViT dalam Klasifikasi Emosi setelah *Fine-Tuning*

Gambar 9 menunjukkan performa model pada tahap evaluasi dengan mAP sebesar 0,5692, jauh lebih rendah dibandingkan hasil *fine-tuning* pada Gambar 8. Ini mengindikasikan bahwa meskipun model belajar baik pada data pelatihan, ia kesulitan menggeneralisasi data baru. Faktor utama kemungkinan adalah ekspresi emosi subtil dalam *dataset*, yang sulit dibedakan secara visual, sehingga menyulitkan klasifikasi akurat pada data *unseen*. Selain itu, nilai *Average Precision* (AP) per kelas emosi (Tabel 2) menunjukkan dampak ketidakseimbangan data. Kelas *Fear*, dengan data paling sedikit, mencatat AP terendah (0,3371), sedangkan kelas *Happy*, dengan data terbanyak, mencapai AP tertinggi (0,8536). Hal ini menegaskan bahwa distribusi data yang tidak merata signifikan memengaruhi performa model pada tiap kelas emosi.

Gambar 10. *Evaluation Metrics* dari Pengujian GPU dan CPU

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1047-1058
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.531

Gambar 10 menampilkan metrik evaluasi model yang diuji pada dua lingkungan komputasi: Google Colaboratory (GPU) dan Miniconda (CPU). Nilai mAP pada kedua perangkat menunjukkan konsistensi, dengan 0,2846 (GPU) dan 0,2842 (CPU), namun waktu inferensi GPU jauh lebih cepat dibandingkan CPU. Dibandingkan dengan evaluasi pada Gambar 9, nilai mAP pada pengujian ini lebih rendah, kemungkinan karena kompleksitas data Sesi 3 yang memiliki lebih banyak wajah per gambar dan ekspresi emosi subtil yang sulit dibedakan. Menariknya, AP untuk emosi *Neutral* lebih tinggi daripada *Happy*, diduga karena model cenderung mengklasifikasikan ekspresi subtil sebagai *Neutral* akibat kemiripan visual. Pada GPU, waktu inferensi rata-rata per wajah adalah 15,35 ms (≈65 FPS), memenuhi kebutuhan *real-time* untuk deteksi emosi per individu. Namun, inferensi per citra mencapai 539,15 ms (≈2 FPS), tidak memenuhi standar *real-time*. Pada CPU, waktu inferensi per wajah adalah 134,94 ms (≈7 FPS) dan per citra 4.737 ms (≈0,2 FPS), keduanya jauh dari kriteria real-time. Kompleksitas citra Sesi 3, dengan rata-rata 35 wajah per gambar (Gambar 13), menyebabkan akumulasi waktu inferensi yang tinggi, terutama pada CPU. Hal ini menunjukkan bahwa jumlah wajah dalam satu citra menjadi faktor utama peningkatan waktu inferensi, terutama pada perangkat dengan kemampuan komputasi terbatas.



Gambar 11. *Confusion Matrix* dari Hasil Pengujian *Pipeline*

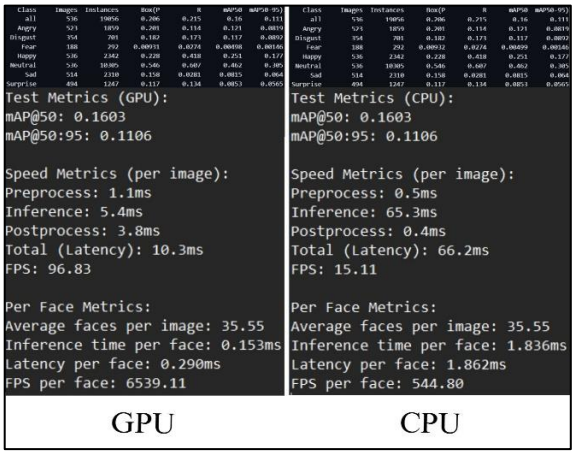
Confusion matrix pada Gambar 11 mengonfirmasi bahwa model memiliki kecenderungan untuk mengklasifikasikan emosi sebagai *Neutral*, kemungkinan besar akibat sifat emosi yang subtil. Kecenderungan model untuk memprediksi emosi sebagai *Neutral* kemungkinan juga disebabkan oleh dominannya jumlah wajah beremosi *Neutral* dalam dataset Facial Expression in Classroom. Hal ini dapat diamati pada cuplikan distribusi *instance* tiap emosi dalam set validasi yang ditampilkan pada Gambar 4.2.4.1. Kondisi ini membuat model lebih terlatih dan terbiasa mengenali ekspresi *Neutral* dibandingkan emosi lainnya yang kurang terwakili. Sementara itu, kelas emosi yang dalam dataset RAF-DB memiliki lebih dari 1.000 gambar, seperti yang ditunjukkan pada Tabel 2, cenderung mampu diklasifikasikan dengan lebih akurat, dengan nilai *True Positive* terendah sebesar 0.3826 pada kelas *Surprise*.

Gambar 12. Hasil Pengujian FER pada Dua Gambar dari Sesi 3

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1047-1058
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.531

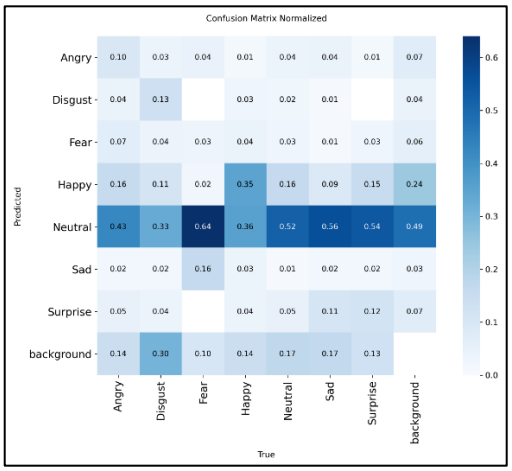
Gambar 12 menunjukkan bahwa proses deteksi wajah dalam *pipeline* berjalan dengan baik. Dengan *confidence threshold* sebesar 0,25, hampir seluruh wajah dalam citra berhasil terdeteksi secara konsisten. Namun, prediksi emosi cenderung kurang akurat, yang kemungkinan besar disebabkan oleh sifat emosi yang subtil dan sulit dibedakan secara visual, sehingga menyulitkan model dalam melakukan klasifikasi secara tepat. Meskipun begitu, skor *confidence* rata-rata berada di atas 0,5, dengan beberapa wajah mencapai nilai *confidence* sebesar 0,9 atau lebih.

3.2.4 Model YOLOv11s (Deteksi Emosi)



Gambar 13. Hasil Evaluasi dan Metrik-Metrik Performa Model GPU dan CPU pada *Dataset* Sesi 3

Gambar 13 menunjukkan bahwa model YOLOv11s (Deteksi Emosi) mencapai mAP@50 sebesar 0,1603 pada pengujian di Google Colaboratory (GPU) dan Miniconda (CPU), menandakan konsistensi performa lintas perangkat, serupa dengan pipeline YOLOv11s (Deteksi Wajah) dan HybridViT (Pengenalan Ekspresi Wajah). Namun, hasil ini jauh lebih rendah dibandingkan mAP@50 sebesar 0,967 pada dataset Sesi 1 dan 2 (Gambar 4.2.4.1), mengindikasikan kegagalan model dalam menggeneralisasi data Sesi 3 yang memiliki lebih banyak wajah dan ekspresi lebih subtil.Performa deteksi emosi model ini juga di bawah *pipeline* FER (mAP ≈0,28), kemungkinan karena *fine-tuning* hanya menggunakan data Sesi 1 dan 2, sehingga kesulitan mengklasifikasikan emosi kompleks pada data Sesi 3 yang bersifat *unseen*. Dari sisi inferensi, model unggul dibandingkan *pipeline* FER (Gambar 9). Pada GPU, waktu inferensi rata-rata per wajah adalah 0,153 ms dengan latensi 0,290 ms (≈6.539 FPS), sementara per citra 5,4 ms dengan latensi 10,3 ms (≈97 FPS), memenuhi kriteria *real-time*. Pada CPU, inferensi per wajah mencapai 1,836 ms dengan latensi 1,862 ms (≈545 FPS), dan per citra 65,3 ms dengan latensi 66,2 ms (≈15 FPS). Performa per wajah pada CPU masih mendukung real-time, tetapi performa per citra hanya optimal pada citra dengan sedikit wajah. Pada skenario intens, performa per citra di CPU dapat menyebabkan pengalaman visual yang kurang mulus.

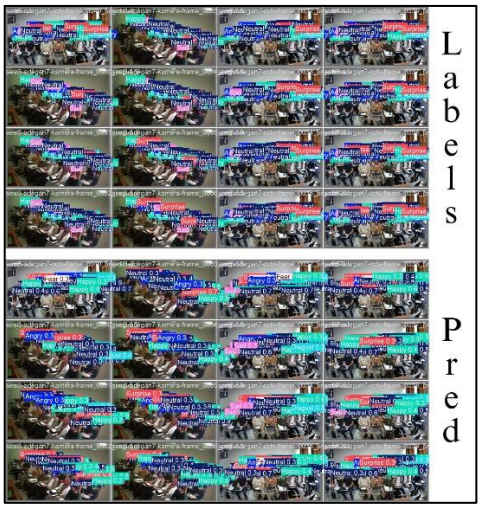


Gambar 14. *Confusion Matrix* dari Hasil Pengujian Model

Confusion matrix pada Gambar 14 menunjukkan pola yang serupa dengan *confusion matrix* pada pipeline FER (Gambar 11), yaitu adanya kecenderungan model untuk mengklasifikasikan berbagai emosi sebagai *Neutral*. Pola ini kemungkinan disebabkan oleh dua faktor utama: dominasi kelas emosi *Neutral* dalam dataset Facial Expression in Classroom serta sifat emosi-emosi lain yang cenderung subtil, sehingga sulit dibedakan secara visual oleh model. Namun, berbeda dengan

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1047-1058
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.531

pipeline FER, model YOLOv11s untuk Deteksi Emosi menunjukkan performa yang lebih lemah dalam mendeteksi emosi secara spesifik. Hal ini dapat disebabkan oleh keterbatasan data pelatihan, di mana model hanya di-*fine-tune* menggunakan *dataset* Facial Expression in Classroom Sesi 1 dan Sesi 2, tanpa proses pelatihan tambahan menggunakan *dataset* umum seperti RAF-DB. Tidak adanya eksposur terhadap data yang lebih bervariasi dan representatif kemungkinan menjadi salah satu penyebab rendahnya kemampuan generalisasi model ini, serta menjelaskan mengapa mAP yang dihasilkan lebih rendah dibandingkan *pipeline* FER.



Gambar 15. Visualisasi Label (Atas) dan Hasil Prediksi (Bawah) Model YOLOv11s (Deteksi Emosi) pada *Batch* Gambar Dataset Facial Expression in Classroom (Sesi 3)

Untuk memberikan gambaran mengenai performa model YOLOv11s (Deteksi Emosi), salah satu hasil deteksi dari *batch* evaluasi yang dihasilkan secara otomatis oleh Ultralytics setelah pengujian dengan dataset Sesi 3 ditampilkan pada Gambar 15. Hasil menunjukkan bahwa model masih kesulitan dalam mengklasifikasikan emosi secara akurat. Pada area dengan dominasi wajah beremosi *Neutral*, model cenderung mampu mendeteksi dengan benar, namun skor *confidence* yang dihasilkan relatif rendah, dengan rata-rata di 0,5 atau lebih rendah dan berkisar antara 0,3 hingga 0,7.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan dan mengimplementasikan sistem pengenalan ekspresi wajah secara *real-time* di lingkungan ruang kelas, dengan dua pendekatan sistem: *dual-stage* menggunakan kombinasi model deteksi wajah YOLOv11s dan model pengenalan ekspresi wajah HybridViT (ResNet-50), serta *single-stage* menggunakan model YOLOv11s yang langsung mendeteksi emosi dari citra wajah. Dari segi kinerja klasifikasi (mAP), sistem *dual-stage* menunjukkan performa yang lebih baik dengan perolehan mAP sebesar 0,2846, dibandingkan sistem *single-stage* yang memperoleh mAP sebesar 0,1603. Hal ini menunjukkan bahwa pemisahan antara deteksi wajah dan klasifikasi ekspresi memberikan hasil yang lebih akurat. Di sisi lain, dari aspek efisiensi inferensi, sistem *single-stage* lebih unggul. Model ini mencatat waktu rata-rata latensi per wajah sebesar 0,290 ms (6.539 FPS) pada GPU dan 1,862 ms (545 FPS) pada CPU, jauh lebih cepat dibandingkan sistem *dual-stage* yang mencatat 15,35 ms (65 FPS) pada GPU dan 134,94 ms (7 FPS) pada CPU. Hal ini menjadikan pendekatan *single-stage* lebih cocok untuk aplikasi *real-time*, terutama pada perangkat dengan keterbatasan sumber daya seperti CPU. Evaluasi terhadap *Average Precision* (AP) per kelas emosi menunjukkan ketidakseimbangan performa klasifikasi antar emosi. Hal ini disebabkan oleh distribusi data yang tidak merata pada *dataset* pelatihan dan *fine-tuning*, di mana emosi tertentu memiliki jumlah data yang jauh lebih banyak dibandingkan emosi lainnya, sehingga memengaruhi kemampuan generalisasi model. Meskipun demikian, kedua pendekatan yang dikembangkan menunjukkan potensi yang menjanjikan untuk aplikasi pengenalan ekspresi wajah di ruang kelas. Keduanya masih dapat ditingkatkan lebih lanjut, baik dari sisi akurasi, generalisasi antar emosi, maupun efisiensi komputasi, melalui peningkatan kualitas dan keseimbangan *dataset*, serta eksplorasi teknik pelatihan lanjutan.

REFERENCES

[1] C. Frith, “Role of facial expressions in social interactions,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3453–3458, 2009, Accessed: Jul. 31, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC2781887/pdf/rstb20090142.pdf

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1047-1058
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.531

[2] M. Batty and M. J. Taylor, "Early processing of the six basic facial emotional expressions," *Cognitive brain research*, vol. 17, no. 3, pp. 613–620, 2003, Accessed: Jul. 31, 2025. [Online]. Available: https://www.ece.uvic.ca/~bctill/papers/facerec/Batty_Taylor_2003.pdf

[3] S. Julika and D. Setiyawati, "Kecerdasan emosional, stres akademik, dan kesejahteraan subjektif pada mahasiswa," *Gadjah Mada Journal of Psychology (GamaJoP)*, vol. 5, no. 1, pp. 50–59, 2019, Accessed: Jul. 31, 2025. [Online]. Available: https://journal.ugm.ac.id/gamajop/article/download/47966/24933

[4] Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression recognition," in *Handbook of face recognition*, Springer, 2011, pp. 487–519. Accessed: Jul. 31, 2025. [Online]. Available: http://www.cs.usfca.edu/~byuksel/affectivecomputing/readings/facial_expression/tian2011.pdf

[5] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," *Symmetry (Basel)*, vol. 11, no. 10, p. 1189, 2019, Accessed: Jul. 31, 2025. [Online]. Available: https://www.mdpi.com/2073-8994/11/10/1189

[6] J. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. Lester, "Automatically recognizing facial expression: Predicting engagement and frustration," in *Educational data mining 2013*, 2013. Accessed: Jul. 31, 2025. [Online]. Available: https://cise.ufl.edu/research/learndialogue/pdf/LearnDialogue-Grafsgaard-EDM-2013.pdf

[7] H. Hikmatiar, N. Sya'bania, and B. Hamsa, "Relation of Facial Expressions and Student Learning Outcomes in Face Recognition-Based Online Learning Article Info," *Jurnal Kajian Teknologi Pendidikan*, vol. 9, no. 1, pp. 1–13, Apr. 2024, doi: 10.17977/um039v9i12024p1.

[8] D. Canedo and A. J. R. Neves, "Facial expression recognition using computer vision: A systematic review," *Applied Sciences*, vol. 9, no. 21, p. 4678, 2019, Accessed: Jul. 31, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/9/21/4678

[9] D. Bhatt *et al.*, "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics (Basel)*, vol. 10, no. 20, p. 2470, 2021, Accessed: Jul. 31, 2025. [Online]. Available: https://www.mdpi.com/2079-9292/10/20/2470

[10] S. Sunardi, A. Fadlil, and D. Prayogi, "Sistem Pengenalan Wajah pada Keamanan Ruangan Berbasis Convolutional Neural Network," *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, vol. 6, no. 2, pp. 636–647, 2022, Accessed: Jul. 31, 2025. [Online]. Available: https://tunasbangsa.ac.id/ejurnal/index.php/jsakti/article/viewFile/480/453

[11] R. Nurhawanti, "Sistem Pendeteksi Sepeda Motor Pelanggar Marka Jalan Menggunakan Metode Convolutional Neural Networks (CNNs)," Universitas Pendidikan Indonesia, Bandung, 2019. Accessed: Aug. 17, 2025. [Online]. Available: http://siad.cs.upi.edu//assets/files/5cd8ea27df49828119.pdf

[12] M. D. L. Yudha, "Deteksi Sepeda Motor di Jalan Raya Menggunakan Faster R-CNN Berbasis VGG16," Universitas Pendidikan Indonesia, Bandung, 2020. Accessed: Aug. 17, 2025. [Online]. Available: http://siad.cs.upi.edu//assets/files/5f4bca8cdf1a121971.pdf

[13] F. A. Febriyanti, "Image Processing Dengan Metode Convolutional Neural Network (Cnn) Untuk Deteksi Penyakit Kulit Pada Manusia," *Kohesi J. Sains dan Teknol*, vol. 3, no. 10, pp. 21–30, 2024, Accessed: Jul. 31, 2025. [Online]. Available: https://ejournal.warunayama.org/index.php/kohesi/article/view/4088/3803

[14] M. R. M. A., "Pemetaan dan Identifikasi Kesiapan Petik Tanaman Teh Berdasarkan Citra Drone Menggunakan Mask Region-Based Convolutional Neural Network (Mask R-CNN) dan Green Leaf Index (GLI)," Universitas Pendidikan Indonesia, Bandung, 2023. Accessed: Aug. 17, 2025. [Online]. Available: http://siad.cs.upi.edu//assets/files/65a4b203c369329936.pdf

[15] S. K. Wulandari and J. Jasmir, "Penggunaan Resnet-50 Untuk Deteksi Penyakit Ikan Air Tawar di Akuakultur Studi Kasus pada Akuakultur Asia Selatan," in *Prosiding Seminar Nasional Bisnis, Teknologi Dan Kesehatan (SENABISTEKES)*, 2024, pp. 17–24. Accessed: Jul. 31, 2025. [Online]. Available: https://www.ejournal.ummuba.ac.id/index.php/SENABISTEKES/article/download/2205/1113

[16] A. Anggraini and H. Zakaria, "Penerapan Metode Deep Learning Pada Aplikasi Pembelajaran Menggunakan Sistem Isyarat Bahasa Indonesia Menggunakan Convolutional Neural Network (Studi Kasus: SLB-BC Mahardika Depok)," *JURIHUM: Jurnal Inovasi dan Humaniora*, vol. 1, no. 4, pp. 452–464, 2023, Accessed: Jul. 31, 2025. [Online]. Available: https://jurnalmahasiswa.com/index.php/Jurihum/article/view/723/432

[17] A. J. Moshayedi, A. S. Roy, A. Kolahdooz, and Y. Shuxin, "Deep learning application pros and cons over algorithm," *EAI Endorsed Transactions on AI and Robotics*, vol. 1, no. 1, p. e7, 2022, Accessed: Jul. 31, 2025. [Online]. Available: https://www.academia.edu/download/115942971/2022_JR_Deep_Learning_Application_Pros_and_Cons_Over.pdf

[18] S. Fakhar *et al.*, "Smart classroom monitoring using novel real-time facial expression recognition system," *Applied Sciences*, vol. 12, no. 23, p. 12134, 2022, Accessed: Jul. 31, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/12/23/12134

[19] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020, Accessed: Jul. 31, 2025. [Online]. Available: https://arxiv.org/pdf/2010.11929/1000

[20] K. Han *et al.*, "A survey on visual transformer," *arXiv preprint arXiv:2012.12556*, 2020, Accessed: Jul. 31, 2025. [Online]. Available: https://arxiv.org/pdf/2012.12556

[21] R. J. Gunawan, B. Irawan, and C. Setianingsih, "Pengenalan Ekspresi Wajah Berbasis Convolutional Neural Network Dengan Model Arsitektur VGG16," *eProceedings of Engineering*, vol. 8, no. 5, 2021, Accessed: Jul. 31, 2025. [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/16400/16113

[22] A. L. S. Guntoro, E. Julianto, and D. Budiyanto, "Pengenalan Ekspresi Wajah Menggunakan Convolutional Neural Network," *Jurnal Informatika Atma Jogja*, vol. 3, no. 2, pp. 155–160, 2022, Accessed: Jul. 31, 2025. [Online]. Available: https://ojs.uajy.ac.id/index.php/jiaj/article/download/6790/2839

[23] D. V. Sang, N. Van Dat, and others, "Facial expression recognition using deep convolutional neural networks," in *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, 2017, pp. 130–135. Accessed: Jul. 31, 2025. [Online]. Available: https://www.researchgate.net/profile/Dinh-

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 2 September 2025
Hal: 1047-1058
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i2.531

Sang/publication/321257241_Facial_expression_recognition_using_deep_convolutional_neural_networks/links/5b12a7824585150a0a619d6c/Facial-expression-recognition-using-deep-convolutional-neural-networks.pdf

[24] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021, Accessed: Jul. 31, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/21/9/3046

[25] N. R. Faikar, "Pengenalan Emosi Manusia Menggunakan Log-Gabor Convolutional Networks Melalui Pendekatan Facial Region Segmentation," Universitas Pendidikan Indonesia, Bandung, 2020. Accessed: Aug. 17, 2025. [Online]. Available: http://siad.cs.upi.edu//assets/files/5f49e8fe6fec438740.pdf

[26] S. M. Wahyono, "Evaluasi Kepuasan Pelanggan Berdasarkan Ekspresi Wajah Menggunakan Real Time Detection Transformer (RT-DETR)," Universitas Pendidikan Indonesia, Bandung, 2025. Accessed: Aug. 17, 2025. [Online]. Available: http://siad.cs.upi.edu//assets/files/67a34fb4799d446045.pdf

[27] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER: facial emotion recognition with vision transformers," *Applied System Innovation*, vol. 5, no. 4, p. 80, 2022, Accessed: Jul. 31, 2025. [Online]. Available: https://www.mdpi.com/2571-5577/5/4/80