Reconocimiento automático de emociones faciales mediante una arquitectura híbrida CNN–Transformer con fusión adaptativa por atención multi-cabeza

Fabrizio Miguel Mattos Cahui

Orientador: Mg Percy Maldonado Quispe

Plan de Tesis presentado la Escuela Profesional Ciencia de la Computación como paso previo a la elaboración de la Tesis Profesional.

UNSA - Universidad Nacional de San Agustín de Arequipa
Octubre de 2025

Abreviaturas

Índice

Índice de cuadros

Índice de figuras

1. Motivación y Contexto

El reconocimiento automático de emociones en imágenes se ha consolidado como un campo fundamental en la visión por computadora, con aplicaciones en vigilancia, salud mental, interacción humano-computadora y moderación de contenido digital. Las emociones humanas, manifestadas a través de expresiones faciales, gestos y posturas, son indicadores clave del estado emocional, lo que permite desarrollar sistemas inteligentes que respondan de manera adaptativa a contextos específicos. Por ejemplo, en vigilancia, detectar emociones como enojo o miedo puede prevenir conflictos (1); en salud mental, identificar emociones negativas puede apoyar diagnósticos tempranos (2); y en interacción humano-computadora, reconocer emociones permite crear interfaces más intuitivas (3).

La digitalización de la sociedad ha incrementado la disponibilidad de datos visuales provenientes de cámaras de seguridad, redes sociales y dispositivos personales, lo que demanda soluciones automatizadas para el análisis emocional. Los métodos tradicionales de reconocimiento de emociones dependen de la observación humana, que es susceptible a errores debido a la fatiga, la subjetividad o el volumen de información visual (4). Las redes neuronales convolucionales (CNNs) han demostrado ser altamente efectivas para analizar patrones visuales complejos, como los asociados a expresiones faciales, superando los enfoques tradicionales basados en características manuales (5). Su capacidad para procesar imágenes en tiempo real y con alta precisión las posiciona como una herramienta clave para abordar estos desafíos.

Este trabajo busca contribuir al desarrollo de sistemas de visión por computadora que automaticen la detección de emociones, mejorando la seguridad, el bienestar psicológico y la experiencia del usuario en entornos digitales. Al aprovechar los avances en CNNs y optimizar su implementación para diversos contextos, esta investigación pretende ofrecer una solución escalable y eficiente que responda a las necesidades de la sociedad moderna.

2. Definición del Problema

El reconocimiento de emociones en imágenes presenta desafíos significativos debido a la variabilidad en las condiciones de captura, como iluminación, ángulos y oclusiones, así como la complejidad de interpretar expresiones faciales que pueden ser ambiguas o similares entre sí, por ejemplo, miedo y sorpresa (6). Los sistemas tradicionales de análisis emocional dependen en gran medida de la supervisión humana o de algoritmos basados en características predefinidas, lo que resulta en limitaciones en precisión, escalabilidad y capacidad de generalización (4). Estos métodos son insuficientes para manejar grandes volúmenes de datos visuales en tiempo real, especialmente en aplicaciones que requieren respuestas rápidas, como vigilancia, salud mental o interacción humano-computadora.

Por lo tanto, existe la necesidad de desarrollar un sistema automático basado en redes neuronales convolucionales (CNNs) que pueda detectar y clasificar emociones en imágenes de manera eficiente, precisa y adaptable a diferentes contextos. Este sistema debe abordar desafíos como el desbalance de clases en los conjuntos de datos, la robustez ante condiciones variables de captura y la interpretabilidad de las decisiones del modelo, para

garantizar su aplicabilidad en escenarios reales (5). El objetivo es superar las limitaciones de los enfoques tradicionales y proporcionar una solución escalable que pueda integrarse en sistemas de vigilancia, herramientas de diagnóstico psicológico y dispositivos interactivos.

3. Objetivos

3.1. Objetivo General

Desarrollar y evaluar un modelo híbrido basado en la combinación de una red convolucional y Swin Transformer Tiny, empleando un módulo de fusión adaptativa con atención multi-cabeza para mejorar la precisión en el reconocimiento automático de emociones faciales.

3.2. Objetivos Específicos

- Preprocesar las imágenes de entrada y aplicar técnicas de aumento de datos para mejorar la generalización del modelo y garantizar la compatibilidad entre ambas ramas de la arquitectura híbrida.
- Implementar modulo de extracción de características de dos etapas en paralelo para extraer características locales y globales.
- Desarrollar un modulo de redimensionamiento que transforme los embeddings generados a un espacio común de dimensión fija, garantizando. compatibilidad para el módulo de fusión.
- Integrar un módulo de fusión basado en atención multi-cabeza que combine de manera adaptativa los embeddings generados por ResEmoteNet y Swin Transformer Tiny.

4. Justificación

El reconocimiento automático de emociones faciales en imágenes tiene un impacto significativo en múltiples dominios, como la vigilancia, la salud mental y la interacción humano-computadora. En vigilancia, detectar emociones como enojo o miedo permite prevenir conflictos de manera proactiva (1). En salud mental, identificar emociones negativas puede facilitar diagnósticos tempranos y monitoreo de pacientes (2). En interacción humano-computadora, el reconocimiento de emociones mejora la experiencia del usuario mediante interfaces adaptativas (3). La creciente disponibilidad de datos visuales provenientes de cámaras de seguridad, redes sociales y dispositivos personales subraya la necesidad de sistemas automatizados que procesen estas señales de manera eficiente y precisa.

5. Trabajos Relacionados

La detección automática de emociones en imágenes es un desafío clave en visión por computadora, con aplicaciones en seguridad, vigilancia, salud mental y moderación de contenido digital. Este problema implica identificar emociones, como enojo, miedo, felicidad o tristeza, a través de expresiones faciales, gestos o posturas, lo que requiere analizar patrones visuales complejos. Los desafíos incluyen la variabilidad en condiciones de captura (iluminación, ángulos, resolución) y la dificultad para distinguir emociones ambiguas, como miedo y sorpresa. Los trabajos fundamentales en redes neuronales convolucionales (CNNs), como (7), establecieron las bases para el procesamiento de imágenes mediante características jerárquicas, mientras que (8) demostró con AlexNet el potencial de las CNNs en la clasificación de imágenes a gran escala. Estos avances son esenciales para los enfoques modernos en detección de emociones.

(9) propusieron un modelo de detección y reconocimiento facial que utiliza Face Mesh para extraer puntos clave faciales (landmarks) y una red neuronal profunda para comparar rostros con una base de datos. Utilizaron el dataset Labeled Faces in the Wild (LFW) y imágenes capturadas en tiempo real, logrando una precisión de 94.23 % en reconocimiento facial. Este trabajo es relevante para la detección de reacciones violentas porque el uso de Face Mesh permite manejar imágenes no frontales y condiciones variables de iluminación y fondo, lo cual es crucial para identificar expresiones faciales asociadas con emociones intensas. Además, la reconstrucción 3D de rostros mejora la robustez ante poses no frontales. Sin embargo, el modelo se centra en reconocimiento de identidades más que en emociones específicas, y su dependencia de Face Mesh y redes profundas puede aumentar los requisitos computacionales, lo que podría limitar su uso en dispositivos de baja potencia.

(3) revisaron enfoques de reconocimiento de emociones multimodales, combinando señales faciales, de voz y fisiológicas (e.g., ECG, EEG) con técnicas de aprendizaje profundo como CNNs, LSTMs y SVMs. Evaluaron datasets como IEMOCAP y SEED, logrando precisiones de hasta 97 % (SEED) con CNNs y múltiples algoritmos de clasificación. Este trabajo es relevante por su enfoque en emociones intensas como enojo y miedo, detectadas en imágenes faciales y señales fisiológicas, que son indicadores de reacciones violentas. Sin embargo, su enfoque multimodal requiere integrar múltiples fuentes de datos, lo que aumenta la complejidad computacional, y la variabilidad en datasets dificulta la generalización a imágenes estáticas en escenarios reales.

(6) propusieron una Red de Máscaras Residuales (Residual Masking Network) para reconocimiento de expresiones faciales, basada en ResNet34 con bloques de máscaras que ponderan regiones faciales clave (e.g., ojos, nariz, boca). Evaluaron su método en los datasets FER2013 y VEMO, logrando precisiones de 76.82 % (FER2013, con ensamblaje) y 65.94 % (VEMO). Este trabajo es relevante por su enfoque en emociones como enojo y miedo, que indican reacciones violentas, y su capacidad para procesar imágenes en tiempo real (100 fps). Sin embargo, enfrenta desafíos con datasets desbalanceados y emociones complejas (e.g., miedo, tristeza), y su alto número de parámetros (142.9M) puede limitar su implementación en dispositivos con recursos restringidos.

(10) propusieron una red AlexNet-Emotion optimizada con una combinación de pérdida Softmax y pérdida de Isla mejorada para reconocimiento de expresiones faciales, abordando variaciones intra-clase e inter-clase. Evaluaron su método en los datasets CK+ y MMI, alcanzando precisiones de 97.14 % (CK+) y 78.68 % (MMI). Este trabajo es relevante por su enfoque en emociones como enojo y miedo, esenciales para detectar reacciones violentas, y su capacidad para manejar oclusiones mediante redes generativas antagónicas. Sin embargo, su enfoque en imágenes controladas (e.g., CK+) limita su aplicabilidad a escenarios reales con condiciones variables, y la complejidad del modelo puede requerir hardware de alto rendimiento.

(11) propusieron un modelo basado en MobileNet para el reconocimiento en tiempo real de siete emociones faciales (felicidad, tristeza, enojo, miedo, sorpresa, disgusto, neutralidad) utilizando los datasets FER2013 y un conjunto aleatorio de imágenes. Lograron una precisión de 97.9 % en entrenamiento y 100 % en validación, destacando la eficiencia de MobileNet para dispositivos con recursos limitados. Este trabajo es relevante porque se centra en emociones como enojo y miedo, indicadores de reacciones violentas, y aborda desafíos en imágenes reales, como iluminación variable y similitudes visuales entre emociones. Sin embargo, el modelo mostró limitaciones en imágenes reales debido a confusiones entre emociones similares (e.g., miedo vs. sorpresa), y su dependencia de datasets preprocesados como FER2013 puede limitar la generalización a entornos de vigilancia con imágenes no controladas.

(12) propusieron un método para mejorar el reconocimiento del miedo combinando imágenes visibles (procesadas con CNNs) e imágenes térmicas (procesadas con ResNet), utilizando un dataset propio de imágenes sincronizadas. Lograron una precisión de 99.17 % para el miedo, mejorando un 4.54 % respecto a solo imágenes visibles. Este trabajo es relevante porque aborda la baja precisión en el reconocimiento del miedo, un indicador clave de reacciones violentas, y demuestra la robustez de las imágenes térmicas ante variaciones de iluminación. Sin embargo, la dependencia de cámaras térmicas especializadas aumenta los costos y limita su implementación en sistemas de vigilancia convencionales. Además, la falta de datasets públicos con imágenes térmicas y visibles sincronizadas restringe la reproducibilidad.

(13) developed a transfer learning-based FER system using pre-trained DCNNs (e.g., VGG-16, DenseNet-161) fine-tuned with a pipeline strategy. Evaluated on KDEF (including profile views) and JAFFE (frontal views), it achieved 96.51 % and 99.52 % accuracy, respectively, in 10-fold cross-validation. This work is relevant for its high accuracy in recognizing emotions like anger and fear, especially in diverse views, which is critical for surveillance applications. The pipeline fine-tuning strategy enhances performance on small datasets like JAFFE. However, reliance on controlled datasets and high computational demands of deep models like DenseNet-161 limit its deployment on resource-constrained devices, and profile view misclassifications (e.g., fear as surprise) highlight challenges in uncontrolled settings.

(13) desarrollaron un sistema de reconocimiento de expresiones faciales basado en aprendizaje por transferencia con DCNNs preentrenadas (e.g., VGG-16, DenseNet-161), logrando un 96.51 % (KDEF) y 99.52 % (JAFFE) de precisión. Es relevante por su alta

precisión en emociones como ira y miedo, especialmente en vistas de perfil. Sin embargo, su dependencia de conjuntos de datos controlados y alta demanda computacional limitan su despliegue en dispositivos con recursos limitados.

(14) propusieron un sistema de reconocimiento de expresiones faciales basado en Vision Transformers (ViT-B/16), ajustado en un conjunto híbrido AVFER (FER2013, AffectNet, CK+48), logrando un 53.10 % de precisión para ocho emociones y 56.94 % para siete (excluyendo desprecio). El uso de Sharpness-Aware Minimizer (SAM) mejora la generalización en datos ruidosos como AffectNet, relevante para vigilancia en condiciones variables. El aumento de datos equilibra la distribución de clases. Sin embargo, su precisión moderada, alto costo computacional (17.5G FLOPS) y dependencia de vistas frontales o casi frontales limitan su efectividad en escenarios diversos, y su gran tamaño (86.5M parámetros) es impráctico para dispositivos de borde.

(2) propusieron PM-ViT, un marco basado en ViTs ajustado en AffectNet y CK+ aumentado, logrando un 83.78 % de precisión y un 84.0 % de F1-score en AffectNet, y un 99.7 % en CK+ para clasificación binaria de emociones (positivas vs. negativas) y clasificación ternaria de emociones negativas (leve, moderada, severa). Es altamente relevante para detectar emociones negativas intensas como ira y miedo, indicadores de reacciones violentas, y evaluar su intensidad para aplicaciones de salud mental. Su robustez ante etiquetas ruidosas y datos aumentados apoya su uso en vigilancia. Sin embargo, su alta complejidad computacional (arquitectura ViT-Large, 40 GB de RAM GPU), dependencia de imágenes frontales y necesidad de datos etiquetados limitan su aplicabilidad en dispositivos con recursos limitados y vistas diversas, lo que nuestro enfoque busca superar mediante optimización e integración multi-vista.

(15) propusieron modelos CNN para el reconocimiento de emociones faciales en el conjunto Emognition, que incluye diez emociones: diversión, asombro, entusiasmo, agrado, sorpresa, ira, disgusto, miedo, tristeza y neutral. Utilizando transferencia de aprendizaje con Inception-V3 y MobileNet-V2, y un modelo desde cero optimizado con el método Taguchi, lograron una precisión del 96 % y un F1-score de 0.95 en datos de prueba. El preprocesamiento convirtió videos en 2535 imágenes faciales, con aumento de datos para mejorar la robustez. Es relevante por su enfoque en emociones negativas como ira, miedo y disgusto, aplicables a la detección de reacciones violentas en vigilancia. Sin embargo, su dependencia de imágenes estáticas omite el contexto temporal de las expresiones, y la falta de interpretabilidad del modelo limita su uso en aplicaciones críticas. Nuestra propuesta aborda estas limitaciones mediante la integración de características temporales y técnicas de inteligencia artificial explicable (XAI).

(16) propusieron un modelo de reconocimiento de emociones faciales basado en una arquitectura VGG-19 con segmentación, que identifica y aísla regiones faciales clave antes de la clasificación. Evaluaron su método en los datasets FER2013 y CK+, alcanzando una precisión del 92.3 % en FER2013 y 98.7 % en CK+ para siete emociones (felicidad, tristeza, enojo, miedo, sorpresa, disgusto, neutral). Este trabajo es relevante por su enfoque en la segmentación para mejorar la precisión en la detección de emociones, especialmente en condiciones de iluminación variable y poses no frontales. La segmentación permite enfocarse en características faciales específicas, reduciendo el impacto de fondos ruidosos.

Sin embargo, el modelo requiere un preprocesamiento adicional para la segmentación, lo que incrementa el costo computacional, y su rendimiento en entornos no controlados, como imágenes de vigilancia, puede estar limitado por la necesidad de imágenes faciales bien definidas.

(17) introdujeron ResEmoteNet, una arquitectura que combina CNNs, bloques Squeeze-Excitation (SE) y bloques residuales, optimizando la representación de características faciales y reduciendo la pérdida. Evaluaron el modelo en FER2013 (79.79 %), RAF-DB (94.76 %), AffectNet-7 (72.93 %) y ExpW (75.67 %), superando a modelos estado del arte como Ensemble ResMaskingNet y S2D. Este trabajo es relevante por su enfoque en mejorar la precisión y eficiencia computacional, aunque su implementación depende de hiperparámetros específicos y datasets variados, lo que puede requerir ajustes para entornos no controlados.

(18) propone un modelo híbrido ligero para el reconocimiento de expresiones faciales en tiempo real, combinando ShuffleNet V2, una CNN optimizada para eficiencia computacional, con EfficientViT-M2, una variante de Vision Transformer diseñada para mantener bajo costo de inferencia. Ambos modelos extraen características locales y globales de manera complementaria, que luego son fusionadas en un único vector de representación alimentado a un clasificador denso con normalización y dropout, alcanzando 97.3 % de exactitud en KMU-FED y 92.44 % en KDEF, superando a varios métodos state-of-the-art con un tiempo de inferencia de solo 3.3 ms por imagen, lo que lo hace atractivo para aplicaciones en sistemas embebidos y automoción. No obstante, una posible limitación del enfoque es que la fusión de características se realiza de manera estática mediante concatenación, lo que puede restringir la capacidad del modelo para adaptarse dinámicamente a diferentes contextos emocionales; además, el énfasis en la eficiencia podría implicar un menor poder de representación en comparación con arquitecturas más pesadas diseñadas para escenarios donde el tiempo real no es crítico.

Cuadro 1: Cuadro comparativo de trabajos relacionados para la detección de reacciones violentas

Trabajo	**Enfoque del Modelo**	**Conjunto de Datos**	**Emociones/ Comportamientos**	**Precisión/ F1-Score**	**Limitaciones**
(18)	Modelo híbrido que combina ShuffleNet V2 (CNN eficiente) con EfficientViT-M2 (Transformer liviano). Fusión de características por concatenación.	Expresiones faciales básicas y gestos de conductores.	Expresiones faciales básicas y gestos de conductores.	97.3 % (KMU-FED), 92.44 % (KDEF). Tiempo de inferencia: 3.3 ms/imagen.	La fusión es estática (concatenación), lo que limita la adaptabilidad. Posible pérdida de capacidad representativa en escenarios no tiempo real.
(17)	CNNs con Squeeze-Excitation y Residual Blocks (ResEmoteNet)	FER2013, RAF-DB, AffectNet-7, ExpW	7 emociones (ira, disgusto, miedo, etc.)	79.79 % (FER2013), 94.76 % (RAF-DB), 72.93 % (AffectNet-7), 75.67 % (ExpW)	Dependencia de hiperparámetros, ajustes necesarios para entornos no controlados
(14)	Vision Transformers (ViT-B/16) con SAM	AVFER (FER2013, AffectNet, CK+48)	8 emociones (ira, miedo, disgusto, etc.)	53.10 % (8 clases), 56.94 % (7 clases)	Baja precisión, alto costo computacional (86.5M parámetros)
(15)	CNNs (InceptionV3, MobileNetV2, modelo propio con Taguchi)	Emognition	10 emociones (ira, miedo, disgusto, etc.)	96 % / F1: 0.95	Imágenes estáticas, falta de interpretabilidad

6. Propuesta

6.1. Preprocesamiento

- Todas las imágenes (RAF-DB en color y FER2013 en escala de grises) son unificadas en un formato estándar de 224×224 píxeles y 3 canales (RGB).
- Aplica técnicas de aumento de datos (data augmentation) como rotaciones leves, volteo horizontal, cambios de brillo y contraste, con el fin de mejorar la capacidad de generalización.

6.2. Arquitectura híbrida de dos ramas

- ResEmoteNet: CNN especializada para reconocimiento de emociones, encargada de extraer características locales y texturas faciales finas (microexpresiones, arrugas, bordes).
- Swin Transformer Tiny: Transformer jerárquico que captura relaciones globales entre regiones del rostro mediante auto-atención en ventanas locales y desplazadas.

6.3. Extracción de embeddings

Se eliminan las capas de clasificación de cada red para quedarse con los vectores de características intermedios de la rama convolucional y la rama Swin Transformer.

6.4. Proyección y normalización de dimensiones

- Ambos embeddings se proyectan a una misma dimensión mediante capas lineales.
- Esto asegura compatibilidad para la fusión.

6.5. Módulo de fusión con Atención Multi-Cabeza (MHA)

- Los embeddings proyectados se tratan como una secuencia de tokens.
- Se aplica un bloque de Multi-Head Attention, que aprende a asignar pesos adaptativos a cada representación según la muestra.

La arquitectura del modelo propuesto se muestra en la Fig. 1

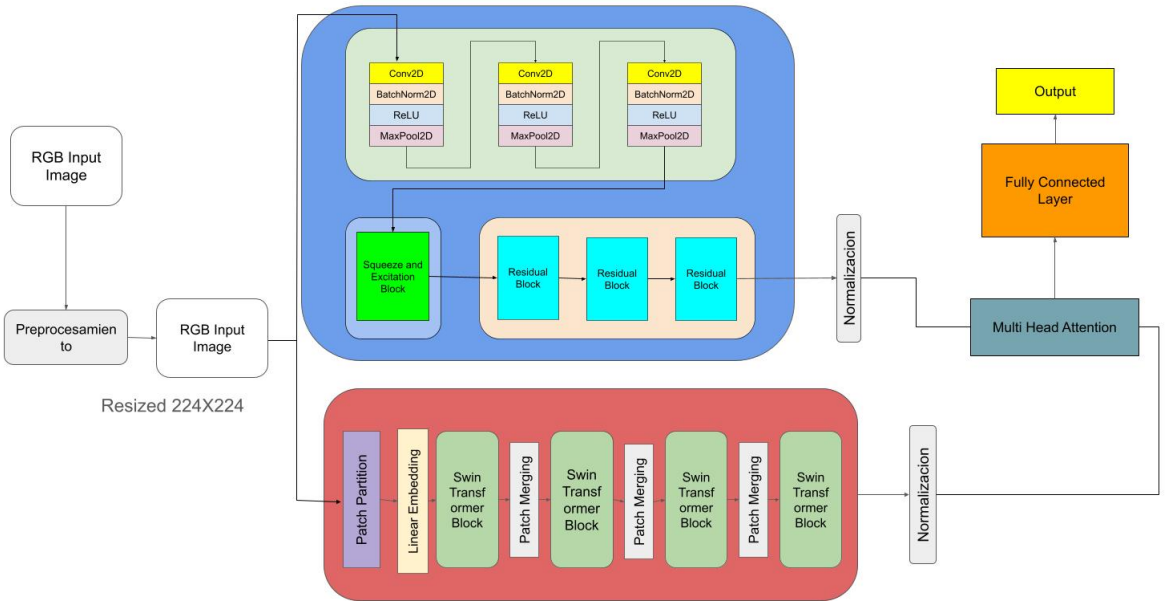

Figura 1: Arquitectura del método propuesto

Referencias

[1] J. Chen, H. Yang, Y.-G. Jiang, and A. G. Hauptmann, "Violence detection in videos: A survey," *Computer Vision and Image Understanding*, vol. 199, p. 103390, 2020.

[2] P. R. Jain, S. M. K. Quadri, and A. Khattar, "Pm-vit: A framework for the recognition of emotions and proclivity toward mental illness using facial expressions," *Journal of Computer Science*, vol. 21, no. 3, pp. 479–493, 2025.

[3] S. M. S. Abdullah, S. Y. Ameen, M. A. M. Sadeeq, and S. R. M. Zeebaree, "Multimodal emotion recognition using deep learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 73–79, 2021.

[4] H. T. Rauf, M. I. U. Lali, S. Zahoor, and S. Kadry, "Vision-based violence detection techniques: A comprehensive survey," *Multimedia Tools and Applications*, vol. 80, pp. 23929–23964, 2021.

[5] M. Hasan, M. M. Islam, J. M. Kim, and Y.-K. Lee, "Facial emotion recognition using convolutional neural networks: A review," *Sensors*, vol. 22, no. 3, p. 903, 2022.

[6] L. Pham, T. H. Vu, and T. A. Tran, "Facial expression recognition using residual masking network," pp. 4513–4519, 2021.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[9] S. Hangaragi, T. Singh, and N. Neelima, "Face detection and recognition using face mesh and deep neural network," *Procedia Computer Science*, vol. 218, pp. 741–749, 2023.

[10] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, "Facial expression recognition based on deep learning," *Computer Methods and Programs in Biomedicine*, vol. 215, p. 106621, 2022.

[11] H. B. U. Haq, W. Akram, M. N. Irshad, A. Kosar, and M. Abid, "Enhanced real-time facial expression recognition using deep learning," *Acadlore Transactions on AI and Machine Learning*, vol. 3, no. 1, pp. 24–35, 2024.

[12] J.-M. Lee, Y.-E. An, E. Bak, and S. Pan, "Improvement of negative emotion recognition in visible images enhanced by thermal imaging," *Sustainability*, vol. 14, no. 22, p. 15200, 2022.

[13] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial emotion recognition using transfer learning in the deep cnn," *Electronics*, vol. 10, no. 9, p. 1036, 2021.

[14] 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.

[15] E. S. Agung, A. P. Rifai, and T. Wijayanto, "Image-based facial emotion recognition using convolutional neural network on emognition dataset," *Scientific Reports*, vol. 14, no. 1, p. 14756, 2024.

[16] S. Vignesh, M. Savithadevi, M. Sridevi, *et al.*, "A novel facial emotion recognition model using segmentation vgg-19 architecture," *International Journal of Information Technology*, vol. 15, pp. 1777–1787, 2023.

[17] A. K. Roy, H. K. Kathania, A. Sharma, A. Dey, and M. S. A. Ansari, "Resemotenet: Bridging accuracy and loss reduction in facial emotion recognition," *arXiv preprint arXiv:2409.10545*, 2024.

[18] M. A. Saadi, M. Khachab, T. Elhaddad, and M. Al-Ayyoub, "Shufflevit-dfer: A lightweight dual-branch cnn-transformer for driver facial expression recognition," *arXiv preprint arXiv:2409.05996*, 2024.