ABSTRAK

Pengenalan ekspresi wajah majemuk (*compound expression*) menghadirkan tantangan signifikan karena kompleksitas dan ambiguitasnya yang melebihi ekspresi dasar, menuntut model yang mampu menangkap nuansa halus dan dependensi spasial global. Berbeda dengan arsitektur *Convolutional Neural Network* (CNN) yang cenderung fokus pada fitur lokal, *Vision Transformer* (ViT) dengan mekanisme *self-attention* globalnya menawarkan potensi untuk memodelkan hubungan kontekstual tersebut secara lebih efektif. Meskipun ViT unggul pada emosi dasar di *dataset* RAF-DB, penerapannya secara spesifik untuk klasifikasi 11 kelas ekspresi majemuk pada *dataset* yang sama masih menjadi celah penelitian. Oleh karena itu, penelitian ini bertujuan untuk mengevaluasi secara komprehensif bagaimana performa arsitektur ViT standar dalam tugas klasifikasi ekspresi wajah majemuk pada *dataset* RAF-DB, serta menganalisis kelebihan dan keterbatasannya dibandingkan dengan pendekatan sebelumnya.

Penelitian ini menggunakan pendekatan kuantitatif eksperimental untuk mengevaluasi model *Vision Transformer* (ViT). Data yang digunakan adalah *subset* ekspresi majemuk dari *dataset* RAF-DB, yang terdiri dari 3.954 gambar yang dibagi menjadi 80% data latih, 10% data validasi, dan 10% data uji. Tahap *preprocessing* meliputi *resize* gambar ke ukuran 224x224 piksel, normalisasi nilai piksel menggunakan *mean* dan standar deviasi dari *ImageNet*, serta augmentasi data pada set pelatihan yang meliputi *Horizontal Flip*, *Random Brightness Contrast*, dan *ShiftScaleRotate* untuk meningkatkan variasi data. Model ViT yang digunakan adalah *vit_base_patch16_224* dengan bobot *pre-trained* dari *ImageNet*, lapisan klasifikasi akhir disesuaikan untuk 11 kelas. Pengujian dilakukan melalui 12 skenario dengan kombinasi *hyperparameter* yang berbeda (*epoch*: 30/50, *learning rate*: 1e-4/1e-5, *batch size*: 16/32/64) untuk menemukan konfigurasi paling optimal.

Hasil penelitian menunjukkan performa puncak pada skenario pengujian ke-11 dengan akurasi sebesar 63.13%, menggunakan konfigurasi 50 *epoch*, *learning rate* 1e-5, dan *batch size* 32. Performa ini secara signifikan melampaui metode sebelumnya seperti DLP-CNN (44.55%) pada tugas yang sama, membuktikan keunggulan mekanisme *self-attention* ViT dalam menangkap hubungan kontekstual global pada wajah. Analisis kualitatif melalui *attention map* mengonfirmasi bahwa model berhasil mempelajari kombinasi fitur yang relevan, seperti alis dan mulut, untuk membedakan ekspresi. Namun, penelitian ini juga mengungkap keterbatasan fundamental, yaitu model kesulitan mengenali kelas-kelas ambigu yang sampelnya sedikit, serta mengembangkan bias dengan belajar dari artefak visual non-wajah seperti area gelap pada pakaian atau latar belakang untuk membuat keputusan klasifikasi. Kontribusi penelitian ini adalah menyediakan evaluasi komprehensif dan *baseline* performa yang solid untuk ViT standar pada klasifikasi ekspresi majemuk, sekaligus menyoroti tantangan bias dan ketidakseimbangan data sebagai area krusial untuk pengembangan di masa depan.

Kata Kunci: Klasifikasi Ekspresi Wajah, Ekspresi Wajah Majemuk, *Vision Transformer*, *Real-World Affective Faces* (RAF-DB), *Deep Learning*

ABSTRACT

The recognition of compound expressions presents a significant challenge due to their complexity and ambiguity, which surpasses that of basic expressions, demanding a model capable of capturing subtle nuances and global spatial dependencies. In contrast to the Convolutional Neural Network (CNN) architecture, which tends to focus on local features, the Vision Transformer (ViT) with its global self-attention mechanism offers the potential to model these contextual relationships more effectively. Although ViT has shown superiority in recognizing basic emotions on the RAF-DB dataset, its specific application for classifying the 11 compound expression classes on the same dataset remains a research gap. Therefore, this study aims to comprehensively evaluate the performance of the standard ViT architecture in the task of compound facial expression classification on the RAF-DB dataset, as well as to analyze its advantages and limitations compared to previous approaches.

This research uses a quantitative experimental approach to evaluate the Vision Transformer (ViT) model. The data used is the compound expression subset from the RAF-DB dataset, consisting of 3,954 images, which were split into 80% training data, 10% validation data, and 10% test data. The preprocessing stage included resizing images to 224x224 pixels, normalizing pixel values using the ImageNet mean and standard deviation, and applying data augmentation to the training set, which included Horizontal Flip, Random Brightness Contrast, and ShiftScaleRotate to increase data variance. The ViT model used was vit_base_patch16_224 with weights pre-trained on ImageNet, where the final classification layer was adapted for 11 classes. The evaluation was conducted through 12 scenarios with different hyperparameter combinations (epoch: 30/50, learning rate: 1e-4/1e-5, batch size: 16/32/64) to find the most optimal configuration.

The research results show that peak performance was achieved in the 11th testing scenario with an accuracy of 63.13%, using a configuration of 50 epochs, a learning rate of 1e-5, and a batch size of 32. This performance significantly surpasses previous methods like DLP-CNN (44.55%) on the same task, proving the superiority of ViT's self-attention mechanism in capturing global contextual relationships on the face. Qualitative analysis via attention maps confirmed that the model successfully learned relevant feature combinations, such as eyebrows and mouth, to distinguish expressions. However, this study also revealed fundamental limitations: the model struggled to recognize ambiguous classes with few samples and developed a bias by learning from non-facial visual artifacts, such as dark areas on clothing or backgrounds, to make classification decisions. The contribution of this research is the provision of a comprehensive evaluation and a solid performance baseline for the standard ViT on compound expression classification, while also highlighting the challenges of bias and data imbalance as crucial areas for future development.

Keywords: *Facial Expression Classification, Compound Facial Expression, Vision Transformer, Real-World Affective Faces (*RAF-DB*), Deep Learning*