BAB V
SIMPULAN DAN SARAN

5.1 Simpulan

Berdasarkan hasil analisis dan evaluasi yang telah dilakukan dalam penelitian ini, dapat ditarik beberapa simpulan yang menjawab rumusan masalah sebagai berikut:

1. Pengembangan model *Hybrid* MobileNetV3-Vision Transformer dengan *Token Downsampling* berhasil dilakukan dengan mengintegrasikan arsitektur CNN yang efisien (MobileNetV3) sebagai *backbone* untuk ekstraksi fitur lokal, dan *Vision Transformer* (ViT) untuk pemodelan konteks global. MobileNetV3 menggantikan mekanisme *patchify stem* pada ViT standar. Untuk mengatasi tingginya kompleksitas komputasi pada ViT, diimplementasikan teknik *Token Downsampling* berbasis skor (*score-based*) yang secara progresif mengurangi jumlah token pada setiap blok *transformer*. Model ini kemudian dioptimalkan menggunakan *Sharpness-Aware Minimization* (SAM) untuk meningkatkan kemampuan generalisasi.
2. Performa model *Hybrid* MobileNetV3-Vision Transformer dengan *Token Downsampling* menunjukkan hasil yang sangat kompetitif. Pada pengujian dengan set data FER-2013, model ini berhasil mencapai akurasi sebesar 71.24% dan F1-*score* sebesar 70.81%. Performa ini melampaui model *baseline* lain yang diuji, termasuk *pretrained* ViT standar. Penerapan *token downsampling* terbukti efektif mengurangi kompleksitas komputasi (FLOPs) dari 9.01 G menjadi 5.48 G dan mempercepat waktu pelatihan hingga 3 kali lipat, tanpa mengorbankan akurasi.
3. Performa sistem *end-to-end* pada pengujian demonstrasi data ekspresi wajah peserta didik di ruang kelas pada awalnya menunjukkan kinerja yang rendah (AP@0.5 sebesar 16.0) akibat adanya pergeseran domain (*domain shift*) antara data pelatihan (FER-2013) dan data pengujian (ruang kelas). Namun, setelah diterapkan strategi adaptasi berupa pelatihan pada set data gabungan menggunakan set data FER-2013 dan data ruang kelas, performa model meningkat secara signifikan hingga mencapai AP@0.5 sebesar 19.1.

Ditemukan juga bahwa model bekerja lebih optimal pada sudut pandang kamera dari depan dibandingkan dari samping. Meskipun demikian, nilai presisi yang masih rendah menunjukkan bahwa model masih sering melakukan kesalahan klasifikasi pada kondisi nyata.

5.2 Saran

Berdasarkan temuan dan keterbatasan yang ada dalam penelitian ini, berikut adalah beberapa saran yang dapat menjadi pertimbangkan dan acuan untuk penelitian selanjutnya:

1. Perlu dilakukan pembangunan set data ekspresi wajah peserta didik di lingkungan kelas yang lebih besar dan beragam, mencakup lebih banyak contoh ekspresi yang subtil serta variasi sudut pandang wajah (terutama dari samping) untuk meningkatkan ketahanan model.
2. Untuk mengatasi masalah latensi pada *pipeline* dua tahap (*two-stage*) saat ini, disarankan untuk menggunakan dan mengintegrasikan model deteksi wajah yang cepat namun tetap akurat atau melakukan penelitian dan pengembangan model deteksi ekspresi wajah *one-stage*. Model *one-stage* ini dapat melakukan deteksi wajah dan klasifikasi ekspresi secara simultan, yang berpotensi meningkatkan kecepatan inferensi secara signifikan dan lebih cocok untuk penerapan waktu nyata (*real-time*).
3. Model saat ini hanya memproses gambar statis. Untuk meningkatkan akurasi, terutama dalam mengenali ekspresi yang ambigu, penelitian selanjutnya dapat memanfaatkan informasi temporal dengan memproses sekuens video. Penggunaan arsitektur seperti *Video Vision Transformer* (ViViT) atau CNN-LSTM dapat menangkap dinamika perubahan ekspresi dari waktu ke waktu.
4. Mengingat tantangan dalam menangani ketidakseimbangan kelas pada set data FER-2013 yang tidak sepenuhnya teratasi dengan *class weight*, *weighted random sampler*, dan augmentasi data *offline*, penelitian selanjutnya dapat mengeksplorasi teknik *resampling* tingkat lanjut seperti penggunaan *Generative Adversarial Networks* (GANs) untuk menghasilkan citra wajah sintetis yang realistis untuk menyeimbangkan distribusi kelas.

5. Selain ekspresi dasar emosi, penelitian selanjutnya juga disarankan untuk menggunakan set data dengan label yang lebih relevan terhadap lingkungan kelas. Misalnya, label yang merepresentasikan kondisi belajar peserta didik seperti fokus, tidak fokus, mengantuk, atau kebingungan. Label semacam ini akan lebih relevan untuk mendukung tujuan praktis dari sistem pengenalan ekspresi wajah dalam pembelajaran, sehingga hasil pengembangan dapat lebih bermanfaat sebagai alat bantu bagi pendidik dalam memantau keterlibatan dan kondisi kognitif peserta didik secara lebih tepat.