

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Dalam satu dekade terakhir, bidang *computer vision* telah berkembang pesat dan memainkan peranan penting dalam berbagai aplikasi, seperti industri manufaktur, transportasi otonom, pengenalan wajah, deteksi objek medis, hingga pengolahan citra lingkungan [1], [2]. Salah satu tugas mendasar dalam *computer vision* adalah klasifikasi citra, yaitu melatih model untuk mengenali dan membedakan objek tertentu berdasarkan pola visual yang direpresentasikan dalam piksel. Kemajuan dalam klasifikasi citra tidak hanya didorong oleh ketersediaan dataset berskala besar, tetapi juga peningkatan kapasitas komputasi, serta inovasi dalam arsitektur *deep neural networks* yang memungkinkan model mengekstraksi fitur kompleks secara efisien [3].

Salah satu penerapan klasifikasi citra yang semakin mendapat perhatian adalah pengelolaan sampah [4]. Dengan volume sampah yang terus meningkat setiap tahunnya, proses pemilahan sampah menjadi tantangan global yang penting. Klasifikasi jenis sampah, seperti kertas, plastik, logam, kaca, organik, hingga baterai, menjadi krusial dalam mendukung efisiensi daur ulang [4]. Proses pemilahan sampah secara manual memiliki beberapa kelemahan, antara lain membutuhkan tenaga yang besar, memakan waktu, serta rentan terhadap kesalahan manusia. Oleh karena itu, pengembangan sistem otomatis yang mampu mengklasifikasikan sampah secara tepat dan konsisten akan sangat mendukung efisiensi pengelolaan sampah dan pengurangan dampak negatif bagi lingkungan.

Sebelum kemunculan pendekatan *deep learning* yang masif, metode konvensional dalam pengenalan jenis sampah umumnya menggunakan teknik *machine learning* tradisional dengan fitur buatan (*hand-crafted features*), seperti menggunakan ekstraksi ciri tekstur, warna, atau bentuk, yang kemudian diproses oleh algoritma, seperti *Support Vector Machine* (SVM) atau *K-Nearest Neighbors* (KNN) [4], [5]. Meskipun pendekatan ini berhasil dalam kasus tertentu, keterbatasan kemampuan model untuk secara adaptif mempelajari representasi fitur tingkat tinggi menghambat kinerjanya dalam lingkungan yang kompleks dan beraneka ragam [5].

Seiring berkembangnya *deep learning*, *Convolutional Neural Network* (CNN) mulai mendominasi bidang klasifikasi citra karena kemampuan ekstraksi ciri berbasis

konvolusi yang efektif. Arsitektur turunan dari CNN, seperti VGG19, ResNet50, dan InceptionNet telah menunjukkan kinerja yang sangat baik dalam berbagai kompetisi dan aplikasi industri [6]. Pada klasifikasi jenis sampah, CNN telah diaplikasikan untuk mempelajari pola visual yang membedakan jenis sampah satu dengan lainnya [7]. Namun, meskipun CNN sangat efektif dalam mengekstraksi fitur spasial, pendekatan ini memiliki keterbatasan dalam menangkap konteks global serta hubungan jangka panjang antar-elemen dalam citra [7]. Keterbatasan ini dapat menyebabkan kesalahan klasifikasi pada sampah dengan latar belakang kompleks atau struktur yang serupa. Hal ini menunjukkan perlunya pendekatan yang mampu memodelkan hubungan spasial secara lebih luas.

Dalam upaya untuk memahami konteks global yang lebih luas dalam citra, pendekatan Transformer mulai diperkenalkan, salah satunya adalah Vision Transformer (ViT) [8]. ViT mengadopsi konsep arsitektur Transformer yang sebelumnya terbukti efektif dalam pemrosesan bahasa alami [9] dengan menggantikan konvolusi berlapis dengan mekanisme perhatian (*attention mechanism*) sehingga dapat memodelkan hubungan global antar-piksel. Hasilnya, ViT mampu mencapai kinerja setara atau bahkan melebihi CNN pada tugas klasifikasi citra standar, seperti ImageNet [8]. Selain klasifikasi, ViT juga telah berhasil diterapkan dalam berbagai tugas visi komputer lain, seperti segmentasi objek serta deteksi objek.

Meskipun demikian, ViT murni memiliki keterbatasan dalam memodelkan fitur lokal dan stabilitas pelatihan ketika data terbatas [10]. Dalam menghadapi tantangan klasifikasi jenis sampah yang memiliki variasi visual tinggi dan jumlah data yang terbatas, penerapan pendekatan hibrida dapat menjadi potensi solusi yang efektif untuk meningkatkan kinerja ViT.

Salah satu pendekatan hibrida yang menjanjikan adalah penggunaan arsitektur ResNet50 yang dikenal mampu memitigasi masalah *vanishing gradient* melalui *skip connections* [11]. ResNet50 dapat bertindak sebagai ekstraktor fitur awal, menghasilkan fitur-fitur yang kaya yang kemudian diteruskan ke blok Transformer. Pada tahap ini, mekanisme perhatian global dalam Transformer membantu model untuk memahami konteks yang lebih luas dari citra sehingga memperkuat kemampuan ViT dalam menangkap hubungan visual yang kompleks.

Berdasarkan analisis terhadap kelebihan dan kekurangan masing-masing arsitektur tersebut, penelitian ini mengusulkan model baru yang diberi nama "HR-ViT: Hybrid

ResNet50-Vision Transformer" untuk klasifikasi jenis sampah. Model ini mengintegrasikan kekuatan ResNet50 dalam mengekstraksi fitur spasial dan struktur lokal dengan kemampuan Vision Transformer dalam memodelkan hubungan global. Melalui pendekatan ini, diharapkan HR-ViT dapat mengatasi keterbatasan pendekatan terdahulu serta meningkatkan akurasi dan ketahanan model dalam mengklasifikasikan enam jenis sampah, yakni baterai, plastik, kertas, logam, organik, dan kaca.

1.2 Rumusan Masalah

Terlepas dari peningkatan kinerja yang telah dicapai oleh metode klasifikasi sampah berbasis *machine learning* tradisional maupun CNN, kedua pendekatan ini masih menghadapi tantangan dalam memahami konteks global dan mengatasi variasi visual yang tinggi. Di sisi lain, Vision Transformer (ViT) dapat memodelkan hubungan global antarpiksel, tetapi performanya terkendala dalam mempelajari fitur lokal serta stabilitas pelatihan ketika data terbatas. Dengan mempertimbangkan kekurangan masing-masing pendekatan, rumusan masalah dalam penelitian ini adalah bagaimana mengintegrasikan arsitektur ResNet50 sebagai *backbone* dengan ViT menjadi suatu model hibrida (HR-ViT) yang mampu meningkatkan akurasi, *robustnes*, dan efisiensi klasifikasi enam jenis sampah (baterai, plastik, kertas, logam, organik, dan kaca) secara efektif.

1.3 Tujuan dan Manfaat Penelitian

1.3.1 Tujuan

Penelitian ini bertujuan untuk mengembangkan model hibrida yang mengintegrasikan arsitektur ResNet50 sebagai *backbone* dengan Vision Transformer guna meningkatkan akurasi klasifikasi enam jenis sampah.

1.3.2 Manfaat

Implementasi HR-ViT diharapkan memberikan beberapa manfaat berikut:

1. Memperkaya pengetahuan akademik tentang pengelolaan sampah menggunakan teknologi berbasis *deep learning*.
2. Memberikan pendekatan yang dapat dijadikan dasar untuk eksplorasi lanjutan dalam pengembangan model hibrida lainnya, serta mendukung inovasi dalam tugas klasifikasi dengan kondisi dataset terbatas dan variasi visual tinggi.

3. Meningkatkan pemahaman peneliti tentang pengembangan arsitektur jaringan hibrida, teknik pelatihan, dan evaluasi model *deep learning* yang kompleks.

1.4 Batasan Penelitian

Penelitian ini memiliki beberapa batasan yang diterapkan untuk mengarahkan ruang lingkup dan memastikan konsistensi dengan tujuan utama. Batasan-batasan tersebut adalah sebagai berikut:

1. Penelitian ini menggunakan algoritma klasifikasi berbasis pendekatan *hybrid*, yang mengintegrasikan arsitektur ResNet50 sebagai *backbone* dengan Vision Transformer untuk mengembangkan dan mengevaluasi model.
2. Dataset yang digunakan merupakan hasil penggabungan dari dua sumber data publik di Kaggle, yaitu dataset 12 kelas [12] dan dataset 6 kelas [13].
3. Total dataset terdiri dari 4.650 gambar yang terdistribusi secara merata ke dalam enam kelas, yaitu baterai, kaca, logam, organik, kertas, dan plastik, dengan masing-masing kelas memiliki 775 gambar.
4. Proses pelatihan dan pengujian model dilakukan menggunakan bahasa pemrograman Python dengan pustaka pendukung, seperti Pandas, TensorFlow, Matplotlib dan NumPy. Untung meningkatkan efisiensi selama pelatihan, eksperimen dijalankan dengan memanfaatkan GPU.

1.5 Keterbaruan

Dalam beberapa tahun terakhir, penelitian terkait klasifikasi sampah menggunakan pendekatan berbasis CNN dan Transformer semakin berkembang. Wang *et al.* [14] memperkenalkan Swin Transformer, yang mencapai akurasi hingga 99,75% dalam klasifikasi sampah plastik melalui kombinasi citra optik dan spektral inframerah. Sementara itu, Alrayes *et al.* [15] mengusulkan Vision Transformer berbasis Multilayer Hybrid CNN (VT-MLH-CNN) dengan akurasi 95,8% pada dataset TrashNet. Namun, kedua pendekatan ini belum mengeksplorasi tantangan dataset yang memiliki variasi visual tinggi dan kelas sampah yang lebih luas.

Model *hybrid* yang mengombinasikan CNN untuk ekstraksi fitur lokal dan Vision Transformer (ViT) untuk fitur global semakin banyak digunakan guna mengatasi keterbatasan kedua metode tersebut. Liu *et al.* [16] mengembangkan Garbage Classification Net (GCNet), mengintegrasikan EfficientNetV2, ViT, dan DenseNet

dengan akurasi mencapai 97,54%. Meskipun unggul dibanding model individualnya (DenseNet 96,40%, ViT 96,75%, dan EfficientNetV2 96,12%), GCNet memiliki beban komputasi tinggi sehingga belum optimal dalam aplikasi *real-time* atau perangkat dengan daya terbatas. Selain itu, GCNet hanya terbatas pada empat kategori sampah utama.

Pendekatan lain, CT-Net dari Cai *et al.* [17], berhasil meraih akurasi 96,55% pada dataset Huawei Cloud tetapi tidak secara jelas mengungkapkan beban komputasi yang dihadapi saat diterapkan dalam skenario industri yang kompleks. Wang *et al.* [18] melalui Garbage FusionNet (GFN) yang mengombinasikan ResNet, ViT, dan modul tambahan, seperti Pyramid Pooling Module (PPM) serta Convolutional Block Attention Module (CBAM), mampu mencapai akurasi hingga 96,54%, tetapi peningkatan beban komputasi akibat modul tambahan menjadi kelemahan utama pendekatan ini.

Berbeda dengan penelitian sebelumnya, penelitian ini mengusulkan model HR-ViT yang secara khusus mengintegrasikan keunggulan ResNet50 sebagai *backbone* CNN dan Vision Transformer tanpa tambahan modul kompleks yang meningkatkan beban komputasi. Pemilihan ResNet50 dilakukan karena efisiensinya dalam ekstraksi fitur lokal dibandingkan model CNN lain, seperti DenseNet atau EfficientNet yang relatif lebih kompleks. Integrasi Vision Transformer memperkuat kemampuan model dalam mengenali fitur global dari dataset dengan variasi visual tinggi, tantangan yang belum sepenuhnya teratasi oleh penelitian terdahulu. Lebih lanjut, penelitian ini mencakup klasifikasi yang lebih komprehensif dengan enam kelas sampah (baterai, kaca, logam, organik, kertas, dan plastik), sekaligus mengatasi permasalahan umum terkait keterbatasan data dan risiko *overfitting*.