# ABSTRAK

Perkembangan pesat teknologi *Generative AI* telah memunculkan tantangan serius terhadap integritas media digital. Model generatif seperti Stable Diffusion, BigGAN, dan VQ-Diffusion kini mampu menghasilkan citra sintetis dengan tingkat fotorealisme tinggi sehingga sulit dibedakan dari citra asli, bahkan oleh manusia. Penelitian ini mengusulkan pendekatan deteksi citra kecerdasan buatan (AI) berbasis Vision Transformer ViT-Base/16 dengan strategi *fine-tuning* berupa *partial freezing* dan *Focal Loss* pada domain spesifik citra burung. Penelitian dirancang dalam bentuk studi ablasi dengan enam konfigurasi: *full fine-tuning* dengan *Cross-Entropy Loss* sebagai *baseline* , tiga variasi *partial freezing* dengan pembekuan 4, 6, dan 8 blok transformer, *full fine-tuning* dengan *Focal Loss* , serta kombinasi partial freezing terbaik dan Focal Loss. Dataset yang digunakan adalah Unbiased Tiny GenImage (23.328 citra, 8 sumber) untuk konteks data umum, dan dataset citra burung (11.314 citra; ImageNet, Stable Diffusion, BigGAN, VQ-Diffusion) untuk konteks data spesifik dengan pembagian 70:15:15.

Evaluasi dilakukan dalam empat skenario: (1) model umum diuji pada data GenImage umum, (2) model umum diuji pada seluruh data burung, (3) model umum diuji pada split uji *stratified* data burung, dan (4) model baru yang dilatih khusus pada data burung diuji pada split uji yang sama. Hasil Skenario 1 dan 1.1 menunjukkan keterbatasan generalisasi model umum pada domain spesifik dengan akurasi sekitar 53%, mengonfirmasi bahwa representasi fitur umum tidak cukup untuk deteksi lintas domain. Skenario 2 membuktikan bahwa fine-tuning berbasis data spesifik secara signifikan meningkatkan performa; konfigurasi S3 (Focal Loss Only) mencapai akurasi tertinggi sebesar 98,65% dan F1-Macro 98,64% dengan jumlah kesalahan hanya 23 dari 1.698 data uji.

Analisis per-generator menunjukkan bahwa VQ-Diffusion merupakan generator paling sulit dideteksi ( *recall* 0,9231), sementara Stable Diffusion dan BigGAN dapat dideteksi denganbaik. Hasil juga menunjukkan bahwa *partial freezing* berkorelasi negatif dengan performa pada domain spesifik, bertentangan dengan hipotesis awal karena ukuran data latih yang memadai memungkinkan *full fine-tuning* berjalan optimal. Focal Loss terbukti efektif meningkatkan *recall* kelas AI tanpa mengorbankan precision, terutama pada sampel yang sulit dideteksi.

.



**Kata Kunci:** citra kecerdasan buatan generatif, Vision Transformer, partial freezing, Focal Loss, deteksi citra AI, domain spesifik, fine-tuning

## ABSTRACT

*The rapid advancement of Generative AI technology has posed a serious challenge to the integrity of digital media. Generative models such as Stable Diffusion, BigGAN, and VQ-Diffusion are now capable of producing synthetic images with a high level of photorealism, making them difficult to distinguish from real images, even by humans. This study proposes an AI-generated image detection approach based on a Vision Transformer (ViT-Base/16), using fine-tuning strategies in the form of partial freezing and Focal Loss in the specific domain of bird images. The study is designed as an ablation study with six configurations, full fine-tuning with Cross-Entropy Loss as the baseline, three partial-freezing variants with 4, 6, and 8 transformer blocks frozen, full fine-tuning with Focal Loss, and a combination of the best-performing partial-freezing setting with Focal Loss. The datasets used are Unbiased Tiny GenImage (23,328 images, 8 sources) for the general-domain context, and a bird-image dataset (11,314 images; ImageNet, Stable Diffusion, BigGAN, VQ-Diffusion) for the domain-specific context, split 70:15:15.*

*Evaluation is conducted in four scenarios, (1) the general model is tested on the general GenImage data, (2) the general model is tested on the entire bird dataset, (3) the general model is tested on the stratified test split of the bird dataset, and (4) a new model trained specifically on the bird dataset is tested on the same test split. The results of Scenarios 1 and 1.1 show the limited generalization of the general model to the domain-specific setting, with an accuracy of around 53%, confirming that general feature representations are insufficient for cross-domain detection. Scenario 2 demonstrates that domain-specific fine-tuning significantly improves performance; configuration S3 (Focal Loss Only) achieves the highest accuracy of 98.65% and an F1-Macro of 98.64%, with only 23 errors out of 1,698 test samples.*

*Per-generator analysis shows that VQ-Diffusion is the most difficult generator to detect (recall 0.9231), while Stable Diffusion and BigGAN can be detected well. The results also indicate that partial freezing is negatively correlated with performance in the domain-specific setting, contrary to the initial hypothesis, because the sufficient size of the training data allows full fine-tuning to perform optimally. Focal Loss is proven effective in increasing recall for the AI class without sacrificing precision, especially on hard-to-detect samples.*



***Keywords:*** *generative AI images, Vision Transformer, partial freezing, Focal Loss, AI image detection, specific domain, fine-tuning*