ABSTRAK

Proses produksi animasi konvensional masih sangat bergantung pada pengerjaan manual, di mana sebagian besar visual terkait dibuat berdasarkan referensi pemandangan dunia nyata. Namun, transformasi adegan dunia nyata ke dalam gaya visual animasi merupakan proses yang kompleks dan memerlukan keahlian teknis tinggi. Oleh karena itu, diperlukan teknik yang secara otomatis mampu mentransformasi adegan dunia nyata menjadi gambar bergaya animasi. Salah satu solusi terhadap permasalahan ini adalah implementasi *Latent Diffusion Model* (LDM) melalui pendekatan *image-to-image translation*. Namun terdapat keterbatasan dalam model difusi, yaitu distorsi bentuk dan inkonsistensi struktur objek pada hasil generasi. Untuk mengatasi keterbatasan tersebut, diperkenalkan ControlNet yang memungkinkan model difusi untuk menambahkan pengondisian spasial untuk mempertahankan struktur utama dari citra referensi.

Dalam penelitian ini, *Latent Diffusion Model* (LDM), melalui model terlatih Stable Diffusion, dikombinasikan dengan ControlNet berbasis peta tepi Canny, digunakan untuk meningkatkan ketahanan struktur konten hasil *image-to-image translation* bergaya animasi. Metodologi penelitian dimulai dengan pengumpulan data latih, yang terdiri dari 270 gambar tangkapan layar film buatan Studio Ghibli diambil dari situs web Fanscap. Data kemudian diproses melalui tahap eliminasi duplikat dan kurasi secara manual. Selanjutnya, dilakukan pelatihan *Low-Rank Adaptation* (LoRA) agar model dapat mengerjakan tugas sesuai dengan objek penelitian yang diangkat. Pada tahap inferensi, model dasar Stable Diffusion digabungkan dengan ControlNet terlatih berbasis citra tepi Canny dan LoRA yang sudah dilatih sebelumnya.

Evaluasi model dilakukan terhadap 20 citra fotografis yang diambil dari situs web Unsplash sebagai citra referensi. Pengujian dilakukan menggunakan tiga metrik pengukuran, yaitu *Structural Similarity Index Measure* (SSIM) dan *Learned Perceptual Image Patch Similarity* (LPIPS) sebagai metrik pengukuran ketahanan struktural, dan penilaian oleh pakar sebagai metrik pengukuran kualitas tekstur. Hasil Inferensi tanpa ControlNet menghasilkan rata-rata nilai SSIM 0,37, LPIPS 0,588, dan skor penilaian pakar 8,5. Sementara itu, inferensi menggunakan ControlNet menghasilkan rata-rata nilai SSIM 0,461, LPIPS 0,475, dan skor penilaian pakar 6,7. Berdasarkan hasil pengujian tersebut, ditemukan bahwa terdapat peningkatan kualitas yang konsisten pada nilai SSIM dan LPIPS, yang mengindikasikan ketahanan struktur konten yang lebih baik pada citra output. Namun, dibalik peningkatan ketahanan struktural tersebut, terdapat trade-off yaitu turunnya kualitas tekstur dari citra output, yang tercermin dari penurunan rata-rata skor penilaian pakar.

Kata Kunci: image-to-image translation, stable diffusion, image generation, canny edge detection

ABSTRACT

The conventional animation production process remains heavily reliant upon manual execution, wherein the majority of related visuals are created based on references to real-world scenery. Nevertheless, the transformation of real-world scenes into animation visual styles constitutes a complex process requiring highly specialized technical expertise. Consequently, there exists a necessity for techniques capable of automatically transforming real-world scenes into animation-styled imagery. One solution to address this problem is the implementation of Latent Diffusion Model (LDM) through an image-to-image translation approach. However, there are limitations inherent in diffusion models, namely shape distortion and structural inconsistency of objects in the generated results. To overcome these limitations, ControlNet has been introduced, which enables diffusion models to add spatial conditioning to preserve the principal structure from the reference image.

In this research, the Latent Diffusion Model (LDM), using the pre-trained Stable Diffusion model, combined with ControlNet based on Canny edge maps, is employed to enhance the structural robustness of content in animation-styled image-to-image translation results. The research methodology commenced with the collection of training data, consisting of 270 screenshot images from films produced by Studio Ghibli, obtained from the Fanscap website. The data was subsequently processed through duplicate elimination and manual curation stages. Following this, Low-Rank Adaptation (LoRA) training was conducted to enable the model to perform tasks in accordance with the research objectives being addressed. During the inference stage, the base Stable Diffusion model was integrated with the pre-trained ControlNet based on Canny edge imagery and the previously trained LoRA.

Model evaluation was performed on 20 photographic images obtained from the Unsplash website serving as reference images. Testing was conducted utilizing three measurement metrics, namely Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) as structural robustness measurement metrics, and expert assessment as a texture quality measurement metric. Inference results without ControlNet yielded average values of SSIM 0.37, LPIPS 0.588, and an expert assessment score of 8.5. Meanwhile, inference utilizing ControlNet produced average values of SSIM 0.461, LPIPS 0.475, and an expert assessment score of 6.7. Based upon these testing results, it was discovered that there exists a consistent quality improvement in SSIM and LPIPS values, indicating superior content structural robustness in the output images. However, behind this structural robustness enhancement, there exists a trade-off, namely the degradation of texture quality in the output images, reflected in the decrease of average expert assessment scores.

Keywords: image-to-image translation, stable diffusion, image generation, canny edge detection