UNIVERSITAS
GADJAH MADA
Pengembangan Model Diffusion Transformer untuk Menghasilkan Desain Sepatu Berbasis Teks dan Sketsa
Muhammad Alif Raher, Moh. Edi Wibowo, S.Kom., M.Kom., Ph.D.; Prof. Dr. Azhari, MT.;
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

PENGEMBANGAN MODEL *DIFFUSION TRANSFORMER* UNTUK MENGHASILKAN DESAIN SEPATU BERBASIS TEKS DAN SKETSA

Oleh

Muhammad Alif Raher
23/524726/PPA/06577

Pengembangan desain produk di industri alas kaki secara tradisional bergantung pada proses iteratif manual yang memerlukan keterampilan ilustrasi dan waktu yang signifikan. Kehadiran model generatif berbasis difusi membuka peluang untuk mempercepat proses ideasi visual, namun sebagian besar penelitian sebelumnya pada dataset *image-to-image translation* seperti *edges2shoes* terbatas pada pengkondisian satu modalitas karena tidak mempunyai anotasi teks. Penelitian ini mengusulkan DiT-DM (*Diffusion Transformer Dual-Modal*), sebuah model generatif yang mampu menerima dua modalitas kondisi secara simultan, yaitu sketsa sebagai panduan struktural dan teks deskriptif sebagai panduan semantik.

Untuk merealisasikan pendekatan tersebut, penelitian ini membangun pipeline *auto-captioning* berbasis BLIP guna menghasilkan anotasi teks otomatis pada dataset *edges2shoes* yang semula tidak memiliki *caption*. Representasi *edge* kemudian diintegrasikan dengan *noise* melalui strategi *early concatenation* pada ruang laten sebelum patchifikasi, sementara arsitektur *Diffusion Transformer* diadaptasi dengan mekanisme *cross-attention* agar mampu memproses pengkondisian semantik berbasis teks.

Melalui *ablation study* yang telah dilakukan pada sepuluh varian arsitektur, penelitian ini menemukan bahwa mekanisme *Early Concatenation*, *Cross-Attention*, dan *Late Diffusion* menurunkan *LPIPS* dari 0,37 (*baseline*) menjadi 0,16 secara konsisten pada seluruh sampel. Penelitian ini juga menghasilkan bahwa *Early Concatenation* sebagai modifikasi arsitektur yang direkomendasikan pada *Diffusion Transformer* dengan nilai *Fréchet Inception Distance* (FID) sebesar 64,27, *CLIP Score* sebesar 29,75 dan LPIPS sebesar 0,162. Hasil eksperimental ini mengindikasikan kemampuan model dalam menghasilkan gambar sepatu realistis sekaligus mempertahankan kesesuaian semantik dengan *caption*. Penelitian ini membuktikan kelayakan pendekatan *dual-modal conditioning* dan memberikan kontribusi pada pengembangan model *image generation* yang lebih akurat untuk mendukung proses ideasi desain sepatu.

Kata Kunci: *Diffusion Transformer*, *image generation*, *dual-modal conditioning*, *auto captioning*, desain sepatu.

UNIVERSITAS
GADJAH MADA
Pengembangan Model Diffusion Transformer untuk Menghasilkan Desain Sepatu Berbasis Teks dan Sketsa
Muhammad Alif Raher, Moh. Edi Wibowo, S.Kom., M.Kom., Ph.D.; Prof. Dr. Azhari, MT.;
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

DEVELOPMENT OF DIFFUSION TRANSFORMER MODEL FOR SHOE DESIGN GENERATION BASED ON TEXT AND SKETCH

By

Muhammad Alif Raher
23/524726/PPA/06577

Product design development in the footwear industry has traditionally relied on manual iterative processes that demand substantial illustration skills and time. The emergence of diffusion-based generative models offers an opportunity to accelerate visual ideation. However, most prior studies on image-to-image translation datasets such as edges2shoes are limited to single-modality conditioning due to the absence of textual annotations. This study proposes DiT-DM (Diffusion Transformer Dual-Modal), a generative model capable of receiving two conditioning modalities simultaneously: edge sketches as structural guidance and descriptive text as semantic guidance.

To realize this approach, the study develops a BLIP-based auto-captioning pipeline to generate automatic textual annotations for the edges2shoes dataset, which originally lacks captions. The edge representation is then integrated with noise through an early concatenation strategy in the latent space prior to patchification, while the Diffusion Transformer architecture is adapted with a cross-attention mechanism to enable text-based semantic conditioning.

Through an ablation study over ten architectural variants, this study finds that the Early Concatenation, Cross-Attention, and Late Fusion mechanism each raise the LPIPS from 0.37 (baseline) to 0.16 consistently across all samples. This study shows that Early Concatenation as recommended architecture modification of Diffusion Transformer with FID of 64.27, CLIP Score of 29.75 and LPIPS of 0,162. This study demonstrates the feasibility of dual-modal conditioning while characterizing the contribution of image generation modeling to support the footwear design ideation process.

Keywords: *Diffusion Transformer, image generation, dual-modal conditioning, BLIP captioning, shoe design.*