氏　　名	Li Sicheng
学位の種類	博士（情報科学）
学位記番号	博情第 545 号
学位授与年月日	令和 7 年 3 月 21 日
論文題目	CHARACTER DESIGN UNDER MULTIMODAL INPUTS BASED ON GENERATIVE MODELS
論文審査委員	宮田 一乗　北陸先端科学技術大学院大学　教授
	池田　心　北陸先端科学技術大学院大学　教授
	岡田 将吾　北陸先端科学技術大学院大学　教授
	吉高 淳夫　北陸先端科学技術大学院大学　准教授
	佐藤 周平　法政大学　准教授

論文の内容の要旨

In recent years, generative models, especially diffusion models, have made significant advancements, pushing the boundaries of traditional image synthesis and addressing various application-specific demands in fields such as virtual reality, film production, and fashion design. These technologies enable the automated generation of characters with precise specifications, dramatically improving workflow efficiency and reducing production costs. One of the most transformative advancements in generative models is multimodal generation, which leverages diverse inputs like text prompts, image references, and spatial guidance (e.g., sketches or segmentation maps). This approach opens up broader creative possibilities, enabling flexible design workflows and improving model robustness and adaptability across tasks. As technology evolves, multimodal inputs are poised to revolutionize character design, enhancing both efficiency and creativity in various industries. Building on multimodal generation, this dissertation investigates character creation and design through generative models incorporating diverse inputs like text, stroke data, and structural maps. The core challenge lies in balancing manipulability, convenience, and precision—three essential yet often competing elements in the design process:

· Manipulability refers to the level of control users have over the generative outputs, allowing flexibility in shaping design details.

· Convenience emphasizes the efficiency of the design workflow, stream lining the process to reduce repetitive tasks and focus on higher-level decisions.

· Precision indicates the model's ability to produce outputs that accurately align with user inputs, such as text descriptions or sketches.

Achieving an optimal balance among these factors is complex, as enhancing one often limits the others. This study addresses this trilemma by identifying the primary contradiction in each design scenario and resolving it according to user needs, guided by Marx's theory of contradictions. This approach enables tailored solutions across three design applications, each focusing on different primary tensions within the trilemma:

· Drawing Multi-Age Facial Features for Anime Characters: This task primarily balances precision and manipulability. Designing lively, age-specific facial features in anime characters demands both detailed control and real-time interactivity. I developed an interactive painting assistance system that leverages user-inputted strokes to create facial features with age-specific characteristics. This system ensures continuous interaction

between the user's design intent and the generative model, effectively balancing user creativity and model-driven generation.

· Character Pose Design: This task requires balancing precision and convenience. Quickly generating character images that align with specific descriptions is crucial in fields such as advertising and poster design. Building upon conditional diffusion models like ControlNet, I developed an enhanced end-to-end text-to-image (T2I) generation framework that enables efficient, accurate pose creation. Users can generate custom character poses by combining text inputs with spatial conditions, such as skeletons, facial landmarks, and sketches, to produce diverse and precise poses.

· Character Head Motion Design: Here, the primary focus is balancing convenience and manipulability for video-based motion design. Traditional frame-by-frame animation is often labor-intensive, especially for designing head movements in character animation. I introduced a head motion prediction model, integrated into an imageto-video (I2V) generative framework, to streamline the workflow. This model uses multimodal inputs—trajectory strokes, audio, and reference images—to predict head movements, allowing for intuitive user control over the motion trajectory while reducing the manual effort required for animation.

In conclusion, this dissertation uniquely contributes to advancing multimodal generative models in character design, optimizing user-customizable workflows across varied scenarios. By carefully balancing manipulability, convenience, and precision, this research enhances the applicability of generative models in creative domains, fostering both efficiency and creativity.

Keywords: Character Design, Generative Models, Multimodal Generation, Computer Vision, Human-Computer Interaction.

論文審査の結果の要旨

最近の深層学技術、特に生成系 AI 技術では、人がコンテンツ制作に直接的に関与する作業量を大幅に削減し、かつ、高品質な出力結果を得られるようになった。しかしながら、人の制作意図を十分かつ柔軟に反映させた生成結果を得るためには、操作性、利便性、正確性の３要素を満たす必要があり、未だ技術的に解決すべき課題が残されている。

本論文は、多様な入力データ（テキストやストロークデータなど）に基づき、深層学習技術を用いてキャラクタデザインを支援する研究成果をまとめたものであり、上述の 3 要素を向上させることを目的とした。

具体的には、(1)アニメキャラクタの顔画像の経時変化の描画支援、(2)テキストおよび、ポーズや顔の特徴量を表した構造データから適切な画像を生成する手法、(3)上半身画像と音声情報、および頭部の動作軌跡の入力に基づいて、人物がその音声を発話している動画を生成する手法、をそれぞれ提案し、客観的評価と主観的評価を行っている。

(1)に関しては、大域的および局所的な特徴量をスケッチ入力し、対象キャラクタの年齢的特徴をより正確に反映した描画支援を実現しており、操作性と正確性を向上させた。(2)に関しては、対象キャラクタの姿勢や顔の特徴量に関する形状的制約条件を入力とすることで、意図した画像を生成し

ており、利便性と正確性を向上させた。(3)に関しては発話の音声データと頭部の移動軌跡から、人物の発話動画を自動生成してしており、操作性と利便性を向上させた。

いずれも異なる応用事例（様々な年齢のアニメキャラの描画支援、キャラクタのポーズデザイン支援、キャラクタアニメーションの制作支援）の研究成果であるものの、ユーザの潜在的なデザイン能力を支援・拡張するためのインタラクティブシステムに貢献するものである。

以上、本論文は、マルチモーダルな入力データに基づいた深層学習技術によるキャラクタデザインの支援技術について論じたものであり、学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。