Title	手書き立ち絵からの3次元ゲームキャラクターモデル生成
Sub Title	Generating 3D models for game characters from hand-drawn illustrations
Author	エン, ネイキン(Yan, Ningxin) 矢向, 高弘(Yakō, Takahiro)
Publisher	慶應義塾大学大学院システムデザイン・ マネジメント研究科
Publication year	2024
Jtitle	
JaLC DOI	
Abstract	
Notes	修士学位論文. 2024年度システムエンジニアリング学 第372号
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40002001-00002024-0023

修士論文　　2024 年度

手書き立ち絵からの 3 次元ゲームキャラクター モデル生成

エン　ネイキン
(学籍番号：82333103)

指導教員　教授　矢向　高弘

2025 年 3 月

慶應義塾大学大学院システムデザイン・マネジメント研究科
システムデザイン・マネジメント専攻

論　　文　　要　　旨

学籍番号	82333103	氏　名	エン　ネイキン

論　文　題　目：手書き立ち絵からの 3 次元ゲームキャラクターモデル生成

本論文では、キャラクター立ち絵を基にした 3D モデル生成手法を提案し、ゲームキャラクター制作のために特別に設計されたものである。近年、単一視点を基にした 3D モデル生成の研究は著しい進展を遂げているが、単一視点画像が提供する情報には限界があるため、生成された 3D モデルは詳細不足や幾何情報の不完全さといった問題を抱え、高品質な 3D モデルの要求を満たすことが困難である。

これらの課題を解決し、3D モデル生成の品質を向上させるため、本研究では手描きの立ち絵を基にした 3D モデル生成手法を提案する。この手法は、ゲームキャラクター制作における高品質かつ高忠実度な要求を満たすことを目的としている。まず、提案手法では Stable Diffusion を用いて入力された手描きの立ち絵の 3D 情報を強化し、対象キャラクターの三次元的な特性をより良く反映できるようにする。次に、事前学習済みの拡散モデルを使用して、入力画像に対応する法線マップを生成する。最後に、幾何感知法線融合アルゴリズムを利用し、入力された複数視点画像と生成された法線マップの情報を統合することで、高品質な 3D モデルの生成を実現するものである。

実験結果により、提案手法は既存の単一視点ベースの 3D モデル生成手法と比較して、細部再現性や幾何的一貫性の面で優れていることが示された。具体的には、従来の手法と比べて、提案手法はキャラクター立ち絵の入力に含まれる細部特性、例えばキャラクターの体型、姿勢、衣装の質感などをより正確に再現できる。この結果から、提案手法はゲームキャラクター制作において設計図面を正確に再現する要求を満たす上で顕著な優位性を持つことが明らかになった。

さらに、提案手法はその革新性と実用性においても多くの特長を示している。例えば、Stable Diffusion 技術を導入することで、手描きの立ち絵に含まれる芸術的処理による詳細の歪みや幾何関係の不一致を効果的に軽減できる。また、法線マップ生成および幾何融合のプロセスでは、複数視点画像から得られる情報を十分に活用することで、生成された 3D モデルが複数視点においても高い一貫性と忠実度を保つことが可能である。特に、キャラクターの複雑な幾何形状や動的な姿勢の再構築において、提案手法はその独自の優位性を発揮している。

最後に、本研究では革新的な 3D モデル生成品質評価手法を提案し、生成モデルの効果を定量的に評価する基盤を提供した。本手法では、生成された 3D モデルの複数視点からのシルエット画像を取得し、それらを元の入力キャラクター立ち絵のシルエット画像と比較・分析することで、モデル生成の品質を評価した。この方法を通じて、提案手法が生成品質において優れていることを実証し、ゲームキャラクター制作における実用的な価値をさらに強調したものである。

以上より、本研究は革新的な 3D モデル生成手法を提案し、複数視点画像、法線マップ生成、幾何感知融合アルゴリズムを組み合わせることで、ゲームキャラクター制作分野において効率的かつ信頼性の高いソリューションを提供した。本手法は 3D モデル生成品質を大幅に向上させるだけでなく、手描き立ち絵入力の処理をさらに最適化し、方法の適用範囲を拡大するための将来的な研究への道を切り開くものである。

キーワード（5 語）

3D モデル生成，拡散モデル，3D 再構築，法線マップ，ゲームキャラクター制作

SUMMARY OF MASTER'S DISSERTATION

Student Identification Number	82333103	Name	Yan Ningxin

Title: Generating 3D Models for Game Characters from Hand-Drawn Illustrations

Abstract

This paper proposes a 3D model generation method based on character illustrations, specifically designed for game character production. In recent years, research on single-view-based 3D model generation has made significant progress. However, due to the limited information provided by single-view images, the generated 3D models often suffer from a lack of details and incomplete geometric information, making it difficult to meet the demands for high-quality 3D models.

To address these issues and improve the quality of 3D model generation, this study introduces a method based on hand-drawn character illustrations. This method aims to meet the high-quality and high-fidelity requirements of game character production. First, the proposed method enhances the 3D information of the input hand-drawn illustrations using Stable Diffusion, allowing the images to better reflect the three-dimensional characteristics of the target characters. Next, corresponding normal maps are generated using a pretrained diffusion model. Finally, a geometry-aware normal fusion algorithm is employed to integrate the information from the multi-view images and the generated normal maps, enabling the creation of high-quality 3D models.

Experimental results demonstrate that the proposed method outperforms existing single-view-based 3D model generation methods in terms of detail restoration and geometric consistency. Specifically, compared to conventional methods, the proposed method more accurately reproduces the detailed features in input character illustrations, such as body shape, posture, and clothing textures. This indicates that the proposed method has significant advantages in meeting the precise reproduction requirements of design illustrations in game character production.

Furthermore, the proposed method exhibits both innovation and practicality in several aspects. For instance, by incorporating Stable Diffusion technology, it effectively reduces distortions or inconsistencies in geometric relationships caused by artistic processing in hand-drawn illustrations. Additionally, the processes of normal map generation and geometric fusion make full use of the information provided by multi-view images, ensuring that the generated 3D models maintain high consistency and fidelity across multiple perspectives. Notably, the method demonstrates unique superiority in reconstructing complex geometric shapes and dynamic poses of characters.

Finally, this study proposes an innovative evaluation method for 3D model generation quality, providing a quantitative assessment of the effectiveness of generated models. The method involves capturing silhouette images of the generated 3D models from multiple viewpoints and comparing them with the silhouette images of the original input character illustrations to evaluate model quality. Through this method, the superiority of the proposed method in terms of generation quality has been successfully verified, further emphasizing its practical value in game character production.

In summary, this study introduces an innovative 3D model generation method that combines multi-view images, normal map generation, and geometry-aware fusion algorithms, providing an efficient and reliable solution for the field of game character production. This method not only significantly enhances the quality of 3D model generation but also opens up new directions for future research, particularly in further optimizing the processing of hand-drawn illustrations and expanding the applicability of the method.

Keywords (5 words)

Generating 3D model, Diffusion model, Recover 3D geometry, Normal map, Game character production

目次

図目次

第1章 序論

近年，コンピュータグラフィックスの発展と成熟，さらにゲームにおけるよりリアルな映像や物理効果を追求するニーズの高まりにより，3D ゲームはゲーム業界において主流となっている.市場調査によると，世界の 3D ゲーム市場規模は過去 5 年間で年平均 2 桁の成長率を記録しており，今後も力強い成長が続くと予測されている.そして,3D キャラクター制作にかかるコストはゲーム開発全体の中でも非常に高い割合を占める.特に AAA タイトルでは，3D モデルの制作費用がゲーム開発全体の 60%以上を占める場合もあり，制作期間が長引くほど開発コストが増大する.このため，中小規模のゲーム企業にとって，高品質な 3D キャラクターを制作することは大きな負担となり，外部のアウトソーシングに依存するケースも少なくない.しかし，外部委託にはコスト削減の利点がある一方で，品質管理の難しさ，制作スタイルの統一の困難さ，および納期の不確実性といったリスクも伴う.また，3D キャラクター制作の専門人材の需要が高まり続けているが, モデリング技術を習得するには高度な専門知識と長期間の学習が必要であり，即戦力となる人材の確保が困難な状況が続いている.

ゲーム開発技術の革新を促進すると同時に, 開発者に更なる高い要求を突きつけるものである.その中でも，3D キャラクターの制作は開発プロセスにおいて極めて重要な役割を果たしている.

ゲーム開発における 3D キャラクターの制作は，通常，キャラクターデザイナーによるコンセプトデザインから始まるものである.デザイナーはゲームのストーリーやアートスタイルに基づき，キャラクターの外見，衣装，小物，ポーズを表現した立ち絵を作成する.この立ち絵が 3D モデラーにとって基本的な参考資料となっている.その後，3D モデラーは Maya，Blender，ZBrush といったプロのモデリングソフトを使用して立ち絵を元に 3D モデルを構築する.このプロセスでは，モデラーが高度な美的感覚と技術力を持つ必要があるだけでなく，キャラクターのアニメーション需要，物理特性，ゲームエンジンでの最適化にも十分配慮する必要がある.例えば，キャラクターモデルは効率的に動作するためにポリゴン数を厳密に管理する必要があり, さらに異なる照明や視点の下でも一貫した視覚効果を保証する必要がある.このようにプロセスが複雑で，細部に高い要求があるため，通常，膨大な時間と人的リソースが必要.このような高い時間コストと技術的ハードルは，3D キャラクター制作を開発プロセスでの大きな課題となっている.

一方，人工知能技術の進歩に伴い，画像から 3D モデルを生成する研究が注目を集めており, これはゲーム開発における自動化と効率化に新たな可能性をもたらすものである.例えば，NeRF(Neural Radiance Fields)0 や 3D Gaussian Splatting[2]といった技術はシーン再構築において優れたパフォーマンスを発揮するものであるが, 多数のマルチビュー画像に依存し 0，主に現実世界のシーンに特化して設計されているため，ゲームキャラクター制作には直接適用することが困難であるものである.また，text-to-3D や single-image-to-3D モデルなどの手法は，生成効率の向上において顕著な進歩を遂げているものであるが, 生成結果の幾何学的一貫性や細部の再現性において未だ課題が残っているものである.これらの制約は，特にキャラクターの立ち絵に含まれる細部を忠実に再現

する必要があるゲームキャラクター制作の要件を満たす上で, 既存技術の限界を浮き彫りにするものである.

これらの課題を解決するために，本研究ではキャラクター立ち絵に基づく 3D モデル生成手法を提案した.まず，入力されたキャラクター立ち絵を拡散モデルによって事前処理を行い，画像の 3D 情報を強化する.次に，法線マップ生成モデルを導入してキャラクターの幾何学的特性を捉えるものである.最後に，入力した RGB 画像と生成された法線マップを幾何学認識融合アルゴリズムで統合することにより，高品質な 3D モデルを再構築する.従来の単一視点手法と比較して，本手法は生成モデルの幾何学的一貫性と細部再現性を大幅に向上させ, ゲームキャラクターの高精度再現に新たな解決策を提供した.

生成結果の評価において, 本研究は生成されたモデルの複数の視点からスクリーンショットを取得し，それらを入力画像と比較分析を行う.その結果，生成された 3D モデルは幾何学的一貫性と細部再現性の面で既存の手法を上回る性能を示したものである.特に，キャラクターの体型，衣装の質感，姿勢の特徴などが高い忠実度で再現されていることが確認された.また，生成された 3D モデルは異なる視点においても一貫性を保持しており，本研究の提案手法の信頼性と実用性がさらに証明された.この結果は，提案手法がゲームキャラクター制作分野において実用的な価値と応用の可能性を持つことを示している.

本研究の実現により，開発コストが大幅に削減され，生産効率が向上する.従来の 3D モデリングには多くの時間と人員が必要とされていたが，その負担が大幅に軽減され，ゲーム開発の期間が短縮される.特に，大規模なゲームプロジェクトにおいて，3D アセットの生成と反復作業の速度が大幅に向上する.これにより，ゲーム企業は美術制作にかかるコストを削減し, より多くのリソースをゲームの革新や最適化に投入することが可能となる.

本論文の構成は以下の通りである.第 2 章では，3D モデル生成を中心に据え，本研究に関連する基礎知識や最新技術について解説し, 以降の章における理論的な基盤を提供する.第 3 章では，研究の課題を明確化するとともに，本研究の目的について述べる.第 4 章では，本研究が提案する 3D モデル生成手法について，その設計思想や技術的な実現方法を詳しく紹介する.第 5 章では，実験を通じて得られた生成結果および定量的な評価結果を示し, 提案手法が幾何学的一貫性や細部の再現性において優れていることを実証するものである.最後に，第 6 章では本研究の結論を総括するとともに，今後の研究方向や改良の可能性について述べる.

第2章 背景

本章では本研究に関連する背景知識を述べる.2.1 では，拡散モデルの基本原理を詳細に紹介し，その生成メカニズムと数学的枠組みについて説明し，後続の技術応用を理解するための基盤を築いた.2.2 では，Stable Diffusion の動作原理および画像生成分野における利点をさらに探求し，特に高品質な画像生成におけるその優れた性能を強調した 2.3 では，3D 再構築の基本理論を述べ，多視点画像から三次元の幾何情報を再構築するための主要な技術と課題について分析した.最後に 2.4 では，3D モデル生成の原理および関連研究の進展について紹介し，現在主流の手法が持つ特徴と，生成精度および効率の面での性能を議論した.2.5 では、本研究との関連研究を述べる.これらの内容を総括することで，本研究の技術的アプローチの選択と方法設計に理論的な支援を提供した.

2.1 拡散モデル

従来，VAE [3]と GAN77[4]は画像生成分野で最も一般的なモデルであったが，2020 年に提案された拡散モデル DDPM(Denoising Diffusion Probabilistic Model)は，生成モデルに新たな発展方向をもたらした[5]. 拡散モデルは，その生成の安定性と詳細な表現力により注目を集め，高解像度画像生成，画像修復，テキストから画像生成などのタスクにおいて卓越した性能を示している. 特に近年の応用において，拡散モデルは主流な方法となっており，Stable Diffusion[6]や DALL·E などがテキストから画像生成分野で画期的な成果を挙げている.

拡散モデルは，その最適化が進むにつれて，例えば DDIM による生成加速や LDM による潜在空間の操作が挙げられ，その地位をさらに強固なものにしている.これにより，画像生成分野の主導的技術へと進化し，高品質な生成タスクにおいて，GAN や VAE が担っていた役割を部分的に置き換えるに至った.拡散モデルの台頭は，生成モデルの新たな時代の到来を象徴するものである.

拡散モデルは確率生成モデルであり，逐次的な逆方向の復元プロセスを通じて高品質なデータ(例えば画像，音声，テキストなど)を生成する手法である.近年，GAN(生成対向ネットワーク)の次世代モデルとして注目され，特に画像生成分野で顕著な性能を示している.

拡散モデルの発想は主に熱力学における粒子の拡散現象から着想を得たものである. 拡散現象とは，粒子や物質が濃度の高い部分から低い部分へ自然に移動し，全体に均一に広がっていく過程を指す.この現象は液体や気体の中で頻繁に観察され，例えば水中におけるインクの広がりや，空気中に漂う香水の広がりがその代表的な例である.

拡散モデルは，前方プロセス(forward process)と逆プロセス(reverse process)の 2 つのプロセスで構成される.前方向プロセスは「拡散プロセス」とも呼ばれ，どちらのプロセスもパラメータ化されたマルコフ連鎖(Markov chain)[7][8]に基づいている[6].逆方向プロセスはデータサンプルを生成するために使用される，この役割は GAN における生成器

に類似しているが，GAN の生成器は次元の変化を伴うのに対し，DDPM の逆方向プロセスでは次元の変化がない.

2.1.1 前方プロセス

拡散モデルの前方プロセスは，実データ分布から始まり，複数の時間ステップを通じて段階的にノイズを追加し，最終的にデータを純粋なノイズ，つまり標準正規分布に近づけるプロセスである.

標準正規分布：

標準正規分布(Standard Gaussian Distribution)は，特別な正規分布の一種であり，平均が 0，標準偏差が 1 の分布を指す.自然界における多くのランダムな現象の分布特性を表現するために用いられる[9].標準正規分布は，確率論および統計学における重要な基礎ツールであり，データ分析，機械学習，工学などの分野で広く応用されている.その数学的定義は以下の通りである：

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (2.1)$$

標準正規分布は，統計学と確率論の基盤として，そのシンプルな形式と広範な応用により，ランダム現象の研究や数学モデルの構築において重要なツールとなっている.それは理論的に重要な意義を持つだけでなく，実際の応用においても重要な役割を果たしている.

マルコフ連鎖

マルコフ連鎖(Markov Chain)は，確率過程を記述するための数学モデルであり，統計学，物理学，機械学習などの分野で広く応用されている.マルコフ連鎖の核心的な特徴は無記憶性にあり，未来の状態は現在の状態のみに依存し，過去の状態には依存しない[7][8].

マルコフ連鎖とは，マルコフ性を持つ離散的な確率変数の集合であり，以下の条件を満たす:

$$P\left(X_{n+1} = x_{n+1} \middle| X_n = x_n，\ X_{n-1} = x_{n-1}，\ \ldots，\ X_0 = x_0\right) = P(X_{n+1} = x_{n+1} | X_n = x_n) \quad (2.2)$$

この式において，X_nは時刻 n における状態を表す.上式はマルコフ連鎖を定義すると同時に，マルコフ性(Markov Property)を定義している.この性質は「無記憶性(memorylessness)」とも呼ばれ，未来の状態は現在の状態のみに依存し，過去の状態には依存しないことを意味する.具体的には，t+1 ステップ目の確率変数は，時刻 t の確率変数が与えられた条件下で，それ以外の確率変数と条件付き独立(conditionally independent)である.

さらに，マルコフ連鎖は強マルコフ性(strong Markov property)も持つ.これは，任意の停止時刻(stopping time)において，停止前の状態と停止後の状態が相互に独立であることを意味する[9].

以上を踏まえると，マルコフ連鎖はそのシンプルな無記憶性という特徴により，複雑な確率過程を記述・解析するための重要な数学的ツールである.統計学，機械学習，自然科学などの分野で広く応用されており，動的システムや確率現象を研究する上での重要な方法論的基盤となっている.

ガウスノイズ画像の生成：

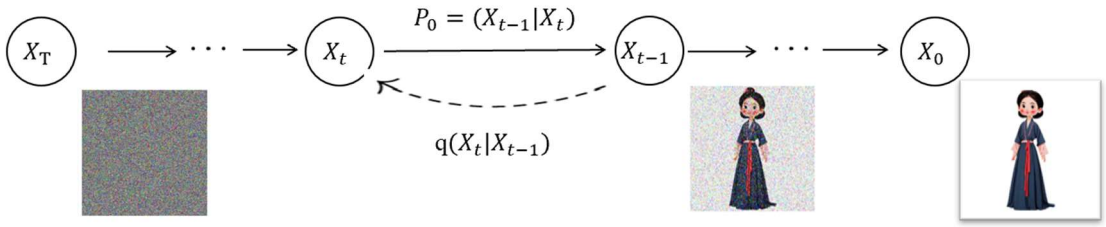

図 2.1 拡散モデルプロセス

図 2.1 に示すように，拡散モデルの前方プロセス(Forward Process)と後向プロセス(Reverse Process)が説明される.

例えば，RGB24 形式の一般的なカラー画像の場合，各ピクセルは RGB の 3 つのチャンネルで表され，それぞれの値は [0，255] の範囲を取り，赤，緑，青の 3 原色の割合を示す.各チャンネルの値を [-1，1] の範囲に正規化した後，標準正規分布に従うランダムなノイズ画像を同じサイズで生成する.このノイズ画像は全てのピクセルが標準正規分布に従う値を持つ[6].次に，生成された高斯ノイズ画像(ガウスノイズ画像)と同サイズの加ノイズ対象画像を混合し，以下の式を用いて混合後の各ピクセルチャンネル値を計算する.

$$\sqrt{\beta} \times \varepsilon + \sqrt{1-\beta} \times x \quad (2.3)$$

ここで，x は入力画像，εはノイズ，βは [0，1] の範囲にある定数であり，係数の生成に使用される.x_{t-1}をx_tで表現したい場合，以下の式が成り立つ：

$$x_t = \sqrt{\beta_t} \times \varepsilon_t + \sqrt{1-\beta_t} \times x_{t-1} (2.4)$$

ここで，ε_tは標準正規分布に従い，すなわち，$\varepsilon_t \sim N(0, 1)$である.また，異なる時間ステップ t における βの値は事前に定義されており，時間 1/T に従って徐々に増加する.増加方法としては，Linear(線形)，Cosine(余弦)などがあり，さらに次の不等式が成り立つ：

$$0 < \beta_1 < \beta_2 < \beta_3 < \beta_{t-1} < \beta_t < 1 (2.5)$$

以降の計算を簡略化するために，ここで$\alpha_t = 1 - \beta_t$と定義する.この結果，x_tは以下の式によって表すことができる:

$$x_t = \sqrt{1-\alpha_t} \times \varepsilon_t + \sqrt{\alpha_t} \times x_{t-1} (2.6)$$

著者は，複数回の反復を避け，x_tを直接x_0で表現する方法を試みている.ここで，ε_tとε_{t-1}は標準正規分布に従う 2 つの変数である.標準正規分布の性質によれば，2 つの正規分布が重ね合わされた場合，その結果も正規分布に従う.このことは次の式で表される :

$$N(\mu_1，\ \sigma_1{}^2) + N\left(\mu_2，\ \sigma_2{}^2\right) = N(\mu_1 + \mu_2，\ \sigma_1{}^2 + \sigma_2{}^2) (2.7)$$

x_tとx_{t-1}次の式で表される：

$$x_t = \sqrt{1-\alpha_t} \times \varepsilon_t + \sqrt{\alpha_t} \times x_{t-1} (2.8)$$

$$x_{t-1} = \sqrt{1-\alpha_{t-1}} \times \varepsilon_{t-1} + \sqrt{\alpha_{t-1}} \times x_{t-2} (2.9)$$

正規分布の性質を代入することで, x_tはx_{t-2}を用いて表現することが可能となる.以下の式がその関係を示している：

$$x_t = \sqrt{1-\alpha_t\alpha_{t-1}} \times \varepsilon + \sqrt{\alpha_t\alpha_{t-1}} \times x_{t-2} (2.10)$$

このように，一層ずつ反復を繰り返すことで，最終的にx_tは直接x_0を用いて表現することが可能となる.具体的には以下の式で示される：

$$x_t = \sqrt{1-\bar{\alpha}_t} \times \varepsilon + \sqrt{\bar{\alpha}_t} \times x_0 (2.11)$$

その中で，$\bar{\alpha}_t = \alpha_t\alpha_{t-1}\alpha_{t-2}\alpha_{t-3} \cdots \alpha_2\alpha_1$となる，この推導過程は「再パラメータ化」(reparameterization)とも呼ばれる.

ここまでが拡散モデルの前方プロセス.

2.1.2 逆プロセス

拡散モデルの逆プロセスでは，まずx_Tからx_0を復元することが主な目的となる.すなわち，認識可能な情報を一切含まないガウスノイズ画像から，元の鮮明な画像を復元するプロセスである.この目標を達成するために，まず後の時刻x_tから前の時刻x_{t-1}の画像をどのように復元するかを検討する必要がある.

ベイズの定理

ベイズの定理は，確率論における基本的な定理であり，ある証拠が既知の場合における事象の条件付き確率を計算するために用いられる.この定理は，イギリスの数学者トーマス・ベイズにちなんで命名されており，現代の確率統計や機械学習における重要な理論的基盤である.データ分析，分類，予測などの分野で広く応用されている.

通常，事象 A が事象 B の発生条件下で発生する確率と，事象 B が事象 A の発生条件下で発生する確率は異なる.しかし，これらの確率には明確な関係があり，ベイズの定理はその関係を述べるものである.ベイズ公式の一つの用途として，3 つの既知の確率から 4 つ目の確率を導き出すことが挙げられる.また，ベイズの定理は確率変数の条件付き確率や周辺確率分布とも密接に関連している[10].

ベイズの定理の数学的表現は以下の通りである：

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (2.12)$$

この式において，A および B はランダム事象を表し，P(B)は 0 ではない必要がある.

ベイズの定理では，各用語には一般的に使用される名称が与えられている：

P(A|B)は事象 B が発生した条件下で事象 A が発生する確率を指し，事後確率(posterior probability)とも呼ばれる.

P(A)は事象 A の事前確率(prior probability)を表し，事象 B の情報を考慮しない確率である.

P(B|A)は事象 A が発生した条件下で事象 B が発生する条件付き確率であり，事象 B の事後確率とも呼ばれる.

P(B)は事象 B の事前確率である.

これらの用語をまとめると，ベイズの定理は次のように表現できる：事後確率 =(尤度 × 事前確率)/ 正規化定数

ノイズ除去プロセス

前方プロセスのアルゴリズムに基づくと，εは標準正規分布に従う一回のランダムサンプリングであることがわかる.このため，ノイズを除去する際には，x_tから x_{t-1}への遷移をランダムなプロセスと見なすことができる.すなわち，x_tの条件下でx_{t-1}が発生する確率として表現することが可能である.以下の式で示される：

$$P\left(x_{t-1}\middle|x_t,\ x_0\right) = \frac{P(x_t|x_{t-1},\ x_0)P(x_{t-1}|x_0)}{P(x_t|x_0)} \quad (2.13)$$

ここで，$P(x_t|x_{t-1},\ x_0)$は，拡散モデルの順方向伝播段階におけるx_{t-1}からx_tへのプロセスに対応する.式 2.13 に基づき，ε_tが標準正規分布に従う乱数であることから，

$P(x_t|x_{t-1},\ x_0)$は $N(\sqrt{\alpha_t}x_{t-1},\ 1-\alpha_t)$に従う正規分布である.

同様に，$P(x_{t-1}|x_0)$は$N(\sqrt{\bar{\alpha}_{t-1}}x_0,\ 1-\bar{\alpha}_{t-1})$に従う正規分布であり，さらに $P(x_t|x_0)$は$N(\sqrt{\bar{\alpha}_{t-1}}x_0,\ 1-\bar{\alpha}_{t-1})$に従う正規分布である.

正規分布の確率密度関数を代入すると，式(2.13)は以下の形式に変形できる：

$$P(x_t|x_{t-1},\ x_0)=\frac{1}{\sqrt{2\pi}(\frac{\sqrt{1-\alpha_t}\sqrt{1-\bar{\alpha}_{t-1}}}{\sqrt{1-\bar{\alpha}_t}})}e^{[-\frac{(x_{t-1}-(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t+\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0))^2}{2(\frac{\sqrt{1-\alpha_t}\sqrt{1-\bar{\alpha}_{t-1}}}{\sqrt{1-\bar{\alpha}_t}})^2}]} \quad (2.14)$$

$$P(x_{t-1}|x_t,\ x_0)\sim N(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t+\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0,\ (\frac{\sqrt{1-\alpha_t}\sqrt{1-\bar{\alpha}_{t-1}}}{\sqrt{1-\bar{\alpha}_t}})^2) \quad (2.15)$$

しかし，この時点で，x_0はまだノイズが加えられていない元の画像を表し，最終的な予測対象であるため，その値は未知である.したがって，x_0を置き換える必要があり，式2.11 に基づいて，x_0は以下の式で表される：

$$x_0=\frac{x_t-\sqrt{1-\bar{\alpha}_t}\times\varepsilon}{\sqrt{\bar{\alpha}_t}} \quad (2.16)$$

x_0を式 2.15 に代入すると，下記の通りになる：

$$P(x_{t-1}|x_t,\ x_0)\sim N(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t+\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\times\frac{x_t-\sqrt{1-\bar{\alpha}_t}\times\varepsilon}{\sqrt{\bar{\alpha}_t}},\ \left(\sqrt{\frac{\beta_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_{t-1}}}\right)^2) \quad (2.17)$$

ここまでで，x_{t-1}はx_tを用いて表現することができるようになり，この式において唯一未知の変数はノイズ はεである.したがって，ニューラルネットワークを利用してノイズを推測することで，元の画像x_0を段階的に復元することが可能となる.具体的には，モデルはノイズが加えられた入力x_tと時間ステップ t を条件として，現在のステップにおけるノイズ成分εを推測する.この方法により，ノイズを段階的に除去し，元のデータx_0に近づけることができる.

図 2.2 は逆プロセスであり，推測されたノイズ値を基に，除去ノイズの公式を適用しながら，ターゲット分布に近いサンプルを段階的に生成する.この逆拡散プロセスの各ステップは，前のステップの出力に依存し，条件付き確率モデルにより構築される.最終的に，複数回の反復を経て，ランダムなノイズから高品質なサンプルを生成することが可能となる.このプロセスの鍵は，ニューラルネットワークの推測精度と，複数ステ

ップの推論を通じて生成結果を最適化する能力にある.拡散モデルは，この特性により，生成タスクにおいて非常に高い柔軟性と生成品質を発揮する.

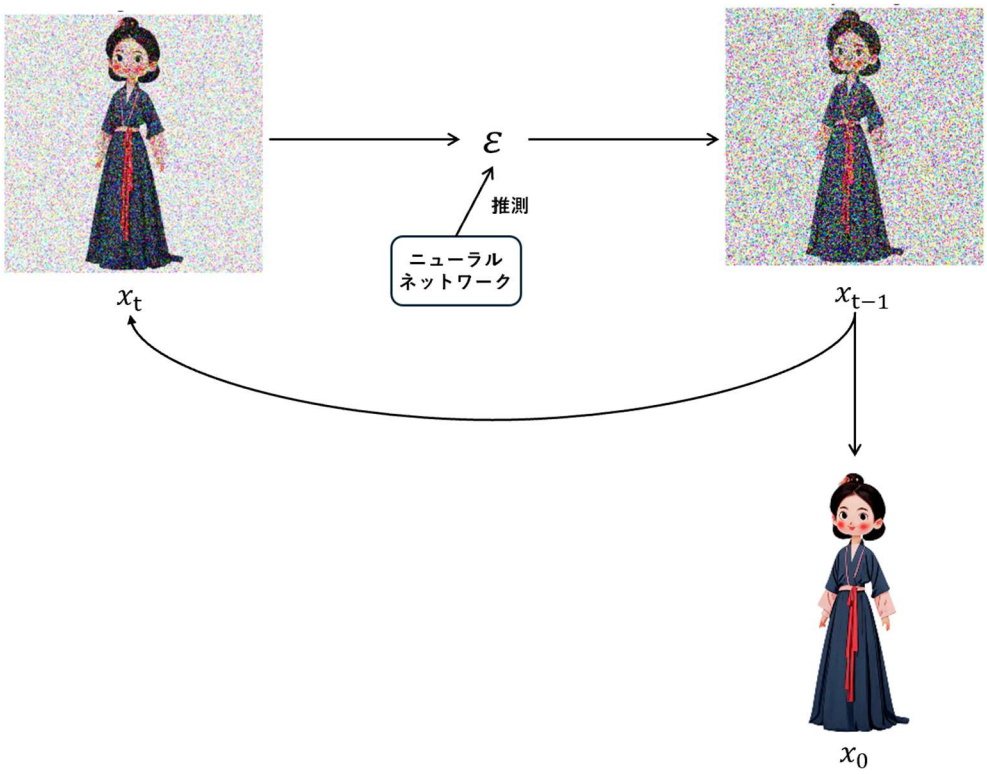

図 2.2 拡散モデルの逆プロセス

2.2 Stable Diffusion

2022 年，Stable Diffusion モデルが突如登場し，AI 業界が従来のディープラーニング時代から AIGC(AI Generated Content)時代へと進化する象徴的なモデルの一つとなった.このモデルは産業界，投資界，学術界，そしてコンペティション分野に新たな AI の想像力を注入した.

Stable Diffusion は，AI による画像生成分野における中心的なモデルであり，テキストから画像を生成するタスク(txt to image)や画像から画像を生成するタスク(img to image)を実現することができる.Stable Diffusion は完全にオープンソースプロジェクトであり，そのおかげで強力かつ繁栄した上下流エコシステム(AI イラストレーションコミュニティ，SD を基盤とした独自学習 AI モデル，多彩な補助 AI ツールやプラグインなど)を迅速に構築することが可能となった.また，多くの AI イラストレーション愛好者を惹きつけ，AI 業界の専門家と共に AIGC 分野の発展と普及を推進している.

拡散モデルでは，筆者が拡散モデルの動作原理を説明したが，それは完全に Stable Diffusion の動作原理を反映しているわけではない.その理由は，拡散プロセスが画像空間で行われるためであり，このプロセスは計算に非常に時間がかかり，単一の GPU 上

では実行がほぼ不可能である.ましてや，ノートパソコンの GPU ではなおさらである.画像空間は非常に巨大であり，Google の Imagen[11]や OpenAI の DALL-E[12]のような拡散モデルはピクセル空間で動作している(ピクセル空間の利点は，生成内容をより正確に制御できる点であり，例えば文字の表示などが可能となる).これらのモデルは動作を高速化するための工夫をいくつか取り入れているが, それでも十分に速いとは言えない.

このような背景を踏まえ，Stable Diffusion は速度問題を解決することを目的としている.それは, 潜在空間で拡散を行うモデル(latent diffusion model)である.このモデルは高次元の画像空間で操作を行うのではなく，まず画像を潜在空間(latent space)に圧縮する.ピクセル空間と比較して，潜在空間のサイズは 48 倍も小さくなっており，処理すべきデータ量が大幅に減少する.その結果，Stable Diffusion は大幅に高速化されている[6].

図 2.3 に示すように，Stable Diffusion モデル全体は End-to-End の構造を持ち，主に VAE(変分オートエンコーダー， Variational Auto-Encoder)，U-Net，CLIP テキストエンコーダーの 3 つのコアコンポーネントで構成されている.以下では，これらの 3 つの要素について順に説明する.

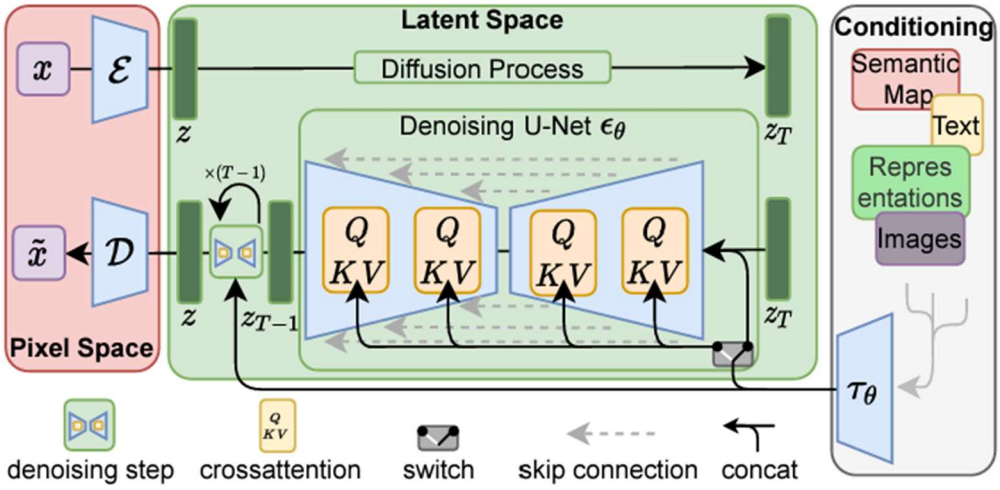

図 2.3 Stable Diffusion 構造図(文献 6 から引用)

2.2.1 VAE モデル

変分オートエンコーダー(Variational Autoencoder， VAE)は，生成型ディープラーニングモデルの一種であり，自動エンコーダー(Autoencoder)のアーキテクチャと変分推論(Variational Inference)を組み合わせることで，複雑なサンプリングプロセスに依存せずに高次元データの潜在表現を学習することが可能である.

SD モデルにおいて，VAE は Encoder-Decoder アーキテクチャに基づく生成モデルで

あり，その核心的なアイデアは，複雑な高次元データの分布(例えば画像)を単純な低次元の潜在空間にマッピングする点にある.そして，この潜在空間からサンプリングを行い，デコードすることでデータを生成する.

例えば，入力データのサイズが H×W×C の場合，VAE のエンコーダーモジュールはこれを低次元の潜在特徴量 h×w×c にエンコードする.このとき，f=H/h=W/w が VAE のダウンサンプリング率(Downsampling Factor)に相当する.一方で, VAE のデコーダーモジュールは同じアップサンプリング率(Upsampling Factor)を持ち，低次元の潜在特徴を復元してピクセルレベルの画像を生成する[6].

Stable Diffusion では，VAE モデルの微調整トレーニングが必要であり，主に L1 回帰損失と感知損失(perceptual loss， Learned Perceptual Image Patch Similarity， LPIPS)が損失関数として採用されている.また，パッチベースの対向訓練戦略も併用されている.

L1 回帰損失は，予測値と目標値の絶対差の合計を計算するもので，次のように定義されます：

$$\mathcal{L}_{L1} = \frac{1}{N}\sum_{i-1}^{N} |y_i - \hat{y}_i| \quad (2.18)$$

ここで，N はデータ内のピクセル総数.y_iはピクセル i の真の値(目標値).$\hat{y}_i$はピクセル i の予測値.

L1 損失は，ピクセルレベルで予測値と真の値の偏差を直接的に計算し，シンプルかつ堅牢な評価指標を提供します.

感知損失は，事前学習された畳み込みニューラルネットワーク(例えば VGG ネットワーク)を用いて，予測画像と目標画像の高次元特徴の差異を測定するものです.その公式は以下の通りです：

$$\mathcal{L}_{perceptual} = \sum_l \frac{1}{C_l H_l W_l} \|\phi_l(y) - \phi_l(\hat{y})\|_2^2 (2.19)$$

ここで，$\phi_l(y)$は事前学習ネットワークの l 層で抽出された真の画像yの特徴マップ.$\phi_l(\hat{y})$は事前学習ネットワークの l 層で抽出された予測画像$\hat{y}$の特徴マップ.$C_l H_l W_l$は l 層の特徴マップのチャンネル数，高さ，幅.$\|\cdot\|_2^2$はユークリッド距離(二乗ノルム).

感知損失は，ピクセルレベルではなく，高次元の特徴マップを比較することで，画像のグローバル構造やセマンティックな一貫性をより正確に捉えることができます.

上述の損失関数に基づいて，VAE モデルの微調整トレーニングを実行することが可能である.

画像解像度 (Image resolution)

画像解像度は，潜在画像テンソルのサイズに反映される.潜在画像のサイズは 4×64×64 と非常に小さく，これは 512×512 ピクセルの画像にのみ適用される.一方，4×96 の縦長画像の場合，テンソルサイズは 64×768×512 となる.このため，より大きな画像を生成するには，より多くの VRAM が必要となる[6].

また，Stable Diffusion v1 は 512×512 ピクセルの画像で微調整されているため，512×512 より大きい画像を生成すると，オブジェクトが繰り返される場合がある.この問題を回避するには，片方の辺を少なくとも 512 ピクセルに維持し，AI アップスケーラーを使用してより高い解像度を得ることを推奨する.

潜在空間での逆拡散

以下は，Stable Diffusion における潜在空間での逆プロセスの実行原理である.

1. ランダムな潜在空間行列を生成する.
2. ノイズ予測器が潜在行列のノイズを推定する.
3. 推定されたノイズを元の潜在空間行列から差し引く.
4. ステップ 2 と 3 を特定のサンプリングステップまで繰り返す.
5. VAE のデコーダーが潜在空間行列を最終的な画像に変換する.

VAE ファイル

Stable Diffusion v1 では，VAE ファイルを使用して目や顔の描画効果を改善している.これは前述の自己符号化器(Autoencoder)のデコーダー部分に相当する.デコーダーをさらに微調整することで，モデルはより精細なディテールを描画できるようになる.

前述では，画像を潜在空間に圧縮しても情報が失われないと述べたが，実際には，画像を潜在空間に圧縮する過程で情報の一部が失われる.これは，元の VAE では精細なディテールを完全に復元できないためである.しかし一方で，VAE のデコーダーはデコード時に精細なディテールを描画する役割を担っている.

VAE は画像を潜在空間へ変換し，また潜在空間から画像へ復元する過程は図 2.4 に示されている.

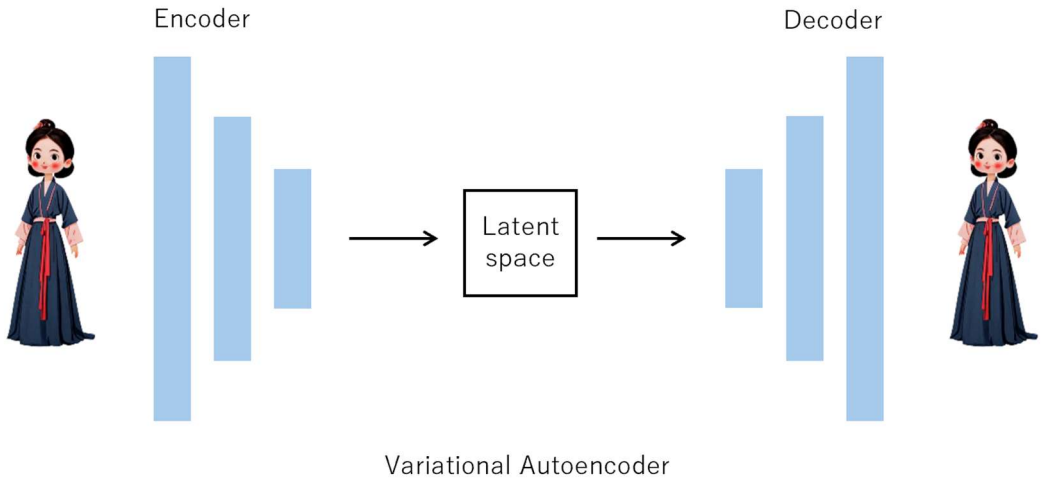

図 2.4 VAE の Stable Diffusion における応用

2.2.2 CLIP テキストエンコーダー

条件付け(conditioning)

前述の説明はまだ完全ではなく，最も重要なステップが欠けている.それはテキストプロンプト(text prompt)である.テキストプロンプトとは，生成モデルに対して出力する画像の内容を指示するための自然言語で構成された指令文のことである[13].テキストプロンプトが存在しない場合，Stable Diffusion はテキストから画像を生成するモデルとはならない.この場合，モデルの出力はランダムであり，例えば猫や犬の画像が生成される可能性があるものの，猫か犬のどちらを生成するかを制御することはできない.ここで必要となるのが条件付け(conditioning)である.条件付けの目的は，ノイズ予測器を誘導し，予測されたノイズを画像から差し引くことで，期待する対象物を得られるようにすることである.Stable diffusion モデルでは，CLIP を使う.

CLIP(Contrastive Language–Image Pretraining)

CLIP は，対照学習に基づくマルチモーダルモデルであり，テキストと画像の意味空間を整列させることを目的としている[14][15].このモデルは主に二つのサブモデルで構成されている：Text Encoder と Image Encoder である.Text Encoder はテキストの意味特徴を抽出するために使用され，通常 NLP 分野で広く利用される Transformer モデルを採用する.一方，Image Encoder は画像の視覚的特徴を抽出するために用いられ，ResNet[16]や Vision Transformer(ViT)[17]などの実装が一般的である.

CLIP のトレーニングは，4 億対の画像とラベルテキストを含むマルチモーダルデータセットを基盤としており，対比学習の手法でモデルを最適化する.具体的には，トレーニングセットからランダムに画像と対応するラベルテキストを 1 組抽出し，それぞれ Image Encoder と Text Encoder を通じて埋め込みベクトルを生成する.そして，余弦類

似度(cosine similarity)を使用して，2 つの埋め込みベクトルの意味的関連性を計算する.トレーニングの目標は，正しいペア(画像とテキスト)の類似度を最大化し，誤ったペアの類似度を最小化することである[14][15].この最適化は，以下の対比損失(contrastive loss)を用いて実現される：

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{\exp\left(sim(z_i^t,\ z_j^i)/T\right)}{\sum_{j=1}^{N}\exp\left(sim(z_i^t,\ z_j^i)/T\right)} \quad (2.20)$$

ここで，$sim(z_i^t,\ z_j^i)$はテキスト埋め込みと画像埋め込みの余弦類似度を表し，T は温度パラメータであり，埋め込みベクトルの分布を制御するために使用される.

Stable Diffusion では，CLIP の Text Encoder がユーザーから入力されたテキスト記述を高次元の意味的埋め込みベクトルに変換し,拡散プロセスにおける条件として利用する.この埋め込みベクトルは拡散モデルが潜在空間でのノイズ除去ステップを実行する際に影響を与え,生成される潜在変数がテキストの意味的記述に徐々に一致するように導く.CLIP の意味的整合性能力により，生成結果が入力されたテキストと一致することを確保するとともに，モデルに多様性を付与し，豊かな視覚コンテンツを生成することを可能にしている.

大規模な事前学習データセットを活用した学習により，CLIP は Stable Diffusion の条件生成モジュールとしての役割を果たすだけでなく,強力なゼロショット分類能力も備えている.この能力により，Stable Diffusion の生成品質と意味的一貫性がさらに向上している.

CLIP を紹介した上で，条件付け(conditioning)を具体的に説明する.トークナイザー(Tokenizer)は，まずプロンプト内の各単語をトークン(token)と呼ばれる数字に変換する.その後，各トークンを 768 次元の値ベクトルである埋め込み(Embedding)に変換する.この Embedding は，テキストエンコーダーによって処理され，ノイズ予測器が使用できる状態に準備される.

2.2.3 Unet モデル

テキストエンコーダーの出力は，U-Net 全体のノイズ予測器によって複数回使用される.U-Net は交差注意(Cross-Attention)メカニズムを通じてこれを活用する.ここが，プロンプトと画像が交わるポイントである.具体的な原理を説明する前に，先ずは Unet について紹介する.

Unet

UNet は，畳み込みニューラルネットワーク(CNN)[18]に基づく深層学習モデルであり，2015 年に Ronneberger らによって提案されたもので[19]，当初は医療画像のセグメンテーション専用に設計された[20].SD モデルでは，UNet は拡散プロセスの核心的なステップを実現する重要なコンポーネントであり，潜在空間で逐次ノイズを除去し，高品質な

潜在表現を生成する役割を担っている.

図 2.5 に示すように，UNet のアーキテクチャは主にエンコーダー(Encoder)とデコーダー(Decoder)の 2 つの主要部分から構成されており，両者はスキップ接続(Skip Connections)によって密接に結び付けられている.拡散モデルの順方向推論プロセスでは，SD モデルは UNet を繰り返し呼び出し，予測されたノイズ残差を元のノイズマトリックスから除去することで，逐次ノイズ除去された画像を得る.その後，VAE のデコーダー構造を通じて，ノイズ除去後の画像をピクセルレベルの画像として再構成する.

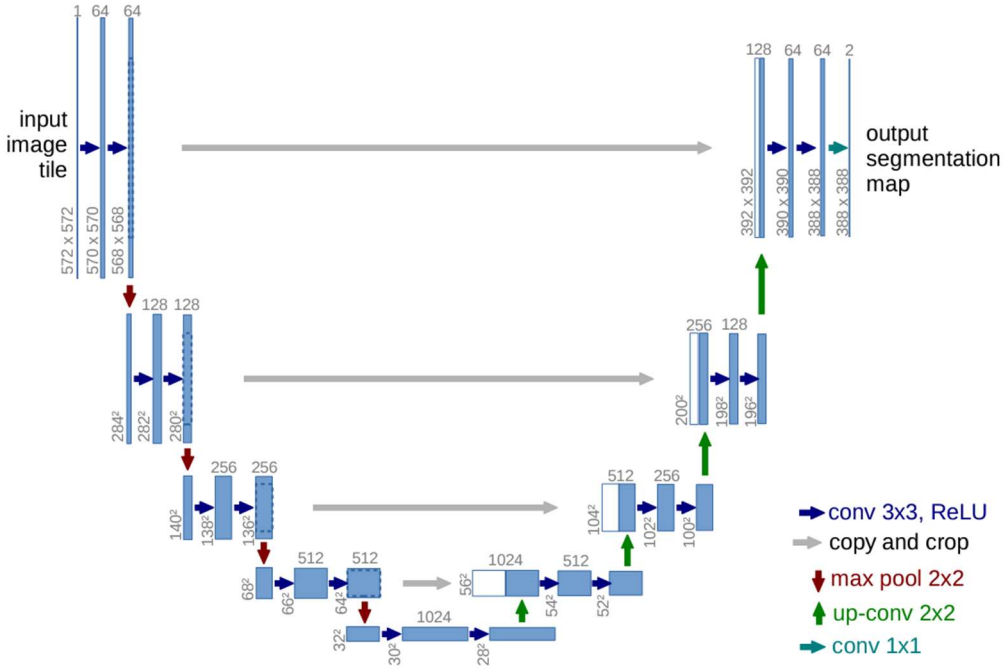

図 2.5 Unet 構造図(文献 19 から引用)

UNet の動作原理

1. 時間埋め込み(Time Embedding)：

UNet は現在の時間ステップ t の埋め込みベクトルを受け取り，それをノイズ除去プロセスに統合する.

時間埋め込みはモデルが現在のノイズ強度を理解するのを助け, 異なる時間ステップでのノイズ除去戦略を指示する.

2. 条件制御(Condition Control)：

Stable Diffusion では，通常 CLIP や他のテキストエンコーダーによって生成されたテキスト埋め込みを条件として使用する.UNet はこれらの条件情報を利用して，テキスト記述と一致する画像を生成する.

3. 多スケール特徴融合：

UNet はスキップ接続を通じて異なる解像度の特徴を統合し，ノイズ除去プロセスがグローバルな意味情報とローカルな細部情報を同時に保持できるようにする.

Stable Diffusion の逆プロセスにおける UNet の具体的な利用フロー

1. ノイズ入力：
現在の時間ステップx_tに対応するノイズ潜在変数を UNet に入力する.
2. 特徴抽出と融合：
エンコーダーを通じて多スケールの特徴を抽出し，時間埋め込みや条件情報と組み合わせて，徐々にノイズが低減された潜在表現を生成する.
3. ノイズ除去の出力：
デコーダーは処理後の特徴を使用して，ノイズが除去された潜在表現 x_{t-1}を生成し，次の時間ステップで使用する.
4. 反復生成：
UNet を繰り返し呼び出し，t=0 になるまで処理を行い，最終的に鮮明な潜在表現を生成する.

以上を踏まえると，プロンプト「青い目の男性」を例にすると，Stable Diffusion は「青い」と「目」という言葉をペアリングする(プロンプト内の自己注意メカニズム)ことで，青い目を持つ男性を生成する.一方で，青いシャツを着た男性を生成することは避けられる.その後，この情報を活用して逆拡散プロセスを青い目を含む画像へと導く(プロンプトと画像間の交差注意メカニズム).

2.2.4 Stable Diffusion 実行プロセス

本節では，前述した Stable Diffusion の基本コンポーネントの原理を総合し，テキストから画像生成(text to image)および画像から画像生成(image to image)の具体的な手順について説明する.

Text to image

Stable Diffusion を使用したテキストから画像生成では，テキストプロンプト(prompt)を提供することで画像が生成される.手順は以下の通り：

ステップ 1：Stable Diffusion は潜在空間でランダムテンソルを生成する.このテンソルは乱数生成器のシードを設定することで制御可能である.同じシード値を設定すれば，常に同じランダムテンソルが得られる.この段階では潜在空間上の画像だが，全てノイズで構成されている.

ステップ 2：ノイズ予測器である U-Net が，潜在ノイズ画像とテキストプロンプトを入力として受け取り，潜在空間(4×64×64 テンソル)内のノイズを予測する.

ステップ 3：潜在画像から予測された潜在ノイズを差し引き，新しい潜在画像を生成

する.

ステップ 2 と 3 を一定回数(例えば 20 回)のサンプリングステップで繰り返す.

ステップ 4：画像のデコード
最後に, VAE のデコーダーが潜在画像をピクセル空間に変換する.これが Stable Diffusion で得られる最終的な画像である.

Image to image

画像から画像生成は，SDEdit という手法で初めて提案された.この手法はすべての拡散モデルに適用可能であり，Stable Diffusion でも同様の機能を備えている.入力画像とテキストプロンプトを提供することで, 生成される画像がこれらの条件に基づいて調整される.手順は以下の通りである：

ステップ 1：入力画像を潜在空間にエンコードする.

ステップ 2：潜在画像にノイズを追加する.ノイズ量は降ノイズ強度(denoising strength)で制御される.降ノイズ強度が 0 の場合はノイズが追加されず，1 の場合は最大量のノイズが追加され，潜在画像が完全なランダムテンソルになる.

ステップ 3： U-Net が潜在ノイズ画像とテキストプロンプトを入力として受け取り，潜在空間(4×64×64 テンソル)内のノイズを予測する.

ステップ 4：潜在画像から予測された潜在ノイズを差し引き，新しい潜在画像を生成する.

ステップ 3 と 4 を一定回数(例えば 20 回)のサンプリングステップで繰り返す.

ステップ 5：最後に，VAE のデコーダーが潜在画像をピクセル空間に変換する.これが画像から画像生成によって得られる最終的な画像である.

2.3 3D 再構築

3D 再構築(3D Reconstruction)は，コンピュータビジョンおよびコンピュータグラフィックスにおける中核的な技術であり，二次元データ(例えば，画像，動画，点群データなど)を利用して物体やシーンの三次元形状，構造，幾何情報を復元するプロセスを指す.この技術は，産業製造，医用画像処理，映像特撮，仮想現実(VR)，拡張現実(AR)，および自動運転など，さまざまな分野で幅広く応用されている.

2.3.1 レンダリングと反レンダリング

レンダリング

レンダリング(Rendering)はコンピュータグラフィックスにおける中核的な技術であり, 三次元のシーンやモデルをアルゴリズムと計算を用いて二次元の画像に生成するプ

ロセスを指す.レンダリングの目的は，デジタル化された三次元データ(幾何形状，マテリアル，照明，テクスチャなど)を写実的な視覚表現に変換すること，または特定のスタイライズされた効果を生成することである.このプロセスは現代のデジタルコンテンツ制作において重要な地位を占めており，仮想現実(VR)，ゲーム開発，映像効果，建築デザインなど，さまざまな分野で不可欠な技術である.

レンダリング技術は主にリアルタイムレンダリング(Real-time Rendering)とオフラインレンダリング(Offline Rendering)の二種類に分類される.リアルタイムレンダリングは速度とインタラクティビティを重視し，ビデオゲーム，仮想現実(VR)，拡張現実(AR)などのシーンで広く利用されている.その主な目標は，限られた時間内に滑らかな映像を生成することであり，通常，毎秒数十フレームから百フレーム以上のレンダリングが要求される.これを実現するために，リアルタイムレンダリングではグラフィックスプロセッシングユニット(GPU)などのハードウェアアクセラレーションや，簡略化された照明モデルが使用され，性能と視覚効果のバランスを取る.

一方で，オフラインレンダリングは画像の高度な写実性に焦点を当て，映画，広告，高品質な静止画生成などに使用される.オフラインレンダリングはリアルタイム性の制約を受けないため，グローバルイルミネーション(Global Illumination)，パストレーシング(Path Tracing)，双方向パストレーシング(Bidirectional Path Tracing)など，光線の伝播を物理的にシミュレートする高度なアルゴリズムを採用することができる.現代のオフラインレンダラー(Blender Cycles，Arnold，RenderMan，V-Ray など)は物理ベースレンダリング(Physically Based Rendering， PBR)を広くサポートしており，材質，照明，環境の相互作用をリアルに表現する能力を持つ 78[21].

レンダリングプロセスは，照明計算，影の生成，反射と屈折のシミュレーション，テクスチャマッピング，アンチエイリアス処理など，複数の中核技術を含む.これらの技術は協調して機能し, 最終的な画像が物理法則に合致しつつ視覚的な美観を満たすことを保証する.例えば，グローバルイルミネーションシミュレーションでは，光線がシーン内で何度も反射や屈折を繰り返すことで，リアルな光影効果を生成する.また，アンチエイリアスアルゴリズムは, エッジの平滑な遷移を提供し，ピクセル化されたギザギザ現象を軽減する.

反レンダリング

反レンダリング(Inverse Rendering)はレンダリングとは逆のプロセスであり, 二次元画像から三次元シーンの構造，材質，照明などの情報を推定することを目的とする.これは, 画像生成プロセスの結果から入力条件を逆に求める典型的な逆問題である[22].反レンダリングの難点は，二次元画像が三次元シーンのすべての情報(深度や幾何学的関係など)を完全には表現できないため，問題の解決が不確定性を伴う点である.

この課題を解決するために, 反レンダリングではコンピュータビジョン, 最適化手法, 機械学習技術が併用される.近年では，深層学習に基づく反レンダリング手法が急速に発展しており，特に畳み込みニューラルネットワーク(CNN)や生成的敵対ネットワーク(GAN)が複雑なシーン情報の推定において顕著な性能を示している.また，反レンダリング技術は拡張現実，画像修復，三次元再構築，ロボットビジョンなどの分野において

も広く応用されている.例えば，反レンダリング技術を用いることで，一般的な写真からシーンの三次元構造を推定し,奥行き感のある仮想ビューを生成することが可能である.

将来的には，ハードウェア性能の向上とアルゴリズムの最適化が進むにつれて，レンダリングと反レンダリング技術は,より写実的かつ効率的な画像生成とシーンの復元を実現し，デジタル世界の構築にさらなる可能性をもたらすと考えられる.

反レンダリングの主要なタスクには以下が含まれる：

1. 三次元再構築：二次元画像から物体の幾何形状を再構築する.
2. 材質およびテクスチャ推定：物体表面の材質と色を推定する.
3. 照明推定：画像中の光源の分布と強度を推定する.

2.3.2 3D データの表現形式

三次元データの表現形式は,コンピュータ上で三次元物体やシーンを記述および処理するための数学的な表現方法である.この表現形式は，用途や計算特性に応じて異なるが，主に以下の種類が存在する.

1. メッシュ(Mesh)

メッシュは，頂点 (Vertices)，辺 (Edges)，面 (Faces) を用いて三次元物体の表面の幾何形状を表現する形式である.特に一般的なのは三角メッシュ (Triangle Mesh) であり，各面が 3 つの頂点によって構成される[23].この表現形式は軽量かつ効率的であり，レンダリングや可視化に適しているため，映画制作，ゲーム，仮想現実 (VR) などの分野で広く使用されている.表面が滑らかで，テクスチャマッピングに対応し，高速なハードウェアレンダリングをサポートするという利点がある.一方で，内部構造の表現には不向きであり，トポロジーの変化を処理する際には複雑さが増す.

2. 点群(Point Cloud)

点群は多数の三次元点(x，y，z 座標)で構成され，各点が物体表面の一部を表す.点群を用いた三次元表現は単純で，シーンや物体のキャプチャに適している.特に自動運転や三次元スキャンなどの分野で広く使用されている[24].点群データは直接的に取得可能である一方，点の分布が疎であり，表面の接続情報を含まない.また，可視化や処理が複雑であるという課題も存在する.

3. ボクセル(Voxel)

ボクセルは三次元空間を均一なグリッドに分割し，各小立方体(ボクセル)がその空間を物体が占有しているかどうかを記録する表現形式である.ボクセルは二次元画像におけるピクセルに類似しているが，深度(Depth)次元が追加されている点が特徴である.主に医療画像(CT/MRI)や物理シミュレーションに使用される.ボクセルは物体の内部構造を表現することが可能であるが,膨大な記憶容量を必要とし,解像度が制限されるため,細部の表現には限界がある[25].

4. 神経放射場(Neural Radiance Field，NeRF)

神経放射場(NeRF)は，暗黙的表現(Implicit Representation)の一種であり，三次元シーンの幾何情報や光照情報を連続的なニューラルネットワーク関数として符号化するものである.具体的な幾何形状やマテリアル情報を直接的に保存するのではなく，ニューラルネットワークを通じてエンドツーエンドで学習を行う点が特徴である.この手法は，従来の表現形式に比べて高度な表現力を持ち，リアルな三次元シーンの生成において注目されている 0.

以上のように，三次元データの表現形式は，用途や要件に応じてさまざまな特徴と利点を持つ.それぞれの形式は特定の課題に対するソリューションを提供し，三次元データの処理および応用の可能性を広げている.

2.3.3 神経放射場(NeRF)

神経放射場(NeRF) は，三次元シーンの表現とレンダリングに用いられる深層学習手法である.NeRF は暗黙的なモデリングの手法を通じて，三次元シーンを連続的な関数として符号化し，多視点画像データから高品質な三次元表現を生成し，新たな視点からリアルなレンダリングを実現することができる.

NeRF は，MLP 多層パーセプトロン[26][27]を用いて静的な三次元シーンを暗黙的に学習することで，複雑なシーンの任意の新しい視点合成(レンダリング)を実現する手法として簡潔に説明できる.

ネットワークを訓練するためには，静的なシーンに対して，カメラパラメータが既知の多数の画像を含むトレーニングセットが必要である.また，それぞれの画像に対応するカメラの三次元座標とカメラの方向も提供される必要がある.多視点のデータを用いて訓練することで，空間内の目標位置に対してより高密度かつ正確な色の予測が可能となり，ニューラルネットワークが連続性の高いシーンモデルを生成することを促進する.

任意のカメラ位置と方向を入力として，訓練済みのニューラルネットワークを通じてボリュームレンダリング(Volume Rendering)を行うことで[28]，画像をレンダリングする結果を得ることができる.ボリュームレンダリングのプロセスは図 2.7 を参考できる.

NeRF のパイプラインは，図 2.6 に示されているように，物体の三次元情報が立方体内に保存されていると仮定する.この立方体内の各座標点には，5 次元の表現が割り当てられており，座標値 (x，y，z)，体密度σ，および観察角度θに基づいて観察された色 C を含む.したがって，NeRF の入力は一連の画像のカメラ姿勢であり，出力は色と不透明度の値の集合である.

NeRF の作業は主に訓練と推論の 2 つの部分に分けられる.以下では，理解を容易にするため，まず NeRF の推論プロセスについて説明する.

推論

推論プロセスのタスクは，入力画像とは異なる視点の画像を生成することである.この異なる視点の画像を生成するには，ピクセルごとに処理を行う必要がある.

まず，ニューラルネットワークが訓練済みであると仮定する.生成する画像のカメラ姿勢に基づいて，その姿勢から 1 本の光線(レイ)を発射する.この光線は神経放射場と交差し，その交差点のすべての点がネットワークに入力され，体密度σと色 RGB が予測される.このとき，ボリュームレンダリング技術を使用して，これらの色彩情報を統合することで，現在のピクセル点の値が得られる.この操作を繰り返すことで，異なる視点の画像をレンダリングすることが可能である.

訓練

訓練プロセスは，推論フェーズと大まかには同じ手順である.入力画像の各ピクセルから光線を発射し，それを神経放射場と交差させる.そして，交差したすべての点をネットワークに入力し，体密度σと色 RGB を推測する.このとき，入力画像の各ピクセル値は既知であるため，ネットワークの推測結果と実際の値との損失を計算する.この損失に基づいて，ネットワークのパラメータを更新する.

このプロセスを繰り返すことで，NeRF は入力画像群を正確に再現できるように学習し，新しい視点での高品質な画像生成を可能にする.

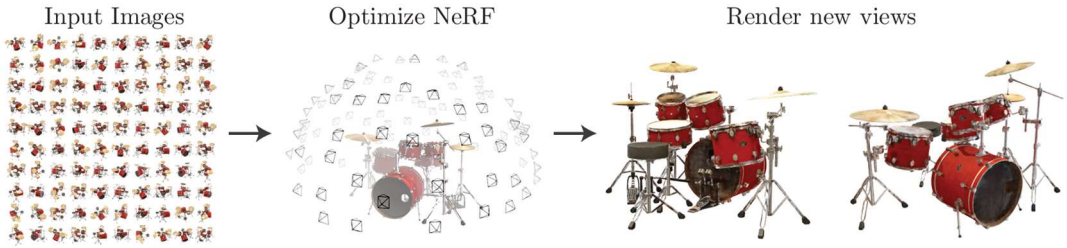

図 2.6 NeRF プロセス(文献 1 から引用)

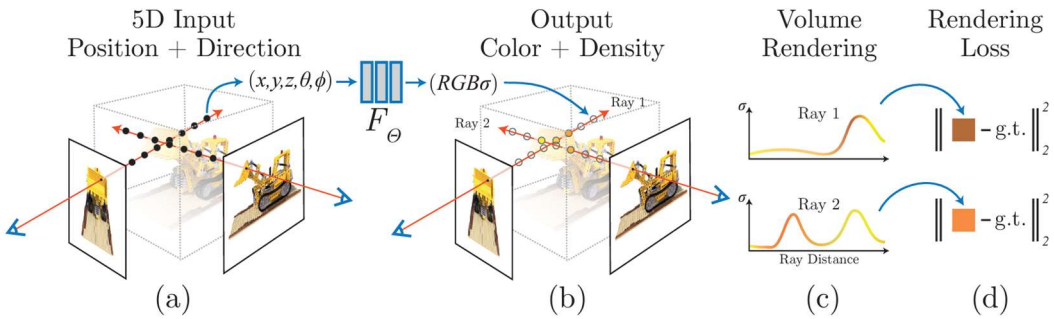

図 2.7 ボリュームレンダリングプロセス(文献 1 から引用)

カメラ姿勢

カメラ姿勢(Camera Pose)は，コンピュータビジョンおよびグラフィックスにおける重要な概念であり，三次元空間におけるカメラの位置と向きを記述するために使用される

[29].カメラ姿勢は，カメラがシーンをどのように観測するかを決定し，生成される画像や映像の視点に影響を与える.

カメラ姿勢は主に以下の 2 つの情報で構成される：位置と向き.

位置(Position)というのはカメラが三次元空間内でどこにあるか, つまりカメラの光心の三次元座標 (x，y，z)である.特定の基準座標系(通常は世界座標系)に対するカメラの空間位置を記述する.

向き(Orientation)はカメラの観測方向を表し，三次元空間内での回転状態を定義する.通常，回転行列または四元数を用いて表現される.

カメラ姿勢の意義は，主に以下の 3 つに分類される.まず，カメラ姿勢は世界座標系とカメラ座標系の変換関係，またはその逆を記述する役割を持つ.また，姿勢情報は，シーン内の物体がカメラの視野においてどのように投影されるかを決定し, 画像の透視効果，スケール，方向に直接影響を与える.さらに，応用例として，三次元再構築では複数の視点から取得した画像を基に物体の三次元構造を推定し, カメラキャリブレーションではカメラの姿勢を推定して歪んだ画像を補正するために使用される.加えて，拡張現実(AR)では，仮想コンテンツを現実のシーンと正確に整列させるために利用される.

カメラキャリブレーション

カメラキャリブレーション(Camera Calibration)は，コンピュータビジョンにおける重要な技術であり, カメラの幾何学的特性および光学的特性を特定するために用いられる[30].これにより, カメラが三次元のシーンをどのように二次元画像に投影するかを正確に記述することが可能となる.キャリブレーションの目的は，カメラの内部パラメータ，外部パラメータ，および歪みパラメータを取得し，三次元再構築，姿勢推定，拡張現実(AR)などの応用に基盤を提供することである.

カメラキャリブレーションのプロセス

1. キャリブレーション対象の選択

キャリブレーションに使用される代表的な対象物には，チェスボードパターン，ドットアレイ，または他の規則的な幾何学的パターンが含まれる.

2. 画像の取得

カメラを使用して，異なる視点からキャリブレーション対象を撮影する.視点や距離に多様性を持たせることが重要である.

3. 特徴点の検出

キャリブレーション対象物上の特徴点(例えばチェスボードの交点)を抽出する.

4. 内部パラメータと歪みパラメータの推定

幾何学的な最適化アルゴリズム(例えば最小二乗法)を使用して，カメラの内部パラメータとレンズ歪みパラメータを推定する.

5. 外部パラメータの推定

世界座標系と画像座標系の対応関係に基づき，カメラの外部パラメータを計算する.

6. 結果の検証

歪み補正後の画像を用いて，キャリブレーションの精度を確認する.

カメラキャリブレーションは，精密な三次元計測やロボットビジョンなど，幅広い分野で欠かせない技術である.

視点の表現(サンプリングポイント)

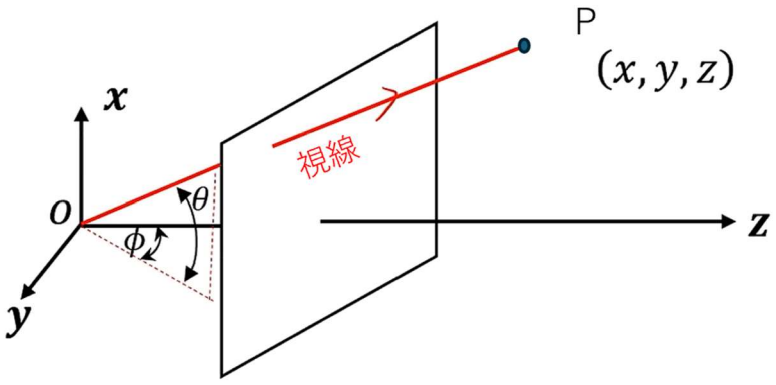

図 2.8 視点の表現方

Nerf において，視線方向は 2 つのパラメータ，図 2.8 に示すように，すなわちθとΦで表現することが可能である.θは視線の仰角を表し，Φは視線の方位角を表す.

視線 r 上の点は次のように表現できる：

$$r(t) = o + td \quad (2.21)$$

ここで，O はカメラの光心が世界座標系における座標を示し，d は直線の方向を表す単位ベクトル(デカルト座標系)である.また，t は実数であり，点 o から視線上の点 r(t) までの距離を示す.

ボリュームレンダリング

ボリュームレンダリング技術は, 三次元ボクセル(Voxel)データを処理して二次元画像を生成し, 物体の内部構造と密度分布を表現する.医療画像, シミュレーションデータ, スキャンデータなどの三次元体データを直接可視化するために使用され, 表面を明示的に構築する必要がない.NeRF において，ボリュームレンダリングは視線 r 上のすべての点を画像に投影してピクセル色 C を形成するプロセスを指す.ボリュームレンダリング

は三次元体データを基盤とし，体データはボクセルで構成されており，ボクセルは三次元空間を均等に分割した小さな立方体である.各ボクセルには物体のある点における属性(密度，色，透明度など)が記録されている.NeRF においては，この属性が体密度σと色RGB を指す.ボリュームレンダリングは，カメラ視点から三次元体データに光線を投射し，その光線に沿ってボクセルの属性を累積することで二次元画像を生成する.また，光線が体データを通過する際の色と透明度の累積モデルに基づいて最終的なピクセル値を計算する，数式は下記の通り：

$$C = \int Color(x) \cdot Opacity(x) \cdot Light(x)dx \quad (2.22)$$

Nerf における色 C の計算式は以下のように示される：

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma\big(r(t)\big)c\big(r(t),\ d\big)dt,\ whereT(t) = \exp\left(-\int_{t_n}^{t} \sigma(r(s))ds\right) (2.23)$$

この式において，r(t)は三次元空間内の一点を表し，d は視線の方向である.三次元シーンの近端および遠端の境界はそれぞれt_nとt_fで示され，$c\big(r(t),\ d\big)$は三次元点 r(t)を方向 d から見たときの色値を示す.$\sigma\big(r(t)\big)$は体密度関数であり，位置 r(t の物理的な材質が光を吸収する能力を表している.また，$T(t)$は光線がt_nから t まで移動する間の累積透過率を示している.

この公式の物理的な意味において，以下の点に注意する必要がある.まず，$\sigma = 0$の場合，最終的なピクセルの色 C に全く影響を与えず，完全に透明な状態を示している.$T(t) = 0$の場合，光の強度が 0 であり，何も見えないことを意味するため，ピクセルの色 C にも影響を与えない.最後に，t およびσが増加するにつれて，ピクセルの色への影響はより大きくなる.

サンプリング方法は以下の通りである：
t_nからt_fを N 個の等間隔な区間に分割し，それぞれの区間からランダムに均一なサンプルt_iを抽出する.

$$t_i \sim \mu[t_n + \frac{i-1}{N}\big(t_f - t_n\big),\ t_n + \frac{i}{N}\big(t_f - t_n\big)] \quad (2.24)$$

ここで，i の値の範囲は 1 から N までである.

体積レンダリングの手法は以下の式を参考とする：

$$\hat{C}(r) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))c_i\ whereT_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j\delta_j\right) (2.25)$$

ここで，$\delta_i = t_{i+1} - t_i$は隣接するサンプリングポイント間の距離を表す.

総括すると，Nerf はニューラルネットワークを活用し，体積レンダリング技術を組み

合わせることで，複数の入力画像から三次元物体を推測する手法である.Nerf では，三次元物体の情報(体密度および各視点に対応する色値)が神経放射場(Neural Radiance Field)に保存される.そして，体積レンダリング技術を通じて入力画像の各ピクセル値を計算することで，新しい視点からの画像生成(レンダリング)を実現する.Nerf の目的は新しい視点の画像を生成することであるが, 神経放射場に物体の三次元情報が保存されるため，Nerf は後続の 3D モデル生成における中核技術の一つとなっている 0.

2.4 3D モデル生成

近年，拡散モデルが AI 絵画分野で活用されるようになったことで，3D 生成分野への応用についての研究が徐々に進められている.本小節では，3D モデル生成分野におけるいくつかの重要な研究について紹介する.

2.4.1 zero123

Zero-1-to-3 は，単一の RGB 画像を基に複数の視点におけるレンダリング結果を生成する革新的なフレームワークである.大規模な拡散モデルと幾何学的先験知識を組み合わせることで，Zero123 は制約のある条件下でも複数の目標視点の画像を生成し，それらの生成結果において幾何学的および外観の一貫性を保証することが可能である[31]. 本モデルは合成データセットを主なトレーニングデータとしているにもかかわらず, 未知のデータセットや自然シーンを含む複雑な入力, 例えば印象派の絵画や芸術的スタイルの画像に対して強力なゼロショット汎化能力を示している.Zero123 は，三次元再構築, 新しい視点の合成, および仮想コンテンツ生成における効果的なソリューションを提供するものである.

Zero123 の核心的なアイデアは，単一の視点入力を基に複数の視点におけるレンダリング画像を生成し, それらの生成された視点間で幾何学的および外観の一貫性を保証することである.このモデルの実現は，いくつかの重要な技術に依存している.まず，幾何学的先験知識の導入により，大規模な訓練データセットから得られる幾何情報(例えばカメラの姿勢や視点間の対応関係)を活用して訓練を行い，生成されたマルチビュー画像の幾何学的一貫性を確保する.次に，拡散モデル(Diffusion Model)を用いて画像を生成し，初期ノイズから段階的にノイズを除去することで，多視点画像の生成品質と一貫性を保証する.最後に，Zero123 は条件付き生成技術を活用して，入力画像の視点変化および対応するカメラパラメータに基づいて適切なレンダリング結果を生成し, マルチビュー画像の精密な制御を実現する.

Zero123 のワークフロー

1. 単一視点入力

モデルは単一の RGB 画像を入力として受け取り，この画像を基に複数の視点におけ

るレンダリング結果を生成するための参照として利用する.

2. 条件付き拡散生成

拡散モデルを用い，条件(例えばカメラの視点変化)に基づいて段階的に新しい視点画像を生成する.

3. 多視点一貫性の制約

モデルは幾何学的先験知識および学習された多視点一貫性のルールを適用することで, 生成された多視点画像が物体の構造や光の表現において一貫性を持つよう保証する.

4. 多視点画像の出力

複数のカメラ視点からのレンダリング画像を出力し, これらの画像は三次元再構築やレンダリングタスクのさらなる処理に活用可能である.

多視点合成のための条件付き拡散モデル

Zero123 の核心は, 拡散モデルに基づく条件付き生成アーキテクチャである.拡散モデルは，ランダム分布から段階的にノイズを除去することで目標画像を生成する.Zero123 は，拡散モデルを条件付き生成に拡張し，入力画像とカメラ外部パラメータ(例えば回転 R と平行移動 T)を条件として新たな視点を生成することを可能にする.その数学的表現は以下の通りである：

$$\hat{x}_{R,\ T} = f(x,\ R,\ T) \quad (2.26)$$

ここで，$\hat{x}_{R,\ T}$は目標視点の生成画像，x は入力画像，R と T はそれぞれカメラの回転および平行移動パラメータを表す.

合成データに基づくトレーニング

Zero123 は Objaverse などの大規模な合成データセットを使用して[32], 視点と画像の対応関係を学習する.トレーニングは以下の損失関数を最小化することで行われる：

$$min_\theta E_{z,\ t,\ \epsilon}\left[||\epsilon - \epsilon_\theta(z_t,\ t,\ c(x,\ R,\ T))||_2^2\right] \quad (2.27)$$

ここで，z_tは時刻 t における潜在変数，ϵ_θはノイズ予測器，$c(x,\ R,\ T)$は条件エンコーディングであり，入力画像とカメラ外部パラメータを組み合わせたものである.このトレーニング方法により，Zero123 は多視点条件下で正確な出力を生成する能力を獲得する

視点条件生成と混合条件生成

Zero123 は視点条件生成を採用し，入力画像とカメラの外部パラメータを条件として用いることで, 生成された画像が幾何学的構造および外観において一貫性を持つことを

保証する.さらに，CLIP による全体的な意味情報と局所的な幾何学的特徴を統合する混合条件生成機構を導入している.その具体的な数式は以下の通りである：

$$c(x,\ R,\ T) = CLIP(x) \oplus (R,\ T) \quad (2.28)$$

ここで，⊕は条件結合演算子を表す.この混合条件生成は，生成された多視点画像の一貫性を高めるものであり，交差注意(Cross Attention)メカニズムを利用して幾何情報と意味情報を効率的に融合している

多視点の一貫性

Zero123 の重要な特長の一つは，多視点における一貫性である.これは，異なる視点の生成結果において幾何学的構造および照明表現が一致することである.モデルはトレーニングプロセスにおいて多視点の関連ルールを学習し，単一視点の入力から完全な多視点レンダリングを推測できる.この特性は，複雑なオブジェクトの再構築や合成において非常に重要である.

3D 再構築への応用

Zero123 は，その多視点生成能力を 3D 再構築タスクに拡張している.ボリュームレンダリング技術や Score Jacobian Chaining(SJC)[33]などのニューラルフィールド最適化技術を組み合わせることで，単一視点から高精細な三次元情報を推測する能力を実現している.SJC の最適化フレームワークは以下の数式で表される:

$$\nabla \mathcal{L} = \mathbb{E}\left[score_{pred} - score_{gt}\right] (2.29)$$

ここで，$score_{pred}$は予測されたスコア，$score_{gt}$は正解スコアを表す.この最適化プロセスにより，生成される三次元幾何学が高い品質と一致性を保つことができる.Zero123 は複数の視点から生成された情報を統合し，単一視点による三次元再構築の新たなソリューションを提供している.

さらに，Zero123 は RGB 画像やカメラの姿勢情報など，複数のモダリティを組み合わせて入力として利用可能であり，Transformer や UNet モジュールを通じてこれらのモダリティ間の関係を効果的に捉えることができる.

Zero123 の利点は，単一視点生成能力の高さ，多視点レンダリングの一貫性，生成結果の高品質さ，そして柔軟性にある.まず，Zero123 は複数の視点入力を必要とせず，1 枚の画像から高品質な多視点結果を生成することができる.幾何学的先験知識と条件付き拡散メカニズムを通じて，生成された視点画像は幾何構造および照明表現において高度な一貫性を保つ.また，拡散モデルの活用により，Zero123 は細部が豊かでリアル感のあるレンダリング結果を実現する.さらに，Zero123 は柔軟性が高く，写真やイラストなど多様な入力タイプをサポートし，異なる視点の画像を生成することが可能である[31].

図 2.9 zero123 出力結果(文献 31 から引用)

図 2. 9 は zero123 の出力結果.総括すると，Zero123 は革新的な単一視点から多視点への生成モデルであり，拡散モデルと幾何一貫性技術を活用することで，1 枚の画像から高品質な多視点レンダリング結果を生成することができる.この技術は，三次元再構築，仮想コンテンツ生成，グラフィックス研究などの分野において広範な応用可能性を持ち, 単一視点から三次元シーンを生成するための効果的なソリューションを提供している.

2.4.2 Wonder3D

Wonder3D は，単一視点の画像から高精細なテクスチャメッシュを効率的に生成するための革新的な方法である.従来の Score Distillation Sampling(SDS)[34]に基づく方法と比較して，このフレームワークは単一視点の三次元再構築タスクにおける品質，一貫性，効率を大幅に向上させている[35].Wonder3D は，クロスドメイン拡散モデルを採用して多視点の法線マップおよび対応するカラー画像を生成し, 多視点クロスドメイン注意メカニズムを通じて各視点間およびモダリティ間の情報交換を実現することで, 生成結果の幾何学的および視覚的な一貫性を保証している.さらに，このフレームワークでは，多視点の 2D 表現から高品質な表面を抽出する幾何認識型法線融合アルゴリズムが提案されている.実験結果から，Wonder3D は再構築の品質，汎化能力，効率の面で従来の方法を上回り，単一視点三次元再構築分野における強力な可能性を示している.

単一画像から 3D 幾何を再構築することは，バーチャルリアリティ，ゲーム，3D コ

ンテンツ制作，ロボット工学など，多岐にわたる分野で重要な課題である.しかし，入力データが 1 枚の画像に限られる場合，3D 形状の推定には可視部分および不可視部分の幾何情報の補完が求められ，非常に難しいタスクである.従来の手法では，以下のような課題があった：

効率性の問題: SDS をベースにした手法は，一つの形状を生成するのに数万回の反復が必要で，数十分から数時間を要する.

一貫性の欠如: 生成される 3D 形状が多視点間で矛盾することが多く，多重顔問題が発生しやすい.

図 2.10 に示すように，Wonder3D は拡散モデルと多視点条件生成に基づくフレームワークであり，入力された 2D 画像から高品質な 3D 表現形式(法線マップや RGB 画像など)を生成することを目的としている.その核心的なメカニズムは以下のように要約できる：

1. 拡散モデル

Wonder3D は拡散モデルを基盤として，ノイズデータを段階的に目標画像に復元する.

拡散モデルは，カメラの姿勢やタスクの種類などの条件埋め込みを用いて，入力画像に対応する多視点出力を生成する.

2. 多視点生成

固定されたカメラ姿勢行列(front，right，back，left など)を使用して，推論により複数視点の出力を得る.単一の入力画像を y とし，カメラの姿勢をπ_1，π_2，π_3，…，π_K，と仮定すると，三次元アセットの結合分布は以下のように表現できる：

$$P_\alpha(z) = P_{nc}(n_{1:K}，\ x_{1:K}|y) \quad (2.30)$$

そのうち，$n_{1:K}$は法線マップ，$x_{1:K}$は RGB 画像を表す.

モデルは訓練過程で多視点条件間の関連性を学習し，単一の入力画像から完全な多視点出力を生成することが可能である.

3. タスク埋め込み

タスクの種類(例えば RGB 画像生成や法線マップ生成)は拡散モデル内に埋め込まれ，特定のタスクラベルを通じてモデルが目標データを生成するよう指導する.

4. モーダル統合

モデルは 2D 画像の特徴と 3D 幾何学的特性を統合し，Transformer や U-Net などのモジュールを通じて，複数のモーダル間の関連性を捉える.

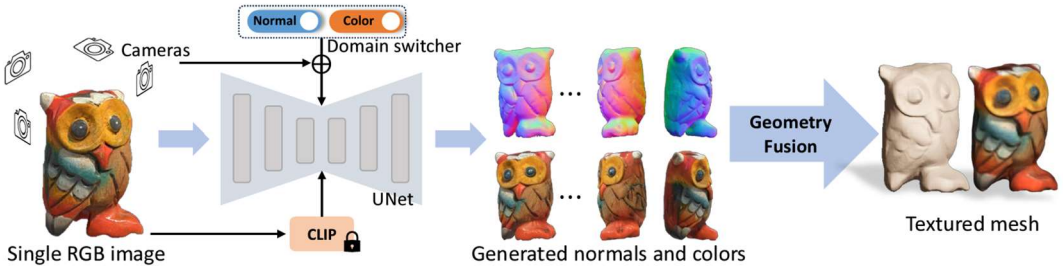

図 2.10 Wonder3D プロセス(文献 35 から引用)

研究の最後に，著者らは Google Scanned Objects データセット[36]を用いて提案手法の効果を検証し，さらに素描やアニメ風イラストなど，異なるスタイルの画像を対象にテストを行った.その結果, Zero123 や SyncDreamer[37]などの手法と比較して, Wonder3D はより高い幾何学的細部および一貫性を示し, 再構築時間を大幅に短縮することが確認された.

2.5 関連研究

近年，3D モデルの研究は大きな注目を集めており，研究の焦点は主に 3D 再構築と 3D 生成の 2 つの主要な領域に集中している.本章では，これらの分野に関連する研究を紹介し，主な貢献と限界について議論することで，本研究の基盤を提供する.

2.5.1 3D 再構築

3D 再構築技術は，物体の三次元情報を二次元の投影や画像を用いて復元するプロセスを指す.第 2.3 節では，本研究に密接に関連する 3D 再構築技術である NeRF について詳しく説明したため，本節では NeRF 以外の技術について簡単に紹介する.

多視点幾何に基づくクラシックな方法

1992 年，Tomasi と Kanade は，三次元形状とカメラの動きを画像列から復元するための順次分解法を提案した論文を発表した.この手法では，特徴点を追跡しながら，奇異値分解(SVD)を用いて頑健で正確な結果を得ることが可能であった.しかし, 従来の分解法はバッチ処理に依存しており，SVD の計算コストが高いため，リアルタイムの応用には適していなかった.この課題を解決するために，著者らは従来の分解法を拡張し，特徴点の位置を時系列ベクトルとして扱う順次処理法を提案した.この新手法では，各フレームごとに形状と動きを推定し，完全な SVD の代わりに 3 つの主要な固有ベクトルの更新計算を用いることで，計算コストを $O(FP^2)$から $O(P^2)$に削減した.この手法は大規模な行列を保存する必要がないため，無限長のシーケンスにも対応可能である.実験結果により, この手法は合成画像と実画像の両方で従来の手法とほぼ同じ精度と頑健

性を達成しながら，計算効率を大幅に向上させたことが示された[38].

2010 年，Furukawa と Ponce らは，Multi-View Stereo (MVS)技術を提案した.この研究では，多視点ステレオビジョンを用いた新しいアルゴリズムが開発され，境界体積の初期化を必要とせず，外れ値や障害物を自動的に検出して除去することが可能となった.この手法は，局所的な光学的一貫性と全体的な可視性制約を基盤としており，「マッチング，拡張，フィルタリング」というプロセスを通じて機能する.まず，疎な特徴点から始め，これを近傍の画素対応に拡張し，可視性制約を利用して誤ったマッチングを除去する.この手法は隣接する特徴間での平滑化を行わないにもかかわらず，複数のベンチマークデータセットで高いカバレッジと精度を実現した.さらに，この論文では，得られたパッチモデルを画像ベースのモデリングに適したメッシュに変換するシンプルで効果的な手法も提案している.この手法は，複雑な表面ディテールや深い凹部，薄い構造を持つ物体，限られた視点から観測される屋外シーン，さらには静的な構造物の複数画像内に移動する障害物が存在する「混雑した」シーンにおいても，その頑健性と高い適応性を示している[39][40].

深層学習に基づく 3D 再構築

近年，深層学習は 3D 再構築に全く新しいアプローチをもたらしている.PointNet は，点群データを直接処理する初の深層学習フレームワークを提案した.この手法は，点群をメッシュやボクセルに変換する必要がなく，点群のグローバル特徴を学習することで，点群の分類とセグメンテーションを実現した[41].その後，DeepSDF により，連続符号付き距離関数 (SDF) に基づく暗黙的表現方法が提案され，ニューラルネットワークを用いて三次元形状の暗黙的な表面表現を学習し，高解像度の三次元モデルを生成することが可能となった[42].

点群に基づく 3D 再構築

点群データは，3D 再構築の重要な表現方法として，多くの効率的なアルゴリズムを生み出してきた.Poisson Surface Reconstruction は，ポアソン方程式に基づく表面再構築手法を提案し，高密度の点群から高品質な三角メッシュを生成する技術として，点群処理のクラシックな技術の一つとなっている[43].さらに，ConvPoint は連続畳み込み技術を提案し，点群データに畳み込み操作を導入することで，点群処理の効率を大幅に向上させた.従来の近傍探索手法と比較して，ConvPoint はリアルタイム応用により適している[44].

3D Gaussian Splatting は，ガウシアンポイントクラウド表現に基づくリアルタイムの 3D レンダリング手法である.この手法では，シーンを三次元のガウス分布ポイントの集合に分解することで，幾何情報と色情報を表現する.これらのガウスポイントは直接レンダリングが可能であり，推論速度が大幅に向上する.しかし，この手法には汎化能力の弱さや高いストレージコストといった欠点がある[2].

ボクセルに基づく 3D 再構築

VoxelNet は，点群データをボクセルに離散化し，深層学習を用いてボクセル構造をモデリングすることで，初めてエンドツーエンドの点群物体検出フレームワークを実現した.一方，OctNet は，ボクセルのストレージと計算をさらに最適化し，八分木階層表現を

利用してボクセル解像度と計算資源のトレードオフを解決した[45][46].

深度マップに基づく 3D 再構築

深度画像は，2D 画像と 3D 幾何情報を結びつける橋渡しとして広く利用されている.DTAM は，深度画像に基づくリアルタイム密な三次元再構築を実現し，リアルタイムアプリケーションの基盤を確立した[47].また，Stereo Matching with Deep Learning は，幾何情報と文脈情報を統合し，深層学習を活用して深度画像の精度を大幅に向上させた[48].

マルチモーダル統合による 3D 再構築

近年，画像，点群，セマンティック情報などの複数の入力形式を融合する多モーダル 3D 再構築技術が注目を集めている.例えば，Pix2Mesh[49]は，ピクセル特徴に基づいてメッシュの頂点位置を予測する深層学習手法を提案し，単一視点または複数視点から三次元メッシュを生成することを可能にした.同様に, オープンソースツールである COLMAP [50]は，SfM と MVS を一つのシステムに統合し，高精度な三次元再構築の実用的な解決策を提供している.

2.5.2 3D 生成

3D モデル生成技術とは，アルゴリズムと計算を用いて，さまざまな入力データから三次元物体やシーンの幾何形状，テクスチャ，材質を自動生成する技術である.近年，拡散モデルや CLIP モデルの研究進展に伴い，テキストから 3D への生成や画像から 3D への生成といったさまざまな 3D 生成技術が大きな成果を上げている.

Shap-E は 3D アセットの条件付き生成モデルであり, 従来の単一出力表現を生成する 3D 生成モデルとは異なり, Shap-E は暗黙的関数のパラメータを直接生成する.このパラメータは，テクスチャ付きメッシュや NeRF としてレンダリング可能である.Shap-E のトレーニングプロセスは 2 つの段階に分けられる.第 1 段階では，エンコーダーを使用して 3D アセットを決定論的に暗黙的関数のパラメータにマッピングする.第 2 段階では，条件付き拡散モデルをトレーニングし，エンコーダーの出力に基づいて 3D データを生成する[51].

Shap-E は大規模な 3D およびテキストデータセットでトレーニングされており, わずか数秒で複雑で多様な 3D アセットを生成できる.点群ベースの明示的生成モデルである Point-E と比較して，Shap-E は高次元かつ多表現の出力空間において，より高速に収束し，より高品質なサンプルを生成することができる.また，著者はモデルの重み，推論コード，サンプルデータを公開し，関連研究の発展を促進している[51].

先駆的な研究である DreamFusion では, テキストから画像を生成するモデルが提案された[34].これらの手法は，3D 表現(例えば NeRF，メッシュ，SDF)を最適化した後，ニューラルレンダリングを用いて異なる視点から二次元画像を生成する.これらの画像を拡散モデルや CLIP モデルに入力して SDS 損失を計算し，3D 形状の最適化を導く.しかし, これらの手法の多くは非効率であり, マルチフェイス問題などの課題を抱えている.

Magic3D は，DreamFusion が抱える課題を克服するために，2 段階の最適化フレームワークに基づく 3D 生成手法を提案した.この手法の第 1 段階では，低解像度の拡散モデルを用いて粗い 3D 表現を生成し，スパースな 3D ハッシュグリッド構造を通じて処理を加速し，後続の最適化における効率的な初期化を提供する.第 2 段階では，高効率な微分可能レンダラーと高解像度の潜在拡散モデルを使用してさらに最適化を行い, 精緻なディテールを持つ高品質な 3D メッシュモデルを生成する.DreamFusion と比較して, Magic3 は生成効率と品質の両面で顕著な向上を実現している.具体的には, Magic3D は平均 40 分での生成時間を達成しており, DreamFusion が約 1.5 時間を要する生成時間の半分に短縮しているだけでなく，生成結果の解像度と細部の品質も向上している.また，ユーザー調査の結果，61.7%のユーザーが Magic3D の生成結果をより好むと評価しており，実用的な優位性がさらに裏付けられた[52].

Magic3D の貢献は，効率的な 2 段階の最適化戦略を提案した点に加え，画像条件付き生成機能を組み合わせることで，3D 生成における柔軟性と制御力を大幅に向上させたことである.この手法の提案により，創造的コンテンツ生成や応用シナリオにおいて新たな可能性が切り開かれ，関連研究における重要な技術的基盤が提供された.

Viewset Diffusion[53]，SyncDream，MVDream[54]などの研究も同様のアプローチを採用し, アテンションレイヤーを活用して一貫性のあるマルチビューの色情報を生成している.しかし，色情報からの再構築では，テクスチャの曖昧性に起因する課題に直面する.その結果，幾何学的な詳細を正確に復元できない，あるいは大規模な計算リソースを必要とする場合が多い.

Zero1to3 に触発され, 一部の手法ではクロスドメイン拡散モデルを導入してマルチビューの RGB 画像と対応する法線マップを生成している.さらに，クロスドメインのアテンションメカニズムを用いて生成の一貫性を向上させ, マルチビューの二次元表現から高品質な表面を抽出するジオメトリ対応法線融合アルゴリズムを設計している.

ゲームモデルの作成において，現存する 3D モデル生成に関する研究は，忠実度の不足，幾何情報の欠損，詳細の不足といった品質上の課題を依然として抱えている.

第3章 課題および目的

本章では，これまで述べてきた 3D モデル生成の背景を踏まえ，本研究が取り組むべき課題を明確にし，その目標および本研究が目指す方向性について述べる.

3.1 ゲーム市場の現状

近年,グローバルなゲーム市場は急速に拡大しており,3D ゲームが主流となっている.それに伴い,開発コストも大幅に増加している.統計によると,3D ゲームの開発コストはプロジェクトの規模や複雑さによって大きく異なる.小規模な 3D ゲームの開発費用は約 3 万～7 万米ドル,中規模の 3D ゲームでは 70 万～300 万米ドルの範囲である.AAA クラスの 3D ゲームの場合,開発コストは 7000 万～1 億米ドルに達することもあり,一部のゲーム（例：「Black Myth: Wukong」）では 1 億米ドルを超える予算が投じられている 81[59].

このコストの中で,3D ビジュアル制作は非常に大きな割合を占める.3D グラフィック関連の費用は,ゲーム開発全体の 60～70%を占めることが一般的である[60].さらに,美術制作の中でも 3D モデリングは最も時間と費用がかかるプロセスであり,美術コスト全体の約 60%を占めるとされている[61].特に AAA タイトルでは,映画レベルの視覚品質を実現するために,1 つのキャラクターモデルの制作に数万ドルのコストがかかり,完成までに数週間から数ヶ月を要する[62].さらに,3D ゲームの制作にはモデリングだけでなく,アニメーション制作,ライティング,物理シミュレーションなどの複雑なプロセスが含まれ,開発コストの増大につながっている.そのため,高品質な 3D ビジュアルを維持しながら,モデリングコストを削減し,生産効率を向上させることが,現在のゲーム業界における重要な課題となっている[63].

3D ゲーム市場の拡大にもかかわらず,3D キャラクター制作には依然として多くの課題が存在する.まず,モデリングプロセスが複雑であり,制作期間が長い.高品質な 3D キャラクターは,コンセプトデザイン,スカルプト,リトポロジー,UV 展開,テクスチャ作成,リギングなどの複数のステップを経る必要がある.経験豊富なモデラーであっても,1 体の詳細なキャラクターモデルの完成には 30～45 日を要する[64].

次に,人件費の上昇がゲーム開発企業にとって大きな負担となっている.特に欧米や日本では,3D モデラーの給与水準が高く,経験豊富なモデラーの年収は 10 万米ドルを超えることもある[65].これは,独立系ゲーム開発者や中小規模のゲーム会社にとって大きな経済的負担となっている.そのため,多くの企業はコスト削減のために 3D モデリング業務を人件費の安い国（例：中国,ベトナム,インド）に外注している[66].しかし,外注によって品質管理の困難さ,アートスタイルの統一の難しさ,コミュニケーションコストの増大といった問題が生じるため,一概にコスト削減の解決策とは言えない[67].

また,3D アート人材の需要が供給を大きく上回っている.3D モデリングを学ぶには,Maya,Blender,ZBrush,Substance Painter などの専門ソフトウェアの習得が必要であり,さらに高度な芸術的センスも求められるため,育成に時間がかかる[68].さらに,

ゲーム業界は長時間労働が常態化しており,頻繁な残業やバーンアウト（燃え尽き症候群）が発生しやすい.特に,3D アート職は過酷なスケジュールと高い要求に晒されることが多く,キャリアの継続が困難なケースも少なくない[69].これにより,業界全体の人材不足が加速し,人件費のさらなる上昇を招いている.

近年,人工知能技術を活用した 3D モデルの自動生成が進展している.一部の研究では,AI を活用することで,従来 30〜45 日かかっていたキャラクターモデル制作を 3 日以内に短縮できる可能性が示されている[70].これにより,開発コストの削減と生産性の向上が期待される.しかし,現在の AI ベースの 3D モデリング技術には,ディテールの不足,制御の難しさ,アートスタイルの統一の困難さといった課題が残っている[71].したがって,高品質な 3D キャラクターを維持しながら,AI を活用して制作の効率を向上させることが,今後のゲーム業界における重要な研究テーマとなる[72].

3.2 課題

近年，3D ゲーム市場の急速な拡大により，高品質なゲームキャラクターを効率的に制作するニーズがますます高まっている.しかし，従来の 3D キャラクターのモデリングプロセスは，専門のモデラーがキャラクターの 2D のイラストや立ち絵を基に，手動で作成する必要がある.このプロセスは膨大な時間を要し，ゲーム会社の元従業員によると，ゲームキャラクターの複雑さに応じて，3D モデルの制作には 1〜3 か月の時間が必要である.そして，3D モデリングには高度な専門スキルを必要とする，モデラーは，Maya や Blender, ZBrush といった 3D モデリングソフトの操作に精通しているだけでなく，解剖学やプロポーション，テクスチャマッピング，マテリアル設計といった分野にも深い理解を持つ必要がある.制作プロセス全体は，モデルの初期設計，メッシュの最適化，リギング(骨格のバインディング)，テクスチャの描画，アニメーション制作といった段階に分かれており，それぞれの段階で膨大な時間と労力を要するのである.このような状況において，新たな自動化技術の導入が求められている.

この課題に対応するため，近年の研究は画像から 3D モデルを生成する自動化手法に注力しており，拡散モデルの台頭により 3D モデル生成に新たな可能性が開かれた.拡散モデルは，テキストから 3D 生成や単一画像から 3D 生成の分野で顕著な可能性を示している.しかし，これらの手法は入力情報の制約により，生成されるモデルに幾何学的一貫性の欠如，細部の再現不足，および編集の困難さといった課題が残っている.一方，NeRF や 3D Gaussian Splatting のような 3D 再構築技術は，実在する物体の多視点再構築において優れた性能を発揮しているが，多数連続な多視点入力を必要とするため，ゲームキャラクターモデリングには適用が難しい.

3D モデル生成の分野では，本研究の要件を十分に満たす 3D モデルの品質を評価するためのシステムは存在していない.このため，生成モデルの品質を評価するための基準や方法が明確ではなく，新たな 3D モデルの評価システムの導入が求められている.

3.3 既存研究における課題に対する取り組み

近年，3D モデルに関する研究は主に 2 つの重要な分野，3D 再構築と 3D 生成に集中している.3D 再構築技術は，二次元の投影や画像を利用して物体の三次元情報を復元する技術である.NeRF および 3D Gaussian Splatting は，この分野における画期的な手法であり，複数の連続した入力画像を用いて 3D シーンを復元する 0[2].しかし，NeRF や 3D Gaussian Splatting は，数十枚の連続した異なる視点の画像を入力として必要とし，さらに現実に存在するシーンを対象とするため，ゲームキャラクターの 3D モデル制作には十分対応できないのである.

一方，CLIP モデルや拡散モデルの発展に伴い，さまざまな 3D 生成技術も顕著な成果を上げている.たとえば, DreamFusion はテキストから 3D を生成するモデルを提案し，wonder3D は単一の RGB 画像から視点合成と 3D 再構築を実現している[34].しかしながら, これらの手法は入力情報の制約により, 生成されるモデルに幾何学的一貫性の欠如，細部の再現不足，さらには編集の困難さといった課題が残されているのである.

3.4 目的

以上の背景に基づき，本研究の目標は，ゲームキャラクター制作に適した 3D モデル生成手法を開発することである.ゲームキャラクターの 3D モデル制作において，制作に時間がかかり，高度な技術を必要とするという課題を解決し，3D モデリングの時間短縮と技術的ハードルの低減を目指す.また，本研究では，現行の単視点手法が抱える幾何学情報の不足や細部再現の課題を克服し，3D モデルの生成品質を向上させることにも取り組む.さらに，新たな 3D モデル生成品質評価アルゴリズムを構築し，生成された 3D モデルに対する数値的評価基準を提供するとともに，生成 AI の結果評価に新しい視点を示すことを目指している.

本研究では，多視点入力を導入することで，拡散モデルを活用して入力されたイラストの 3D 情報を強化し，法線マップを生成する.その上で，ジオメトリ感知型の法線融合アルゴリズムを用いて，多視点の RGB 画像と法線マップを統合し，高品質な 3D モデルの生成を実現する.本研究は，ゲームキャラクター制作における効率的で信頼性の高い技術ソリューションを提供し, 高い忠実度と精度を求める業界ニーズに応えることを目的としている.

本研究の実現により、開発コストが大幅に削減され、生産効率が向上する。従来の 3D モデリングには多くの時間と人員が必要とされていたが、その負担が大幅に軽減され、ゲーム開発の期間が短縮される。特に、大規模なゲームプロジェクトにおいて、3D アセットの生成と反復作業の速度が大幅に向上する。これにより、ゲーム企業は美術制作にかかるコストを削減し、より多くのリソースをゲームの革新や最適化に投入することが可能となる。

次に、ゲーム美術の職種構造が変化する。AI による 3D モデル生成技術の発展に伴い、初級モデラーの需要は減少し、代わりに AI が生成したモデルを最適化し、美術スタイ

ルを調整できる上級アーティストの需要が高まる。ゲームアーティストの役割は、「手作業によるモデリング」から「AI との協働とアートディレクション」へと移行し、ゲーム企業はより総合的なスキルを持つ人材の育成を重視するようになる。

さらに、ゲームコンテンツの多様性と個別性が一層向上する。3D モデルの自動生成技術により、開発者は多種多様な美術スタイルのゲームアセットを迅速に作成できるようになり、ゲーム世界がより豊かで多彩なものとなる。特に、オープンワールドゲーム、サンドボックスゲーム、オンラインマルチプレイヤーゲームにおいて、動的に生成される 3D 環境やキャラクターは、プレイヤーの没入感とインタラクティブな体験を強化する。

同時に、ゲーム業界の市場競争構造にも影響を及ぼす。小規模なゲーム開発チームやインディー開発者は、低コストで高効率な 3D モデル生成技術を活用することで、より大きな市場競争力を獲得できる。一方、大手ゲーム企業は、大規模な自動化アセット生産を通じてコンテンツの拡充を加速させ、市場競争がさらに激化する可能性がある。

総じて、3D モデル生成技術はゲーム開発プロセスを根本的に変革し、生産効率の向上、コスト削減、人材構造の最適化、そしてゲームコンテンツの多様化を促進する。一方で、新たな管理課題や法的課題も生じ、ゲーム業界は新たな発展段階へと移行することになる。

第4章　提案手法

本章では，本研究で提案する手法について述べる.まず 4.1 で提案手法の概要について述べる.次に 4.2 では，3D 効果図生成の必要性，生成方法，および生成結果について紹介する.4.3 では，本研究で提案する法線生成アルゴリズムの生成方法について説明する.4.4 では，本研究で使用した 3D メッシュ合成アルゴリズムについて紹介する.

4.1　概要

キャラクター立ち絵から 3D モデルを作成するために，本研究では先行研究である Wonder3D を基盤とし，新しい法線生成アルゴリズムを構想し，生成された 3D モデルの品質を比較するための新しい生成結果評価方法を設計した.図 4.1 に示すように，生成結果の精度を向上させ，手描き画像が生成結果に与える影響を可能な限り減らすため，本研究ではまず，Stable Diffusion を使用して入力画像の 3D 効果を強化した.次に，事前学習済みの法線生成モデルを使用し，異なる視点のカメラ行列を追加することで，異なる視点に対応した入力画像から法線マップを生成するモデルを完成させた.最後に，稀疎な二次元画像に基づくメッシュ抽出方法を使用し，3D 効果を強化した画像と対応する法線マップを融合して 3D モデルを生成した.

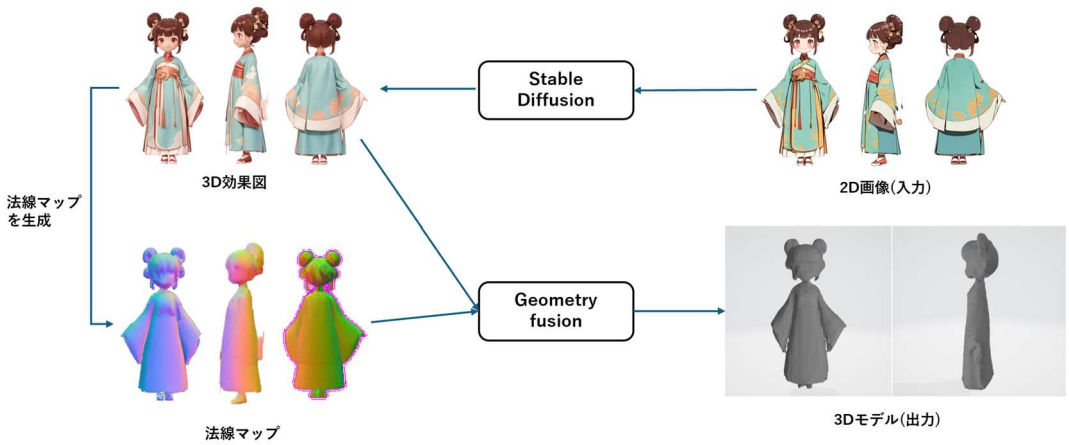

図 4.1 本研究のプロセス

4.2　3D 効果図の生成

本節では，3D 効果図の必要性を主に説明し，3D 効果図生成の主な手法について述べ，さらに 3D モデル生成への影響を分析する.

4.2.1 3D エンハンスの必要性

ゲームキャラクターデザインにおいて，キャラクターデザイナーは通常，芸術的なスタイルや視覚効果を優先するため，キャラクターイラストには三次元情報が不足していることが多い.ゲームキャラクターの立ち絵では，着色段階で基本的な塗りのみが行われることが多く，その結果，光や影の効果が欠如している場合がある.さらに，よりリアルな立ち絵であっても，光や影の効果は手描きによって制作されることが多く，芸術的な目的で簡略化や誇張が施される場合が一般的である.以下の図 4.2 図 4.3 に示すように，このような芸術的な処理により，イラストが現実的な物理光照の効果を正確に反映することが難しくなる.特に，手描きによる影やテクスチャは物理法則に基づいていないため，3D モデルを生成する際に幾何学的な情報や光照効果が欠落し，生成結果の精度や品質に悪影響を与える可能性がある.

この課題に対応するために，本研究では Stable Diffusion を導入し，画像の最適化を行うことで，キャラクターイラストにおける芸術的な光影効果が生成結果に与える負の影響を最小限に抑えることを目的としている.Stable Diffusion は，手描きイラストを高品質な三次元情報を含む画像へと変換する強力なツールであり，特に光影の補正や立体感の強化において優れた性能を発揮する.この処理を通じて，イラストの不足している三次元情報を補完し，生成結果の精度と一貫性を大幅に向上させることが期待される.

4.2.2 3D 効果図の生成方法

本研究において，3D 効果図とは，三次元情報が強化された画像を指す.図 4.2 図 4.3 に示すように，本研究では Stable Diffusion を用いて 3D 効果図を生成した.これらの 3D 効果図は，対応する 2D 画像と比較して，マテリアルや照明がより詳細であり，立体感と写実性が向上している.また，これらの画像には，より豊富な 3D 情報が含まれている.

参考とした先行研究である Wonder3D では，3D モデルを生成する際に，RGB 画像と法線マップを組み合わせて利用している.法線マップには生成する 3D モデルの三次元情報が含まれており，その生成は入力画像に依存している.

本研究では，3D 効果図を生成するために，Stable Diffusion に基づいた生成手法を採用した.具体的には，大規模な事前学習済み拡散モデルである SD1.5 および RevAnimated を使用した.Stable Diffusion 1.5 は，テキストから画像を生成する人気の高いモデルであり，一方，RevAnimated は Stable Diffusion を基盤に派生したモデル(いわゆる LoRA モデルまたは調整版モデル)である.RevAnimated は，Anything V3 や SD1.5 など複数のモデルを統合し，微調整・最適化を行うことで，アニメーション風の画像を生成する特徴を持つ.

SD1.5/RevAnimated は，特にアニメキャラクターの生成において広く利用されており，高品質で手描きのアニメキャラクターに近い画像を生成する能力を持つため，本研究におけるゲームキャラクターを対象としたニーズを満たしている.

また，生成効果をさらに向上させるために，本研究では BlindBox モデルを導入した[55].

BlindBox モデルとは，LoRA(Low-Rank Adaptation)[56]に基づく訓練方法であり，ブラインドボックス風の画像を生成するために設計されたモデルである.この LoRA モデルによって生成された画像は，入力画像の形状を良好に保持し，生成された画像は入力画像の輪郭と高い一致性を示す.また，合理的かつ高品質に 3D 情報を推測し，2D エフェクト画像を 3D 効果図へと変換する能力を持つ.本研究では，BlindBox モデルが複雑なテクスチャや細部の処理において優れた性能を発揮する点を利用し, 生成画像の精密さを強化した.

さらに，2D エフェクト画像から 3D 効果図への変換は，画像から画像を生成する(image-to-image)プロセスであるため, 画像生成過程を精密に制御する目的で, ControlNet コントローラーを導入した.ControlNet は，Stable Diffusion を基盤とした画像生成において形状, ポーズ, 構造などを制御可能にする拡張技術である.ControlNet は, エッジ検出, 深度マップ，ポーズ，スケッチなどの外部条件を導入することで，生成内容に対する精密な制御を可能にし，結果が期待により忠実に応えるようにする.

ControlNet コントローラーのこれらの利点を活用するため，本研究では，ControlNet の使用時に MiDaS による深度情報推定をプリプロセッサとして選択し，モデルとして control_depth を選択した.その結果得られた 3D 効果図は，以下の図 12 に示されている.

また，3D 効果図を生成する過程において，本研究では通常，まずタグ付けツールを使用して入力画像の基本的なタグを取得し，その後，現在の画像に対する詳細な説明を追加する.以下の図 4.2 に示すように，画像内の女の子に対する正向プロンプトは次の通りである：

A cute chibi-style girl with blonde hair holding a stuffed bunny， wearing white pajamas. The character has big blue eyes， a soft smile， and an adorable cartoon-like appearance. The illustration is clean and minimal， with a pastel color palette， set against a transparent or white background. Designed in a kawaii art style.

さらに，Stable Diffusion を使用して画像を生成する際，反向プロンプトも欠かせない要素である.ただし，反向プロンプトは一般的に汎用性が高く，大きな変更を加える必要はない.本研究で使用した反向プロンプトは以下の通りである：

Low quality， bad anatomy， extra fingers， missing fingers， deformed hands， bad proportions，blurry，out of focus，poorly drawn，ugly，distorted，mutated，bad posture，low resolution， JPEG artifacts， overexposed， grainy， oversaturated， unrealistic，overprocessed， extra arms， extra legs， extra eyes， wrong colors， bad perspective， bad lighting， overly dark， too bright， crowded， cluttered background， overly detailed background， busy background， text， logo， watermark， creepy， scary， harsh lighting，sharp shadows， monochrome， horror elements.

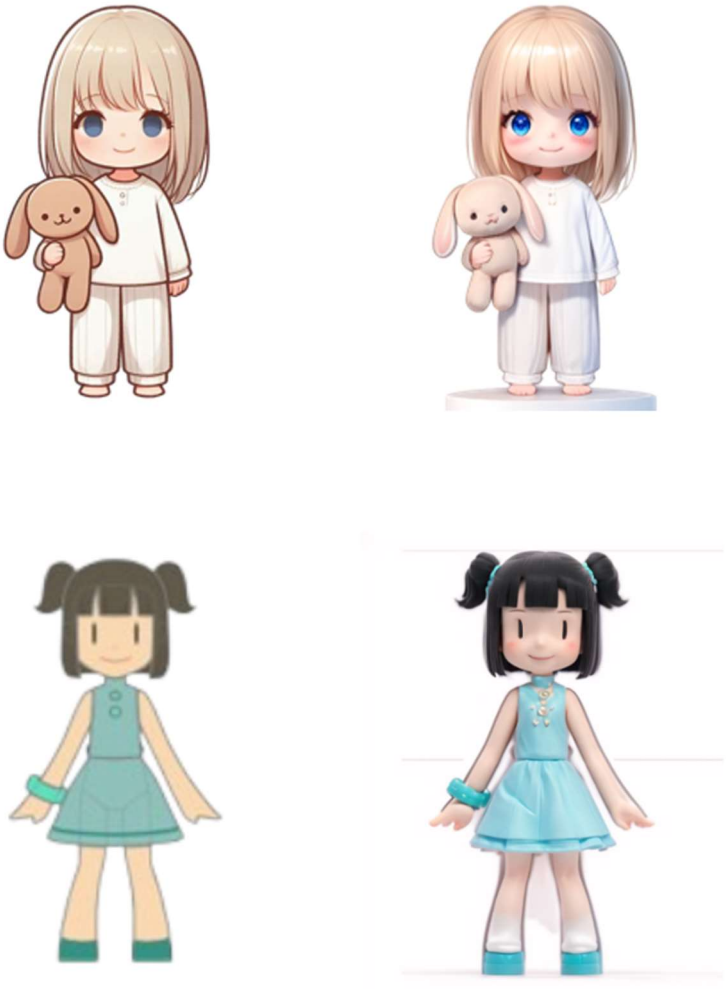

図 4.2 2D 効果図と 3D 効果図の比較

図 4.3 立ち絵の 2D 効果図と 3D 効果図比較

4.2.3 3D 効果図の生成結果

本節では，2D 効果図と 3D 効果図をそれぞれ入力画像として使用した場合に生成結果へ与える影響を示す.

本研究では，Stable Diffusion を使用して複数の 2D 効果図を生成し，それらを入力として使用するとともに，手描きによるキャラクター立ち絵を別の入力として用いた.これらの画像を基に，先述の生成手法を利用して対応する 3D 効果図を生成した.最終的に，2D 効果図と 3D 効果図をそれぞれ先行研究で提案された 3D モデル生成器に入力し，その違いを比較した.

図 4.4 は，元の 2D 画像と 3D 強化後の画像をそれぞれ入力として使用して生成された 3D モデルである.本研究では，生成されたモデルの正面，左右，背面の視点からスクリーンショットを撮影し，結果を観察した.その結果，3D 強化後の画像を入力として使用した場合，細部の再現性が向上していることが確認できた.例えば，以下の図では，肉眼では大きな違いがないように見えるが，キャラクターのスカートのラインがより滑らかになっている点が確認できる.また，キャラクターの足部に余計な 3D 構造がなく，ラインがより滑らかであることが確認できる.

図 4.4 2D 効果図と 3D 効果図の生成結果比較

4.3 法線生成アルゴリズム

本研究は，単一画像の法線生成方法を開発することを目的としており，これは先行研究である Wonder3D の改良として位置付けられる.これにより，ゲーム開発におけるキャラクター3D モデル制作により適した手法を提案するものである.Wonder3D の動作原理は，単一の画像を入力として，Zero123 の手法を利用して 6 つの異なる視点画像(正面，左側，右側，左前，右前，背面)を生成し，さらにクロスドメイン(cross domain)技術を用いて，それぞれの視点に対応する法線図を同時に生成する.本研究はキャラクターの多角度立ち絵から 3D モデルを生成することを目指しており, Wonder3D と比較して，入力データを単一画像からキャラクター立ち絵に変更し, 単一画像から多角度画像を生

成するステップを省略している.このため，各画像ごとに対応する法線図を生成するだけで十分であり，生成時間を短縮し，計算負荷を削減し，過剰な計算による誤差を回避することが可能となる.

単一画像の法線生成アルゴリズムは，高効率な画像生成手法であり，単一の 2D 画像から標準的な RGB 形式の法線マップを生成することに特化している.その核心的な動作原理は以下の 4 つのステップに分けられる.まずは入力前処理の段階である.このアルゴリズムは，単一の 2D 画像を入力として受け取り，一連の前処理を行う.具体的には，設定ファイル内の img_wh パラメータに基づいて画像サイズを調整する.また，入力画像に Alpha チャンネルが含まれている場合, 透明な背景を置き換えることで背景の一貫性を確保する.最終的に，処理済みの画像をテンソル形式に変換し，後続のモデル計算に適した規格化された入力を提供する.

次に条件埋め込みの段階である.生成プロセスを制御し，出力がタスクに関連することを確実にするために，アルゴリズムは条件埋め込みベクトルを定義する.このベクトルには主にタスクの種類とカメラ姿勢のエンコード情報が含まれる.タスク種類の埋め込み(例 : [1,　0])は，法線マップ生成の目的を明示的に指定し，カメラ姿勢(例 : front, right など)は視点情報を提供し，多視点生成タスクに利用される.条件埋め込みと入力画像を融合することで, アルゴリズムはより精度の高い法線マップを生成することが可能となる.

第 3 段階は拡散モデルを用いた逆生成プロセスである.これはアルゴリズムの核心となるステップであり，ノイズの初期化と段階的なノイズ除去が含まれる.まず，潜在ベクトルにガウスノイズを加え，これを標準正規分布に近づける.次に，アルゴリズムは拡散モデルを採用し, 段階的にノイズを除去して条件に適合する潜在法線表現を生成する.生成プロセスを高速化するために，簡易化されたアルゴリズムでは DDIM サンプリング技術を使用し, 従来の拡散モデルの生成ステップ数を 1000 ステップから 20〜50 ステップに削減することで，計算効率を大幅に向上させている.

最後のステップでは法線マップを出力する.ノイズ除去が完了した後，潜在法線表現はデコーダーによって RGB 形式の法線マップに復元される.この復元された画像は，異なる法線方向を色で表現している.たとえば，右側視点の法線は赤，上方向の視点は緑として示される.

4.3.1　法線マップ

法線の定義

法線(Normal Vector)は，曲面や平面のある点において垂直なベクトルを指し，その点の局所的な方向特性を記述するために使用される.法線はコンピュータグラフィックス，物理シミュレーション, 工学計算などの分野で重要な幾何学的意味を持ち, 表面の照明計算, 幾何モデリング，三次元レンダリングにおける重要なツールである[57].

法線マップ

法線マップ(Normal Map)は，三次元表面の法線情報を二次元テクスチャに格納する技術であり，レンダリングにおいて 3D モデルの表面の詳細やリアリズムを強化するために使用される.この技術により，モデルの幾何学的な複雑さを大幅に増加させることなく，光照計算時の法線方向を変更し，表面により豊かな凹凸感を持たせることができる.これにより，モデルのポリゴン数を抑えたまま，視覚的なクオリティを向上させることが可能となる.

法線マップは二次元のテクスチャであり，RGB カラーチャンネルを用いて三次元空間での法線の 3 つの成分(x， y， z)をそれぞれエンコードしている.法線マップを使用することで，モデルの各ピクセルに独立した法線方向を提供し，より精細な表面のディテール表現を実現することができる[57].下記図 4.5 は 3D オブジェクトと法線マップの一例である.

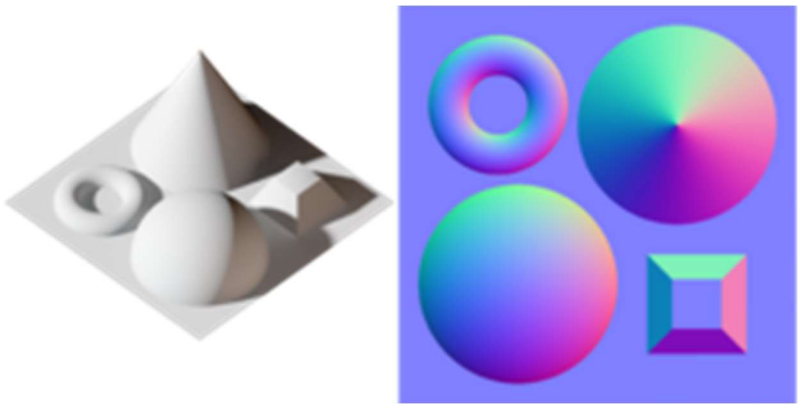

図 4.5 3D オブジェクトと法線マップ

法線方向のエンコード：

法線マップは三次元の法線ベクトルを二次元のテクスチャにマッピングし，以下の方法で表現されるのが一般的である：

R チャンネルは法線ベクトルの x 成分を表す.

G チャンネルは法線ベクトルの y 成分を表す.

B チャンネルは法線ベクトルの z 成分を表す.

通常，法線ベクトルの値域は [−1，1]であるため，これを正規化して [0，255]のピクセル値の範囲にマッピングする必要がある.

レンダリングでの応用：

レンダリングプロセスでは，法線マップが各ピクセルの表面法線を提供し，ジオメトリ計算によって得られるモデル法線の代わりに使用される.光照モデル(例えば，Phong シェーディングや Blinn-Phong シェーディング)は，これらの法線を基に光線の反射や屈折効果を再計算し，より豊かな表面ディテールを表現する.

高効率なディテール強化の実現：
法線マップはジオメトリモデルとは独立しており, 低い計算コストで高いディテールを持つレンダリングを提供できる.この特性は，特にリアルタイムレンダリング(ゲームエンジンにおけるリアルタイム光照計算など)に適している.

図 4.6 は，RGB 画像と法線マップの比較を示している.この法線マップは Blender によって作成されたものである.

図 4.6 RGB 画像と法線マップ

4.3.2 法線マップ生成

まず，データ読み込み段階において，single_image_dataset.py ファイルを調整し，入力視

点を front に固定した上で，画像パスから正面視点の画像データを読み込むように設定した.また，多視点拡張処理の簡略化のため，正面視点に対応するカメラ姿勢行列のみを保持し，多視点データ拡張に関連する操作を削除した.条件埋め込み段階では，特定のタスク埋め込みベクトル(例：[1，　0])を設定し，法線マップ生成タスクを表現した.また，カメラ姿勢の埋め込みを現在の視点情報に固定し，モデルが法線マップ生成タスクに専念できるよう指導した.

次に，推論段階では，test_mvdiffusion_seq.py ファイルを修正し，モデルが法線マップ関連のタスクのみを処理するようにした.また，生成フロー内の条件埋め込みと推論呼び出しパラメータを調整し，モデルの出力を単視点の法線マップ生成に集中させた.さらに，計算効率を向上させるために，設定ファイルを調整し，単視点タスクのみを有効化(例：num_views=1 および view_types=['front'])し，タスクタイプを normal_only に設定した.

モデルのデコード部分では，pipeline_mvdiffusion_image.py ファイルのデコーダーロジックを最適化し，法線マップのデコード出力のみを保持し，RGB 画像の生成を削除した.最後に，結果保存のロジックを簡素化し，生成された法線マップファイルのみを保存し，その他のデータを無視するようにした.

4.3.3　カメラ行列

単一の法線マップの出力を実現するためには，ソースファイル内で出力画像の数を削減するだけでなく，新しいカメラ姿勢行列を追加する必要がある.それを実現するために，先ずは法線の視点変換を行う.

法線の視点変換とは，法線方向を線形変換するプロセスであり，その目的は，法線をある基準視点(例えば正面視点)から別の基準視点(例えば左側視点，右側視点，または背面視点)にマッピングすることである.この変換は通常，回転行列を用いて計算される.回転行列は，目的の視点が初期視点に対してどのように座標系が変化するかを記述するものである.

三次元空間において，法線ベクトルは通常 [x，y，z]の形式で表現される.初期視点が正面視点である場合，その局所座標系の [x，y，z]はそれぞれ画像の水平軸，垂直軸，深度軸に対応する.法線を左側視点に変換する際の核心は，Y 軸を中心に 90°回転させることである.この回転は以下のような回転行列で表される：

$$R_{left}=\begin{bmatrix}\cos(90°) & 0 & -\sin(90°)\\ 0 & 1 & 0\\ \sin(90°) & 0 & \cos(90°)\end{bmatrix} \quad (3.1)$$

ここで$\cos(90°)=0$と$sin(90°)=1$を行列に代入すると，以下の行列になる.

$$R_{left}=\begin{bmatrix}0 & 0 & -1\\ 0 & 1 & 0\\ 1 & 0 & 0\end{bmatrix} \quad (3.2)$$

この行列は，法線の z 成分を新しい x 方向にマッピングし，x 成分を新しい z 方向にマッピングし，y 成分はそのまま維持する.法線ベクトルにこの回転行列を掛けることで，視点変換が完了する.

この法線視点変換は，コンピュータグラフィックスや三次元再構築において広く応用されている.例えば，複数の視点から合成して単一視点の法線マップを生成する際，視点変換を利用して異なる視点の法線表現を統一し，後続の処理の精度を向上させる.また，このプロセスは 3D レンダリングにおいて，物体の照明効果を動的に調整するためにも使用される.精密な行列演算を通じて，法線の方向が常に目標視点と一致することが保証されるため，物体表面の特性を正確に表現することが可能になる.

また，本研究で使用したその他の視点のカメラ行列は以下の通りである：

右側視点：

$$R_{right}=\begin{bmatrix}0 & 0 & 1\\ 0 & 1 & 0\\ -1 & 0 & 0\end{bmatrix} \quad (3.3)$$

后方視点：

$$R_{back}=\begin{bmatrix}-1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & -1\end{bmatrix} \quad (3.4)$$

左前方 45° 視点：

$$R_{front\_left}=\begin{bmatrix}\frac{\sqrt{2}}{2} & 0 & -\frac{\sqrt{2}}{2}\\ 0 & 1 & 0\\ \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2}\end{bmatrix} \quad (3.5)$$

右前方 45° 視点：

$$R_{front\_right}=\begin{bmatrix}\frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2}\\ 0 & 1 & 0\\ -\frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2}\end{bmatrix} \quad (3.6)$$

先行研究である Wonder3D では，著者がクロスドメイン手法を使用したため，出力され

る法線マップは統一された 6 つの視点となり，それぞれの視点に対応するカメラ姿勢行列が設定されている.ソースコードの出力結果を単一視点に変更した場合，出力される法線マップの視点が入力画像の視点と一致しない問題が発生する可能性がある.例えば，左側視点の画像を入力した場合でも，生成された法線マップの色調が正面を基準としたものになるといった問題である.このような問題を回避するために，本研究では，上述の視点行列をコードに追加する必要がある.

図 4.7 カメラ行列を入れた後の法線マップ

図 4.7 は，Wonder3D の法線生成結果と本研究で提案する簡略化された法線生成アルゴリズムの結果を示している.本研究の簡略化された法線生成アルゴリズムでは，クロスドメイン属性を削除し，アルゴリズムが入力画像に対応する単一の法線マップのみを生成できるようにした.しかし，Wonder3D の設定上，出力される法線マップはすべて正面視点に統一されているため，本研究では視点変換のためのカメラ行列を追加し，正面視点の法線マップを左側や右側など他の視点に変換できるようにした.上記の図から，修正後の結果の違いが明確に確認できる.

4.4 メッシュ抽出

生成された法線マップと入力 RGB 画像から 3D モデルを構築し，スパースな視点による予測の不正確さを軽減するために，本研究では Wonder3D 論文で提案されたニューラルインプリシットサインドディスタンスフィールド(SDF)に基づく最適化手法を採用する.本研究では，ジオメトリ対応の法線損失関数を設計しており，その式は以下のように表さ

れる：

$$L = L_{normal} + L_{rgb} + L_{mask} + R_{eik} + R_{sparse} + R_{smoot} \quad (3.7)$$

L_{normal}は法線損失を指し，L_{rgb}は MSE 損失項を示し，L_{mask}はバイナリクロスエントロピー損失を表す.また，R_{eik}はエイコナル正則化項，R_{sparse}はスパース性正則化項，R_{smoot} は3D平滑性正則化項をそれぞれ意味する.この手法は,SDFを最適化することで,法線マップやカラ―画像などの生成された2Dデータを統合し，高品質な明示的3Dモデルを生成することを目的としている[35].

4.5 本研究の新規性

本研究の新規性は以下の三点である.

第一に,3D効果図生成という新たな手法を提案した.従来の単一視点からの3Dモデル生成では,奥行き情報の欠如や幾何学的な一貫性の低さが課題とされていた.本研究では,拡散モデルを活用して入力画像に3D情報を付加することで,より安定した形状の3Dモデルを生成することを可能にした.この手法により,特にキャラクターの輪郭や衣装の細部表現が向上し,視点変更時の形状の破綻を軽減できる.

第二に,先行研究の生成アルゴリズムを修正し,本研究の目的に適した形へと最適化した.先行研究では,法線生成アルゴリズムが12枚のRGB画像と法線マップを生成しており,不必要な計算が発生していた.本研究では,cross domainアルゴリズムを削除し,さらにカメラの姿勢行列を導入することで,法線生成アルゴリズムが対象の視点に対応した1枚の法線マップのみを出力するように最適化した.

第三に,新しい3Dモデル評価システムを提案し,生成されたモデルの品質を定量的に評価する手法を確立した.従来の評価方法では,視覚的な主観評価に依存することが多く,客観的な指標が不足していた.本研究では,生成された3Dモデルの複数の視点のスクリーンショットを取得し,それらを原画像と比較することで,不一致率を数値化する評価手法を開発した.これにより,生成 AI の出力品質を定量的に評価し,より精度の高いモデル生成技術の改良につなげることが可能となった.

第5章 結果および評価手法

本実験で提案した手法の有効性を検証するため，実験を実施し，本章ではその結果を示し，説明するとともに，実験結果に対する考察を述べる．

まず，5.1 では実験に使用したハードウェア環境およびソフトウェア環境について説明し，5.2 では実験データの準備およびデータの前処理について述べる．5.3 では，本研究で提案した手法による生成結果を示し，5.4 **評価手法**では，新たに提案した生成結果の評価手法について説明する．最後に，5.5 で検証プロセスを紹介する．

5.1 実験環境

実験環境では，コード実行時の環境について説明する．

ハードウェア環境

本実験は Linux システム上で実行され，Linux のバージョンは Ubuntu 22.04 である．使用した GPU は nVidia A5000 で，搭載されたビデオメモリ(VRAM)は 24GB である．

開発環境

本研究では，Python 3.8 を使用して開発を行い，PyTorch の深層学習フレームワークを導入した．また，torchvision，OpenCV などの複数のサードパーティライブラリも使用している．その他のサードパーティライブラリに関する情報は付録を参照できる．さらに，CUDA は本研究において必須のツールであり，CUDA 11.7 バージョンでの実行を推奨する．

5.2 データ

本節では，実験データの取得方法について説明し，一部のデータを示した後，データの前処理手順について解説し，関連する画像を提示する．

5.2.1 データの準備

本研究の目的はキャラクター立ち絵から 3D モデルを再現することであるが，著作権問題などの制約により，実験データを容易に取得することは困難である．この問題を解決するために，筆者はキャラクター立ち絵の基準に基づき自ら手描きで立ち絵を作成

し，これを入力データとして使用したものである.また，Stable Diffusion を活用し，一貫性を持ったキャラクター立ち絵を生成し，これを実験に利用したものである.さらに，著作権フリーのキャラクター立ち絵も積極的に収集したものである.以下の**図 5.1 キャラクター**は，インターネット上で公開されている著作権フリーのキャラクター立ち絵である.次に示す図 5.2 と図 5.3 は Stable Diffusion を使用して生成したキャラクター立ち絵である.これらの画像はいずれも 2D 効果図であり，実験を実施する前に 3D 情報の強化が必要である.その具体的な手順については，3D 効果図の生成方法で述べる.

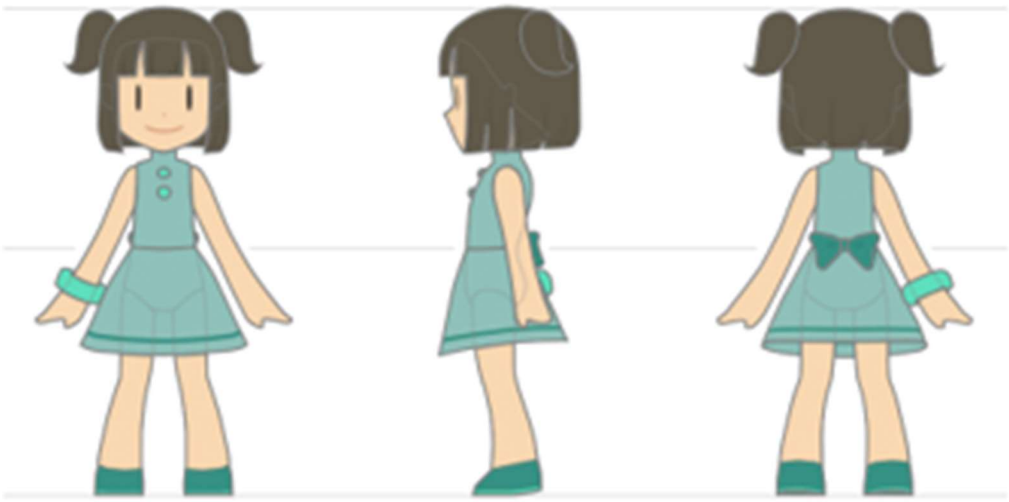

図 5.1 キャラクター立ち絵三面図

図 5.2 Stable Diffusion で生成されたキャラクター立ち絵1

図 5.3 Stable Diffusion で生成されたキャラクター立ち絵 2

上記の画像に加え，実験結果をより正確に評価するために，本研究では追加の入力画像データも準備したものである.これらのデータは，モデリングソフトウェアで制作した 3D モデルからスクリーンショットとして取得したものである.本研究では 4 枚, 6 枚, 8 枚の画像入力をサポートしているが，Stable Diffusion を使用して一貫性のある異なる視点画像を生成する場合，その品質が安定しないことが多く，筆者の要件を十分に満たすことが難しい.このため，モデルの固定視点から画像をキャプチャする方法が非常に有効である.以下の図 5.4 は，サンタクロースのモデルを正面，左側，右側，左前方 45 度，右前方 45 度，後方の 6 つの角度からキャプチャした画像を示している.これらの画像は，元々正確かつ十分な 3D 情報を含んでいるため，3D 効果の強化は必要としない.

図 5.4 サンタクロース六面図

5.2.2 データの前処理

コードのいくつかの制約により，入力画像を使用する前に，画像に対していくつかの前処理を行う必要がある.

背景置換

本実験では，入力画像の枚数を 6 枚に設定し， 提案手法で法線マップの生成に成功した後，メッシュ抽出アルゴリズムの入力として必要とされる RGB 画像および法線マップの白色背景版と透明背景版を準備するため，まず画像の背景を置換する必要がある.

本研究では，Python と OpenCV ライブラリを使用したプログラミングベースの方法を採用した.その核心原理は，色分割技術を用いて画像内の背景領域を認識し，これを白色背景に置き換えることである.まず，画像を読み込み，HSV(色相，彩度，明度)色空間に変換する.この色空間は，色の範囲を定義するのに適している.その後，元の背景色の特性に基づいて上下限の閾値を設定し，cv2.inRange メソッドを用いて二値化マスク(Mask)を生成し，背景と前景を区別する.次に，純白の背景画像を作成し，マスクと元の画像を組み合わせてビット演算を実行することで，背景部分を白色に置換し，前景部分はそのまま保持する.最後に，処理された画像を新しいファイルとして保存する.

また，画像の背景を透明に置き換えるために，本研究では引き続き Python と OpenCV ライブラリを使用して画像を加工した.この手法の核心原理は，色分割技術を用いて背景領域を認識し，背景ピクセルの透明度(アルファチャネル)を完全に透明に設定することである.まず，入力画像を読み込み，アルファチャネルの操作を可能にするために RGBA 形式に変換する.その後，HSV 色空間を使用して画像内の背景色を分離し，上下限の閾値を設定して二値化マスクを生成し，前景と背景を分離する.次に，アルファチャネルを操作し，マスクで認識された背景ピクセルのアルファ値を 0(完全に透明)に設定し，前景領域のアルファ値はそのまま保持する.最後に，処理後の画像を透明背景をサポートする PNG 形式で保存する，結果は図 5.5 に参考.

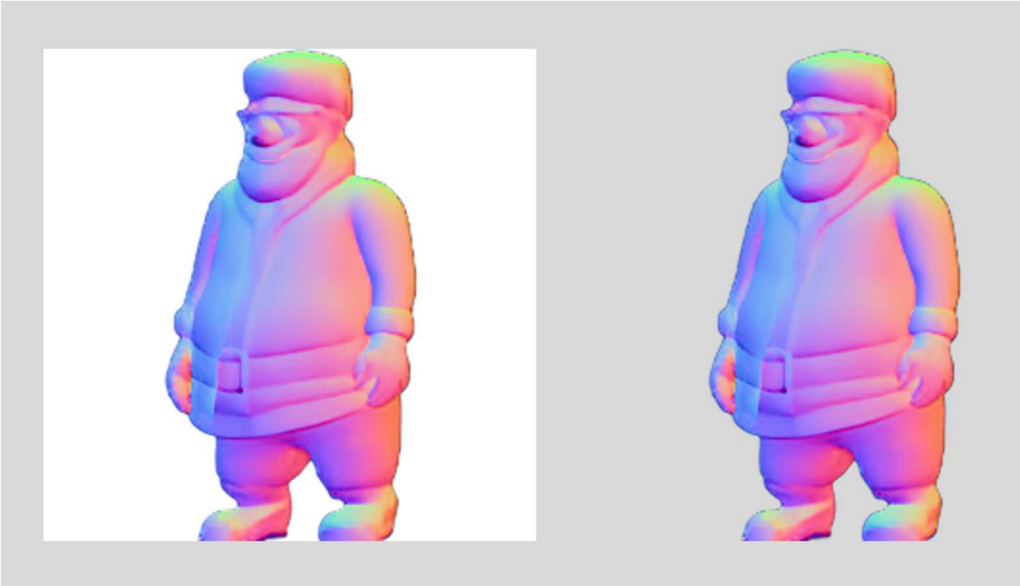

図 5.5 背景置き換え操作

サイズの統一

以下の図 5.7 に示すように，入力データを準備する際，生成された画像や手描きされた画像が複数の異なる視点でキャラクターの一貫性を満たしていても，入力画像をスクリーンショットで分割する過程で，図 5.7 のように画像サイズが統一されていない問題が発生する可能性がある.このため，これらの画像を使用する前に，サイズを統一する必要がある.

準備された画像はキャラクター立ち絵であり，視点を回転させた場合でも横方向のサイズが変化する一方，縦方向のサイズは基本的に一定である.このため，本研究では縦方向の高さを基準としてサイズを統一する.具体的な手順は以下の通りである：

まず，キャラクター全体を切り抜く.この操作には ImageMagick の convert コマンドを使用する.ImageMagick は，非常に強力でオープンソースの画像処理ツールキットであり，形式変換，画像編集，特殊効果の追加など，さまざまな画像処理タスクで広く利用されている.そのクロスプラットフォーム対応と柔軟性により，学術研究や産業用途で幅広く活用されている.ImageMagick は豊富なコマンドラインツールを提供しており，中でも convert コマンドは最も使用されるツールの一つで，主に画像形式の変換や基本的な画像編集操作に使用される.このコマンドは，JPEG，PNG，GIF，TIFF など多くの画像形式をサポートし，トリミング，サイズ変更，回転，フィルター追加などの多様な機能を提供する.

切り抜きの際には以下のコマンドを使用する：

convert before.jpg -trim after.jpg

-trim は ImageMagick のオプションであり，画像中の余白部分を自動的にトリミングするために使用される.具体的には，背景色と一致する画像周囲の領域を削除し，背景以外の内容を保持する.このトリミング後の画像例を図 5.6 に示す.

図 5.6 トリミング操作

次に，縦方向のサイズを基準として画像を補完し，256×256 の正方形に整える.以下のコマンドを使用する：

convert input.png -resize x256 -gravity center -background white -extent 256x256 output.png

このコマンドの役割は，入力画像(input.png)を処理し，サイズの調整，中央揃え，背景の拡張などの操作を通じて, 固定サイズの 256×256 の正方形画像に変換し, output.png ファイルとして出力することである.

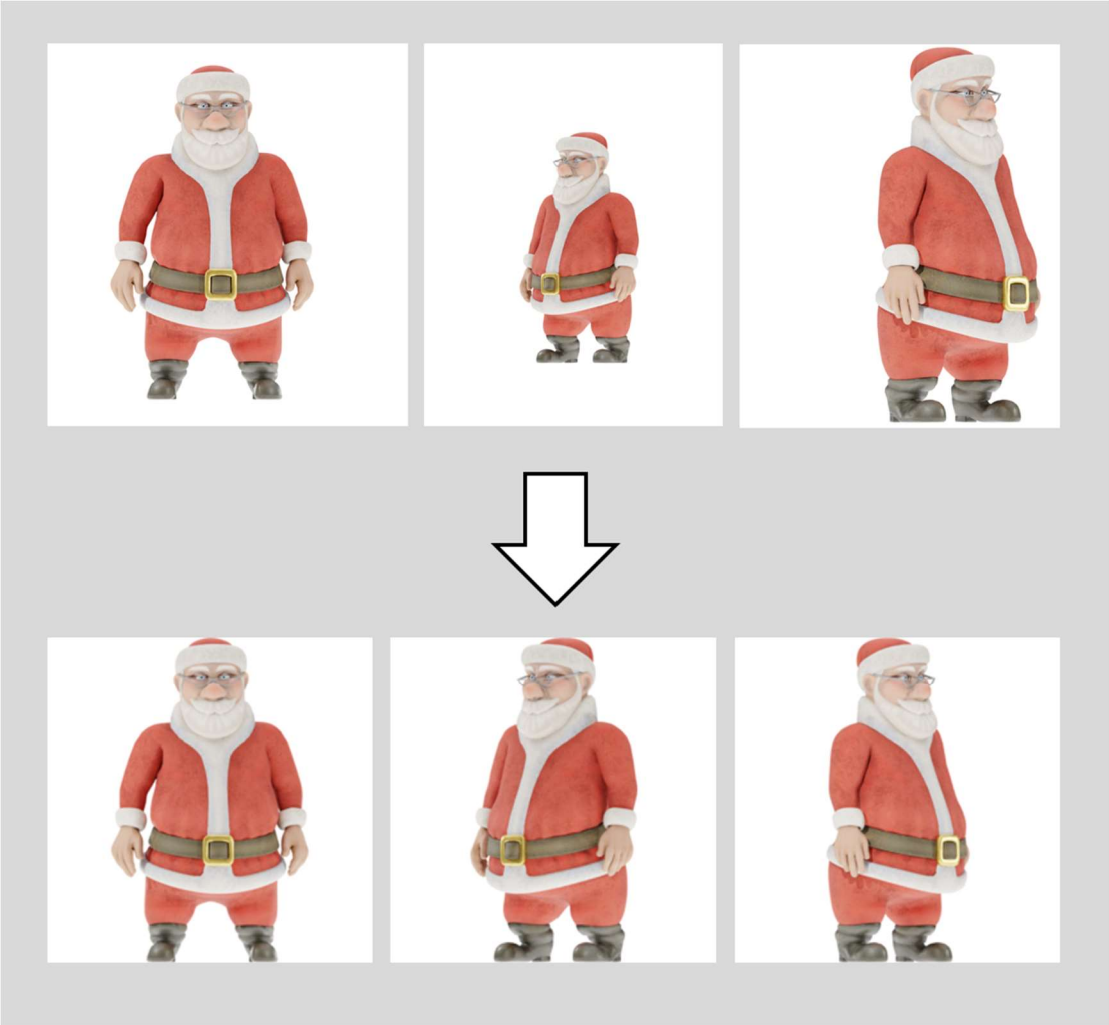

図 5.7 トリミングとサイズ調整操作

5.3 生成結果

本章では，本研究による 3D モデル生成の結果を示す.出力ファイルは 3D データであるため, 結果を分かりやすく視覚化するために, 生成されたモデルの正面, 左前, 左側, 右前の方向からスクリーンショットを撮影して提示する.

本研究の生成結果は入力画像の 3D 構造をより忠実に再現し，入力画像との一致性が高く，ディテールの表現においても従来研究より優れている，結果は図 5.8 に示す.

本研究では,サンタクロース,ウサギ,小さな女の子の 3 組のデータを用いて生成結果を示した。図 5.8 では,入力データとしてサンタクロース,ウサギ,小さな女の子の 6 つの視点から取得した RGB 画像を使用し,出力結果としてはメッシュ形式の 3D モデルを生成した。論文での視覚的な比較を容易にするために,本研究では正面,左側,左前方 45°,お

よび右前方 45°の視点からスクリーンショットを取得した。これらのスクリーンショットから,本研究の生成結果は入力データ全体の再現性が高いことが確認できる。生成された 3D モデルはいずれも形状が明確に示され,さらにキャラクターの体型も入力画像と高い一致性を保っている。たとえば,サンタクロースの大きな腹部,ウサギの体の長さ,小さな女の子の体型などが挙げられる。また,生成されたデータは細部の再現度にも優れており,サンタクロースのヒゲ,鼻,襟元,ウサギの尻尾や目,小さな女の子のアホ毛や髪の先端など,細かい特徴が忠実に表現されている。

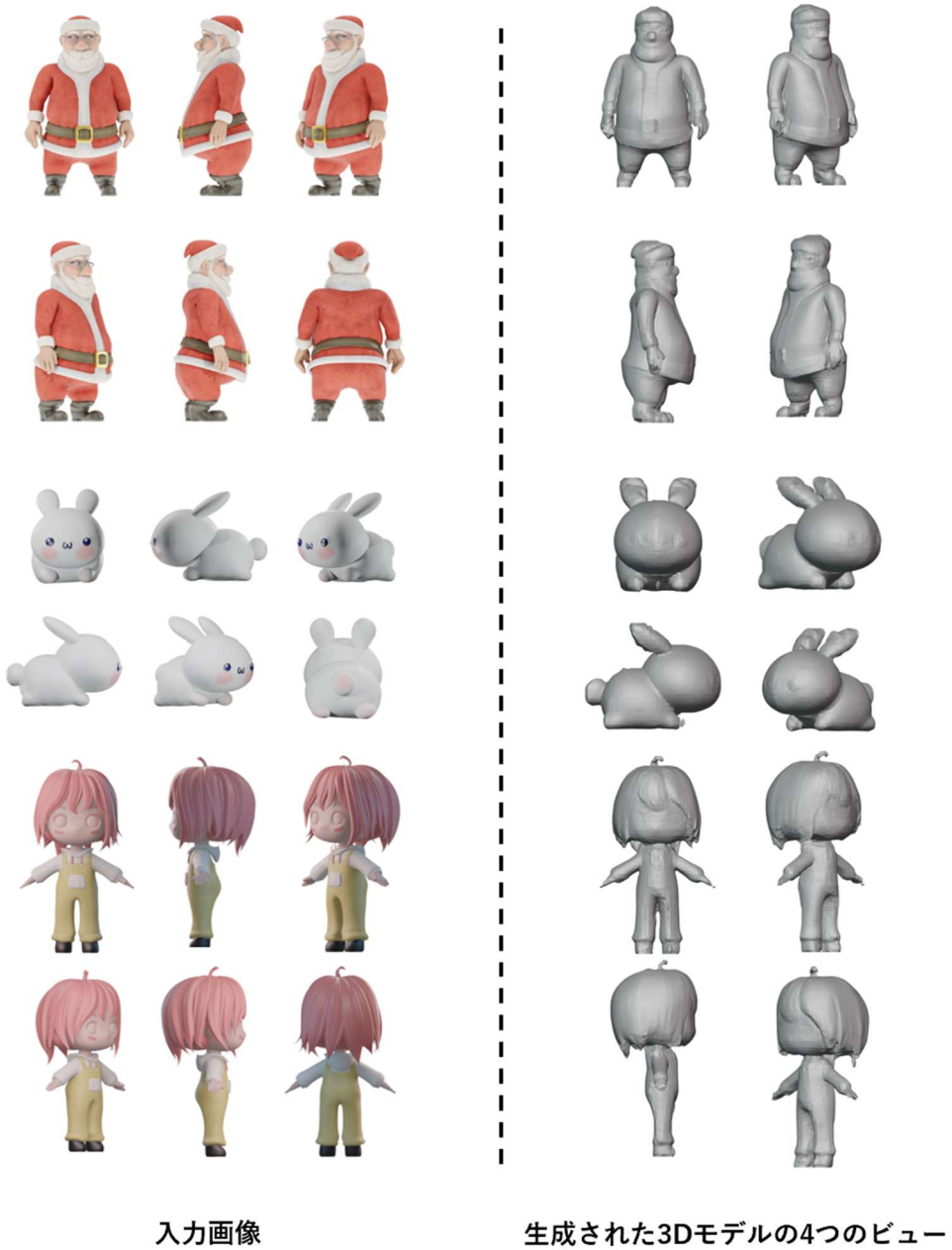

図 5.8 入力画像と生成結果のスクリーンショット

5.4　評価手法

本研究では，新しい 3D モデル生成品質の評価方法を設計し，生成モデルの性能を定量的に評価する手法を提案した.具体的には，生成された 3D モデルの多視点(正面，側面，左前方 45 度，右前方 45 度など)のシルエット画像を取得し，それらを元の入力キャラクター立ち絵のシルエット画像と比較することで，生成モデルの品質を評価する.この評価の客観性と一貫性を確保するため，シルエット画像は二値化され，背景を白，物体部分を黒に設定し，統一された基準を生成した.

その後, 生成されたモデルのシルエット画像と入力画像のシルエット画像をピクセル単位で比較するために, 排他的論理和(XOR)演算を使用した.この結果，不一致ピクセルの割合を計算し，それを評価指標とした.この不一致率の値は，生成モデルと入力画像との幾何学的一致性や細部の再現度を直感的に示している.不一致率が低いほど，生成された 3D モデルが元の立ち絵の形状特徴に近く，再現性が高いことを意味する.

この評価方法は, 本研究で提案した生成方法の科学的な評価ツールとしてだけでなく, 将来の関連研究における 3D モデル生成品質の評価の参考基準としても役立つものである.この方法を通じて，本研究で提案する方法が生成品質において優れていることを実証し，ゲームキャラクター制作における実用的価値をさらに強調した.

5.4.1　二値化

二値化(Binarization) は, 画像処理技術の一つであり, グレースケール画像やカラー画像を，0 と 1 または黒と白の 2 つの画素値だけを持つ二値画像に変換する方法である.この処理により，画像内の画素を 2 つのクラス(対象物＝前景と背景)に分類することが可能となる.

グレースケール画像の場合：

画像内の各画素値は，0 から 255 の範囲を持つ(8 ビットのグレースケール画像).

二値化の目的は，画素値を 2 つの離散的なクラスに分けることである.

値が 1(白)：対象物(前景)を表す.

値が 0(黒)：背景を表す.

閾値分割(Thresholding)：

二値化の核心は，閾値(Threshold)を設定して画素値を分類することである：

$$I'(x,\ y) = \begin{cases} 1, & I(x,\ y) > T \\ 0, & I(x,\ y) \le T \end{cases} \tag{5.1}$$

ここで：

$I(x,\ y)$は元の画像の画素位置 (x, y)におけるグレースケール値.T は二値化の閾値.

$I'(x,\ y)$は二値化後の画像の画素値.

カラー画像の場合：

カラー画像を二値化する際には，通常，最初にカラー画像をグレースケール画像に変換し，その後グレースケール値に基づいて二値化処理を行う.

二値化は，物体の検出や画像の前処理として幅広く使用される基本的な技術であり，そのシンプルさと計算効率の高さから，多くの画像処理アプリケーションで利用されている.二値化操作を行った画像は図 5.9 に示す.

図 5.9 では，左側の 3 枚の画像は上から順に，左前方 45° 視点の入力画像，本研究が生成した 3D モデルの左前方 45° 視点のスクリーンショット，そして先行研究である Wonder3D が生成した 3D モデルの左前方 45° 視点のスクリーンショットを示している。右側には，それぞれ対応する二値化後の画像が示されている。二値化後の画像は，次の工程を進めるために生成されたものである。

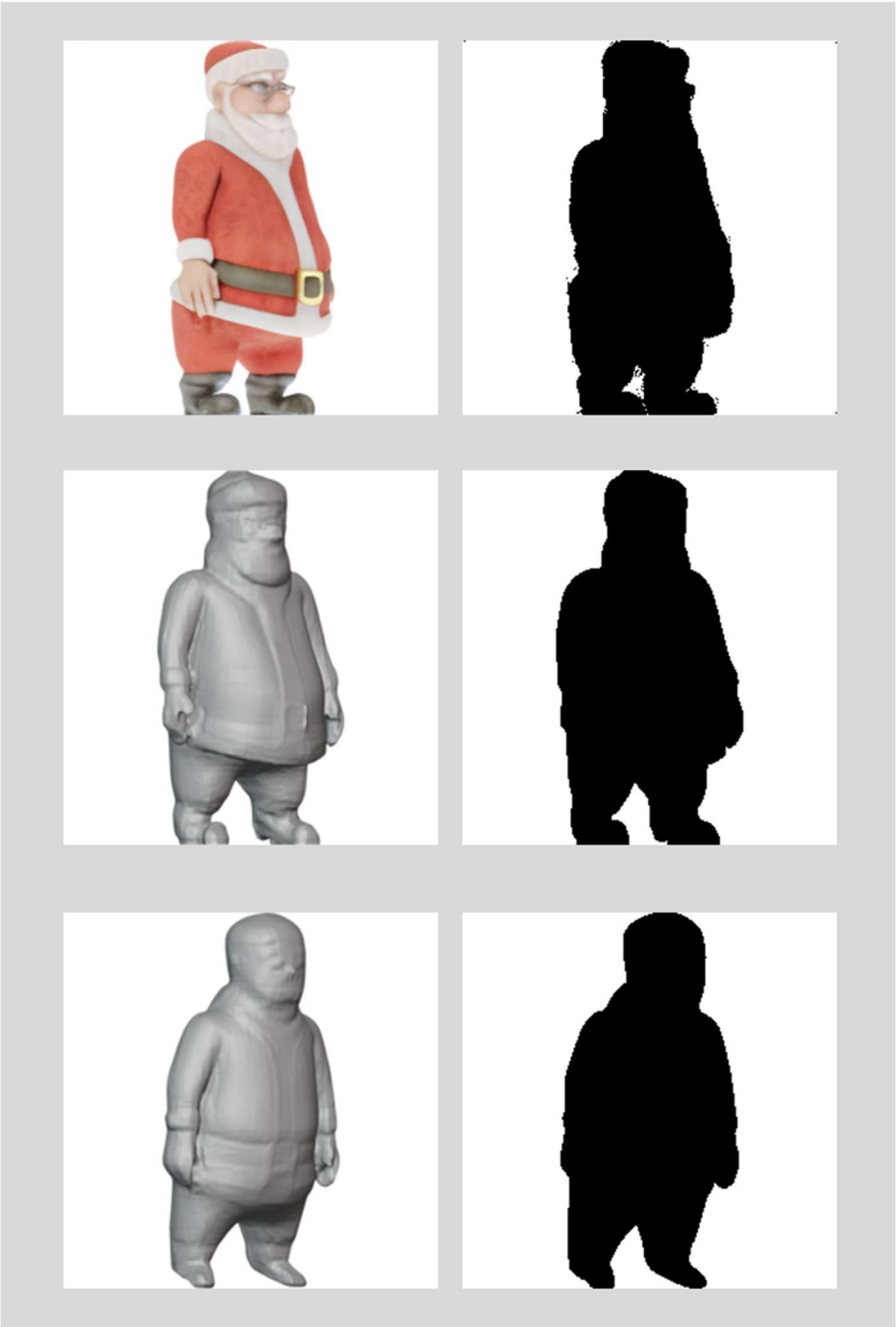

図 5.9 二値化後の画像

5.4.2 不一致率(Non-overlapping Rate)

本研究では画像の XOR(排他的論理和)操作を用いて，2 枚の画像の不一致部分を比較

し，それらの差異を定量化した.具体的な方法は以下の通りである.まず，2 枚の画像をメモリに読み込み，同じサイズに統一することでピクセルの整合性を確保した.その後，画像をグレースケール(灰度)に変換し，輝度情報を抽出して，閾値を 253 に設定して二値化処理を行い，ピクセルを前景(対象物)と背景に分類した.次に，二値化された画像に対してピクセル単位で XOR 操作を実行した.その結果，XOR の出力画像では，ピクセル値が 255(白)の部分が 2 枚の画像で値が一致しない，すなわち不重合部分を示し，ピクセル値が 0(黒)の部分が一致する部分を示した.XOR 結果の白いピクセル数を統計し，それが画像全体のピクセル数に占める割合を計算することで, 不重合領域の大きさを定量化し，2 枚の画像の差異を評価した.また，閾値を 253 に設定することで，二値化処理時に背景領域を明確に分離し，対象領域をより厳密に抽出することが可能となった.最後に，XOR 結果を視覚化した画像として保存し，2 枚の画像の不重合部分の分布を直感的に示すことで，定量的な分析と視覚的な参考を提供した.

以下の図 5.10 に示すように，本研究では 2 枚の画像の不一致率を計算する際，主に(d)で赤色で囲まれた部分のピクセル総数を算出し，それを 256×256 の元の画像と割り算することで割合を計算している.

図 5.13 は，本研究が生成したモデルと先行研究が生成したモデルの不一致率数値および生成結果の比較である.本研究の生成結果は，先行研究に比べて不一致率が低いことが確認できる.このため，本研究の生成品質はより高いものであるといえる.

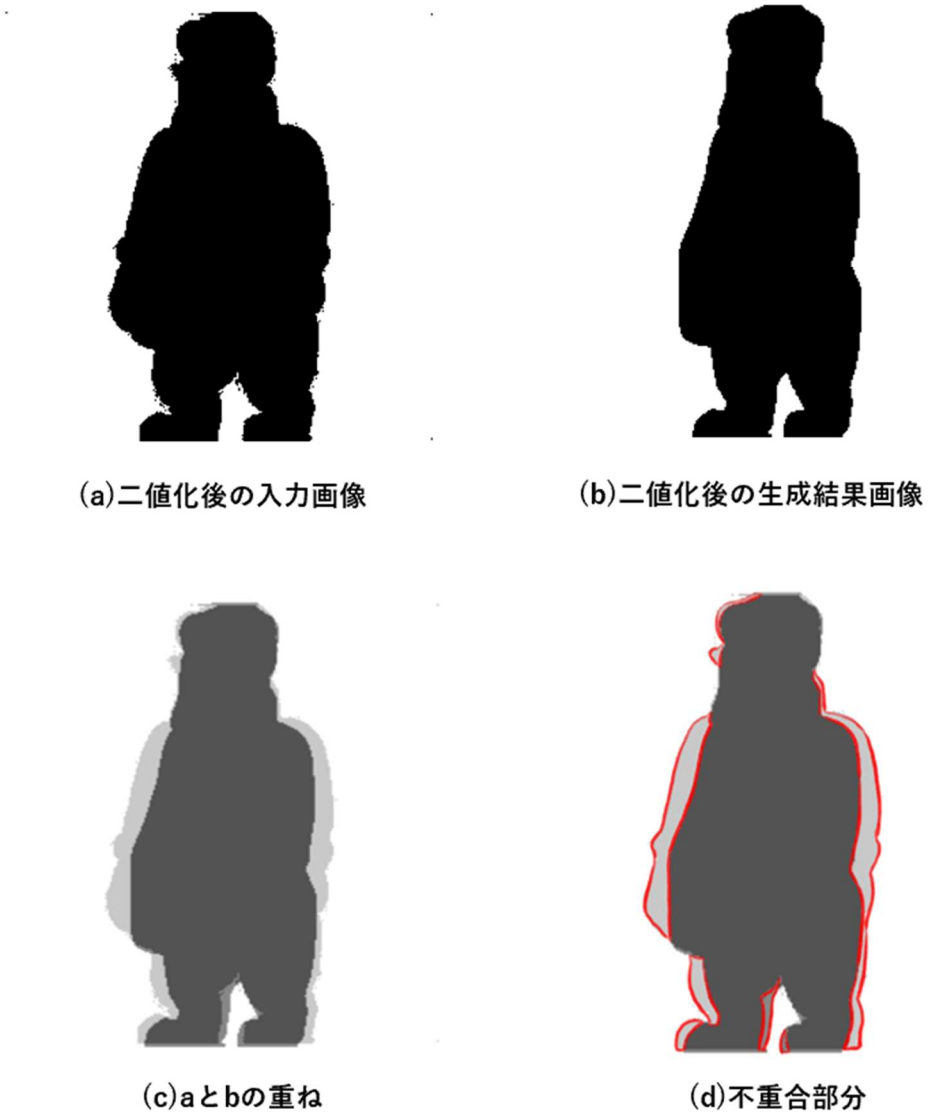

図 5.10 不一致率の計算方法

5.5 評価結果のまとめ

本節では，生成結果のデータ分析を行う.

下図 5. 11 **本研究で生成された 3D モデルの不一致率**は，本研究の生成結果における不一致率を示したものである.左側の第 1 列と第 2 列には，入力画像およびその二値化処理後の画像が示されている.第 3 列には，本研究が生成した 3D モデルの 4 つの視点からのスクリーンショットが示されており，第 4 列には，それらに対応する二値化後の画

像が示されている.不一致率を計算する際には，第 2 列の画像と第 4 列の画像に XOR 演算を適用する必要がある.第 5 列は計算された不一致ピクセルの数を示し，第 6 列には対応する不一致率が示されている図 5. 11 **本研究で生成された 3D モデルの不一致率**から，本研究の生成結果における不一致率が平均して 5%以内に抑えられていることが確認できる.

図 5.11 本研究で生成された 3D モデルの不一致率

下図 5. 12 は，先行研究の生成結果における不一致率を示したものである.同様に，左側の第 1 列と第 2 列には入力画像およびその二値化処理後の画像が示されている.第 3 列には先行研究が生成した 3D モデルの 4 つの視点からのスクリーンショットが示されており，第 4 列には，それらに対応する二値化後の画像が示されている.不一致率を計算する際には，第 2 列の画像と第 4 列の画像に XOR 演算を適用する必要がある.第 5 列は計算された不一致ピクセルの数を示し，第 6 列には対応する不一致率が示されている．図 5. 12 から，先行研究の生成結果における不一致率が約 10%であることが確認できる.

このデータセットの出力結果を比較すると，本研究の生成結果の各視点における不一致率が，先行研究と比較して小さいことが確認できる.そのため，本研究が生成した 3D モデルの品質がより高いといえる.また，本研究と先行研究の不一致率の平均値を算出した結果，本研究の平均不一致率は 3.07%，先行研究の平均不一致率は 7.94%である.本データセットにおいて，本研究の生成モデルは 4.87%の精度向上を達成したことがわかる.

図 5.12 先行研究 Wonder3D で生成された 3D モデルの不一致率

本節では，さらに 3 組のデータを用いて不一致率を計算した.この 4 組のデータにおける本研究の平均不一致率は 3.35%, 先行研究の平均不一致率は 9.26%である.本研究の不一致率は引き続き 5%以内に抑えられており，先行研究の平均不一致率は約 10%であることが確認された.また，本研究は不一致率を 5.91%低減させており，3D モデルの生成品質を向上させていることがわかる.

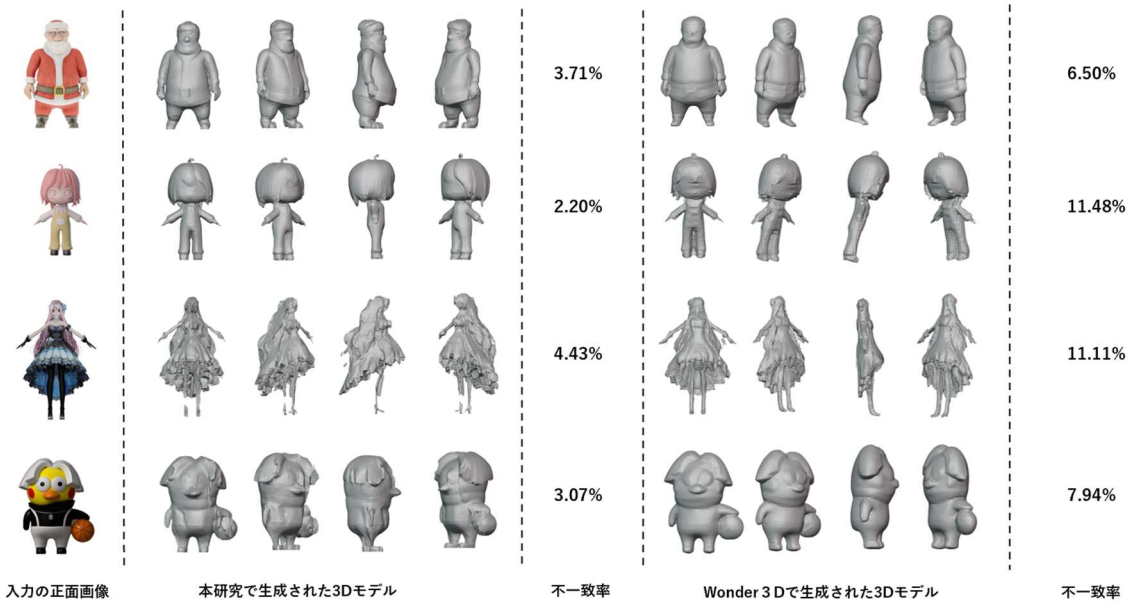

図 5.13 本研究と Wonder3D の生成結果不一致率比較

5.6　妥当性確認

3D モデラーへのインタビューを実施したところ、本技術の結果について「非常に驚くべきものであり、確かにモデリング時間の短縮に大きく貢献する」との評価を得た。また、今後さらなる技術発展が期待され、将来的にはモデラーの補助ツールとして定着する可能性が高いと考えられる。

一方で、現状の生成モデルはそのままでは使用できず、品質のばらつきが見られるため、手動での修正が必要である。具体的には、生成された 3D モデルを実際のゲーム開発に適用するためには、形状の微調整やトポロジーの修正、テクスチャの調整が不可欠であり、これらの作業には約 1 日程度の時間がかかるとされている。

また、今後の 3D モデリング教育にも変化が求められる。従来の教育では、Maya や Blender などの 3D ソフトウェアの使用方法を中心に指導していたが、今後は AI ツールの活用も含めた教育カリキュラムが必要になると考えられる。特に、AI を用いたモデリングの効率化や、AI 生成モデルをどのように最適化・修正するかといったスキルが重要視されるようになるだろう。

さらに、本技術の発展により、3D モデリングに関する人材需要の減少が予測される。AI の導入によって、多くの単純なモデリング作業が自動化されるため、企業にとっては人件費削減につながる。一方で、高品質な 3D モデルを作成するためには、AI が生成したモデルを適切に調整・修正できる熟練したモデラーが依然として求められるため、モデラーの役割は「手作業でのモデリング」から「AI と協働し、最適化するスキルを持つ専門職」へと変化していくことが予想される。

第6章 結論および今後の課題

本章では，本研究の結論及び今後の課題を示す.

6.1 結論

本研究では，キャラクターのイラストから 3D モデルを生成する手法を提案し，ゲームキャラクター制作における独自のニーズに焦点を当てている.3D 効果の強化，法線マップ生成，および幾何学的認識に基づくメッシュ抽出を組み合わせることで，高品質な結果を実現し，忠実度と幾何学的一貫性を向上させた.

本実験では，多視点のゲームキャラクター立ち絵を用いて 3D モデルを生成することが可能であり，1 つの 3D モデルを約 3 分で生成できる.また，生成されたデータは NeRF からメッシュ表現に変換し，3D モデリングソフトにインポートして編集することができる.この手法は，従来の手動による 3D モデル制作と比較して，時間とコストを大幅に削減する効果がある.

実験結果から，本手法は既存の単視点手法を上回る性能を示し，限られた 2D 情報から 3D モデルを再構築する上で顕著な効果を発揮することが確認された.具体的には，本研究の生成結果は入力画像の 3D 構造をより忠実に再現し，入力画像との一致性が高く，ディテールの表現においても従来研究より優れている.また，従来研究で頻発していたモデルの厚みが不足する問題も解決している.

しかしながら，生成された 3D モデルは直接的な利用にはまだ課題が残っている.例えば，生成モデルの表面は手動制作のモデルに比べて滑らかさに欠けており，さらに生成モデルは一体型の構造であるため，部位ごとの生成(例：頭部，体，衣服など)ができない.このため，特定の部位を個別に編集する際には困難が伴う.

加えて，本研究では新たな 3D モデル生成結果の評価方法も提案している.特定の角度でのモデルの輪郭と入力画像の輪郭を比較し，不重複率を計算することで生成モデルの品質を評価する.この評価方法は，生成 AI に関連する研究において，本研究の要件に完全に一致する評価システムが存在しないという課題に対し，新たな視点を提供するものである.

本研究で提案した新たな評価アルゴリズムを用いた結果，本研究が生成した 3D モデルの不一致率は 5%以内に抑えられることが確認された.一方,先行研究における不一致率はおおむね 10%程度であり，本研究の提案手法によって不一致率を約 5%低減することが可能となった.この結果から，提案手法により生成される 3D モデルの全体的な構造品質が向上していることが示された.

6.2 課題

本研究では, キャラクターイラストから 3D モデルを生成する革新的な手法を提案し, 忠実度, 幾何学的一貫性, 生成効率の面で優れた成果を示した.しかし, 実用化に向けてさらに解決すべき課題がいくつか存在する.

まず, 本研究にはいくつかの制約がある.例えば, 入力画像は少なくとも 4 つの異なる視点(4 枚のキャラクターイラスト)を含む必要があり, これらの入力画像は 3D 的一貫性を厳密に満たす必要がある.また, 入力画像に関する要求が多く, 生成モデルを作成する際には, 背景の削除, 画像のトリミング, サイズの統一といった多くの前処理が必要となる.これらの工程は, 使用の複雑性や時間コストをある程度増加させる要因となっている.そのため, 今後の研究では, 自動入力処理手法を探求し, 入力画像の一貫性に対する厳しい要件を緩和することで, データ準備の複雑性を低減し, より汎用的かつ効率的な 3D モデル生成技術の実現を目指すべきである.

次に, 生成された 3D モデルは, 幾何学的一貫性や細部の再現性において既存手法を上回るものの, モデル全体を分割せずに生成するという制限があり, この点がモデルの柔軟性に課題をもたらしている.特に, 特定の部位を局所的に修正する必要がある場合, 困難が伴う.実際の作業を考慮すると, モデラーはモデルの編集時に部分ごとに処理を行うことが多いため, 今後の研究では, 分割生成手法を導入し, モデルを頭部, 身体, 衣服などの部位ごとにモジュール化して生成するアプローチを探る必要がある.この手法は, モデルの柔軟性を向上させるだけでなく, 特定部位の修正を容易にする可能性がある.

さらに, 多モーダルデータ(例：テキスト記述や深度マップ)を補助情報として導入することで, 生成結果の品質や柔軟性の向上が期待できる.また, 生成アルゴリズムのさらなる最適化により, 生成モデルが表面の滑らかさや構造の編集可能性において手動作成モデルに近づくことが可能となる.最後に, 主観的評価への依存を減らし, 生成結果の品質を精密かつ効率的に評価できる, より汎用的で自動化された評価システムの構築も重要である.これらの改善は, より実用的で効率的な 3D モデル生成技術の開発を後押しし, ゲームキャラクター制作などの分野において幅広い応用可能性を提供するものである.

参考文献

[1] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. “NeRF: representing scenes as neural radiance fields for view synthesis”. Commun. ACM 65, 1 (January 2022), 99–106. https://doi.org/10.1145/3503250.

[2] Kerbl, Bernhard, Georgios Kopanas, Thomas Leimkuehler and George Drettakis. “3D Gaussian Splatting for Real-Time Radiance Field Rendering.” *ACM Transactions on Graphics (TOG)* 42 (2023): 1 - 14..

[3] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014. [Online]. Available: https://arxiv.org/abs/1312.6114

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In NeurIPS, 2020. 3

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models”, in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 10674-10685, doi: 10.1109/CVPR52688.2022.01042.

[7] A. A. Markov, "Extension of the limit theorems of probability theory to a sum of variables connected in a chain (in Russian)," *Izvestiya Fiziko-Matematicheskogo Obschestva pri Kazanskom Universitete*, vol. 2, pp. 135–156, 1906.

[8] J. R. Norris, *Markov Chains*. Cambridge, U.K.: Cambridge University Press, 1997.

[9] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Pacific Grove, CA, USA: Duxbury Press, 2002.

[10] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. Boca Raton, FL, USA: Chapman and Hall/CRC, 2013.

[11] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv preprint arXiv:2205.11487*, 2022. [Online]. Available: https://arxiv.org/abs/2205.11487

[12] P. Dhariwal and A. Nichol, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint arXiv:2204.06125*, 2022. [Online]. Available: https://arxiv.org/abs/2204.06125

[13] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Retrieved from https://arxiv.org/abs/2102.12092

[14] Radford, Alec et al. “Learning Transferable Visual Models From Natural Language Supervision.” *International Conference on Machine Learning* (2021).

[15] Openai. (2021, January 5). *CLIP: Connecting Text and Images*. https://openai.com/index/clip/.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations (ICLR)*, 2021.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998. doi: 10.1109/5.726791

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention (MICCAI)*, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[20] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention (MICCAI)*, 2016, pp. 424–432.

[21] M. Pharr, G. Humphreys, and P. Hanrahan, "An introduction to physically based rendering," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 777–786, 2004.

[22] P. Hanrahan and W. Krueger, "Inverse rendering for computer graphics," in *Proc. 20th Annu. Conf. Comput. Graph. Interactive Tech. (SIGGRAPH)*, 1993, pp. 411–418.

[23] Garland, M., & Heckbert, P. S. (1997). Surface simplification using quadric error metrics. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* (pp. 209–216). ACM.

[24] Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 652–660). IEEE.

[25] Kajiya, J. T., & Von Herzen, B. P. (1984). Ray tracing volume densities. In *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* (pp. 165–174). ACM.

[26] F. Rosenblatt, "The Perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958. doi: 10.1037/h0042519

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[28] J. T. Kajiya and B. P. Von Herzen, "Ray tracing volume densities," in *Proc. 11th Annu. Conf. Comput. Graph. Interactive Tech. (SIGGRAPH)*, 1984, pp. 165–174. doi: 10.1145/800031.808594

[29] L. Quan, Z. Lan, and P. Sturm, "A sequential solution of the three-point pose problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 3, pp. 241–246, Mar. 1999. doi: 10.1109/34.750448

[30] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000. doi: 10.1109/34.888718

[31] C. Zheng, C. H. Lin, X. Xie, Z. Li, J. B. Tenenbaum, W. T. Freeman, and J. Wu, "Zero-1-to-3: Zero-shot One Image to 3D Object Generation," *arXiv preprint arXiv:2303.11328*, 2023.

[32] M. Gharbi, C. Wu, D. Salesin, A. Hertzmann, and B. Russell, "Objaverse: A Universe of Annotated 3D Objects," *arXiv preprint arXiv:2302.04847*, 2023. [Online]. Available: https://arxiv.org/abs/2302.04847

[33] Y. Alaluf, O. Patashnik, Z. Wu, S. Shamai, D. Cohen-Or, and A. Geiger, "Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation," *arXiv preprint arXiv:2212.00774*, 2022. [Online]. Available: https://arxiv.org/abs/2212.00774

[34] Poole, B., Jain, A., Barron, J.T., & Mildenhall, B. (2022). DreamFusion: Text-to-3D using 2D Diffusion. *ArXiv, abs/2209.14988*.

[35] X. Long et al., "Wonder3D: Single Image to 3D Using Cross-Domain Diffusion," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 9970-9980, doi: 10.1109/CVPR52733.2024.00951.

[36] W. B. Shen, J. Krause, A. Tao, and others, "Google Scanned Objects: A High-Quality Dataset of 3D Scanned Household Items," *arXiv preprint arXiv:2211.09153*, 2022. [Online]. Available: https://arxiv.org/abs/2211.09153

[37] Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., & Wang, W. (2023). SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. *ArXiv, abs/2309.03453*..

[38] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *Int. J. Comput. Vis.*, vol. 9, no. 2, pp. 137–154, 1992. doi: 10.1007/BF00129684

[39] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multi-view stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010.

[40] C. Bugeau and P. Perez, "PatchMatch stereo: Stereo matching with slanted support windows," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2011, pp. 1–10.

[41] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 652–660. doi: 10.1109/CVPR.2017.16

[42] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 165–174.

[43] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proc. Fourth Eurographics Symp. Geometry Process. (SGP)*, 2006, pp. 61–70. doi: 10.2312/SGP/SGP06/061-070

[44] G. Boulch, "ConvPoint: Continuous convolutions for point cloud processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 68–69. doi: 10.1109/CVPRW50498.2020.00015

[45] Y. Yan, Y. Mao, and B. Li, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4490–4499. doi: 10.1109/CVPR.2018.00472

[46] G. Riegler, A. O. Ulusoy, and A. Geiger, "OctNet: Learning deep 3D representations at high resolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 3577–3586. doi: 10.1109/CVPR.2017.380

[47] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 2320–2327. doi: 10.1109/ICCV.2011.6126513

[48] A. Kendall et al., "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 66–75. doi: 10.1109/ICCV.2017.15

[49] N. Wang et al., "Pix2Mesh: Generating 3D mesh models from single RGB images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 52–67. doi: 10.1007/978-3-030-01252-6_4

[50] J. L. Schönberger and J. M. Frahm, "Structure-from-Motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.

[51] A. Nichol, H. Jun, and P. Dhariwal, "Shap-E: Generating conditional 3D implicit functions," *arXiv preprint arXiv:2305.02463*, 2023.

[52] Lin, C., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M., & Lin, T. (2022). Magic3D: High-Resolution Text-to-3D Content Creation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 300-309.

[53] Szymanowicz, S., Rupprecht, C., & Vedaldi, A. (2023). Viewset Diffusion: (0-)Image-Conditioned 3D Generative Models from 2D Data. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8829-8839.

[54] Shi, Y., Wang, P., Ye, J., Long, M., Li, K., & Yang, X. (2023). MVDream: Multi-view Diffusion for 3D Generation. *ArXiv, abs/2308.16512*.

[55] Civitai. (n.d.). https://civitai.com/models/25995/blindbox

[56] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[57] B. Girod, "A survey of methods for recovering and representing shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 7, no. 3, pp. 257–267, 1985. doi: 10.1109/TPAMI.1985.4767663

[58] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. Pearson, 2017.

[59] Newzoo, "Global Game Market Report," 2023. [Online]. Available: https://newzoo.com

[60] China Venture, "ゲーム開発コスト分析," 2023 年 8 月. [Online]. Available: https://www.chinaventure.com.cn/news/108-20230816-376753.html

[61] S. Finance, "3D モデリングのコスト構造," 2023 年 5 月. [Online]. Available: https://finance.sina.com.cn/blockchain/roll/2023-05-09/doc-imyteaxa8636376.shtml

[62] Bilibili Research, "AAA ゲームの開発費," 2023 年 6 月. [Online]. Available: https://www.bilibili.com/read/cv38442812

[63] The Paper, "AI を活用した 3D モデリングの進展," 2023 年 9 月. [Online]. Available: https://www.thepaper.cn/newsDetail_forward_23905837

[64] GameIndustry.biz, "ゲームアートの制作工程," 2023 年 10 月. [Online]. Available: https://www.gameindustry.biz/articles/2023/10/3d-art-pipeline

[65] Forbes, "3D モデラーの給与調査," 2023 年 7 月. [Online]. Available: https://www.forbes.com/sites/gamedevelopment/3d-modeller-salary-2023

[66] TechNode, "ゲーム業界の外注戦略," 2023 年 11 月. [Online]. Available: https://technode.com/gaming-outsourcing-trends-2023

[67] IGN Japan, "ゲーム開発の人材不足問題," 2023 年 12 月. [Online]. Available: https://jp.ign.com/game-development/shortage-2023

[68] Unity Developer Blog, "3D アーティストのキャリアパス," 2024 年 1 月. [Online]. Available: https://blog.unity.com/career-paths-for-3d-artists

[69] Game Developer Magazine, "労働環境とバーンアウト," 2024 年 2 月. [Online]. Available: https://www.gamedeveloper.com/stress-and-burnout-in-3d-art

[70] AI Game Lab, "機械学習を用いた 3D モデル生成," 2024 年 3 月. [Online]. Available: https://www.aigamelab.com/ai-generated-3d-models

[71] Stanford AI Research, "AI 生成 3D モデルの課題," 2024 年 4 月. [Online]. Available: https://ai.stanford.edu/research/ai-3d-generation

[72] Unreal Engine Blog, "次世代 3D コンテンツ制作," 2024 年 5 月. [Online]. Available: https://www.unrealengine.com/blog/next-gen-3d-content-creation

謝辞

本研究の実施および論文の執筆にあたり,多くの方々からご支援をいただきました.ここに深く感謝の意を表します.

まず初めに,指導教員の矢向高弘教授には,主査として本論文の執筆に至るまで,多大なるご指導とご鞭撻をいただきました.先生との個別面談では,研究の方向性や課題解決の方法について丁寧にご指導いただき,その中で多くの新しい視点や気づきを得ることができました.また,定期的な研究室のゼミでは,私の進捗に対する的確なアドバイスをいただき,研究の精度を高める大きな助けとなりました.さらに,先生の励ましのお言葉や熱意あるご指導が,私の研究活動を支える大きな原動力となりました.心より感謝申し上げます.

副査を担当していただいた当麻教授には,ご多忙の中にも関わらず,研究内容の議論やフィードバックの時間をいただきました.先生の専門的なご助言は,本研究の完成度を大きく向上させるものとなりました.深く感謝申し上げます.

さらに,研究室の関係者の皆さまにも,日ごろのゼミや学習会におけるご協力に感謝申し上げます.特に,小木先生には,ゼミでの研究発表や議論を通じて多くの示唆をいただき,研究を進める上での重要なアドバイスを数多くいただきました.また,ゼミの仲間たちとの議論や情報交換は,研究を進める中で多くの励みとなりました.前輩の皆さまからは,研究の進め方やスケジュール管理の仕方など,多くの経験を共有していただきました.そのおかげで,研究の全体像をつかむことができ,効率的に進めることができました.

加えて,栗田先生には,研究の専門的な内容に関するご意見やフィードバックをいただきました.先生の具体的なご指導がなければ,現在の研究結果に至ることは難しかったと思います.深く感謝いたします.

そして,研究活動の傍らで生活面を支えてくださった家族,友人,そして恋人にも感謝の意を表します.特に両親には,日ごろから温かい応援と励ましをいただき,精神的な支えとなっていただきました.友人たちには,研究の悩みを共有し,時には息抜きの場を提供していただきました.恋人には,忙しい日々の中で励ましてくれたこと,そして日常生活をサポートしてくれたことに深く感謝しています.

最後に,SDM16期生春入学の同期の皆さまとともに,この2年間の学びを乗り越えられたことを誇りに思います.皆さまとの交流や協力が,この研究活動を支える大きな力となりました.心より感謝申し上げます.

以上,多くの方々の支えによって本研究を完成することができました.この場を借りて,改めて御礼申し上げます.