修士論文

複数光源の仮想 3D 空間中の配置を考慮した 2D 線画の陰影生成

Wu, Liudong

主指導教員　　吉高 淳夫

北陸先端科学技術大学院大学
先端科学技術専攻
（情報科学）

令和 7 年 3 月

Abstract

In recent years, image generation technology has evolved from simple encoder-decoder structures to adversarial generative networks and diffusion models. With advancements in technology, both the quality and controllability of generated images have improved, enabling the production of high-quality and diverse images.

In the fields of artistic creation and visual design, image generation technology reduces the time and effort required for production while also providing significant support. Among these advancements, research on shading generation from sketches has gained increasing attention. However, most existing methods for shading generation from sketches assume a single-light-source environment and are not capable of handling shading under multiple light sources. One major challenge is the difficulty in appropriately capturing the interactions between multiple light sources. Furthermore, due to the insufficient recognition of sketch structures, incorrect shading is often generated in empty regions both inside and outside the sketched objects.

To address these issues, this study proposes a diffusion-based shading generation method that considers multiple light sources in a virtual 3D space. By incorporating mask generation using the Segment Anything Model (SAM), the proposed method clarifies the regions where shading should be applied, thereby suppressing unintended shading generation outside the target areas. Additionally, a light embedding module is designed to integrate light source information into the diffusion process. By utilizing ConvNeXtBlock, this method ensures that light source information is appropriately embedded, thereby improving the accuracy of shading generation under multiple light sources.

Experimental results demonstrate that, in the case of a single light source, the proposed method reduces the total area of incorrectly generated shading in empty regions by 54.8% compared to the existing method, ShadeSketch. Furthermore, the similarity between the generated shading images and ground truth images improves by 3.4% over ShadeSketch. The proposed method is also adaptable to multiple light sources. In experiments involving two light sources, the average SSIM reached 0.897, indicating that the method successfully generates shading that reflects multiple light sources.

As a future direction, improving the encoding format of light sources remains an important challenge. The current method projects light source information onto predefined labels and controls image generation through label embedding. To enhance generalization and scalability, further investigation into encoding methods that incorporate additional light source attributes, such as type and intensity, is required.

概要

近年、画像生成技術は、単純なエンコーダ・デコーダ構造から、敵対的生成ネットワークや拡散モデルへと進化してきた。技術の進展に伴い、生成品質および制御能力が向上し、高品質かつ多様な画像の生成が可能となっている。

芸術創作やビジュアルデザインにおいて、画像生成技術は制作に要する時間や労力を削減するとともに、支援を提供する。その中に、線画からの陰影生成技術に関する研究が進められている。しかし、既存の線画から陰影を生成する手法の多くは単一光源環境を前提としており、複数光源下での陰影生成には対応できていない。複数の光源間の相互作用を適切に反映することが困難である。さらに、線画の形状認識が不十分であることから、線画の内部や外部の空白部分に誤った陰影が生成される問題が生じる。

これらの問題を解決するために、本研究では、拡散モデルを基づき、仮想 3D 空間における複数光源の配置を考慮した陰影生成手法を提案する。Segment Anything Model (SAM)による線画のマスク生成を導入することで、陰影を適用すべき領域を明確し、範囲外の陰影生成を抑制する。また、入力する光源情報を組み込むための light embedding モジュールを設計し、ConvNeXtBlock を用いて拡散過程に光源情報を適切に統合する。これにより、複数光源下における物体の陰影生成の精度を向上させる。

実験結果から見ると、単一光源の場合、提案手法は既存研究 ShadeSketch に比べて空白部分に生成される陰影の総面積を 54.8%減少させた。生成した陰影画像の Ground Truth 画像との類似度を ShadeSketch より 3.4%向上させた。また、複数光源の入力に対して適応可能であり、二つの光源の実験データにおいて、平均 SSIM が 0.897 に達し、複数光源を反映する陰影が生成できるようにした。

今後の課題として、光源のエンコード形式の改良が挙げられる。現行の手法では、光源情報を事前に設定したラベルへと投影し、label embedding を通じて画像生成を制御する方式を採用している。より汎用性と拡張性を向上させるためのエンコード手法の検討が必要である。

目次

図目次

表目次

第 1 章 はじめに

1.1 背景

画像生成モデルは、単純なエンコーダ・デコーダ構造から敵対的生成、さらに確率モデルへと進化してきた。技術革新のたびに、生成品質や制御能力が向上し、より高品質で多様な画像生成が可能になっている。

Autoencoder[1]は、Encoder によって入力データを低次元の潜在空間に圧縮し、Decoder を用いて元のデータを再構成する手法であり、主に特徴学習や次元削減に利用される。しかし、生成能力には限界があり、十分な細部や多様性を持つ高品質な画像を生成することは困難であった。この問題を克服するために、VAE[2]が確率モデルを導入し、潜在空間の連続性と制御性を向上させた。しかし、VAE の訓練においては、KL ダイバージェンスによる正則化が課されるため、潜在表現が過度にスムーズになりやすく、結果として生成画像の細部が不鮮明になる傾向がある。その後、GAN[3]の登場により、画像生成の品質は飛躍的に向上した。GAN は、生成器（Generator）と識別器（Discriminator）から構成され、生成器はランダムノイズからリアルな画像を生成し、識別器は生成画像と実画像を区別するように学習する。両者が敵対的に学習を進めることで、生成器はより高品質な画像を生成できるようになる。しかし、GAN には学習の不安定性やモード崩壊（Mode Collapse）などの課題があり、大規模データでの適用が制限される。これに対し、研究者は WGAN[4]や StyleGAN[5]などの改良モデルを提案し、それぞれ損失関数の最適化や制御可能な画像生成の方向から GAN の安定性と実用性を向上させ、高解像度画像の生成を可能にした。

近年、拡散モデル（Diffusion Model）は、確率的な拡散過程に基づく生成手法として注目を集めており、GAN の学習の安定性やデータ分布の多様性の課題を克服している。拡散モデルは、データに対して段階的にノイズを付加し、最終的に標準正規分布に近づける前向き過程を構築した上で、その逆過程を学習することにより、純粋なノイズから高品質な画像を逐次的に復元するモデルである。この逆過程の学習には、ノイズ除去ネットワークを用い、各段階におけるノイズ成分を推定と除去することで、元のデータを再構成する。

拡散モデルは、DDPM[6]によって初めて提案され、その後、Latent Diffusion Models（LDM）[7]などの手法により計算効率や生成品質が向上し、テキストや

深度情報といった多様なモダリティを条件とした高品質な画像生成が可能となった。例えば、Stable Diffusion[7]や DALL·E[8]は、拡散過程をテキストによって制御することで、高度に制御可能なテキストから画像への変換（Text-to-Image Generation）を実現し、AI-Generated Content（AIGC）の発展を加速させた。

画像生成技術は、芸術的創作やビジュアルデザインなどの分野において、創作時間や労力を削減しつつ、高度な支援を提供する。その中で、線画から陰影を生成する技術は、作品の視覚的リアリズムを高める上で重要な役割を担っている。しかし、線画は物体の形状を簡略化した表現であり、照明や陰影などの情報を含まない。そのため、2D 線画一枚から 3D 情報を推定し、物理的に整合性のある陰影を自動生成することは技術的に困難な課題である。

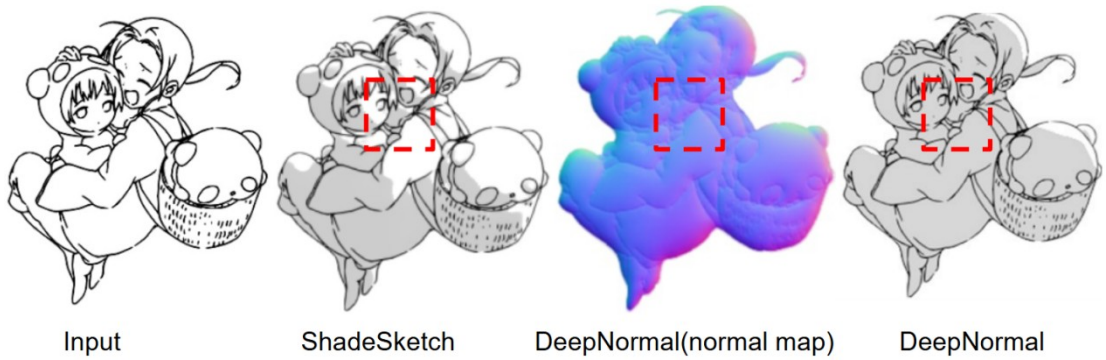

図 1.1：関連研究の陰影生成例（光源数 1、光源位置表側の右上）

既存の手法として、DeepNormal[9]は、CNN[10]に基づく線画から法線マップを推定する手法を提案した。法線マップには幾何情報が反映されており、これを利用することで線画の陰影を計算することができる。しかし、複雑な線画を処理する際には、複数の物体を一つのオブジェクトとして扱ってしまう問題が生じる。また、ShadeSketch[11]は GAN をベースとした手法であり、物体内部や外部の空白部分を正確に学習できないため、範囲外の陰影を誤って生成してしまうという問題が残っている。図 1.1 の赤い点線で示した箇所において、DeepNormal は複雑な線画に対して、指定された光源に応じた適切な陰影を生成することができない。さらに、複数の人物を正しく識別できず、本来は人物間の隙間であるべき領域にも法線情報を生成してしまう。ShadeSketch の生成結果は、光源に対応した陰影の表現において DeepNormal より優れているが、人物間の隙間を正しく識別することはできていない。

新たな生成技術である拡散モデル（Diffusion Model）は、逐次的なノイズ付加と除去のプロセスを通じてデータの確率分布を学習する手法である。GAN と比較して、拡散モデルは多様なデータ分布を学習でき、モード崩壊（mode collapse）の問題を軽減できる。また、逐次的なノイズ除去によって各ステップで小規模な修正を加えながら画像を生成するため、複雑なテクスチャや微細な構造を含む高品質な画像の生成が可能である。特に、本手法を線画からの陰影生成に応用することで、物理的に整合性のある陰影表現を得ることができる。

1.2 目的

DeepNormal や ShadeSketch などの既存の線画から陰影を生成するモデルは、多くの応用シーンにおいて優れた性能を示しているが、依然として無視できない問題が存在する。

顕著な問題の一つは、線画の内部や周辺の空白部分において、形状や構造を正確に認識できず、誤った陰影を生成してしまう点である。この問題の根本的な原因は、線画が高度に簡略化された2次元表現であり、3次元の幾何学情報が欠如していることにある。線画内の空白部分は通常、輪郭線の間の隙間で表現されるため、この情報が特徴抽出時に物体の陰影生成部分と誤って認識されやすい。その結果、モデルは生成過程で空白部分と陰影がある部分を区別できず、特に複雑な線画デザインの場合には陰影のずれがより顕著になる。これにより、精密な陰影生成が求められる応用分野（芸術創作、アニメーション制作、工業デザインなど）において生成効果が参考にならないという問題があり、既存モデルの実用性が制限される。

この問題を解決するために、本研究ではSAMによって生成されたマスクを追加した生成モデルを提案する。従来の生成モデルは通常、線画と光源情報のみを入力として受け取るが、本手法では、陰影生成マスクを導入した。このマスクは、画像内のどの部分に陰影を生成するべきかをモデルに明確に伝えるものである。具体的には、SAM を用いて線画を自動的に識別し、空白部分に陰影を生成しないように正確にマークすることで、誤った陰影生成を防止する。

もう一つの既存研究における問題は、多くのモデルが位置が指定された複数の光源に対する陰影を生成できない点である。現在の多くの陰影生成モデルは、

単一の光源シーンの処理に限定されており、複数の光源を扱う場合、モデルは光源間の空間的な関係を十分に学習できず、それぞれの光源から正確に陰影を生成することが難しい。これは、複数光源下での陰影の生成は、複数の光源と物体との複雑な相互作用を同時に考慮する必要があるためであり、光源間の相互作用や遮蔽効果が生成結果に大きな影響を与える。もしモデルが単一光源のシーンのみで訓練されている場合、複数光源の入力に対処する際の正確さが損なわれる可能性が高い。

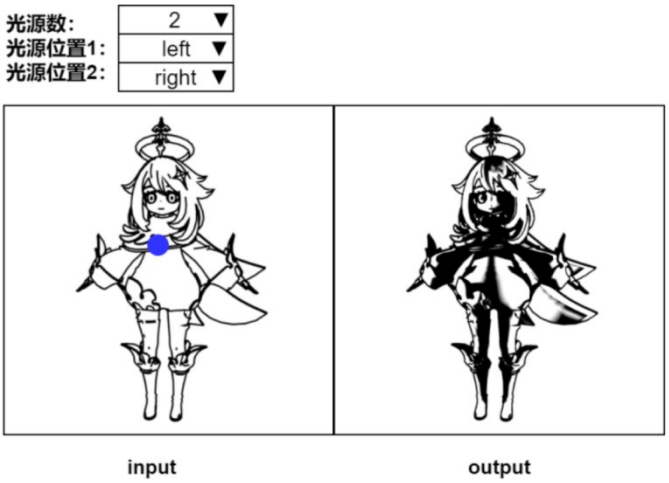

図 1.2：提案手法の機能デモンストレーション

本研究では、複数光源条件下で生成された陰影画像データを作成し、モデルの学習を行うことで、複数光源の間の相互作用を適切に処理できる能力をモデルに付与する。Light Embedding モジュールという光源情報をエンコードするモジュールを設計し、ConvNeXtBlock を用いて拡散過程の U-net に光源情報を導入する。これにより、光源入力で陰影生成をコントロールする。

図 1.2 は本研究で提案する手法の機能を示す図であり、ユーザーは線画とマスクの生成を補助する指示点を入力し、光源の数と光源の位置情報を選択することで、モデルが対応する光照条件下での陰影画像を直接生成する。

1.3 論文の構成

本論文は全六章で構成されている。第一章では、陰影生成に関する課題の背景と研究の動機について述べた。第二章では、関連分野における既存の研究について述べた。第三章では、生成モデルに用いる訓練データの構築プロセスについて詳細に説明している。第四章では、新たな手法を提案し、陰影生成マスクの入力を追加し、複数光源陰影データセットを組み合わせることで、モデルが複雑なシーンに対する生成能力を向上させる方法について詳述している。第五章では、実験の設定および結果を示し、改良されたモデルを定量的に分析している。第六章では、論文の主な貢献を総括する。また、本手法の限界について議論するとともに、今後の発展可能性およびさらなる研究課題について展望を述べる。

第 2 章 関連研究

2.1 線画からの陰影生成モデル

2.1.1 DeepNormal

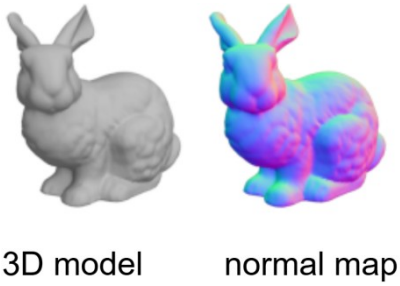

図 2.1：法線マップの例

法線マップ（Normal Map）とは、物体表面の法線ベクトルを RGB 画像として記録したデータ表現である。各ピクセルにおいて、その点における法線方向を視覚的にエンコードすることで、形状や質感に関する詳細な情報を保持する。これにより、3D モデルや 2D 画像に対して、ライティングやシェーディングの処理をより精緻に行うことが可能となる。

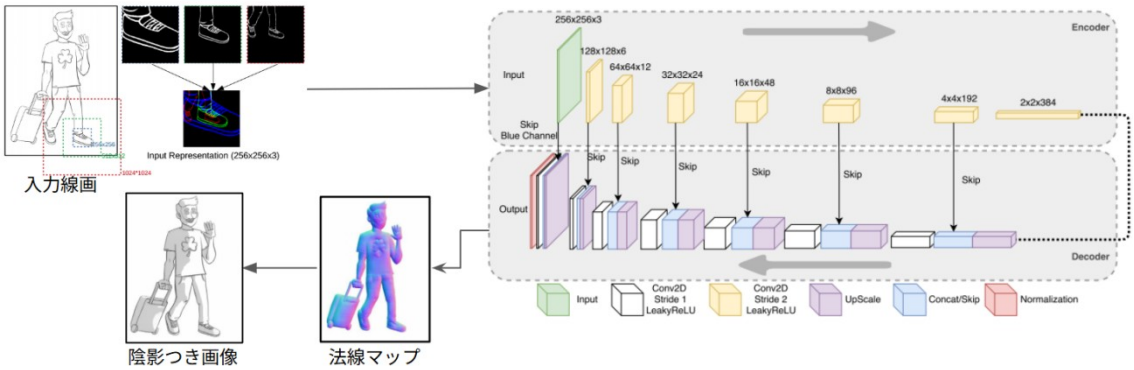

図 2.2：DeepNormal の仕組み[9]

DeepNormal の研究では、CNN に基づく法線マップ生成手法が提案されている。線画を直接 CNN に入力するのではなく、マルチスケール表現へ変換した後にエンコーダへ入力することで、法線マップの推定精度を向上させている。このアプローチにより、従来手法と比較してより高精度かつリアルな法線マップ

の生成が可能となることが示されている。

DeepNormal は、2D スケッチから 3D 情報を持つ法線マップを生成する問題の解決を目的としている。実験結果において、DeepNormal は細部保持および法線推定精度の点で優れた性能を示した。特に、指や衣服のしわ、顔の特徴といった複雑な領域において、従来の幾何学的アプローチ [13,14,15] と比較してより精細な法線表現を実現している。さらに、生成された法線マップを用いることで、線画に対して物理的に整合性のある陰影を描画することが可能となる。

2.1.2 ShadeSketch

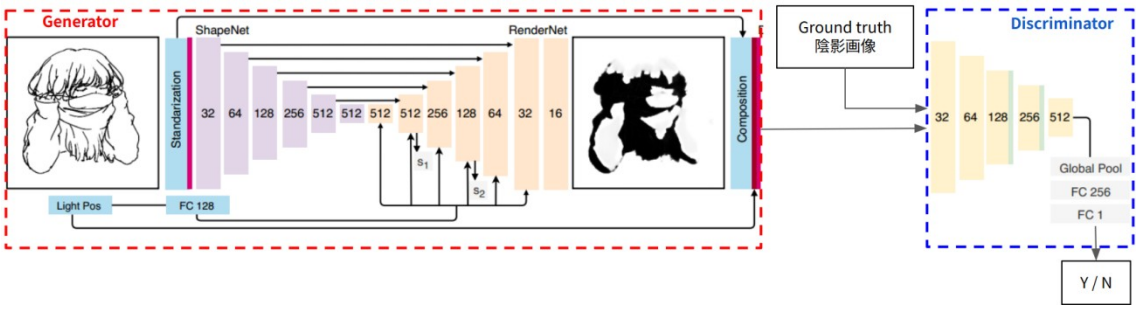

図 2.3：ShadeSktech のネットワーク構造[11]

ShadeSketch は、線画から直接的に陰影を生成する手法に焦点を当てた研究である。陰影は、2D 線画において物体の形状や立体感を表現する重要な要素である。しかし、手描きで陰影を作成する際には、3D 参照モデルが存在しないため、アーティストの直感や想像力に頼る必要がある。特に、構造が複雑な線画では、陰影の調整に時間がかかり、正確な表現が難しい。

ShadeSketch では、GAN に基づく陰影生成手法が提案されている。実際の陰影画像（訓練データ）に近い陰影画像を生成する Generator と、生成画像と実際の陰影画像を識別する Discriminator が対抗的に学習することで、リアルな陰影画像を生成する。

この手法は、入力された線画と指定された光源方向に基づき、陰影画像を生成するものである。そのために、1160 枚の光源方向ラベル付き手描き線画と陰影画像のデータセットを構築し、2D 線画に内在する 3D 情報を学習するネットワークを訓練した。実験結果により、ShadeSketch は単一光源環境下で一貫した

陰影を生成できるだけでなく、建築物、衣服、動物など多様な線画に適用可能であることが示された。さらに、2D アニメーションやデジタルイラスト制作において、効率的な陰影生成ツールとして有用であることが確認された。また、Pix2Pix [16]、U-Net [17]、Sketch2Normal [18] などのベースラインモデルと比較し、陰影のリアリティやディテールの豊かさにおいて、従来手法を上回る性能を示した。

2.2 画像生成モデル

2.2.1 DDPM

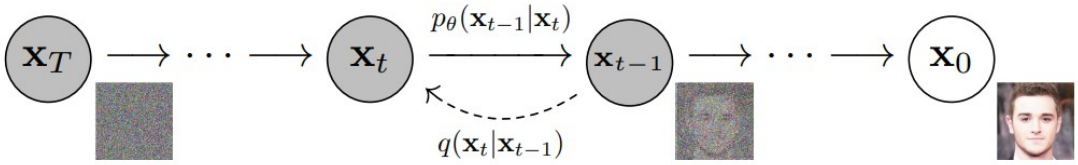

図 2.4：DDPM の生成過程[6]

Diffusion Model の一種である Denoising Diffusion Probabilistic Model (DDPM) は、逐次的なノイズ付加と除去の過程を通じてデータを学習する手法である。画像の生成プロセスを前向き拡散過程（forward diffusion process） と 逆拡散過程（reverse diffusion process） に分けている。

前向き拡散過程では、元画像データ x0 に対して、時間ステップ t ごとにガウスノイズを逐次加え、最終的に標準正規分布に近いノイズ分布へと変換する。この過程は以下のマルコフ過程として記述される。βt は拡散係数であり、通常は線形または指数関数的に増加するように設計される。

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

逆拡散過程では、前向き拡散過程とは逆に、純粋なノイズサンプル xt から元の画像を段階的に復元する。$\mu\theta$ は U-Net によって学習されたノイズ予測ネットワークであり、$\sum\theta$ はノイズ除去の分散を表す。

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

Diffusion Model において、Timestep Embedding は、現在の拡散ステップ t の時間情報を表現するベクトルである。その主な役割は、U-Net などのノイズ除去ネットワークに時間情報を提供し、モデルが異なるノイズレベルを識別し、それぞれの時間ステップに適したノイズ除去を学習できるようにすることである。

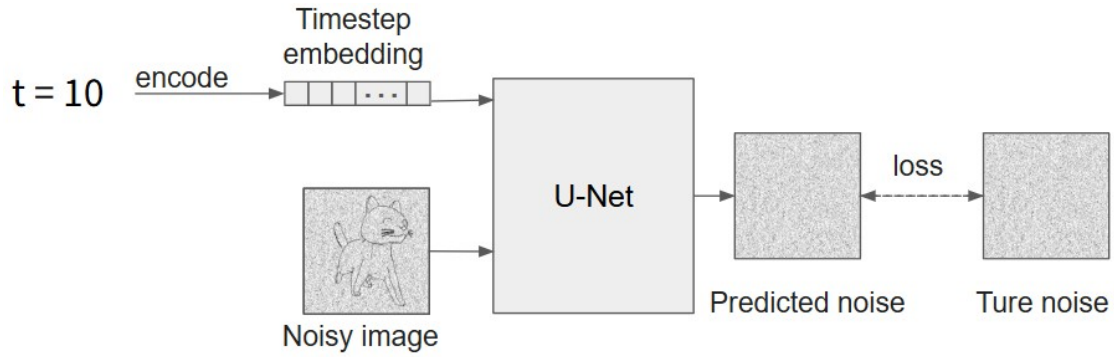

図 2.5：DDPM の U-Net の学習プロセス

図 2.5 の U-Net の学習プロセスから分かるように、Time step は Timestep Embedding としてエンコードされ、U-Net に入力される。

拡散モデルの学習および推論過程では、入力データは異なる時間ステップ t でノイズが付加または除去されるため、モデルは現在の入力データが拡散過程のどの段階にあるのかを明確に把握する必要がある。Timestep Embedding はこの時間情報を適切にエンコードし、モデルが各ステップで最適な特徴表現を学習できるようにサポートする。

2.2.2 LDM

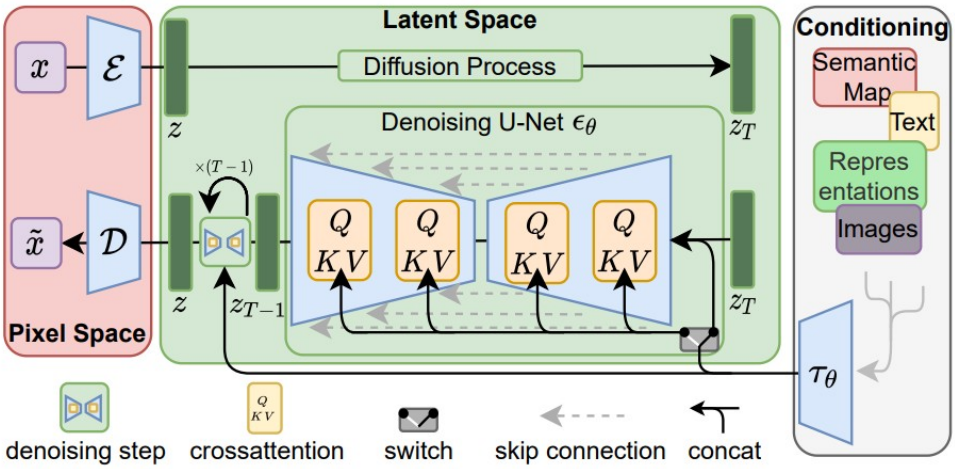

図 2.6：LDM のアーキテクチャ[7]

従来の DDPM は、複数の時間ステップでノイズ除去を行うため、計算量が非常に大きい。特に高解像度画像の場合、ステップ数が増えることで処理時間が長くなり、計算コストが高くなるという課題があった。処理時間を削減するため、Latent Diffusion Model（LDM）[7]は、拡散プロセスを潜在空間で行うことで計算効率を大幅に向上させた。LDM は、まずオートエンコーダ（Autoencoder）を用いて高次元の画像を低次元の潜在表現に圧縮し、その潜在空間上で拡散モデルを適用する。これにより、従来のピクセル空間での拡散に比べ、計算コストを削減しつつ、依然として高品質な画像生成を実現することが可能となった。

さらに、LDM は条件付き生成（Conditional Generation）の柔軟性も向上させた。例えば、テキストからの画像生成（Text-to-Image Generation）では、潜在空間における拡散過程をテキスト情報で制御することで、意味的に整合性のある画像を生成できる。このアプローチは Stable Diffusion において採用され、高解像度かつ多様な画像生成を可能にした。

本研究では、この LDM の特性を活用し、2D 線画から物理的に整合性のある陰影を生成するためのモデルを構築する。

2.3 分割モデル

2.3.1 Segment Anything

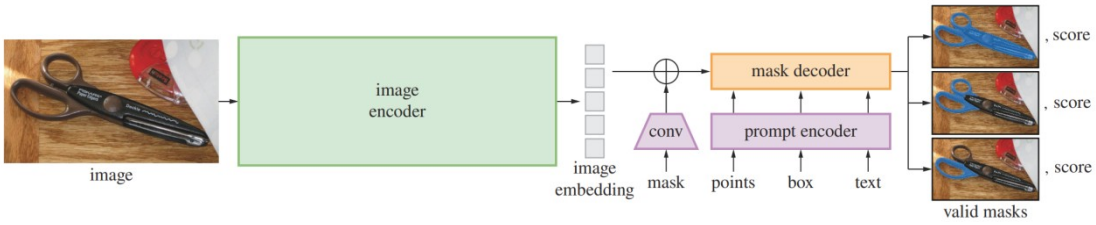

図 2.7：SAM のアーキテクチャ[12]

物体の領域抽出やセグメンテーションは、画像理解やコンピュータビジョンにおいて極めて重要な課題であり、従来のセグメンテーション手法では、大量のラベル付きデータを用いた教師あり学習が主流であった。しかし、特定のタスクに特化したモデルでは、新たなデータセットやタスクへの適応に制約があるという問題があった。これに対し、Segment Anything Model（SAM）は、大規模なセグメンテーションデータセット（SA-1B）を用いて学習された汎用的なインタラクティブセグメンテーションモデルであり、多様な入力に対して適応可能な特徴を持つ。

本研究では、線画から陰影を生成する拡散モデルに SAM を組み込むことで、陰影を適用すべき領域を明確的にモデルへ伝達する手法を提案する。従来の陰影生成モデルは、エッジ情報や畳み込みネットワークを用いた特徴抽出に依存していたが、エッジだけでは物体の領域を正確に識別することが難しく、陰影の生成範囲が不確定になるという問題があった。SAM を導入することで、この問題を解決する。

第 3 章 データセット構築

3.1 データの構成

物理的に正確な陰影生成を実現するため、本研究ではレンダリングエンジンから得たデータを用いてモデルを学習させた。具体的には、25 個の 3D プレフ

ァブモデル[19]をダウンロードし、Blender にインポートして合計 52,500 セットのデータを生成した。図 3.1 に示すように、データには対応する線画、マスク、陰影画像が含まれている。各陰影画像に対応する光源情報は画像の命名に記録している。

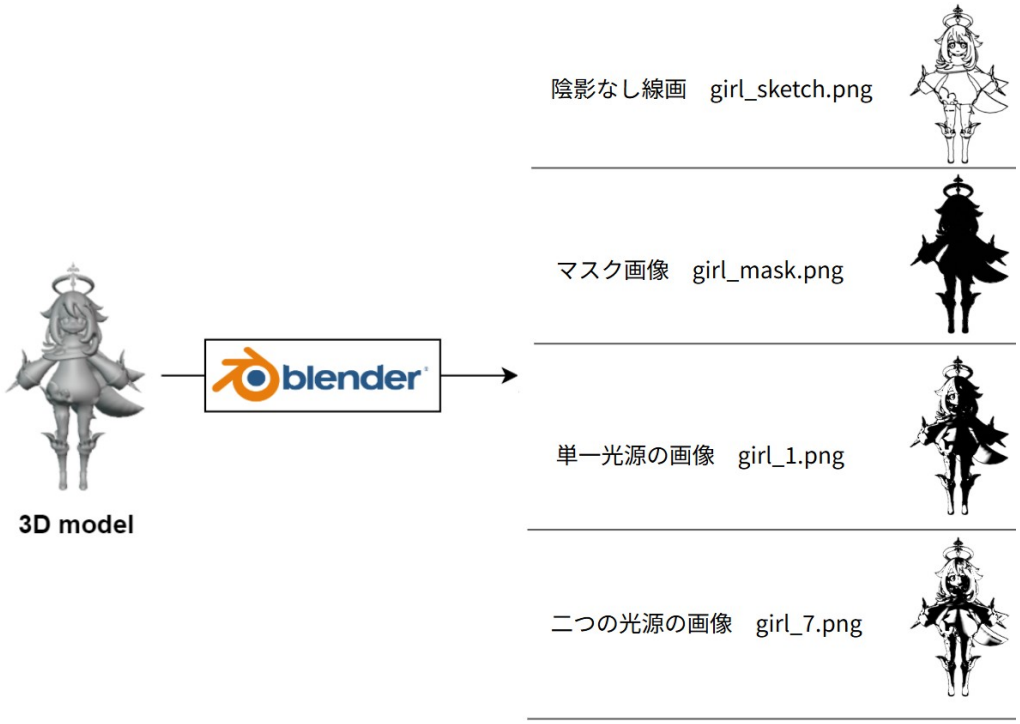

図 3.1：データの例

3D プレファブモデルの種類について、異なる種類の線画を識別し、それに応じた陰影画像を生成できるようにするため、人物、動物、植物、建物、機械の 5 カテゴリの 3D プレファブモデルを使用した。また、同一画面に複数の物体が配置される場合にも対応できるように、複数のプレファブモデルを同一レンダリング画面に配置し、画像を生成した。

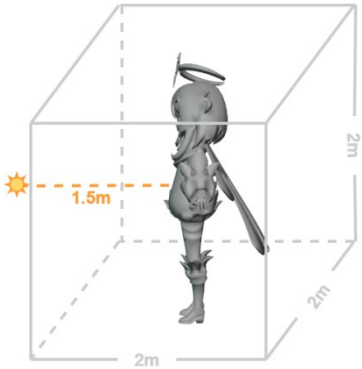

図 3.2：仮想 3D 空間

被照射物体のサイズと角度について、異なる種類の 3D プレファブモデルは現実世界においてサイズが異なるが、データセットの一貫性を保つため、すべてのプレファブモデルのサイズを統一した。各物体を図 3.2 に示したような 2m×2m×2m の仮想空間にスケーリングし、小さい物体は拡大し、大きい物体は縮小することで、仮想 3D 空間をできるだけ埋めるようにした。

また、異なる観察角度における線画と陰影の対応関係を学習させるため、まず、z 軸を中心に 36 度回転させて最初の画像を生成し、その後、y 軸を中心に 36 度回転させて次の画像を取得する。y 軸方向の回転を 36 度ずつ繰り返した後、再度 z 軸を中心に 36 度回転させて次の画像を生成する。この操作を繰り返すことで、最終的に 1 つの 3D プレファブモデルから 100 枚の異なる角度のレンダリング画像を得ることができる。

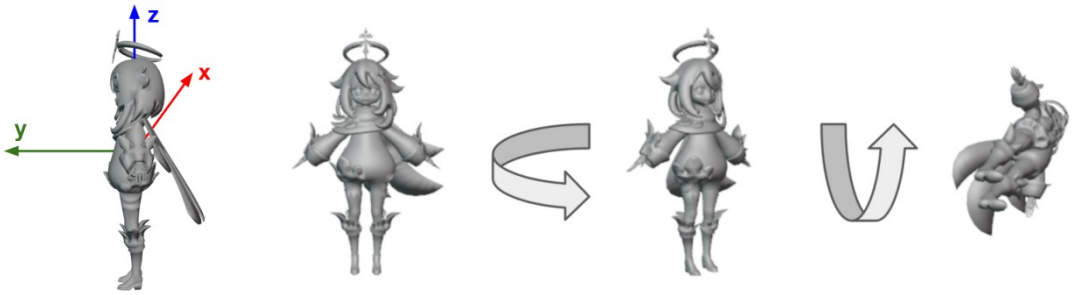

図 3.3：座標系とプレファブモデルの回転

光源位置と照明方向について、本研究における光源は中心から全方向に均等に光を発する点光源として設定した。光源の位置は正面、左右、背面、上方、下方の 6 つの方向に限定し、光源の数は 1 から 2 個とした。本研究では、光源と物体の間の距離は研究範囲外とするため、各光源は被照射物体の幾何学的中心から 1.5 メートルの位置に設定されている。

第 4 章 提案手法

4.1 ネットワーク構造

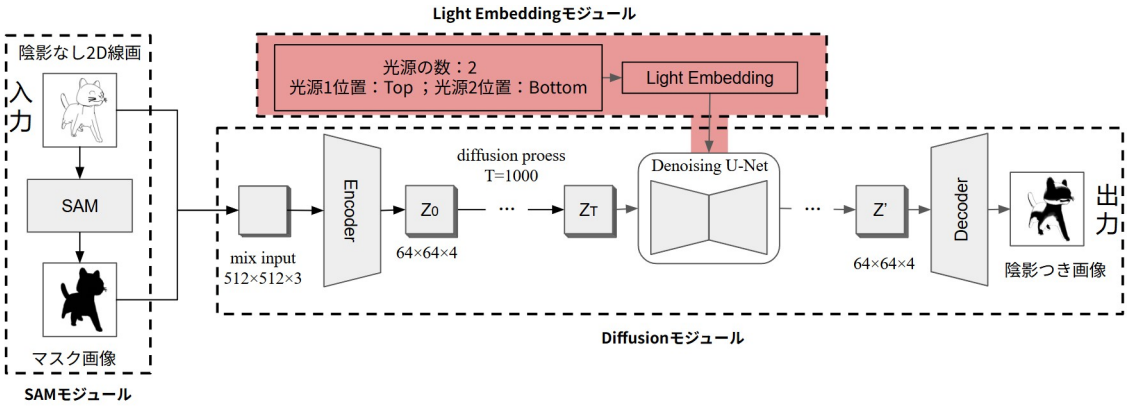

図 4.1：提案手法のフレームワーク

本研究で提案する手法は、その機能に基づき 3 つのモジュールに分類される。

- SAM モジュール：線画の内容を認識し、それに対応するマスクを生成することで、陰影を適用すべき領域を明確化する。
- Light Embedding モジュール：ユーザが指定した光源情報を受け取り、陰影生成の制御信号へ変換する。
- Diffusion モジュール：入力データの次元削減を行った後、light embedding に基づき画像特徴に拡散操作を適用し、最後にデコーダによって、Denoising U-Net で生成された特徴ベクトルを画像へ復元する。

4.2 SAM モジュール

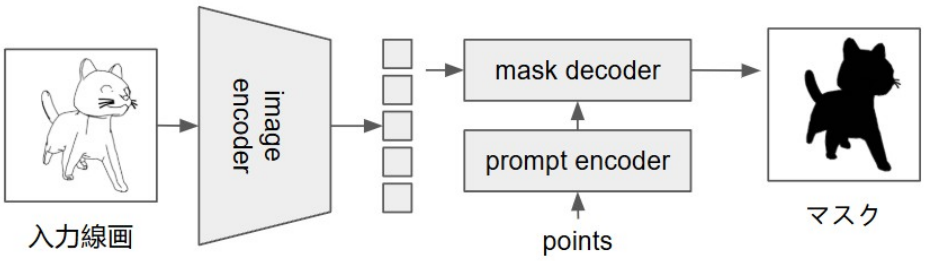

図 4.2：SAM モデルの構造

本研究では、点入力（point prompt）を用いた手法を採用し、分割精度の向上を図った。それに、SAM モデルは高精度な画像分割能力を有するが、線画の分割タスクにおいては、より精確な線画マスクを生成するために追加の訓練が必要である。

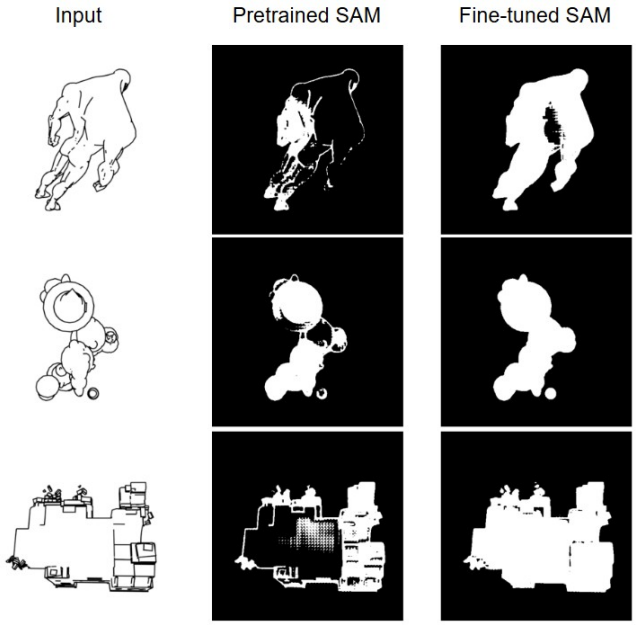

図 4.3：Fine-tuned 前後のマスク生成効果

Method	Mean IoU ↑	Mean Dice ↑
Pretrained SAM	0.45	0.54
Fine-tuned SAM	0.93	0.96

表 4.1: SAM 定量比較

図 4.4 は、SAM モデルが線画に対して微調整を行う前後の分割結果を比較したものであり、微調整によって線画の内部をより正確に捉えたマスクが生成されることが確認できる。さらに、表 4.1 には、微調整前後のモデルによる分割性能の定量的な比較を示しており、平均 IoU および平均 Dice スコアにおいて、微調整後のモデルが顕著な性能向上を達成したことが示されている。

4.3 light embedding モジュール

本研究では、光源の位置および光源の数を限定しているため、図 4.5 に示すように、全 21 種類の照明条件が存在する。これらの 21 種類の照明条件に対して一意のラベルを割り当て、それぞれを識別可能にした。

Label	Light Direction
1	left
2	front
3	right
…	…
20	back,bottom
21	top,bottom

図 4.4：Label と照明条件の対応関係

図 4.5：light embedding の構成

Timestep Embedding は、時間ステップの情報を持つ埋め込みとして U-Net 内で出力を制御する役割を果たす。そのため、ユーザーの光源入力をエンコードし、Timestep Embedding と同じサイズの特徴量として表現した Label Embedding を作成し、これを Timestep Embedding と加算することで light embedding を生成する。この light embedding を U-Net に入力することで、光源情報を考慮した陰影画像の生成が可能となる。

4.4 Diffusion モジュール

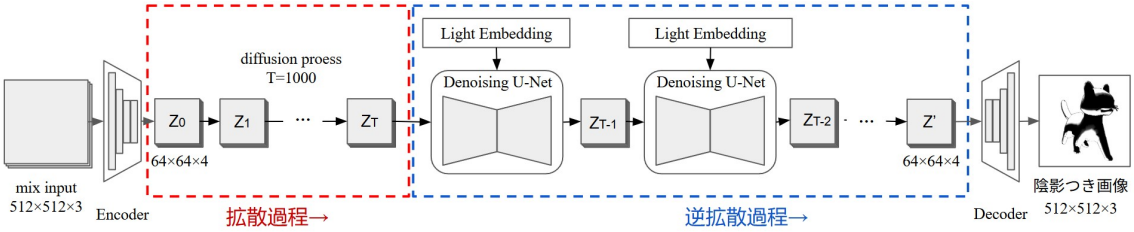

図 4.6：Diffusion モジュールの構造

提案手法の Diffusion モジュールでは入力データにノイズを段階的に加える拡散過程と、ノイズを取り除き、データを再構築する逆拡散過程という二つの拡散過程がある。拡散過程を効率的に実行するため、高次元の入力データを潜在空間に適切にエンコードする必要がある。512×512×3 の mix input を 64×64×4 の潜在特徴へと圧縮し、拡散プロセスの後に 512×512×3 の画像として復元する構造を採用した。

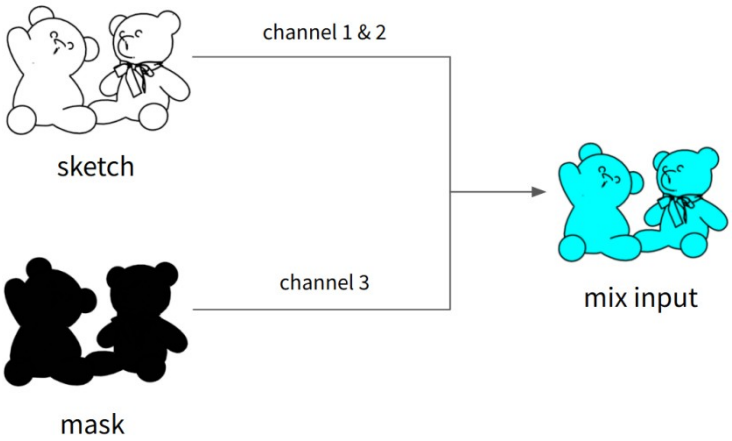

図 4.7：mix input の構成

Mix Input は入力線画とマスク画像から構成されている。具体的には、入力線画の第 1 および第 2 チャネルと、マスク画像の第 3 チャネルを統合し、3 チャネルの入力を作成する。

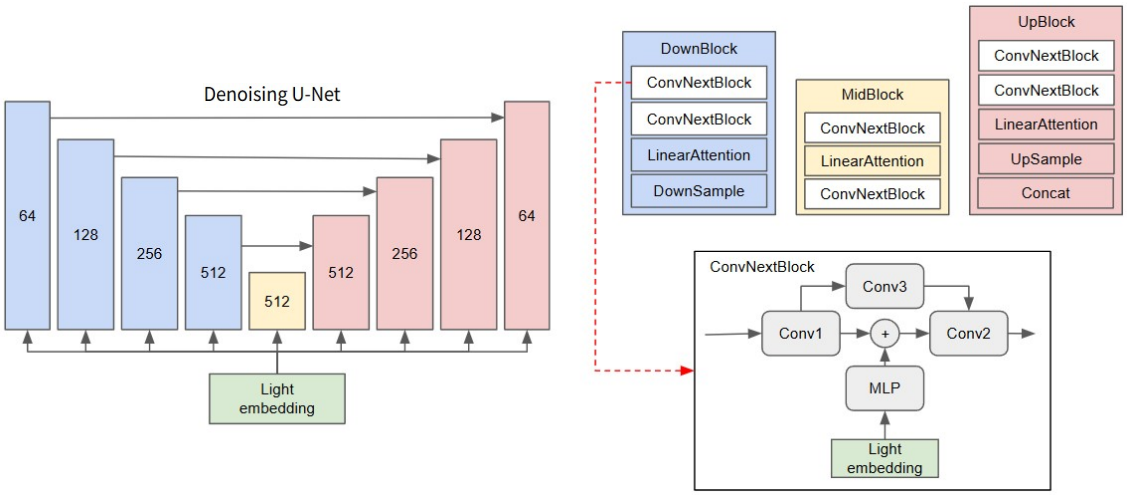

図 4.8：Denoising U-Net の構造

Denoising U-Net には DownBlock、MidBlock、UpBlock の 3 種類のブロックが含まれている。各 Block では、ConvNeXtBlock[20]を用いて、light embedding を U-Net の各レイヤーに入力する。LinearAttention は、グローバルな関係を捉え、情報の損失を抑制する。UpBlock では、エンコーダの対応する出力とスキップ接続（skip connection）を行い、情報を再利用する。

4.5 Loss Function

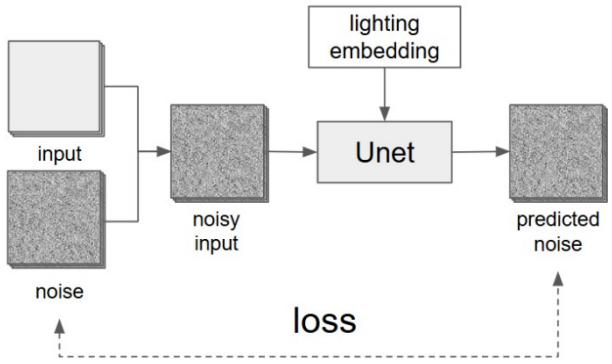

図 4.9：提案手法の学習プロセス

U-Net のトレーニングでは、異なる時間ステップ t におけるノイズを正確に推定できるように、以下の Loss を最適化する:

$$Loss = \mathbb{E}_{z,\epsilon,t}\left[\lambda(t)\left\|\epsilon - \epsilon_\theta(z_t, t, y)\right\|_2^2\right]$$

ここで、各記号の定義は以下の通りである。

- Z_t：入力する潜在変数 Latent Input。
- ϵ：標準正規分布からサンプリングされたガウスノイズ noise。
- $\epsilon_\theta(z_t, t, y)$：U-Net が推定する Predicted noise。
- $\lambda(t)$：時間ステップ t に応じた重み関数。
- y：陰影生成を制御する light embedding。

第 5 章 実験・評価

5.1 実験の詳細

モデルの学習には、NVIDIA A40 GPU（48GB）×2 基を使用し、合計 100,000 イテレーションを実施した。最適化手法としては Adam オプティマイザを採用し、学習率は 1e-4 に設定した。また、バッチサイズは 16、入力画像の解像度は 512×512 ピクセルとした。

データ拡張については、入力画像に対するスケーリングや回転などの操作が光源情報と画像内容の整合性を損なう可能性があるため、本研究ではデータ拡張を適用せず、元のデータをそのまま使用した。この方針により、光源情報の整合性を維持し、適切な光源条件のもとで陰影を正確に生成できるようにした。

実験データには、訓練データに含まれない 1,050 セットのデータを使用した。その内訳は、単一光源データが 300 セット、二つの光源を持つデータが 750 セットである。

客観的評価には、「人間の視覚に基づいた画像類似性評価」SSIM を用いた。単一光源のデータについては、提案手法と ShadeSketch を比較し、平均 SSIM を算出した。一方、二つの光源を持つデータに関しては、比較対象となる既存研究がないため、提案手法の平均 SSIM のみを計算した。

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

SSIM は、以下の記号を用いて定義される。

- x,y：比較対象となる 2 つの画像。
- $\mu x, \mu y$：各画像の平均値。
- $\sigma x, \sigma y$：各画像の分散。
- σxy：2 つの画像の共分散。
- C1, C2：数値が 0 にならないようにする安定化定数。

5.2 結果評価

5.2.1 単一光源

本小節では、提案手法と ShadeSketch との比較を行う。評価には、訓練データには含まれない独立した実験データを使用し、全ての手法を同じ条件のもとで比較する。ShadeSketch の出力は、提供された訓練済みモデルを用いて生成し、いかなる変更が加えていない。

図 5.1 の赤枠で示した部分のように、提案手法は既存研究である ShadeSketch と比較して、線画の形状をより正確に認識し、空白領域への不要な陰影生成を抑制している。

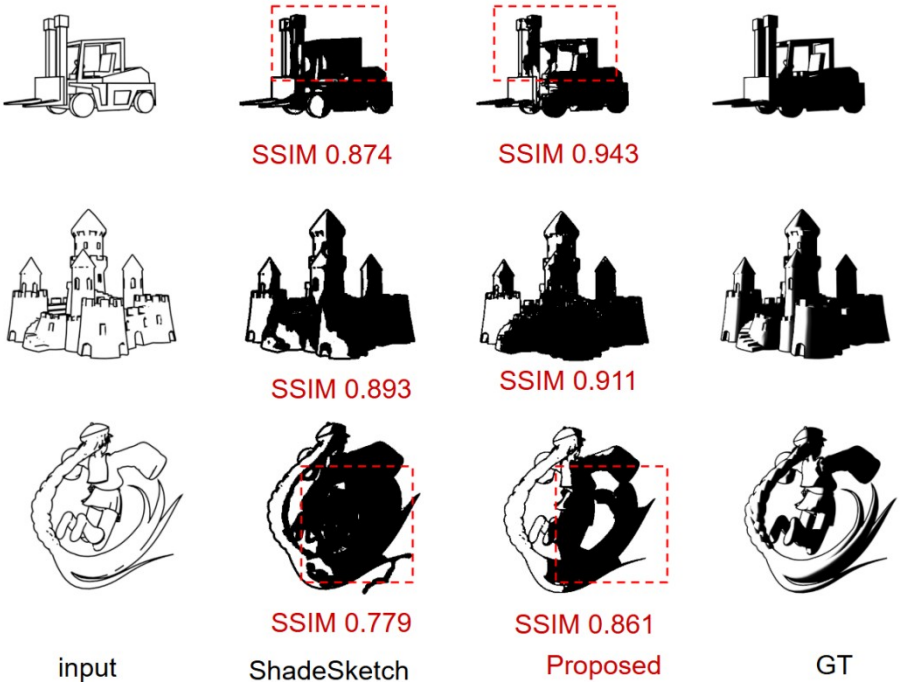

図 5.1：既存研究との比較（光源の数は 1 、光源位置は左）

単一光源条件下において、実験データセットの 300 セットを用いた評価では、ShadeSketch が生成した陰影画像の平均 SSIM は 0.879、提案手法の平均 SSIM は 0.909 となった。数値上、提案手法が生成した陰影画像の Ground Truth 画像との類似度を ShadeSketch より 3.4%向上させた。

図 5.2：空白部分の陰影の獲得方法

実験データにおいて、提案手法は ShadeSketch に比べて空白部分に生成される陰影の総面積を 54.8%減少させた。空白部分の陰影面積の計算は図 5.2 に示したように、生成した陰影画像から Ground Truth の線画マスクを差し引くことで、線画範囲外の陰影を得る。

5.2.2 複数光源

既存研究で処理できない複数光源の入力に対して、本研究のモデルは適用可能である。本実験では、二つの光源を持つデータを入力とし、その生成結果を図 5.3 に示す。

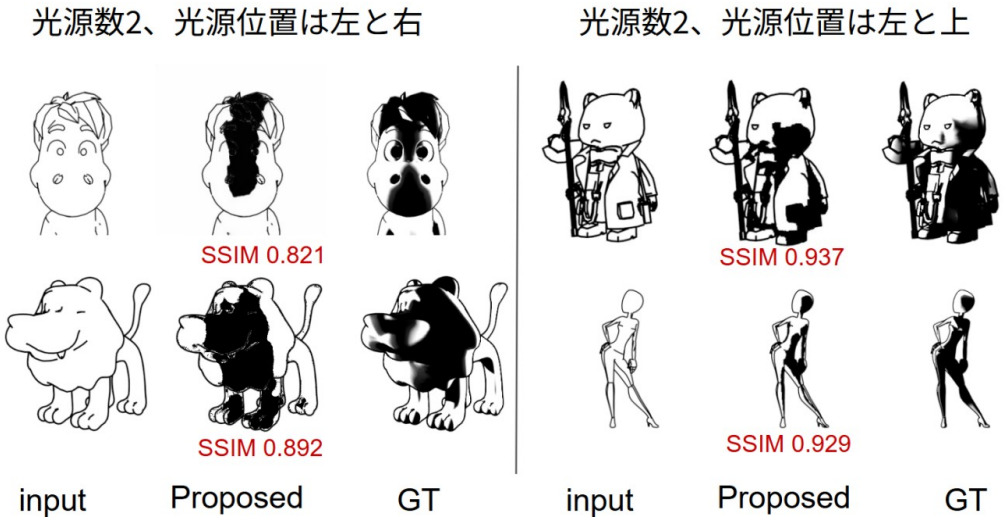

図 5.3：複数光源下の生成効果

実験に使用した 750 セットのデータに対して評価を行った結果、生成された陰影画像の平均 SSIM は 0.897 となった。数値上、単一光源条件と比較すると、多光源条件下での陰影生成精度は低下していることが確認された。

さらに、図 5.2 右上に示すように、SSIM の数値が高いにもかかわらず、視覚的に適切でない陰影が生成されるケースも存在することが分かった。

5.2.3 失敗例の分析

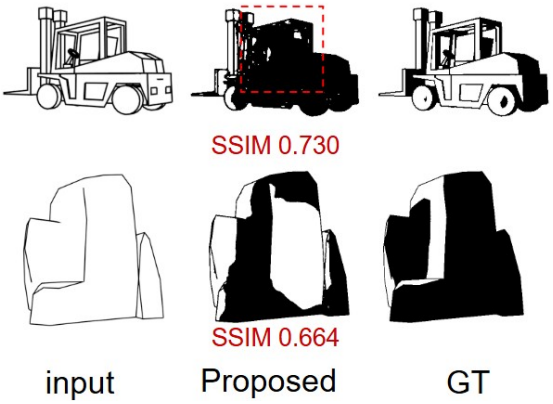

図 5.4：失敗例（光源の数は 1 、光源位置は左）

マスク入力を追加としても、空白部分に陰影を生成する問題を完全的に解決することができない。失敗例から見ると、フォークリフトのように構造が複雑で、または多くのホローを含む物体に対して、構造認識能力が不十分であることがわかる。また、岩のような線画が簡単な物体においても、空間関係を正確に認識できないという問題が存在する。

第 6 章 おわりに

本研究の実験結果により、陰影生成マスクの導入が、特に内部構造が複雑なオブジェクトに対して、物理的整合性のある陰影の生成に有効であることが確認された。また、Latent Diffusion Model（LDM）と Segment Anything Model（SAM）を組み合わせることで、2 次元線画から 3 次元的な構造を推測し、複数の光源条件下の陰影を直接生成することが可能であることを示した。

一方で、本手法は複数の光源が存在する環境下では、単一光源の場合と比較して陰影生成の品質が低下するという問題がある。具体的には、異なる光源間の相互作用を適切に学習することが難しく、ground truth と比べて生成される陰影の一貫性が低下する傾向が見られた。

今後の展望として、さらに実用化に向けて、光源の種類や配置など、より多様な制御要素の導入が求められる。現在のモデルでは点光源のみを考慮しており、スポットライトや面光源、環境光などの異なる照明条件に対応することができない。このため、現実世界での照明環境により忠実に対応できるよう、光源の種類に応じた陰影生成を実現することが必要である。これにより、生成する陰影の表現力を一層高め、より多様なシーンや状況に対応できるモデルを構築することができると期待される。

参考文献

[1] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504−507, Jul. 2006.

[2] D. P. Kingma, "Auto-encoding variational Bayes," arXiv preprint, arXiv:1312.6114, Dec. 2013.

[3] I. Goodfellow, et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems, vol. 27, 2014.

[4] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2017.

[5] T. Karras, "A style-based generator architecture for generative adversarial networks," arXiv preprint, arXiv:1812.04948, 2019.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems, vol. 33, pp. 6840-6851, 2020.

[7] R. Rombach, et al., "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022.

[8] A. Ramesh, et al., "Zero-shot text-to-image generation," in Proc. Int. Conf. Mach. Learn. (ICML), PMLR, 2021.

[9] M. Hudon, M. Grogan, and A. Smolic, "Deep normal estimation for automatic shading of hand-drawn characters," in Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops, 2018.

[10] Y. LeCun, et al., "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.

[11] Q. Zheng, Z. Li, and A. Bargteil, "Learning to shadow hand-drawn sketches," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020.

[12] A. Kirillov, et al., "Segment anything," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023.

[13] B. Henz and M. M. Oliveira, "Artistic relighting of paintings and drawings," Vis. Comput., vol. 33, no. 1, pp. 33-46, 2017.

[14] O. Wang, et al., "Video relighting using infrared illumination," Comput. Graph. Forum, vol. 27, no. 2, Oxford, UK: Blackwell Publishing Ltd, 2008.

[15] T. P. Wu, et al., "Interactive normal reconstruction from a single image," ACM Trans. Graph. (TOG), vol. 27, no. 5, pp. 1-9, 2008.

[16] P. Isola, et al., "Image-to-image translation with conditional adversarial networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), vol. 18, pp. 234-241, 2015.

[18] W. Su, X. Yang, and H. Fu, "Sketch2normal: Deep networks for normal map generation," in SIGGRAPH Asia 2017 Posters, pp. 1-2, 2017.

[19] Sketchfab, "Sketchfab - Your 3D content online," Available: https://sketchfab.com/feed. [Accessed: Jan. 30, 2025].

[20] Z. Liu, et al., "A convnet for the 2020s," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022.