空間的な編集指示に忠実な画像編集モデル

石井里奈[1]　宮森恒[2]

本稿では，空間操作を伴う画像編集の問題に取り組む．近年，Stable DiffusionやDALL-Eをはじめとする画像生成モデルの進化により，ユーザから与えられたテキストに基づいて画像を編集する画像編集技術の研究が大きく進展している．しかし，従来の画像編集モデルでは，色やスタイルの変更に焦点を当てたものが多く，特に空間操作を伴う画像編集において課題が残されている．例えば，オブジェクトの指定した位置への移動や拡大縮小などの空間操作は，従来のモデルでは十分生成結果に反映できないことが多い．そこで，本研究では画像編集モデルにおける空間操作に焦点を当て，ユーザの指定した空間操作を反映する画像編集モデルの実現を目指す．提案手法では，オブジェクトの平行移動や回転，拡大縮小といった空間操作を反映するデータセットを構築し，Stable Diffusionで学習を行う．実験では，モデルに対して編集タスクを行い，編集された画像の評価を行った．

1　はじめに

近年，画像生成技術は目覚ましい発展を遂げ，特にDALL-E [1]やStable Diffusion [2]などの大規模生成モデルが注目を集めている．画像生成モデルとは，テキストを入力すると，それに基づいて画像を生成する技術であり，生成したい画像の内容やスタイルを指定することで，リアルで独創的な画像を作り出すことが出来る．これらの技術は，アート，デザイン，広告，教育といった幅広い分野で応用されている．このような背景から，画像生成技術は今後さらに重要な役割を担うと考えられており，その応用範囲はますます拡大していくことが期待される．

さらに，画像生成技術の応用として，画像編集モデルが注目を集めている．画像編集モデルとは，ユーザから編集指示と画像を入力として受け取り，編集指示に基づいて画像の編集を行うモデルである．画像生成モデルを用いることで，自由度の高い編集が可能になるだけでなく，編集作業の効率化や柔軟性という点で優れた特徴を持ち，クリエイティブな作業を支援するツールとして注目されている．例えば，オブジェクトの色や形状を変える，背景を差し替えるなどの多様な編集が可能である．

しかし，既存の画像編集モデルには，ユーザの意図を正確に反映しきれないという課題が挙げられる．具体的には，複数のオブジェクトの位置関係や空間的な操作を正確に行うことが非常に困難である．そのため，オブジェクトの位置の移動や回転に関する指示に対して，生成される画像が期待通りの結果を示さない場合があり，技術の信頼性や利用価値が損なわれる可能性がある．このような問題を解決することで，ユーザにとって使いやすく，満足度の高いツールを提供することに繋がる．また，画像編集の作業が効率化されるとともに，より広範な分野での利用が可能になると考える．

そこで，本研究では，ユーザからの空間的な編集指示を正確に反映するための新たな画像編集モデルを提案する．具体的には，空間操作を反映するデータセットを構築し，画像編集モデルの追加学習を行った．データセットは，編集前の画像，編集後の画像，編集指示のテキストの3つ組の集合とする．編集後の画像では，オブジェクトの切り抜きを行うため，SAM2（Segment Anything Model） [3]を利用する．画像からオブジェクトを正確に切り出した後，アフィン変換を用いてオブジェクトの空間移動を行い，移動したオブジェクトを背景画像に合成する．ここで，オブジェクトが移動したことによって生じた空白部分を補完するため，Stable Diffusion Inpaintを使用することで空白部分の修復を行い，オブジェクトの空間移動後の自然な画像を生成する．最終的に，編集前の画像，編集後の画像，編集指示のテキストの3つ組を集合としたデータセットを作成し，画像編集モデルInstructPix2Pix [4]に対して追加学習を行う．

実験では，画像編集モデルが空間的な編集指示をどの程度，正確に反映できるかを明らかにすることを目的とする．定性実験と定量実験を実施し，提案手法と従来手法の比較を行った．

本稿の貢献は，以下の通りである．

1. 空間操作を反映した画像編集のためのデータセットを新たに構築した．

2. 構築したデータセットで追加学習することによる画像編集モデルの性能の変化を分析した．

3. 提案モデルは，空間操作を伴う画像編集の性能を向上させることを確認した．

2　関連研究

2.1　画像生成モデルに関する研究

近年，画像生成技術はコンピュータビジョンの分野で重要なトピックとして注目を集めている．これらの画像生成技術は，ユーザからの入力に基づいて高品質な画像を生成することを可能とし，社会や文化にも大きな影響を与えている．近年の画像生成モデルでは，深層学習を利用した生成モデルに関する研究が多く行われており，代表的な3つの生成モデルに分類される．

1つ目は，VAE（変分オートエンコーダ）である．VAEは，訓練データの特徴を学習し，似たような画像を作成する生成モデルである．これらのモデルは，データ内の隠れたパターンや構造を発見し，それを用いて全く新しいデータを創出することができるため，非常に柔軟性が高く，様々な種類のデータに適用できる．

2つ目は，GAN（敵対的生成ネットワーク）である．GANとは，擬似画像を生成するモデルと，擬似画像の正誤を判定するモデルを敵対させて画像を生成するモデルである．GANの応用例として，入力された画像に対応する画像を生成できるPix2Pixや，超高精度な画像を生成することができるStyleGAN [5]が挙げられる．

[1] 学生会員　京都産業大学情報理工学部
i.rinazzzzz00@gmail.com
[2] 正会員　京都産業大学情報理工学部
miya@cc.kyoto-su.ac.jp

3つ目は，拡散モデルである．拡散モデルとは，元の画像にノイズを加え，画像全体をノイズの状態に変換させた後，ノイズ化したデータを逆に適用させるモデルである．元の画像のデータを復元したり，新たな画像を生成したりすることができ，拡散モデルを学習すれば，1つのデータから高品質で多様な画像データを生成することができる．応用技術として，Text-to-Image [6]や，Image-to-Image [7]が挙げられる．また，潜在拡散と呼ばれるアルゴリズムを採用しているStable Diffusionやクリエイティブ性に富んだDALL-E，生成される画像の細部を細かく調整することができるMidjourneyなど，拡散モデルを駆使してユーザの意図を反映する画像生成モデルが多く存在し，高精度な画像生成や変換を実現している．

これらのモデルは，新しい画像をゼロから生成することが目的であり，高解像度でリアルな画像生成や，入力テキストと生成する画像の一致性向上を目指している．しかし，本研究では，ユーザから与えられた画像とテキストに基づく画像編集を行うことに焦点を当て，空間操作を正確に反映することを目的とする．これにより，画像生成技術がさらに実用的なツールとして進化し，様々な分野での応用が期待される．

2.2 テキスト指示に基づく画像編集モデルに関する研究

画像生成技術の急速な発展に伴い，ユーザの編集指示に基づく画像編集に関する研究も盛んに行われている．これまで多くの画像編集手法では，潜在空間操作を利用した編集が行われていた [8, 9]．例えば，StyleGANの潜在空間に画像を変換し，潜在ベクトルを操作することにより，特定のスタイル変換や画像内の要素の変更が可能であった．また，拡散モデルをベースとし，CLIP埋め込みを使ったテキストベースの画像編集も提案されている．これらのモデルは，Text2Live [10]，InstructPix2Pix，CLIPDraw [11]が代表的なモデルであり，画像の意味的な特徴の操作に焦点を当てている．CLIPを使用してテキストと画像を共通の特徴空間に埋め込むため、テキストの意味を反映させた画像操作が可能となる．

これらの研究では，オブジェクトのスタイルに焦点を当てたものが多く，オブジェクトの移動など，空間的な推論を必要とする編集タスクでは十分な成果を挙げられていない [4, 12]．そこで，本研究では，位置関係の調整やオブジェクト間の関係性の編集といった空間操作に焦点を当て，編集タスクの性能向上を目指す．

2.3 空間的な制御を利用した画像生成モデルに関する研究

近年，画像生成において，空間的な制御に注目した研究も盛んに行われている．これらの研究では，テキストから画像を生成する画像生成モデルに空間的な制御を加える方法を提案している．具体的には，オブジェクトの位置を操作するため，視覚的なレイアウトを入力として扱う特定のlocalization componentを訓練する手法 [12, 13]があり，画像内のオブジェクトの配置や位置をより精密に制御できることが示唆されている．また，訓練不要で空間的な条件付けを加える手法も存在する [14]．拡散モデルによるサンプリング過程に空間的な情報を組み込むことで，画像生成時に位置情報を適切に反映させることができるため，ユーザが指定した位置にオブジェクトを配置することが可能になり，空間的な操作を行いやすくなることが示唆されている．

画像編集の分野でも空間的な制御が重要な役割を果たしており，拡散モデルを用いた画像編集タスクでは，空間的推論を必要とするタスクの精度向上を目指す研究が進んでいる．例えば，GLIGEN [13]は，特にオブジェクトの移動や編集に優れた性能を発揮しているが，スタイルや美的分布がシフトしてしまう場合があり，改善の余地があるとされている．

また，画像全体の位置関係を保ちながらオブジェクトを新しい位置へ移動させるモデルとして，Diffusion self-guidance [15]やDragonDiffusion [16]，DiffEditor [17]などの研究が行われている．いずれのモデルも，オブジェクトの移動や編集に優れた性能を発揮しているが，オブジェクトの形状や背景との整合性を保つことが難しい場合がある．

そこで，本研究では空間的な制御に焦点を当て，オブジェクトの空間操作に優れた画像編集を実現するための新たな手法を提案する．また，ドラッグや移動によるオブジェクト操作ではなく，編集指示をテキストとして入力する．

3 提案手法

本稿では，空間操作を伴う画像編集モデルの実現を目指し，空間操作を反映したデータセットの構築を行う．従来の画像編集モデルにおいて不足している空間操作に関する学習データを増加させることで，空間操作を伴う編集能力の向上が期待できる．

3.1 問題設定

本稿では，空間操作を含む画像編集タスクに取り組む．画像編集タスクとは，ユーザから編集指示と画像を入力として受け取り，編集指示を反映した画像を出力するタスクである．この画像編集タスクにおいて，本研究で取り扱う空間操作の範囲を表1に，具体例を図2にそれぞれ示す．空間操作は，与えられた画像に対して2次元の空間操作を行うものとし，それぞれの単位はピクセル（平行移動），操作前を1としたときの倍率（拡大縮小），度（回転）とする．

表1: 本稿で扱う空間操作の範囲

空間操作	操作量
平行移動（x軸）	50 〜 100ピクセル
平行移動（y軸）	50 〜 100ピクセル
拡大	1.1 〜 2.0 倍
縮小	0.5 〜 0.9 倍
回転	-90 〜 90度

3.2 データセットの構築

データセットは，編集前の画像，編集後の画像，編集指示の3つ組の集合とする．データセットの構築手順を図1に示す．編集前の画像として，COCO [18]の画像を使用し，各画像に対してオブジェクトの空間操作を行った画像を編集後の画像とする．

3.2.1 編集後の画像生成

編集後の画像は，オブジェクトの切り抜き，移動と合成，修復の3段階の手順で作成する．まず，SAM2（Segment Anything Model） [3]を用いて画像内の移動対象となるオブジェクトを切

図1: データセットの構築手順

図2: 本稿で扱う空間操作の種類

り出す．SAM2とは，画像や動画のオブジェクトを一貫してセグメント化するセグメンテーションモデルである．本稿では，COCO [18]で提供される移動対象オブジェクトのbbox情報と画像をSAM2に与え，移動対象オブジェクトのみの領域分割結果を取得し，オブジェクトの切り抜きを行った．

次に，切り抜いたオブジェクトの空間移動と合成を行う．表 1 で示す空間操作の範囲内でアフィン変換を適用することでオブジェクトを移動させ，背景画像と合成する．

最後に，合成画像に対してインペインティング（inpainting）を行う．合成後の画像には，移動したオブジェクトが移動前の位置で占めていた領域に，多くの場合，空白が残る．この空白部分を補うため，画像内の欠損部分を予測し，再構成する技術であるインペインティングを行う．本研究では，Stable Diffusionベースのインペインティングモデル [19]を使用し，画像の修復を行う．

3.3 編集指示の作成

編集指示は，空間操作の種類と操作量を指定し，自動的に生成する．また，多様な編集指示を作成するため，画像内に存在する移動対象のオブジェクト数に基づいて編集指示を決定する．

例えば，移動対象のオブジェクトが画像内に1つしかない場合，対象のオブジェクトを特定する必要がないため，空間操作の種類，操作量のみを使用したシンプルな編集指示を作成する．例えば，「クマを右に移動させて」といった，簡潔な指示を作成する．画像内に存在するオブジェクトが1つの場合における編集指示の例を表2に示す．一方，移動対象のオブジェクトが複数存在する場合，どのオブジェクトを移動させるのかを明確にし，複数のオブジェクト間での位置関係を考慮した編集指示を作成する必要がある．例えば，画像内にクマが3匹いた場合，「真ん中のクマを右に移動させて」というように，オブジェクトの座標位置を考慮した編集指示を作成した．

3.4 画像編集モデルの訓練

本稿では，構築したデータセットを用いて画像編集モデルInstructPix2Pix [4]に対して追加学習を行う．InstructPix2Pixとは，テキストの指示により画像を編集する，拡散モデルベースの画像編集モデルである．このモデルは，オブジェクトの色や形状，スタイルの編集を得意としており，ユーザーの意図により忠実な編集を実現することができる．

追加学習ではInstructPix2Pixと同様に，テキストによる編集指示を潜在空間に組み込み，その指示に対応した画像が生成されるようにStable Diffusion [2]の学習を行う．学習に使用する画像生成モデルStable Diffusionは，画像生成の際に潜在空間を利用することで，生成された画像とテキストの関係を強化することができる．そのため，学習においてテキストに基づく編集が正確に反映される可能性が高い．また，編集指示と編集後の画像を潜在空間における制約条件として取り入れ，生成される画像が編集指示に従ったものとなるようにする．これにより，単なる画像生成ではなく，ユーザの意図に即した編集を行うことができる．

本研究で構築した，空間操作に関するデータセットを以後，Spatial-COCOと呼ぶこととする．Spatial-COCOデータセットの内訳を表3に示す．このデータセットを用いてInstructPix2Pixの追加学習を行った．

4 実験

4.1 実験目的

本実験では，画像編集モデルが空間的な編集指示をどの程度正確に反映できるかを明らかにすることを目的とする．

4.2 実験設定

Spatial-COCOデータセットにより，事前学習済みInstructPix2Pixモデルの追加学習を行った．学習率は 10^{-4} とし，バッチサイズは32，エポック数は100とした．また，

表2: 空間操作に基づく画像編集指示の分類

空間操作	操作量	編集指示
平行移動（x軸）	pixel > 0	Move the [object name] (a little) to the right. Shift the [object name] (a little) to the right.
	pixel < 0	Move the [object name] (a little) to the left. Shift the [object name] (a little) to the left.
平行移動（y軸）	pixel < 0	Move the [object name] (a little) down. Shift the [object name] (a little) downward.
	pixel > 0	Move the [object name] (a little) up. Shift the [object name] (a little) upward.
拡大		Magnify the [object name] [multiple] times. Scale up the [object name] by a factor of [multiple] times.
縮小		Reduce the [object name] by [multiple] times. Scale down the [object name] by a factor of [multiple] times.
回転	degree > 0	Rotate the [object name] [degree] degrees counterclockwise.
	degree < 0	Rotate the [object name] [degree] degrees clockwise.

※ pixelはピクセル数，multipleは倍数，degreeは角度を表す．なお，画像サイズの10%未満の平行移動を行う際は‘a little’を加えて指示を与えた．

表3: 構築したSpatial-COCOデータセットの内訳

	平行(x)	平行(y)	拡大	縮小	回転	total
train	1,715	928	908	212	434	4,197
test	215	87	93	28	43	466
total	1,930	1,015	1,001	240	477	4,663

学習時の解像度は 256×256とし，推論時は解像度が 512×512 となるように生成した．その他の学習設定は，公開されているStable Diffusionおよび先行研究のInstructPix2Pixと同じにした．samplerには，Karrasらによって提案されたEuler ancestral sampler [20]を100ステップ使用した．

4.3 比較手法

実験では，Spatial-COCOデータセットにより，事前学習済みInstructPix2Pixモデルの追加学習を行った提案手法によるモデルと，従来手法InstructPix2Pixによるモデルを比較する．InstructPix2Pixは，テキストの指示により画像を編集する拡散モデルベースの画像編集モデルであり，色や質感，スタイルの変更を得意としている．

4.4 評価指標

定性実験

定性実験では，20代の男女10名で，「生成された画像が編集指示を守れているか」「生成された画像が自然であるか」「編集前の画像と生成された画像で，スタイルに一貫性があるか」の3項目で「あてはまる」「ややあてはまる」「どちらとも言えない」「あまりあてはまらない」「まったくあてはまらない」の5段階評価を行った．

定量実験

定量実験では，評価指標としてCLIP Image SimilarityとCLIP Text-Image Direction Similarityを使用する．CLIP Image Similarityは，編集後の画像と編集前の画像のコサイン類似度を測る指標である．これにより，編集後の画像と編集前の画像がどの程度一致しているかを評価する．一方，CLIP Text-Image Direction Similarity [21]は，編集による画像の変化と画像キャプションの変化のコサイン類似度を測る指標である．これにより，指示文に対する画像の変化が，どれだけ指示通りに行われたかを測定する．画像キャプションの変化は，編集前画像，編集後画像のそれぞれの画像に対するキャプション埋め込みの差である．編集前後画像のキャプションは，clip-filtered-dataset [4]で提供されるキャプションを使用する．2つの指標は，互いにトレードオフの関係にあり，編集後の画像が編集指示を正確に反映するほど，編集前の画像との類似度は低下する傾向にある．これらの指標により，モデルが編集前の画像の特徴を保ちながら，意図した編集を正確に反映できているか，評価を行った．また，実験では，InstructPix2Pixで使用されているclip-filtered-dataset [4]を使用した．このデータセットには，313,010件の画像が含まれており，編集前の画像，編集後の画像，編集指示の3つ組で構成されている．これらのデータは，オブジェクトのスタイル変更に関する編集指示が多く含まれており，空間的操作を含まないものが多い．本実験では，提案手法と従来手法による編集能力の違いを定量的に分析するため，ランダムに選んだ，全体の1%にあたる3,130件を使用し，評価を行った．

5 結果と考察

定性実験の結果を図3に，空間操作ごとの定性実験の結果を表4に示す．図3の結果から，提案手法は「編集指示への忠実さ」において，従来手法より20ポイント以上正確に反映できるよう改善されたことがわかる．また，それぞれのモデルで編集を行った結果を図4に示す．これらの結果から，提案手法は編集前の画像の特徴を保ちつつも，空間的な操作を反映した編集が行えている

表4: 空間操作ごとの定性実験の結果

空間操作	評価項目	従来手法	提案手法
平行移動（x軸）	編集指示への忠実さ	1.52	**2.32**
	画像の自然さ	**3.89**	3.46
	スタイルの一貫性	4.30	**4.31**
平行移動（y軸）	編集指示への忠実さ	1.59	**2.47**
	画像の自然さ	**3.29**	2.98
	スタイルの一貫性	3.69	**3.83**
拡大	編集指示への忠実さ	1.90	**2.62**
	画像の自然さ	**3.93**	3.85
	スタイルの一貫性	3.62	**3.84**
縮小	編集指示への忠実さ	1.65	**2.64**
	画像の自然さ	**4.11**	3.8
	スタイルの一貫性	4.06	**4.24**
回転	編集指示への忠実さ	1.32	**1.66**
	画像の自然さ	**3.04**	2.64
	スタイルの一貫性	2.69	**2.87**
全て	編集指示への忠実さ	1.60	**2.37**
	画像の自然さ	**3.72**	3.39
	スタイルの一貫性	3.89	**3.99**

ことがわかる．一方で，提案手法における「画像の自然さ」の評価が，従来手法よりも低い傾向にある．表4の結果からも，全ての空間操作における「画像の自然さ」の評価項目が，従来手法を下回る結果となった．このことから，提案手法は，編集指示の忠実な反映を重視するあまり，画像全体の自然さや調和を損なった可能性が考えられる．

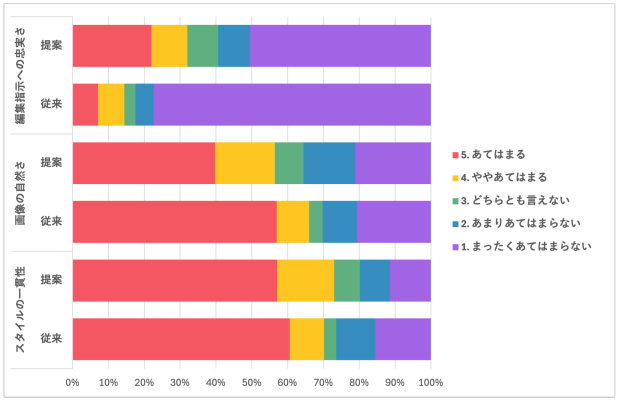

図3: 定性実験の結果

まず，提案手法の「画像の自然さ」が低下した原因として，学習データの質が影響しているのではないかと考える．学習データにおけるインペインティングの失敗例を図5に示す．図5のように，オブジェクトの移動により生じた空白部分をインペインティングした画像に，残像や黒い影が残る場合が多かった．このような質の悪いデータにより学習を行ったことで，画像全体の自然さや調和を損なったと考えられる．

図4: 編集結果の例

次に，従来手法が「画像の自然さ」において高評価を得た理由として，従来手法が入力画像を複製したものを出力する傾向にあったことが挙げられる．従来手法では，図6に示すように，入力された画像を単に複製し，同じ内容の画像が編集後の画像として出力されることが多かった．

編集指示の種類ごとの高品質な画像・低品質な画像・複製した画像の分布を表5と表6に示す．ここでは，評価者の平均点が全ての評価項目で3点以上であった場合を“高品質な画像”，評価者

図5: Spatial-COCOデータセットの例

の平均点が全ての評価項目で3点未満であった場合を“低品質な画像”，評価者の平均点が忠実さで2点以下，自然さで4点以上，スタイルの一貫性で4点以上の場合を“複製した画像”と定義している．この結果から，従来手法は，複製した画像枚数が提案手法よりも多いことがわかる．編集前の画像を複製し，編集結果として出力することで，編集前後の画像間に大きな差異が生じることが少なく，結果として「画像の自然さ」の評価が高くなったと考えられる．

図6: 編集結果の例（従来手法）

表5: 編集指示の種類ごとの画像品質（従来手法）

	平行（x）	平行（y）	拡大	縮小	回転
高品質な画像	0	0	2	0	0
低品質な画像	34	30	18	1	18
複製した画像	110	29	25	13	3

表6: 編集指示の種類ごとの画像品質（提案手法）

	平行（x）	平行（y）	拡大	縮小	回転
高品質な画像	67	23	22	9	0
低品質な画像	12	12	9	2	23
複製した画像	30	5	11	5	0

定量実験の結果を図7に示す．従来手法では，編集前後の画像間の類似度が低く，主にスタイル変更を伴う編集指示に従って画像が変換されていることが確認された．また，編集前後の画像変化とキャプション変化との類似度が高いことからも，スタイル変更が反映された画像生成が行われていることが伺える．

一方，提案手法においては，編集前後の画像間の類似度が高く，スタイル変更に関する編集指示に従った画像変換が十分に行われていないことが示唆された．さらに，編集前後の画像変化とキャプション変化との類似度が低いことからも，スタイル変更の指示への追従性が低下していると考えられる．

今後は，clip-filtered-datasetで提供される編集前後のキャプションを利用した評価ではなく，空間的操作を伴う編集をより適切に扱える評価手法を検討する必要がある．

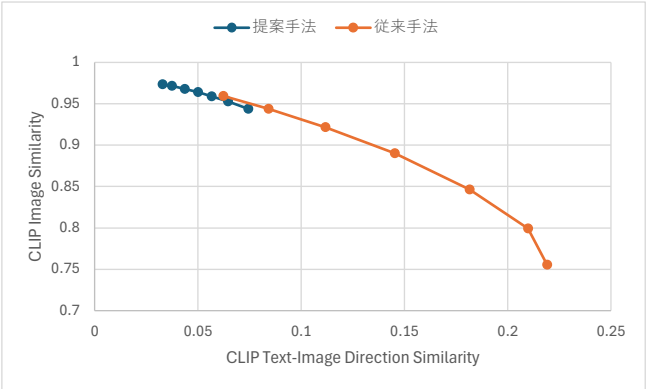

図7: 定量実験の結果

6 まとめ

本稿では，空間操作を伴う画像編集の問題に取り組んだ．従来の画像編集モデルでは，色やスタイルの変更に焦点を当てたものが多く，特に空間操作を伴う画像編集において課題が残されていた．そこで，本研究では画像編集モデルにおける空間操作に焦点を当て，オブジェクトの平行移動や回転，拡大縮小といった空間操作を反映するSpatial-COCOデータセットを新たに構築した．そして，構築したデータセットを用いてInstructPix2Pixの学習済みモデルに対して追加学習を行った．

実験では，画像編集モデルが空間的な編集指示をどの程度，正確に反映できるかを明らかにすることを目的とした．定性実験と定量実験において，従来手法と提案手法の比較を行った．定性実験の結果として，提案手法は「編集指示への忠実さ」と「スタイルの一貫性」において，従来手法より正確に反映できるよう改善されたが，「画像の自然さ」の評価が，従来手法よりも低い傾向にあった．定量実験の結果，提案手法は，空間的操作の学習に伴い，主にスタイル変更に関する従来の編集指示に従った画像変換能力は低下したことを確認した．一方，空間的操作の定量的評価において，編集前後のキャプション変化を用いるCLIP Text-Image Direction Similarityは十分な指標とは言えず，今後はこの点を改良する必要があることがわかった．

今後は，データセットの品質向上や他の画像編集モデルでの検証，評価指標の改良を行う予定である．

謝辞

本研究では，CC BY 4.0 で提供されているMS COCOデータセット [18]の画像を使用している．また，本研究の一部は科研

費23K11342の助成を受けたものである.

参考文献

[1] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, June 2022. IEEE Computer Society.

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023.

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023.

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228, December 2021.

[6] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, In So Kweon, and Junmo Kim. Text-to-image diffusion models in generative ai: A survey, 2024.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

[8] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[9] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H. Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing, 2021.

[10] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723. Springer, 2022.

[11] Kevin Frans, Lisa B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *ArXiv*, abs/2106.14843, 2021.

[12] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18370–18380, June 2023.

[13] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22511–22521, 2023.

[14] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *Proceedings of Machine Learning Research*, 202:1737–1752, 2023. 40th International Conference on Machine Learning, ICML 2023 ; Conference date: 23-07-2023 Through 29-07-2023.

[15] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. 2023.

[16] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8839–8849, 2023.

[17] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. *arXiv preprint arXiv:2402.02583*, 2023.

[18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[19] Stable diffusion web ui. `https://github.com/AUTOMATIC1111/stable-diffusion-webui`, 2023.

[20] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022.

[21] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators, 2021.