U-Net 内の中間表現操作による生成画像の一貫性を確保する手法の検討

Investigation of a Method to Ensure Consistency of Generated Images Through Intermediate Representation Operations in U-Net

山田　涼太
Ryota Yamada

広島市立大学 言語音声メディア工学研究グループ
Language and Speech Research Group, Hiroshima City University

概要: Stable Diffusion の登場から画像生成が急激に注目を集め始めている. その中でも画像生成を用いた動画制作への関心が高まっている. 画像生成を用いた動画制作の大きな壁の一つに画像の一貫性の確保が難しいことが挙げられる. これを改善する手法の多くが入力やタスクに応じた再学習をしており既存の技術との併用が難しい場合もある. そこで本研究では再学習を必要とせず画像の一貫性を確保する手法について検討した. 本研究の提案手法は, 画像の生成中の U-Net 内の中間表現を保存し, その結果を別の画像の生成時に直接利用し操作を行うものである. 異なる二種のボトムスを着用した画像を生成するタスクを行い画像の一貫性を検証した. 視覚評価と数値評価の結果, 構図や背景の変化を抑制することができた.

1. はじめに

Stable Diffusion は, Latent Diffusion Model [Rombach 21]という生成モデルを基にして学習した, 高精度かつ高速で画像の生成が可能なオープンソースソフトウェアである. その Stable Diffusion の公開から画像生成技術が急激に注目を集めており, 広告やゲームなど, 画像生成を活用した事例を見かけることも多くなってきている.

最近ではそのような画像生成という分野の中でも, 特に動画制作への関心が高まっている. 画像生成を用いた動画制作は, 従来の動画制作と比べ, 大幅なコスト削減を期待できる. しかし課題の一つとして画像の一貫性の確保が難しいことが挙げられる. これを改善する手法の多くが, 入力やタスクに応じた再学習をしており, 既存の技術との併用が難しい場合もある. そこで本研究では, 再学習を必要とせず画像の一貫性を確保する手法について検討した.

本稿で提案する手法は, 画像生成中の中間表現を保存し, その保存済み表現を別の画像の生成時に直接利用し操作を行うものである. 異なる 2 種の衣服を着用した画像を生成するタスクを行い, 画像の一貫性の検証を行う.

2. Stable Diffusion

Stable Diffusion は 2022 年に公開された Latent Diffusion Model を基にした画像生成モデルである. コードと重みが一般公開されており, 無制限に利用できる画像生成モデルとして大変話題となった.

Latent Diffusion Model は従来の Diffusion Model [Sohl-Dickstein 15]の欠点であった膨大な計算量を改善するため, 変分オートエンコーダ[Kingma 13]を用いている. 変分オートエンコーダを用いることで, 従来のものでは, ピクセル単位で必要であった計算を Latent Space に圧縮した上で行うことで計算量を減らすことができた.

Stable Diffusion は, ガウシアンノイズから段階的にノイズを除去することで画像を生成する. この除去するノイズを出力するモデルが U-Net [Ronneberger 15]である. 図 1 は Stable Diffusion における U-Net の構造である. 赤色で示した層はテキスト情報の入力がある層を示しており, 青色で示した層は入力がない層を示している. このテキスト情報は CLIP [Radford 21]でエンコードしたものであり, Cross Attention 機構で取り込まれる.

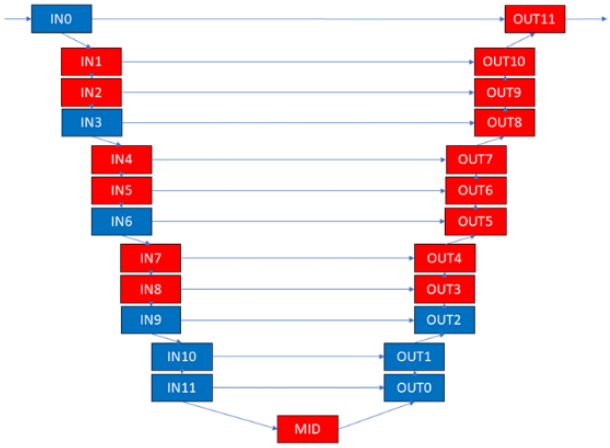

図 1 Stable Diffusion における U-Net の構造

3. 提案手法

本稿で提案する手法は, 複数のプロンプトを入力とする. まず第 1 のプロンプトで画像の生成を行った後, 第 1 のプロンプトで生成した画像との一貫性を保ちながら, 第 2 以降のプロンプトで画像を生成する. 本手法の手順を以下に示す.

1. 第 1 のプロンプトでの画像生成時に, 各ステップにおける U-Net 内の各層での中間表現を保存する.
2. 第 2 以降のプロンプトで画像生成する際の中間表現を, 保存済みのステップおよび層が対応する中間表現と入れ替え生成を行う.

本稿の実験では, 入れ替える中間表現のステップ, 層, 範囲について比較検討を行う.

4. 実験

本研究の目的は, 生成画像の一貫性の確保である. そのため 本実験では, 異なる 2 種の衣服を着用した画像を生成するタスクを行い, 画像の一貫性の検証を行った.

連絡先: 山田涼太, 広島市立大学,
mj67014@e.hiroshima-cu.ac.jp

4.1 実験設定

本実験では，diffusers ライブラリ[1]を用いて画像生成を行った．画像生成の設定を表 1 に示す．第 2 のプロンプトが，第 1 のプロンプトと違っている点は“school uniform，blue sailor collar，blue skirt”が” pants，dress shirts”となっている点のみである．

表 1 実験設定

項目	設定
使用モデル	waifu-diffusion
ステップ数	50
画像サイズ	640*896
guidance scale	7.5
シード値	1111
第 1 のプロンプト	masterpiece, best quality, girl, white hair, twintail, green eyes, school uniform, blue sailor collar, blue skirt, outdoor
第 2 のプロンプト	masterpiece, best quality, girl, white hair, twintail, green eyes, **pants, dress shirts,** outdoor
共通のネガティブプロンプト	nsfw, worst quality, low quality, medium quality, deleted, lowres, comic, bad anatomy, bad hands, text, error, missing fingers, extra digit, fewer digits, cropped, jpeg artifacts, signature, watermark, username, blurry

4.2 予備実験

本研究では予備実験として，U-Net 内のテキスト情報の入力がある層のうち，中間表現の入れ替えにより生成画像に特に大きな影響を与える層について調査した．調査対象の層を絞り込むのは，全てのテキスト情報の入力がある層で調査を行うと 2^16 枚の生成画像を出力する必要があり，全ての画像に対し目視による評価（視覚評価）することが困難なためである．

図 2 は，実験設定で提示した第 1 プロンプトで生成した画像とそのプロンプトの“green eyes”を“red eyes”に変更して生成した画像である．以降では，この変更後プロンプトを“予備実験プロンプト”と記述する．図 2 より，眼の色の指定を変更しただけにも関わらず，手の構図まで大きく異なっていることが確認できる．そこで予備実験プロンプトでの画像生成中に，全てのステップにおいて“U-Net の各層での中間表現”を“第 1 プロンプトでの画像生成中の対応する層の中間表現”に順次入れ替え，画像生成を行う．そして生成した画像の手の構図が第 1 プロンプトで生成した画像の手の構図と同じであった場合，入れ替えを行った層での影響が大きいと定める．

図 3 は予備実験の結果である．IN1, 2, 4, 5, 7, MID で入れ替えを行い生成した画像は，予備実験プロンプトで生成した画像と大きく異なっている点はない．また，OUT7, 8, 9, 11 で入れ替えを行い生成した画像は，第 1 プロンプトで生成した画像で手の描かれていた位置が少し白くなり，その位置の背景が予備実験プロンプトで操作なしに生成した画像と異なっている．しかし，手の構図は予備実験プロンプトで生成した画像と同じであるため，本実験では影響の大きな層としては扱わない．一方，IN8, OUT3, 4, 5, 6, 10 で入れ替えを行い生成した画像は第 1 プロンプトで生成した画像と手の構図が同じであるため，本実験ではこの 6 層を“生成画像の一貫性維持に影響の大きい層”とみなし，調査対象層として定める．

図 2 中間表現操作なしに生成した画像

図 3 予備実験結果

4.3 評価実験

本稿では中間表現を入れ替える範囲を，①全体，②視覚的に選択した範囲，③閾値で選択した範囲の 3 パターン設定して実験を行った．本節，次節では各実験について説明する．

①は，ステップ 1～5, 10, 15, 20, 25, 30, 35, 40, 45, 50 で，調査対象層（IN8, OUT3, 4, 5, 6, 10）の全ての組み合わせで，中間表現全体の入れ替えを行い，画像を生成し調査を行う．

②では，①の実験結果に基づいて設定したステップおよび層に対して中間表現全体の入れ替えを行った後，それ以降のステップにおいて，一貫性を維持したい画像領域に限定して調査対象層における中間表現を入れ替えることで画像を生成する．そして生成した画像について比較調査を行う．図 4 は作業者 1 名が目視にて選択した“中間表現の入れ替える範囲に対応する位置”を生成画像上に示している．黒く塗りつぶされていない範囲が第 1 プロンプトの中間表現と入れ替える範囲である．10*14 に分割して選択しているのは，640*896 のサイズで画像を生成した場合，最下層である MID での中間表現のサイズが 10*14 であり，どの層でも同じ範囲で選択できるようにするためである．

③の閾値での選択は，「第 1 のプロンプトでの生成と第 2 以降のプロンプトでの生成とで変更したい要素の存在している位置は中間表現の値の差も大きい」という仮説の下行った．保存済みの中間表現と生成中の中間表現との差をとり，その差が閾値未満である位置の中間表現の要素のみ入れ替えを行う．③は，①の結果に基づいて設定したステップおよび層に対して中間表

[1] https://huggingface.co/docs/diffusers/index

現全体の入れ替えを行った後，それ以降のステップにおいて②の結果に基づいて設定した層で閾値に基づき選択した範囲の入れ替えを行い，画像を生成し調査を行う．閾値の設定は 0.01 から 1.00 まで 0.01 刻みで実験を行った．

また，評価は，まず視覚評価を行い，評価のポイントを満たした生成画像に対し，数値評価を行った．視覚評価のポイントは，「画像内の人物が大きな違和感のないズボンを着用しているか」と「人物のポーズが第 1 のプロンプトで生成した画像と比べ変わっていないか」の 2 点とした．数値評価は，生成画像の背景のみを対象に MSE（平均二乗誤差）を求めることで行った．背景の検出には，Rembg ライブラリ[1]を利用した．

図 4 視覚的に選択した範囲

4.4 実験結果

前節で述べたように，本研究では中間表現を入れ替える範囲を①全体，②視覚的に選択した範囲，③閾値で選択した範囲の 3 つに設定して実験を行った．

①で視覚評価のポイント 2 点を満たした生成画像に対し，数値で評価した結果を表 2 に示す．縦軸は中間表現を入れ替えた U-Net の層を，横軸は中間表現を入れ替えたステップの範囲をそれぞれ示している．①で最良の結果となったのは，ステップ 20 まで OUT3, 4 で中間表現全体を入れ替えて生成した画像であり，数値評価の結果は，74.5（背景の MSE）となった．生成された画像を図 5 に示す．

②では全ての生成画像が視覚評価のポイント 2 点を満たしたため，全ての生成画像に対して数値で評価した結果を表 3 に示す．「入れ替える層」に書かれたビット列は，左から順に IN8, OUT3, 4, 5, 6, 10 にあてており，1 の時はその層で入れ替えを行っており，0 の時はその層での入れ替えを行っていないことを示している．②で最良の結果となったのは，ステップ 20 まで OUT3, 4 で中間表現全体を入れ替え，それ以降のステップでは，OUT3, 4, 5, 6, 10 で視覚的に選択した範囲の中間表現を入れ替え生成した画像であり，数値評価の結果は，62.0（背景の MSE）となった．生成された画像を図 6 に示す．

③で視覚評価のポイント 2 点を満たした生成画像に対し，数値で評価した結果を表 4 に示す．③で最良の結果となったのは，ステップ 20 まで OUT3, 4 で中間表現全体を入れ替え，それ以降のステップでは，OUT3, 4, 5, 6, 10 で，閾値 0.13 で選択した範囲の中間表現を入れ替え生成した画像であり，数値評価の結果は，67.0 となった．生成された画像を図 7 に示す．

[1] https://github.com/danielgatis/rembg

表 2 ①の数値評価の結果（MSE）

中間表現を入れ替えた U-Net の層	中間表現を入れ替えたステップの範囲		
	1～15	1～20	1～25
OUT3, OUT4	81.5	**74.5**	74.9
IN8, OUT3, OUT4	81.1	79.7	78.4

表 3 ②の数値評価の結果（MSE）

入れ替える層	背景のMSE		
000001	69.2	100000	75.2
000010	70.3	100001	68.1
000011	65.4	100010	72.0
000100	71.5	100011	64.9
000101	65.7	100100	74.4
000110	68.8	100101	65.0
000111	63.9	100110	70.5
001000	74.9	100111	64.4
001001	65.5	101000	74.9
001010	69.4	101001	65.1
001011	64.8	101010	70.5
001100	72.3	101011	64.6
001101	65.3	101100	73.6
001110	71.0	101101	65.2
001111	64.2	101110	70.4
010000	76.3	101111	64.7
010001	67.8	110000	75.5
010010	71.0	110001	68.0
010011	66.4	110010	72.6
010100	73.7	110011	66.3
010101	65.2	110100	74.4
010110	70.3	110101	66.5
010111	65.3	110110	72.0
011000	74.8	110111	66.3
011001	67.0	111000	76.1
011010	72.5	111001	66.4
011011	64.0	111010	73.0
011100	73.5	111011	64.7
011101	64.2	111100	73.3
011110	69.9	111101	64.2
011111	62.0	111110	70.8
		111111	62.7

表 4 ③の数値評価の結果（MSE）

閾値	背景のMSE
0.01	73.1
0.02	72.2
0.03	71.0
0.04	70.1
0.05	69.1
0.06	69.2
0.07	68.7
0.08	68.8
0.09	68.9
0.10	69.4
0.11	68.4
0.12	68.9
0.13	67.0
0.14	67.3
0.15	68.3
0.16	69.0
0.17	67.5
0.18	68.3
0.19	68.0

図 5　①で最良の生成画像

図 6　②で最良の生成画像

図 7　③で最良の生成画像

以上の結果より，1 枚絵として見た時のポーズや背景などの一貫性は，①，②，③のどの実験で生成した画像にも確保できているものは存在することがわかった．また，MSE による数値評価の結果，②の手法が最も背景の一貫性のある画像を生成することができた．

5. まとめと今後の課題

本稿では，画像生成中の中間表現を保存し，その保存済み表現を別の画像の生成時に直接利用し操作を行うことで，画像の一貫性を確保する手法を検討した．

視覚的に選択した範囲で中間表現を入れ替える手法が最も良い結果となった．しかし，これは手動による操作が必要なため，大量の画像を生成しなければならない場面では有効な手法とは言えない．

今後は，閾値で選択した範囲で中間表現を入れ替える手法について改良を行う予定である．

参考文献

[Kingma 13] Kingma, D. P., and Welling, M.: Auto-Encoding Variational Bayes, arXiv: 1312.6114v11, 2013.

[Radford 21] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision, arXiv: 2103.00020v1, 2021.

[Rombach 21] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models, arXiv: 2112.10752v2, 2021.

[Ronneberger 15] Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp.234-241, 2015.

[Sohl-Dickstein 15] Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics, arXiv: 1503.03585v8, 2015.