SHONAN INSTITUTE OF
TECHNOLOGY JOURNAL
Vol. 58, No. 1, 2024

拡散モデルに基づく画像領域分割タスク高精度化の検討

安藤　慎吾*

Improvement of Image Semantic Segmentation Using Diffusion Models

Shingo ANDO

Abstract:
Semantic segmentation is the task of assigning each pixel in an image to the object or region it represents. Despite the significant improvement in the accuracy of semantic segmentation by the progress of deep learning technologies in recent years, further technological ingenuity is required to accurately segment the boundaries of objects in complex backgrounds. In this paper, we propose a method to improve the inference results of semantic segmentation by introducing diffusion models, which are known as fundamental technologies for image generation AI.

Keywords : Semantic segmentation, U-Net, Diffusion model, DDPM

要旨:
セマンティック・セグメンテーションは，画像内の各ピクセルを，そのピクセルが表す物体または領域に割り当てるタスクである．近年の深層学習技術の進歩により，セマンティック・セグメンテーションの精度は大きく向上したが，複雑背景下で対象物体の輪郭を正確に分割するには，より一層の技術的工夫が必要となる．本論文では，画像生成AIの基幹技術として知られる「拡散モデル」を導入することで，セマンティック・セグメンテーションの推論結果を改善する手法を提案する．

キーワード : セマンティック・セグメンテーション，U-Net，拡散モデル，DDPM

1. はじめに

セマンティック・セグメンテーションは，画像内の各ピクセルを，そのピクセルが表す物体または領域に割り当てるタスクである．具体的には，さまざまなシーンを撮影した画像から，道路，歩行者，車両，標識などの物体や，建物，木，草などの自然物などを分類し，それらの領域を詳細に切り出すことを目的としており，これまでにさまざまな手法が検討されてきた．近年では深層学習技術の進歩が目覚ましく，セマンティック・セグメンテーションの精度は深層学習の発展とともに大きく向上してきた経緯がある．そして，その汎用性の高さから，自動運転，医療画像診断，衛星画像解析，ロボティックビジョンなど幅広い分野での応用が期待されている．

しかし，セマンティック・セグメンテーションの精度は，現状でも未だにユーザーの満足するレベルに至っているとは言い難い．車の自動運転を例に挙げると，車の影となる部分のタイヤとアスファルトの境界線など，位置関係が明確でなく，なおかつ色合いの似たものなどを正確に判別して切り出すことが難しいとされている．正確な切り出しができないと，その車がどのくらい離れているかを推測する際に大きな誤差を生んでしまう可能性がある．そのため，安全な運転を保証するにあたり，正確な切り出しは重要な問題であると考えられる．

より明確な技術課題としては，以下の 2 つの項目に整理することができる．

・複雑背景下での物体の細かい部分の識別
・物体同士の重なりが存在するシーンでの識別

*湘南工科大学
情報学部　情報学科　准教授

前者は，例えば草むらの中に隠れている動物などの複雑な輪郭線を精度良く検出することが具体的な課題になる．後者は，道路と歩行者が重なっている場合や，歩行者の境界を正確に認識する場合などで，物体が重なっている部分の境界が不明瞭になり，識別が困難になるケースが存在する．このような状況でも安定した物体切出しが求められることがある．

本稿では，セマンティック・セグメンテーションの高精度化に対するアプローチとして生成モデルを活用することを検討する．生成モデルとはテキストや静止画，動画像などさまざまなデータをより自然なかたちで自動生成するための技術である．実用例として ChatGPT や Stable Diffusion などのアプリを挙げることができ，近年飛躍的な進歩が見られた技術として一般にも広く認知されるようになった．

生成モデルは識別モデルとは対照的な概念と位置づけることができる．識別モデルは，与えられたデータがどのクラスに属するかを判断する技術である．一方で生成モデルは，データがどのように生成されるのかを解析する技術である．そのため，セマンティック・セグメンテーションは従来から識別モデルを用いて推論する方式が検討されてきた．しかし，セマンティック・セグメンテーションはその出力結果が画像の形態をなすため，生成モデルとの相性も大変良いと言える．そのため，セマンティック・セグメンテーションを高精度化するにあたり，従来から使われる識別モデルに生成モデルを組み合わせることは自然なことと考えられる．ところが意外なことに，セマンティック・セグメンテーションで識別モデルと生成モデルを組み合わせて推論する方式は，未だにほとんど検討されていない．そこで本稿では，生成モデルの一つである「デノイジング確率拡散モデル（DDPM）」[1]をセマンティック・セグメンテーションに導入することで，明確ではないピクセルを従前より高い精度で分類させることを検討する．

以降の章では，生成モデルを活用してセマンティック・セグメンテーションを高精度化させる手法について，より詳細に説明していく．2 章では，従来の識別モデルに基づくセマンティック・セグメンテーションの手法について概観する．3 章では，デノイジング確率拡散モデルの概要について説明する．4 章では，本稿の提案手法である，識別モデルと生成モデルを組み合わせたセマンティック・セグメンテーションの高精度化手法について述べる．5 章では，提案手法の実験結果について述べる．6 章は本稿の結論である．

2. 識別モデルベースの手法

セマンティック・セグメンテーションの従来手法は，大きく分けて以下の 2 つの手法に分類される．

・ルールベースの手法
・機械学習の手法

ルールベースの手法は，画像の特徴を人間がルールとして渡して物体の境界を検出する手法である．最もシンプルなものとしては，色の差やエッジの検出などを用いて物体の境界を検出する手法が考えられる．ルールベースの手法は，シンプルで高速なため実用化に非常に向いているが，画像の状況に応じてルールを調整する必要があり，一般に高い精度は期待できない．

一方，機械学習の手法は，画像とラベルデータのセットを用いて学習を行う．機械自身で画像から特徴を抽出し，その特徴に基づいて物体を分類する．機械学習の手法は，ルールベースの手法に比べて，精度が高いという特徴がある．しかし，学習において画像とラベルデータが多く必要であることと，計算コストが高いというデメリットがある．ただし近年では，機械学習を行うのに十分な規模のデータセットが数多く公開されるようになったため，広範なタスクへの応用というユーザーのニーズも後押しして，機械学習手法を用いるケースが急増している．

機械学習の手法で最もよく用いられるのがディープ・ニューラルネットワーク（DNN）による手法である．初期に提案されたのが Fully Convolutional Network（FCN）[2]と呼ばれるモデルである．FCN は，DNN の全結合層を畳み込み層に置き換えることで，セマンティック・セグメンテーションに適用した手法である．FCN は，従来の機械学習ベースの手法に比べて，大幅な精度向上を達成した．このモデルの成功により，DNN によるセマンティック・セグメンテーション手法が数多く提案されることとなった．

その後，FCN の改良である U-Net[3]というモデルが提案され，大きな注目を浴びた．U-Net は，エンコーダとデコーダからなるネットワーク構造を特徴としている．具体的な構造を図 1 に示す．エンコーダは画像の特徴を抽出し，デコーダはエンコーダで抽出した特徴を元に，画像の各画素を特定のクラスに分類する．さらに，スキップ接続を導入することで，エンコーダで抽出した特徴を，デコーダでより直接的に利用できるようにした．これにより，画像全体の情報と，小さな物体や細かい部分の情報を，

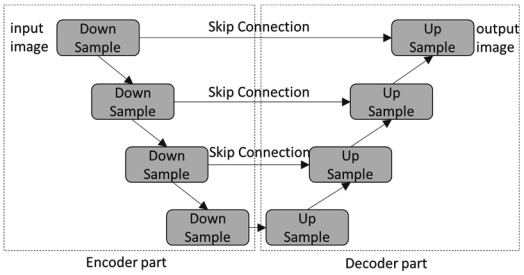

図 1 U-Net のモデル構造

より効果的に統合させることができるようになった．その革新的な構造により，後のセグメンテーション研究に大きな影響を与えている．

U-Net 以後は，異なるカーネルサイズを用いることでコンテキスト情報を効率的に学習する PSPNet[4]や，Atrous Spatial Pyramid Pooling（ASPP）モジュールを導入し，その中の dilation rate をピラミッド状に変化させることでより広範囲の関係性を学習する DeepLab v3+[5]などが提案され，セマンティック・セグメンテーションの性能は年々向上してきた．さらに最近では，Transformer を導入したセマンティック・セグメンテーション手法として UNetFormer[6]，SegFormer[7]なども提案されている．しかし，現状でもその識別性能は画素単位で見ると未だ十分とは言えず，今日も継続的に研究が行われている．

3. デノイジング拡散確率モデル

本章では生成モデルの代表的な手法の一つである，デノイジング拡散確率モデル（Denoising Diffusion Probabilistic Models: DDPM）[1]について説明する．デノイジング拡散確率モデルは，より一般化したモデルとして「拡散モデル」と呼ばれることもあり，近年の生成 AI の急激な進化と多彩な実応用に貢献した極めて重要なモデルである．

デノイジング拡散確率モデルは，ある画像に対してランダムノイズを徐々に加えるマルコフ過程を考え，完全にノイズになったものを逆向きに推定した際に，ノイズ除去後の画像と元の画像の差分を少なくするように学習することで，より自然な画像を生成可能にする手法である．図 2 にその概念を図化して示す．画像$\boldsymbol{x}_0$に対してランダムノイズを繰り返し加えていき，遂には時刻Tに完全なノイズ画像$\boldsymbol{x}_T$に至る．これを拡散過程と呼び，離散的な時刻$t = 0, 1, \ldots, T$に対し，

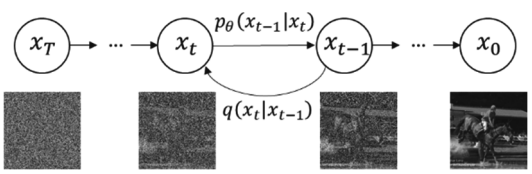

図 2 デノイジング拡散確率モデルの基本概念

$$q(\boldsymbol{x}_{0:t}) := q(\boldsymbol{x}_0)\prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \quad (1)$$
$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}\left(\boldsymbol{x}_t;\ \sqrt{1-\beta_t}\boldsymbol{x}_{t-1},\ \beta_t\boldsymbol{I}\right) \quad (2)$$

と定式化できる．ここで，$\mathcal{N}$は正規分布，β_tは時刻tに加えるノイズの強さを表すハイパーパラメータである．このノイズを与える過程を逆に遡るのを逆拡散過程と呼び，その時に元の画像と近い画像が得られるようにモデルを最適化するのが DDPM の考え方である．そして，逆拡散過程は，

$$p_\theta(\boldsymbol{x}_{0:t}) := p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \quad (3)$$
$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) := \mathcal{N}\left(\boldsymbol{x}_{t-1};\ \mu_\theta(\boldsymbol{x}_t, t),\ \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)\right) \quad (4)$$
$$p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{0}, \boldsymbol{I}) \quad (5)$$

と定式化できる．ここで，θは最適化の対象となるモデルパラメータを表す．この逆拡散過程は必ずしも正規分布で表せないのだが，β_tが十分小さい時には，拡散過程と逆拡散過程で同じ関数形を持つため，正規分布の利用が数学的に正当化されている．したがって DDPM では，画像$\boldsymbol{x}_0$を生成するために，式(4)の中の$\mu_\theta(\boldsymbol{x}_t, t)$に相当する部分を推論するモデルをディープ・ニューラルネットワークで構築する（なお，$\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$はモデルパラメータ$\theta$に依存しない$\beta_t^2\boldsymbol{I}$などの固定値を用いることが多い）．直感的には，画像$\boldsymbol{x}_t$からどのようなノイズを除去したら画像$\boldsymbol{x}_{t-1}$に戻すことができるか，そのノイズ成分を機械学習によって導出することになる．実際のデータ生成では，初期値として用意するランダムノイズ$\boldsymbol{x}_T$に対し，時刻tで条件付けされたデノイジング処理，およびサンプリングを意図した（式(4)の正規分布に基づく）ノイズ重畳処理，という操作を交互に繰り返すことで最終的な画像$\boldsymbol{x}_0$が生成される．

さて，ここまでの説明では初期値としてランダムノイズを想定していたが，これと異なり，何らかの画像を生成モデルに入力することで，新たな画像を生成することも可能である．この操作は「Image2Image」と呼ばれている．これを行うことにより，入力画像に類似したパターンやレイアウトの下で，あらかじめ学習した画像群の特徴を反映し

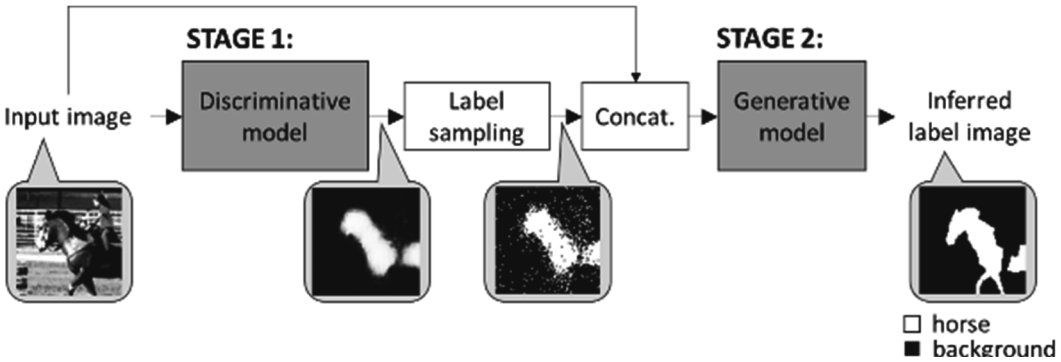

図 3　提案手法の処理フロー

た画像に変換することが可能となる．Image2Image は，通常の方法で学習された DDPM のモデルがあれば，非常に簡単な手順で実現可能となる．すなわち，画像を入力として，拡散過程をある程度実行し，得られたノイジーな画像を逆拡散して戻せば良い．この際，どこまで戻るかを示すパラメータとして Denoising Strength というパラメータが定義される．Denoising Strength は[0.0, 1.0]の範囲で指定し，逆拡散にかかる総ステップ数とこの数値の積で実際に戻るステップ数を決定する．例えば Denoising Strength を 0.75 とした場合，実際のステップ数に対し 75%まで拡散するという意味になる．

さらに，DDPM は条件付き生成モデルに拡張することも比較的容易である．詳細は省略するが，分類器ガイダンスや分類器無しガイダンスといった手法を用いることで条件付き生成を達成することができる[8]．また，条件付けにはテキストや画像，音声など任意形式のデータを適用することができる．

4. 提案手法

2 章で説明した従来のセマンティック・セグメンテーション法は，全て識別モデルをベースにした手法であった．一方，本稿ではこれまでと異なり，従来の識別モデルに生成モデルを組み合わせて利用することで，切り出し精度のさらなる向上を目指す．識別モデルと生成モデルとでは学習の方法が根本的に異なるため，両者を組み合わせることで，より高い精度でのセグメンテーションが期待できる．

さて，識別モデルと生成モデルの組み合わせにもさまざまな構成が考えられるが，本稿で提案する手法は，前段に識別モデル，後段に生成モデルを配置した 2 ステージによる疎結合方式である．つまり，前段の識別モデルによる出力結果をさらに改善させるために，生成モデルを介して出力結果を変換するアプローチである．このシンプルなアプローチの最大のメリットは，識別モデルとして過去に提案されたあらゆる手法が無条件で適用可能な点にある．

提案手法の詳細を，図 3 を用いて説明する．前段の識別モデルでは，n種類のラベルに対し，識別結果としてnチャネルの 2 次元データが出力される（図 3 では単純化のため，$n = 2$として図示している）．各チャネルにはラベル推定の確信度がマッピングされている．これらは推定された事後確率と考えることもできる．そこで，この確率値を基に，各画素のラベルをランダム選択し，選択されたラベルのチャネルのみを1に，それ以外を−1に置き換える（後段の生成モデルの入力範囲である[-1.0, 1.0]に合わせるため）．つまり，確率的に One-hot ベクトルの形式に変換する．この方式で変換されたnチャネルの画像データを，後段の生成モデルに転送する．

後段の生成モデルは，デノイジング拡散確率モデルを利用する．ここでは，生成モデルに入力する前に，前段の識別モデルの出力データと，識別モデルの入力データである元画像を，チャネル方向に Concatenate（結合）する．これにより，生成モデルでも元画像を条件付けてデータ生成できるようになる．なお，事前に実施する「生成モデルの学習」においても，正解ラベル画像とその元となる画像のペアを Concatenate したデータを使って学習することに注意されたい．また，生成の際は，3 章で述べた Image2Image とは少し異なる操作を行う．すなわち，初期値として先ほど Concatenate したデータを用いるが，拡散過程はスキップして，逆拡散過程のみを実施する．最後に，生成したデータから元画

像にあたるチャネルを分離し，ラベルにあたるチャネルのみを出力する．これらのデータから，それぞれの画素において最大の確信度となるラベルを選択して One-hot ベクトル化すれば，セマンティック・セグメンテーションのタスクは完了である．

5. 実験

提案手法の有効性を検証するための実験を行った．実験用のデータとしては，セマンティック・セグメンテーションの研究において広く使用されている公開データセットである Pascal VOC 2012[9]を利用した．Pascal VOC 2012 には 20 クラスの物体カテゴリが含まれており，9,993 枚の正解セグメンテーション付き画像が用意されている．これらには，本研究で着目している複雑背景下の物体や物体同士の重なりなどのシーンが豊富に含まれている．本実験ではこれらの中から特にカテゴリ内の種類が豊富，かつ個体の変動も比較的大きい“動物”関連のカテゴリ 5 クラス（bird, cat, dog, horse, sheep）を対象とした．また，画像は全て正方形となるよう左右の端を切り取り，全体を 128×128pixel にリサイズして用いた．

提案手法は，識別モデルのアーキテクチャに依存しない枠組みであるが，本実験においてはスタンダードなアーキテクチャを用いることを重視し，識別モデルに U-Net を採用した．なお，DDPM のハイパーパラメータである Denoising Strength は 0.5 とした．

クラス分類については，動物クラス 1 種類と背景 1 クラスの 2 クラスを分割するセグメンテーションタスクを評価対象とした．また，従来手法である U-Net 単独でのセグメンテーションに対して，DDPM による改善処理を施した結果，どの程度切り出し精度が向上するかを評価した．その際の評価指標として，IoU（Intersection over Union）を用いた．IoU は，推定された領域と真の領域の重なり率を計算することによって，推定結果の正確さを評価する指標である．IoU が 1.0（100%）に近いほど，推定結果と真の領域の重なり率が高く，推定結果が正確であることを示す．また，IoU は画像 1 枚に対して 1 個の数値が算出されるため，評価には，使用した画像の枚数で平均化した数値を用いた．

評価を集計した結果を表 1 に示す．平均的には，どのカテゴリも従来手法である U-Net（識別モデルのみ）より提案手法の性能が上回る結果となった．ただし，画像ごとに見ると，精度が上がったものと下がったものが混在していた．図 4 に精度が上がった出力例，図 5 に精度が下がった出力例を挙げる．図 5(c)の右側面に見られるように，大幅に過検出してしまった領域をキャンセルできるだけの性能は備わっていないと考えられる．一方で図 4 のように，検出漏れしてしまった部分を提案手法によって補完する効果があることが数多く確認されている．このことが，全体として IoU 向上に大きく寄与していた

表 1　評価結果

Category name	IoU(%)	
	Only U-Net	Proposed
1. bird	41.4	**48.3**
2. cat	60.2	**64.8**
3. dog	53.6	**58.8**
4. horse	51.7	**53.6**
5. sheep	51.4	**55.2**
TOTAL(1～5)	52.1	**56.8**

(a)入力画像　(b)正解　(c)従来手法 (d)提案手法
図 4 推定結果（精度が上がった例）

(a)入力画像　(b)正解　(c)従来手法 (d)提案手法
図 5　推定結果（精度が下がった例）

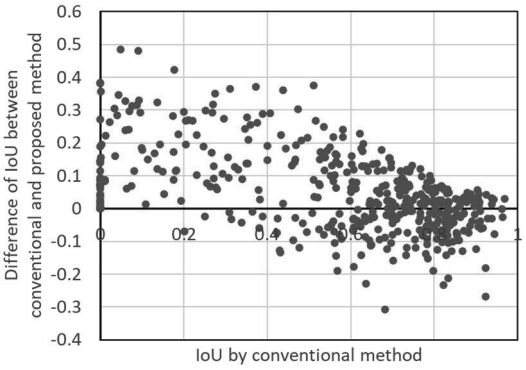

図 6　従来手法の IoU と IoU 上昇度の関係

と思われる．

さらに，どういう画像が提案手法により改善されたのかを詳細に調べるため，従来手法の IoU 値と，提案手法による IoU 上昇度（提案手法 IoU と従来手法 IoU の差分）の関係を示す散布図を作成した．図 6 にその散布図を示す．これを見ると，もともとの IoU が小さければ小さいほど，提案手法によってより大きく改善できたということが分かる．つまり，生成モデルは識別モデルの推定結果の細部を微修正することよりも，むしろ識別モデルの苦手な面をサポートすることにおいてより効果的だった，と解釈することもできる．

6. おわりに

本稿では，セマンティック・セグメンテーションタスクを高精度化するため，従来の識別モデルに生成モデルを組み合わせる手法を提案した．実証実験の結果，生成モデルの導入により，従来手法のセグメンテーション結果を高精度化できることが確かめられた．ただし今回の実験は 2 クラス分類の検証に留まっているため，多クラス分類（画像のシーン理解）においても同様に有意な結果が表れるかどうかは，さらに検証を進める必要がある．また，今回生成モデルとして採用した DDPM は処理時間とメモリ効率にやや問題があるため，より高速に処理できる DDIM[10]や，高解像度の画像も処理可能な Latent Diffusion Model[11]などを導入することも併せて検討したい．さらに，動画像や 3 次元点群など，静止画以外のデータについても同様に切り出し精度を向上できる手法の検討も今後の課題と考えている．

参考文献

[1] Jonathan Ho, Ajay Jain, Pieter Abbeel, "Denoising Diffusion Probabilistic Models", Proceedings of NeurIPS, 2020.
[2] Jonathan Long, Evan Shelhamer, Trevor Darrell, "Fully Convolutional Networks for Semantic Segmentation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431-3440, 2015.
[3] Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation." International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2015.
[4] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia, "Pyramid Scene Parsing Network", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881-2890, 2017.
[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, Hartwig Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation", Proceedings of the European Conference on Computer Vision (ECCV), pp. 801-818, 2018.
[6] Libo Wang a b, Rui Li h, Ce Zhang c d, Shenghui Fang a, Chenxi Duan e, Xiaoliang Meng a b, Peter M. Atkinson c f g, "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery", ISPRS Journal of Photogrammetry and Remote Sensing, Volume 190, pp.196-214, 2022.
[7] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers", Proceedings of NeurIPS, 2021.
[8] 岡野原大輔, "拡散モデル ―データ生成技術の数理", 岩波書店, 2023.
[9] Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A. "The PASCAL Visual Object Classes Challenge: A Retrospective", International Journal of Computer Vision, 111(1), 98-136, 2015.
[10] Jiaming Song, Chenlin Meng, Stefano Ermon, "Denoising diffusion implicit models", International Conference on Learning Representations (ICLR), 2021.
[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684-10695, 2022.