2S-09

フォントスタイルを指定可能なテキストからの画像生成

XIA XIN　　遠藤　結城　　金森　由博
筑波大学

1. はじめに

グラフィックデザインにおいて、文字のデザイン、すなわちタイポグラフィは視覚的に重要である。パッケージやブックカバーなどの製品に含まれるタイポグラフィが魅力的で適切であるかどうかは、宣伝の効果に大きな影響を及ぼす。しかし、タイポグラフィデザインには高度な専門知識と多大な労力が必要である。既存研究では、テキストレンダリングができる画像生成手法がいくつか提案された。しかし既存手法では、フォントスタイルを明示的に指定できない課題がある。なお本稿でフォントスタイルとは、Serif（セリフ体）, Gothic（ブラックレター）, Cute（かわいい）など、文字の形や印象を指す。

そこで本研究では、拡散モデルに基づいた、フォントスタイルを制御可能なテキストからの画像生成手法を提案する。この目的を達成するために、フォントスタイルに関するデータセットを構築し、訓練済みの画像生成モデル [1] を文字を含む画像とフォントスタイルに関するデータで Fine-tuning し、フォントスタイルに関する入力を反映できるようにする。また、画像とテキストの共通の特徴を抽出できる既存手法である CLIP [2] をフォントスタイルに関するデータセットで Fine-tuning し、定量評価に用いる。定性・定量評価を通じ、提案手法が既存手法に比べ、指定したフォントスタイルをより反映した画像生成を実現できることを示す。

2. ベース手法

本研究のベースとなる TextDiffuser について説明する。TextDiffuser は、拡散モデルである Stable Diffusion をベースとしたテキストからの画像生成モデルである。従来の拡散モデルが文字を正しく生成できない問題を解決し、生成画像内の文字を指定できるようにした。入力されたプロンプトからユーザが生成したい文字をモデルが推定し、テキストレイアウトをレンダリングし、正解の文字位置情報として得る。そして拡散モデルを通じて画像を生成し、文字の形と位置に関する損失関数を追加することによって文字の形が崩れないように制御する。しかし前述の通り、TextDiffuser ではフォントスタイルを指定することはできない。

3. 提案手法

提案手法の概要を図 1 に示す。提案手法は、フォントスタイルに関するデータセットを作成し、TextDiffuser を Fine-tuning する。学習段階では、正解画像のスタイルラベルと文字の内容を連結し、TextDiffuser に与える。文字の内容は引用符で囲むことで指定する。それに加え、正解画像、各文字の位置情報、一部または全ての文字部分のマスクとマスク後の画像のエンコードした特徴マップを連結してネットワークに入力し、Stable Diffusion における条件付け生成を行う。損失関数は、ネットワークの出力との正解の特徴マップのノイズ除去損失 l_d と文字位置損失 l_s が含まれる。また提案手法では、フォントスタイルが適切であるかどうかを評価するために、損失関数に Fine-tuning された CLIP（以下で説明）のスコアの逆数を l_c として追加する。

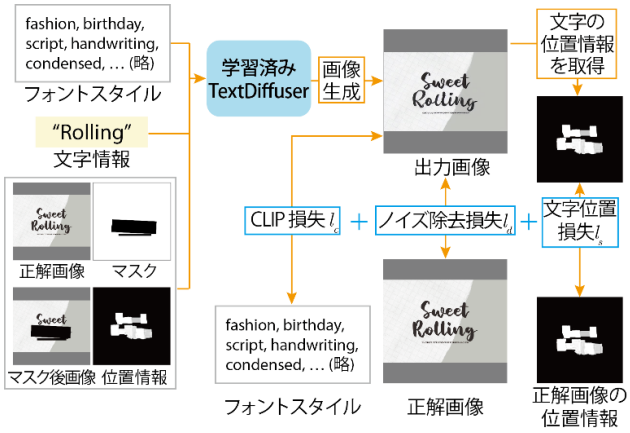

図 1: 提案手法の概要。

3.1 CLIP の Fine-tuning

CLIP とは、テキストと画像の共通の特徴を抽出する手法である。出力画像のフォントスタイルが適

切であるかどうかを評価するために、作成したデータセットで Fine-tuning された CLIP を用いる。本研究では CLIP を、結果の定量評価と学習時の損失関数に用いる。同じデータで Fine-tuning された CLIP モデルを学習と評価の両方に使用するのは適切でないと考え、TextDiffuser の Fine-tuning 用のデータと、それとは別のデータで、それぞれ CLIP を Fine-tuning した。前者の CLIP モデルを CLIP-Train、後者のモデルを CLIP-Eval と表記する。CLIP-Train は損失の計算に使用し、CLIP-Eval を定量評価で使用する。文字の表現に注目させたいため、学習の際は、画像データの文字部分のみを残し、学習の対象とする。

4. 実験と結果

提案手法を Python および PyTorch を用いて実装し、NVIDIA RTX 6000 を用いて実験した。オプティマイザには AdamW を用い、学習率は 4e-6、バッチサイズを 12 とした。画像生成の高速化のために xformers を使用した。実験で用いた提案手法のモデルは、15 エポックまで学習したものである。学習には 2 日程度かかった。推論にかかる時間は、512 × 512 画素の画像 1 枚の出力に 12 秒程度であった。

4.1 定量的評価

3.1 節で説明した CLIP-Eval を用い、TextDiffuser と提案手法の生成画像を比較した。各手法において、36 枚の生成画像の文字部分のみに対して、CLIP-Eval スコアの平均を計算した。その結果、TextDiffuser の平均スコアが 12.67 であり、提案手法の平均スコアが 13.97 であった。スコアが高い提案手法が、与えられたフォントスタイルをより反映できたといえる。

4.2 定性的評価

同じプロンプトを入力とし、文字が含まれる画像を生成できる既存手法である DeepFloyd IF、SD-XL、TextDiffuser、及び TextDiffuser の最新版 TextDiffuser-2 と、提案手法の生成画像を定性的に比較した。

図 2, 3 に示す通り、既存手法では指定された文字が出力されなかったり、文字を間違えたりする問題がある。提案手法のベースとなる TextDiffuser では文字は正しかったが、フォントスタイルが反映されていない。既存手法と比べ、提案手法の方が文字は正しく、かつフォントスタイルが一番反映された。

図 2: ラベル「Sans-Serif, Display, Funny, Comic, Script, Graffiti, Bold, Cool, Cursive」から得られた画像。

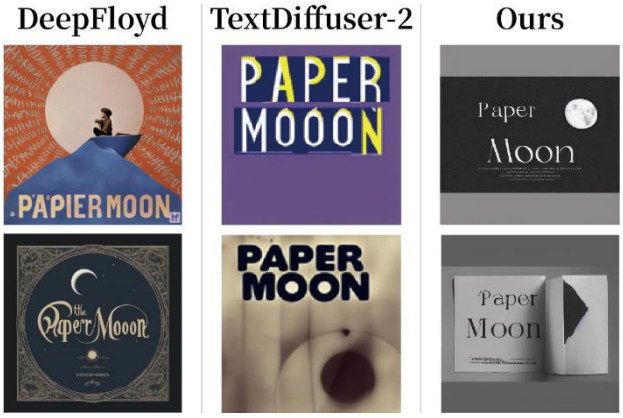

図 3: ラベル「serif, elegant, modern, vintage, elegant, classy」から得られた画像。

5. まとめ

本研究は、テキスト入力からフォントスタイルを指定可能な、文字を含む画像の生成手法を提案した。具体的には、フォントスタイルに関するデータセットを作成し、TextDiffuser にフォントスタイルをより明示的に学習させた。

参考文献

[1] Chen et al., TextDiffuser: Diffusion Models as Text Painters. *arXiv preprint arXiv:2305.10855, 2023.*

[2] Radford et al., Learning Transferable Visual Models From Natural Language Supervision. *ICML 2021, pp. 8748-8763, 2021.*