コンピュータビジョン分野の国際会議 WACV2024 にて拡散モデルを用いたテキスト超解像に関する論文が採択

トヨタ自動車株式会社
社会システムＰＦ開発部
InfoTech-AS データ解析１グループ
野口千尋
福田竣
山中正雄

トヨタ自動車株式会社は、テキスト画像の文字認識を行うための画像生成ＡＩ技術を開発しました。コンピュータビジョン分野の国際会議である WACV2024（Winter Conference on Applications of Computer Vision 2024）にて、当社の研究論文 ”Scene Text Image Super-resolution based on Text-conditional Diffusion Models”が採択され、2024 年 1 月 4 日の本会議にて発表しました。

背景

私たちは身の回りにおいて、さまざまな形でテキスト情報を視認し、それに基づいた意思決定を行なっています。そのため、画像を用いたアプリケーションを構築する上で、これらの情報を正確に読み取ることは非常に重要です。例えば、車載カメラによって撮影された画像には、交通標識やナンバープレートなどのテキスト情報が含まれています。自動運転や地図生成サービスを実現するために、これらを高精度に認識することは極めて重要です。しかし、画像中のテキスト領域はしばしば小さく、ぼやけやノイズの影響を強く受けるため、これらを高精度で読み取るためには様々な課題があります。

テキスト超解像は、テキスト画像に特化した超解像技術です。条件の悪いテキスト画像（低解像、強いぼやけ・ノイズを含む、など）から劣化前のクリアな画像を復元することを目的とします（図１）。この技術の主な利用用途の１つは文字認識モデルの前処理です。入力画像をより可読性の高いテキスト画像に変換することで、後段の認識モデルの認識精度を向上させることができます。この技術は、文字認識モデルを直接チューニングすることが難しい状況（例えば、他社製の API を利用している、また、すでに製品に組み込まれていて頻繁にアップデートが出来ない、など）にて特に有用です。

図 1 テキスト画像の例。画像下に文字認識モデルの認識結果を示す。赤文字は誤認識文字を示す。（上）入力テキスト画像を線形補間でリサイズした画像。劣化が激しく文字を読み取ることが難しい。（中）テキスト超解像適用後の画像。可読性が大きく改善し、高解像画像をうまく復元できている。（下）正解の高解像度画像。

技術概要

テキスト超解像分野における大きな課題の 1 つは、学習データセットの収集コストが高い点です。学習のためには、ピクセル単位で位置関係が一致した低解像・高解像ペア画像が必要ですが、これらを大量に用意するには大きな人的コストを必要とします。人工的なぼやけやノイズを組み合わせることで、実画像における画像劣化パターンを再現する手法も提案されていますが、実環境で生じる画像劣化を再現することは難しく、特に劣化の激しいテキスト画像の復元にはまだまだ課題があります。そこで、提案法では、限られた枚数の既存の低解像・高解像ペア画像から劣化パターンを学習し、テキスト画像に適用することで、より実画像に近い人工ペア画像を大量に生成します。この拡張データセットを既存のテキスト超解像モデルの学習に用いることで、飛躍的な性能向上を達成しました。

提案法では、画像生成モデルとして拡散モデルを用います。拡散モデルは、その学習の安定性から、様々なドメインにおける条件付き生成モデルとして用いられています。拡散モデルは、非常に高い表現能力がある一方で、推論時間の長さが大きな欠点として知られています。提案法では、生成されたペア画像を既存のテキスト超解像手法の学習データセットとして用いるため、推論時間の遅さは大きな問題にはなりません。一方で、生成されるテキスト画像の品質は学習データセットを構築する上で非常に重要です。論文中ではまず、既存のテキスト超解像モデルの自然な拡張としての拡散モデルベースの手法を考えます。我々の実験では、この単純な拡張にもかかわらず、既存手法を大きく上回る結果が得られています。しかし、提案法では、こうして学習された拡散モデルにより生成された画像を既存モデルの学習データセットして用いるというアプローチを考えるため、ここでの性能向上は満足の

いくものではありません。

テキスト超解像には、2種類の正解データが存在します。1つは、生成する目標である高解像画像です。これは、上述の様に、コスト面から大量に入手することが難しいデータです。もう1つの正解データは、テキスト画像内のテキストです。こちらは比較的入手が容易で、テキスト認識モデルの学習データセットなどから大量に入手可能です。提案法では、低解像画像だけでなく、正解テキストによっても条件付けされた拡散モデルを考え、高解像画像の生成分布を学習します。正解テキストを条件付けとして用いる効果は極めて大きく、我々の実験では、正解の高解像画像と遜色ないほどの高品質な超解像画像が復元出来るという結果が得られています（図 3）。提案法では、正解テキストによって条件付けされた拡散モデルを用いてペア画像を生成し、生成画像を既存モデルの学習データセットとして用いることで、飛躍的な性能向上を達成しました。

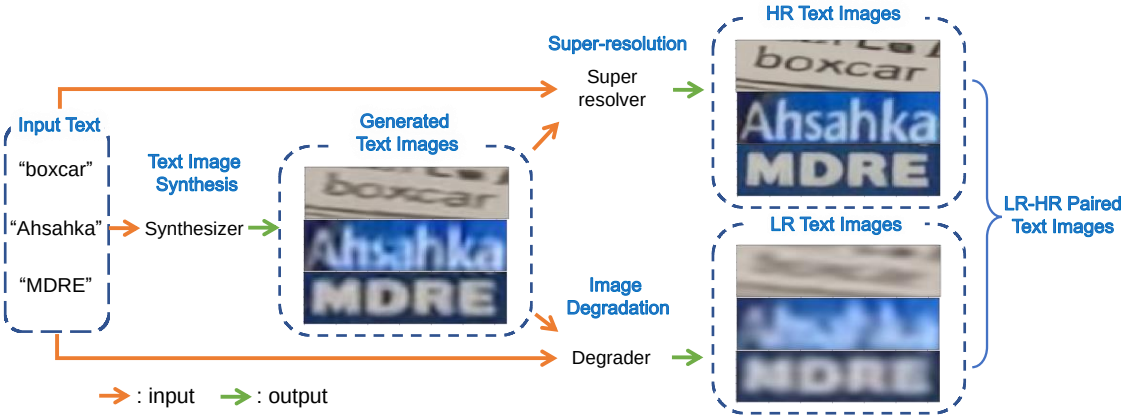

図 2　提案方法の概略図

図 1 は、我々が提案する低解像・高解像ペア画像生成フレームワークの具体的な構成を示します。提案法は、3 つの拡散モデルにより構成されます。1 つ目は、テキスト画像生成モデル（Synthesizer）です。このモデルは入力されたテキストに対応するテキスト画像を生成します。ただし、ここでの目的は、テキスト画像と対応するテキストを用意することであり、代わりに既存のテキスト認識モデルの学習用データセットを利用可能です。論文中の実験では、既存のテキスト認識モデル学習用のデータセットを用いた場合と、それを用いて学習したテキスト画像生成モデルからの生成画像を用いた場合の 2 パターンを検証しました。2 つ目は、テキスト超解像モデル（Super-resolver）です。上述の様に、正解テキストを条件付けとして用いることで、入力テキスト画像を高品質な超解像画像へと変換します。3 つ目は、画像劣化モデル（Degrader）です。入力テキスト画像を劣化画像へと変換します。テキスト超解像モデルと画像劣化モデルは、TextZoom という既存のペア画像データセットを用いて学習を行いました。また、既存のテキスト認識モデル学習用のデータセットは比較的劣

化の少ない画像が多く、画像劣化モデルを適用することでより超解像モデルの学習に適した画像生成が可能です。

Method	GT Text	SSIM ($\times 10^{-2}$)	PSNR	Acc. (%)
Bicubic		69.61	20.35	26.8
HR		-	-	72.4
SRCNN [10]		72.27	20.78	29.2
SRResNest [25]		74.03	21.03	35.1
TSRN [51]		76.90	21.42	41.4
STT [5]		76.14	21.05	48.1
PCAN [61]		77.52	21.49	47.4
TG [6]		74.56	21.40	48.9
TATT [29]		79.30	21.52	52.6
C3-STISR [62]		77.21	21.51	53.7
DDPM*		79.50	22.70	55.0
TCDM		**79.58**	**22.83**	**55.7**
TATT	✓	79.34	22.35	61.0
TCDM	✓	**80.25**	**22.86**	**68.1**

図 3 定量評価の結果

図 3 には、拡散モデルを用いたテキスト超解像の定量評価の結果を示しています。SSIM と PSNR は一般的な超解像分野においても広く用いられる指標で、生成した超解像画像と正解の高解像画像との平均的な差分を示しています。Acc.は学習済み文字認識モデルによる認識精度を示しています。提案法で用いた拡散モデルを用いたテキスト超解像モデル（TCDM）は、正解テキスト（GT Text）を用いない場合でも、既存手法を大きく上回る性能（55.7%）を達成していることが分かります。これに正解テキストを入力として用いると性能が飛躍的に向上（68.1%）し、正解の高解像画像における認識精度に匹敵する認識精度（72.4%）を達成していることが分かります。

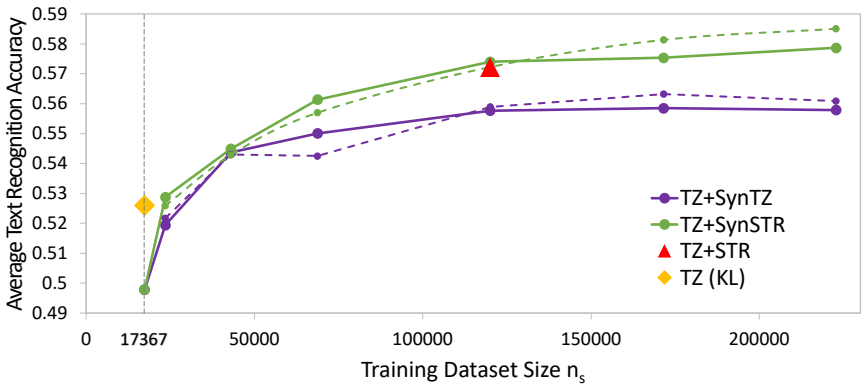

図 4 生成したペア画像データセットを用いて既存のテキスト超解像モデル（TATT）を学習した結果

図４は、提案法により生成したペア画像を、既存のテキスト超解像モデル（TATT）の学習データとして用いた際の評価結果を示しています。横軸は学習データセットの枚数、縦軸は学習済みテキスト認識モデルを用いて評価した認識精度を示しています。学習データセットを増やすほど、認識精度が向上しており、提案法の有用性を確認することができます。

まとめと今後の課題

本稿では、当社のテキスト超解像に関する研究論文の技術解説を行いました。提案法では、拡散モデルを用いて、既存の低解像・高解像ペアデータセットから劣化パターンを学習し、既存データセットの拡張データセットを構築しました。テキスト超解像の問題設定に拡張した拡散モデルは、既存のテキスト超解像手法と比べても高品質な超解像画像を生成可能ですが、追加の条件付けとして正解テキストを活用することで、より正確に元の高解像画像を復元することが可能である事が分かりました。提案法では、この性質を用いて拡張データセットを構築し、既存のテキスト超解像モデルの学習に用いることで、テキスト超解像の飛躍的な性能向上を達成しました。

提案法は、劣化パターンの学習のために低解像・高解像ペアデータセットを必要としました。しかし、公開データセットとして利用可能なペアデータセットは非常に限られており、現状では、英語のテキスト画像に限られます。そのため、提案法の適用範囲も英語のテキスト画像に限られることになり、英語圏以外での実用性に課題があります。今後の発展として、ペア画像に依存しない形で劣化パターンを学習し、テキスト画像に適した形でそれを適用することで、非ラテン文字に対しても適応可能な手法の検討を進めています。