



動的対象への追従投影のための Image-to-Image Translation の高速化

野崎亮汰¹⁾, 渡辺義浩¹⁾

1) 東京科学大学 工学院 情報通信系 (〒 226-8501 神奈川県横浜市緑区長津田町 4259 G2-31, nozaki.r.c7ab@m.isct.ac.jp)

概要: 動的対象への追従投影と Image-to-Image Translation (img2img) を組み合わせることで、モデル化が困難な変換をリアルタイムに実行し、画像を生成して投影することが可能になる。しかし img2img は画像生成に時間を要するため、対象と投影像の位置がずれる問題がある。そこで本稿では、Laplacian Pyramid による画像分解を用いる img2img の処理を並行化し、一部処理の頻度を低下させ、同処理の最新の計算結果を他処理で利用することで、生成画像の劣化を知覚されにくいように抑制しつつ処理時間を削減する手法を提案する。

キーワード: ダイナミックプロジェクションマッピング, Image-to-Image Translation, 高速化

1. はじめに

動的対象に映像を追従させて投影することで対象物体の見た目を操作する Dynamic Projection Mapping (DPM) が注目されている [1, 2, 3]。特に顔のような非剛体変形を伴う物体を対象とする DPM では、対象を撮像して 3 次元形状を推定し、見た目を操作するために事前に用意したテクスチャを形状に合わせて変形することで投影を行っている [1, 3]。他にも、投影像生成のために光線の反射屈折をシミュレートすることでガラスの透過や鏡面反射物体の映り込みを再現した DPM が提案されている [2]。

一方、2 つの画像ドメイン間の対応関係を学習することで、入力画像を異なる見た目の画像に変換可能な Image-to-Image Translation (img2img) がある [4, 5, 6]。img2img は写実的な画像生成や複雑な変換が可能かつ、学習データセット次第で異なる変換が可能であるため、様々な応用利用されている。また、img2img は入力画像の形状を保持したまま変換を行い、変容した画像を出力できる。このため、対象を撮像し img2img を用いることで、複雑な変換をリアルタイムに行って生成した画像を DPM に用いることができる。しかし、img2img は 1 枚の画像の変換に数 ms から数十 ms 以上要するため、動的対象への追従投影において対象と投影像の位置がずれる問題がある。

上記を踏まえ、本稿では img2img を用いる DPM を提案する。従来の DPM ではテクスチャの変形や物理現象のモデル化を基に投影画像を生成していたが、img2img によってモデル化が難しい変換をリアルタイムに行い、投影に用いることが可能になる。上記の実現に向けて img2img を高速化するために、Laplacian Pyramid (LP) による画像分解を用いた img2img の処理を並行化し、一部処理の頻度を低下させ、同処理の最新の計算結果を他処理で利用することで、生成画像の劣化を知覚されにくいように抑制しつつ処理時間を削減する手法を提案する。さらに、img2img を用

いる DPM では人の顔の年齢感を操作する応用がある [3]。同 DPM を想定し顔を老けさせるような変換について、本手法の生成画像の品質と処理時間の評価を行う。

2. 関連研究

2.1 顔に対する DPM

顔に対する DPM を用いた応用には化粧 [1] や年齢感の操作 [3] がある。特に、年齢感操作の応用では対象に合った、皺のような複雑なテクスチャを事前に 1 枚作成して用いていた。そのため、多様な表情に合わせた年齢感の変化を表現することが難しく、没入感の低下につながっていた。上記に対しリアルタイムに img2img を用いる DPM では、年齢感の操作のような複雑な変換においても対象の位置姿勢・表情に合わせたリアルタイムなテクスチャ生成が可能である。

一方、DPM では対象の撮像から投影までの間に遅延が生じ、観察者にずれが知覚されると没入感が低下する。上記の遅延が知覚される時間として、顔に対する DPM では平均 3.87 ms を超えると遅延が知覚されると報告されている [7]。それゆえ、DPM に用いることができる高速な img2img が必要である。

2.2 img2img

img2img とは 2 つの画像ドメイン間の対応関係を学習し、入力画像を別ドメインの画像に変換する技術である。img2img は十分な学習データセットが与えられれば、モノクロ画像やエッジ画像といった多様な入力形式から写実的な画像の生成や複雑な変換を行うことが可能である [4]。

また、映像制作分野における応用例として、FRAN と呼ばれる手法がある [6]。同手法では、同一人物の幼年期から老年期までの幅広い年齢における合成顔画像データセットを作成し、img2img モデルの学習に用いていた。その結果、年齢の変化に伴う顔の印象の変化を写実的に再現している。

上記のように img2img は写実的な変換や様々な応用が可

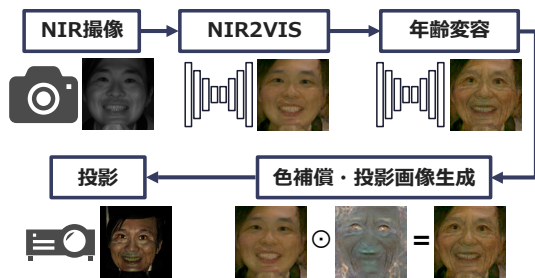


図 1: img2img を用いる DPM の処理のパイプライン。図中の年齢変容に該当する箇所は学習させる変換内容によって変化する。

能である一方、計算負荷が大きく処理時間が長い課題があった。この課題に対し、LP を用いて画像を周波数分解し段階的に変換を行うことで、計算負荷を低減し img2img の高速化を実現した Laplacian Pyramid Translation Network (LPTN) が提案されている [5]。同手法は 512×512 pixels の画像を数 ms で写実的に変換可能である。しかし依然として DPM に用いるためには高速性が不十分であった。

3. img2img を用いる DPM

3.1 概要

本稿では、img2img を用いる DPM を提案する。また、図 1 に処理の流れを示す。まず、投影中の対象から白色照明下の画像を取得する。ここでは Near-Infrared (NIR) 画像から白色照明下の可視光画像を推定する処理を行う。同処理は NIR 画像と可視光画像の対応関係を学習することで、img2img による実現が可能である。

次に img2img を用い、可視光画像から目標画像へと変換を行う。同処理は DPM によって実現したい変容を表すデータセットを作成し img2img に学習させることで実現できる。その後、投影結果を目標画像へ近づけるために色補償を行うことで投影画像を生成する。色補償についても同様に、対象と投影結果の対応関係を学習することで img2img による変換が可能だと考えられる。上記のように、全ての処理で img2img を用いる DPM によって対象の変容を実現できると考えられる。

上記より、先に述べた DPM の実現に向けて、img2img の高速化が重要である。本稿では、img2img の処理の中でも可視光画像から目標画像への変換に注目し、同変換の目標画像の劣化を抑えつつ img2img の処理時間を削減する手法を提案する。

3.2 img2img の高速化

本稿では LP を用いて周波数成分ごとに変換を行う LPTN [5] に注目し高速化を図る。また、図 2 に本手法の処理の流れを示す。

まず、LP とは入力画像をダウンサンプリングし、得た画像と同画像をぼかした画像から差分画像を生成する過程を繰り返す画像分解処理である。繰り返しの回数によって画像 Pyramid の層数が増加するため、上記の繰り返しの回数は層数と表現される。同処理によって解像度の異なる複数

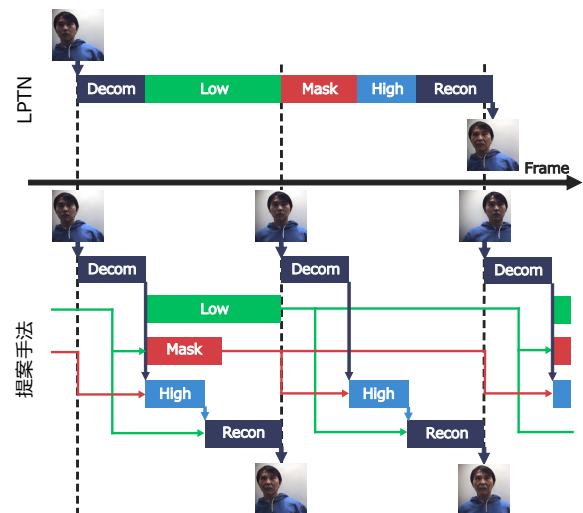


図 2: img2img の高速化の概要図。上段が LPTN [5]，下段が本手法を表す。また Decom, Low, Mask, High, Recon はそれぞれ画像分解、低周波成分の変換、マスクの生成、高周波成分の変換、画像再構成を表す。本図で例示している提案手法では Decom, High, Recon が 500 fps のとき、Low, Mask は 250 fps となる。

枚の差分画像と最も低解像度のぼけた画像が得られる。同処理は疑似的な周波数分解に相当し、差分画像群は高周波成分、ぼけた画像は低周波成分が表現されている。

次に、LPTN の処理の流れについて説明する。最初に LP による画像の分解を行い、低周波成分を変換する。その後、低周波成分と変換結果を基に、高周波成分で変換を行う領域を示すマスクの生成処理を行う。次にマスクを基に高周波成分の変換を行い、最後に、画像の再構成によって結果画像が得られる。上記より LPTN の処理は明確に分離しており、図 2 上段に示すように逐次的に行われるので、高速化のために画像分解、低周波成分の変換、マスクの生成、高周波成分の変換、画像再構成の 5 つの処理を並行化する。

また、上記のような処理を高フレームレート下で行うと、低解像度のため連続フレーム間の低周波成分の差は小さいと考えられる。これは低周波成分から生成されるマスクについても同様である。また、画像の再構成時に低周波成分は主に色成分を示す一方、高周波成分は輝度の変化が大きいエッジに相当する。上記に対し、人間は色と輝度を別々の要素として知覚し、色よりも輝度に対して高い感度を持っていることが報告されている [8]。以上から、高周波成分に対して低周波成分がずれていても、同ずれは知覚されにくいと考えられる。そのため、図 2 に示すように低周波成分の変換とマスクの生成処理の頻度を低下させ、他処理では最新の計算結果を用いることで、画像の劣化を抑えつつ画像入力から変換結果出力までの処理時間の短縮を図る。

4. 実験

4.1 学習条件

本手法の img2img を用いた DPM の応用例として、人の顔の年齢感を操作する DPM がある。そこで本稿では同



図 3: 500 fps, 平行移動の動画について変換した動画を 60 fps に変換した結果. 左から crop 前の入力画像, Reference, Low=83.3 fps かつ Mask=166.7 fps の変換結果, Low=25 fps かつ Mask=50 fps の変換結果.

DPM を想定して img2img を学習させ, 画質・処理時間について評価を行う. そのため, 23 歳 → 75 歳の変換を img2img に学習させた. データセットは FRAN [6] と同様の方法で作成した, 2,000 人, 2 年齢, 512×512 pixels の合成画像ペアデータセットである. この内, 1,800 枚を訓練データ, 200 枚を評価データとして学習を行った.

また, 生成器は LPTN の 3 層のものをを用いた. さらに, 式 (1) に FRAN [6] を基に構成した損失関数を示す.

$$\mathcal{L} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{LPIPS}\mathcal{L}_{LPIPS} + \lambda_{adv}\mathcal{L}_{adv} \quad (1)$$

$\mathcal{L}_{L1}$ は目標画像と生成画像の L1 ノルム, $\mathcal{L}_{LPIPS}$ は同じく目標画像と生成画像を入力とした LPIPS, $\mathcal{L}_{adv}$ は識別器を用いた敵対的損失である. 本稿では識別器に PatchGAN Discriminator を用いた. また, λ_{L1} , λ_{LPIPS} , λ_{adv} はそれぞれの損失関数の重みを表しており, $\lambda_{L1} : \lambda_{LPIPS} : \lambda_{adv} = 5 : 10 : 2$ と設定した. 加えて, 学習における iteration 数は 300,000, 学習率は $1e^{-4}$ とした.

4.2 評価

img2img の高速化を目的とする本手法の評価のため, 変換された画像の画質の評価と処理時間の計測をそれぞれ行った. まず画質の評価において, 入力は 512×512 pixels, 500 fps, 5 秒間の 2,500 枚からなる, 顔の動作についての動画とした. また, 同動画は平行移動 (約 1 m/s), 回転運動, 複数の表情への変化, 喋りの動きを取る 4 種類の動画である. 上記の 500 fps の動画について, 全フレームの画像に対し LPTN の全処理を行って変換した動画を Reference とする. これに対し, 低周波成分の変換とマスクの生成処理の fps を設定し, 他の処理を 500 fps で変換した動画を本手法による出力動画として扱った. 本稿では顔の変換に注目するため, 顔領域を中心に 360×360 pixels に切り出し, 先に述べた動画と Reference を比較することで定性・定量評価を行った. ここで出力動画は 500 fps の動画であるため, フレームごとに評価すると人間の知覚とは異なる条件で評価されることになる. そのため, 500 fps の出力動画を基に 60 fps の動画を作成した.

まず図 3, 図 4 に出力動画を 60 fps に変換した動画の一部を示す. 図 3 では, Reference と Low=83.3 fps かつ Mask=166.7 fps の変換結果は大きな違いが無かった. 一方, Low=25 fps かつ Mask=50 fps の変換結果では顔上でノイズによる劣化が確認された. 上記に対し, 図 4 では複数の表情に対し図 3 と同条件で変換を行った結果を示しているが, 大きな劣化は確認されなかった.



図 4: 500 fps, 表情変化の動画について変換した動画を 60 fps に変換した結果. 左から crop 前の入力画像, Reference, Low=83.3 fps かつ Mask=166.7 fps の変換結果, Low=25 fps かつ Mask=50 fps の変換結果.

表 1: 4 種類の動画の変換結果に対する PSNR, SSIM, LPIPS. 表内上段の数値は低周波成分の変換の fps, マスクの生成処理の fps を表す.

		125,250	100,200	$\frac{250}{3}, \frac{500}{3}$	50,100	25,50
平行移動	PSNR (↑)	49.119	46.925	45.602	41.381	36.771
	SSIM (↑)	0.9962	0.9942	0.9924	0.9808	0.9522
	LPIPS (↓)	0.0022	0.0040	0.0055	0.0165	0.0448
回転運動	PSNR (↑)	51.504	49.469	48.331	44.345	39.904
	SSIM (↑)	0.9969	0.9953	0.9942	0.9870	0.9718
	LPIPS (↓)	0.0019	0.0039	0.0051	0.0147	0.0354
表情変化	PSNR (↑)	55.631	55.209	54.914	53.202	50.827
	SSIM (↑)	0.9985	0.9984	0.9983	0.9979	0.9968
	LPIPS (↓)	0.0002	0.0003	0.0003	0.0005	0.0012
喋り	PSNR (↑)	55.296	54.752	54.266	51.990	48.398
	SSIM (↑)	0.9984	0.9983	0.9982	0.9975	0.9954
	LPIPS (↓)	0.0003	0.0003	0.0004	0.0007	0.0020

次に, 表 1 に上記の変換結果に対し PSNR, SSIM, LPIPS を用いた評価を示す. 表 1 から Low, Mask の処理 fps の低下に伴って画像が劣化していることが読み取れる.

また, 人間の視覚特性をモデル化した動画の評価指標の FovVideoVDP [9] を用いて 500 fps の出力動画の評価を行った. FovVideoVDP は Reference となる動画と比較対象の動画を入力とし, Just-Objectable-Difference (JOD) が出力される. JOD は 10 を最大とし, JOD が 9 以下のとき Reference は比較対象よりも好ましい動画だと区別がつくことを示すスコアである. つまり, JOD が 9 以下のとき, 対象動画が劣化していると見なすことができる. ここで, 表 2 に FovVideoVDP を用いて評価した結果を示す. 同表から低周波成分の変換とマスクの生成処理の fps を低下させることで JOD が低下する傾向を確認した.

また, 表 1, 表 2 に共通して平行移動・回転運動と表情変化・喋りの間で数値の傾向に違いがある. これは動画の種類によって顔の動きの大きさが異なることに起因しており, 本手法では対象の動きの大きさによって劣化の大小が変化することが読み取れる.

次に, 本手法による処理時間について評価を行った. 低周波成分の変換とマスクの生成処理の fps を他処理よりも低く設定することで同処理に用いられるリソースを削減しつつ, 他処理と合わせて並行処理を行い, 処理時間を算出した. ここでは画像の分解, 高周波成分の変換, 画像の再構

表 2: FovVideoVDP [9] による評価. 表内上段の数値は低周波成分の変換の fps, マスクの生成処理の fps を表す.

JOD (↑)	125,250	100,200	$\frac{250}{3}, \frac{500}{3}$	50,100	25,50
平行移動	9.5397	9.3704	9.2242	8.6969	7.8986
回転運動	9.6118	9.4623	9.3634	8.9545	8.3578
表情変化	9.8735	9.8816	9.8509	9.7331	9.4777
喋り	9.8510	9.8323	9.7878	9.6093	9.2273

表 3: 平均処理時間. 表内の括弧で囲われた数値は低周波成分の変換 fps, マスクの生成処理の fps を表し, 並行は fps を特別に設定せずに実行する条件を表す. また表に示す数値の単位は ms である.

	LPTN [5]	並行	(125,250)	(100,200)	(50,100)	(25,50)
RTX 3090	3.053	2.516	2.209	2.135	2.136	2.128
RTX 5090	1.462	1.843	1.238	1.171	1.059	1.019

成に要した時間の合計が処理時間に当たる. 評価に用いた GPU は GeForce RTX 3090 (VRAM 24 GB) と GeForce RTX 5090 (VRAM 32 GB) である. 表 3 に上記の結果を示す. 同表から低周波成分の変換が 100 fps, マスクの生成処理が 200 fps の設定において, LPTN [5] の従来処理に対して RTX 3090 で 1.43 倍, RTX 5090 で 1.25 倍の高速化を確認した.

5. 考察

表 2 に示した FovVideoVDP による評価と表 3 に示した処理時間の評価から, 画像の劣化を抑えながら img2img を高速化できることを確認した. 平行移動や回転運動のような動きの大きい動画についても, JOD を 9 以上に保ちながら RTX 3090 で 1.43 倍, RTX 5090 で 1.25 倍の高速化が可能である.

一方, DPM における遅延知覚の閾値として報告されている 3.87 ms [7] に対しカメラの露光時間やプロジェクタの投影時間を差し引くと, 本稿で示した img2img の処理時間は十分ではない. そのため, img2img のさらなる高速化が必要である. 加えて, 本稿では低周波成分の変換, マスクの生成について処理 fps を設定して定量評価を行ったが, 評価を行った条件は離散的であるため, 処理時間および画像品質の観点から適切な処理 fps について調査する必要がある.

また, 図 4 上段の Smile の入力画像と Reference を観察すると, 入力画像に含まれているほうれい線上の皺に重なるような皺の生成が確認された. 他にも, 図 4 の Smile や Sad と比べて, 図 3 ではゴルゴ線 (Mid cheek line) 上に皺が強く生成されていることが確認された. 上記の結果は, 対象が持つ皺や表情に合わせて異なる変容が行われた結果だと考えられる.

6. まとめ

本稿では, LP による画像分解を用いた img2img の処理を並行化し, 一部処理の頻度を低下させ, 同処理の最新の計算結果を他処理で利用することで, 生成画像の劣化を知覚されにくいように抑制しつつ処理時間を削減する手法を

提案し, 生成される画像の品質や処理時間について定量評価を行った. その結果, 生成画像の劣化を抑えつつ処理時間の削減が可能であることを確認した.

一方, img2img を用いる DPM の実投影を実現できていない課題や, DPM の遅延を知覚されないための高速性が達成されていない課題が残る. また, 本稿では図 1 にて処理を個別に示しているが, すべての処理を統合し NIR 画像から投影画像を生成する img2img を提案することで処理の簡略化や短縮が望める.

参考文献

- [1] Hao-Lun Peng, Kengo Sato, Soran Nakagawa, and Yoshihiro Watanabe. Perceptually-Aligned Dynamic Facial Projection Mapping by High-Speed Face-Tracking Method and Lens-Shift Co-Axial Setup. *IEEE Transactions on Visualization and Computer Graphics*, 2025.
- [2] Takashi Nomoto, Ryo Koishihara, and Yoshihiro Watanabe. Realistic dynamic projection mapping using real-time ray tracing. In *ACM SIGGRAPH 2020 Emerging Technologies*, 2020.
- [3] 袁璐, 彭浩倫, 林隆介, 渡辺義浩. プロジェクションマッピングによる顔の年齢変容に関する検証. 第 68 回複合現実感研究会, MR2023-4, 2023.
- [4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5967–5976, 2017.
- [5] Liang Jie, Zeng Hui, and Zhang Lei. High-Resolution Photorealistic Image Translation in Real-Time: A Laplacian Pyramid Translation Network. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9387–9395, 2021.
- [6] Gaspard Zoss, Prashanth Chandran, Eftychios Sifakis, Markus Gross, Paulo Gotardo, and Derek Bradley. Production-Ready Face Re-Aging for Visual Effects. *ACM Transactions on Graphics*, Vol. 41, No. 6, 2022.
- [7] Hao-Lun Peng, Shin'ya Nishida, and Yoshihiro Watanabe. Studying User Perceptible Misalignment in Simulated Dynamic Facial Projection Mapping. In *2023 IEEE International Symposium on Mixed and Augmented Reality*, pp. 493–502, 2023.
- [8] 吉澤達也. 視覚モダリティにおける色情報の役割とその機序. *視覚の科学*, Vol. 36, No. 1, pp. 12–14, 2015.
- [9] Rafał K. Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. FovVideoVDP: a visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics*, Vol. 40, No. 4, 2021.