長岡技術科学大学大学院
工学研究科修士論文

題 目

拡張畳み込みニューラルネットワークを用いたフレーム補間に関する研究

指導教員

准教授　　杉田 泰則

著 者

電気電子情報工学専攻

16315091　　小林　新

令和 2 年 2 月 7 日

ABSTRACT

A Study on Frame Interpolation using Dilated Convolutional Neural Network

Author : 16315091 Arata KOBAYASHI
Supervisor : Assoc. Prof. Yasunori SUGITA

Frame interpolation is one of the issues in computer vision, and is applied to high frame rate conversion and slow motion creation. Frame interpolation generate a new intermediate frame from consecutive frames of the existing video to convert to a higher frame rate.

In traditional frame interpolation method, the optical flow between input frames is estimated, and pixels are synthesized along the flow vector to interpolate the frames. However, since this method depends on the accuracy of the optical flow, it is difficult to estimate in the area with blur and brightness change. Therefore, the generated image has a distortion. In recent years, frame interpolation methods based on direct image generation using convolutional neural network have been proposed. However, there are problems such as blur and disturbance in the generation result and high computational cost. In order to solve these problems, the method combining the optical flow method and the neural network method has been proposed. However, the image generated by this method has a distortion because the estimation accuracy of the optical flow decreases for large motion between frames.

This paper proposes a method using dilated convolution that can learn features in a wider area in order to cope with large motion between frames. In the conventional network structure using convolutional layer and pooling layer, a wide area can be learned by downsampling the feature map with pooling layers, however spatial information is lost by downsampling. In the proposed method, dilated convolution is used in place of the standard convolutional layer and pooling layer, and the structure is configured to exponentially increase the receptive field of the network while maintaining the resolution of the feature map. In addition, skip connection is used to keep the feature map of each layer behind the network and to suppress learning for only wide receptive fields.

The images generated by the proposed method and the conventional method were evaluated with PSNR and SSIM. As a result, the proposed method had results equal to or better than the conventional method in various examples including large motion between frames. From the above results, it was confirmed that dilated convolution is effective for interpolation between frames with large motion.

目次

第 1 章

はじめに

1.1 研究背景

フレーム補間とは，コンピュータビジョン分野における課題の 1 つであり，高フレームレート変換やスローモーション生成などに適用される技術である．フレーム補間では，既存ビデオの連続するフレームから新たな中間フレームを生成することで高フレームレートへ変換する．

従来のフレーム補間手法として，フレーム間のオプティカルフローを推定し，そのフローベクトルに沿うようにピクセルを合成しフレームを補間する方法がある [1,2]．しかし，これらの方法はオプティカルフローの精度に大きく依存しており，オクルージョンやぼやけ，輝度変化を伴う領域では推定が困難となる．近年では，畳み込みニューラルネットワーク (CNN) を用いたオプティカルフロー推定も行われているが [3,4]，これらの手法ではオプティカルフロー画像のような特別な学習データが必要となり，一般的なシーンなどを学習させることが困難である．

また，近年では CNN によって直接補間画像を生成するフレーム補間手法も研究されている．敵対的生成ネットワークによるフレーム補間手法 [5] は，フレーム補間に敵対的生成ネットワークを初めて適用した手法であり，マルチスケールのネットワーク構造で補間フレームを生成するが，生成結果にはぼやけや乱れが含まれている．CNN により各ピクセルの空間適応型畳み込み 2D カーネルを推定し，近傍ピクセルから補間フレームを生成する手法 [6,7] も提案されているが，大きな動きに対しては推定カーネルサイズが大きくなり計算コストが非常に高くなるといった問題がある．

これらのオプティカルフローベース手法とニューラルネットワークベース手法の利点を組み合わせた手法も提案されている [8]．この手法では，CNN によってフレーム間のオプティカルフローと時間成分で構成されるボクセルフローを推定し，その推定結果から中間フレームを合成する．ボクセルフローの推定結果は直接評価されず，推定結果から

合成された中間フレームが評価されるため，任意の動画を学習させることが可能となっている．また，合成フレームはボクセルフローに基づくピクセルのコピーであるため，補間フレームを直接生成する手法よりぼやけが少ない結果が得られる．しかし，この手法ではフレーム間の動きが大きい場合にはボクセルフローの推定が困難となり，生成画像に乱れが生じる．

1.2　研究目的

本論文では，フレーム間の大きな動きに対応させることを目的とし，より広い領域での特徴を学習可能な拡張畳み込み [9] を用いる手法を提案する．従来のような畳み込み層とプーリング層を用いたネットワーク構造では，プーリングにより特徴マップをダウンサンプリングすることで広域を学習できるが，ダウンサンプリングにより情報の欠落が生じる．提案法では，通常の畳み込み層とプーリング層の代わりに拡張畳み込みを用いることで，特徴マップの解像度を保持しつつネットワークの受容野を指数関数的に増加させる構造とする．また，ネットワークに skip connection を用いることで，各層の特徴マップをネットワーク後方に保持し，広い受容野のみに対する学習を抑える．

1.3　本論文の構成

本論文の構成は以下の通りである．第 1 章では，本論文の研究背景及び目的を述べた．第 2 章では，提案法に用いる基礎理論について述べる．第 3 章では，従来法とその問題点について述べる．第 4 章では，提案法である拡張畳み込みを用いたフレーム補間について述べる．第 5 章では，提案法と従来法の比較実験を行い，提案法の有効性を示す．第 6 章では，本論文を通してのまとめと今後の課題を述べる．

第 2 章

基礎理論

2.1　畳み込みニューラルネットワーク

畳み込みニューラルネットワーク (Convolutional Neural Network: CNN) とは，畳み込みを用いたニューラルネットワークの 1 つであり，主に画像認識や物体検知などのコンピュータビジョン，画像生成などの画像処理に使用される．一般的に CNN は，畳み込みと活性化関数を含む畳み込み層と入力のダウンサンプリングを行うプーリング層を複数組み合わせたネットワーク構造となっている．

2.1.1　畳み込み層

畳み込み層の例を図 2.1 に示す．入力としてサイズ H × W で C チャンネル (H × W × C) の画像が与えられた場合，畳み込み層では同チャンネル数であるサイズ FH × FW × C の畳み込みフィルタを用いて畳み込みを行う．畳み込みは入力と畳み込みフィルタの対応する各要素の乗算・総和であり，フィルタを一定間隔でスライドさせながら入力全体に対して行うことで特徴マップが得られる．また，FH × FW × C の畳み込みフィルタ 1 つにつき特徴マップが 1 つ生成され，畳み込みフィルタを OC 個適用することでサイズ OH × OW × OC の特徴マップが得られる．得られた特徴マップは，活性化関数を適用することで畳み込み層の出力となる．

2.1.2　プーリング層

プーリング層では，プーリング処理により入力をダウンサンプリングし出力する．プーリング処理の例を図 2.2 に示す．例では，入力に対して各範囲内の最大値を出力とする処理を行っており，最大値プーリング (Max pooling) と呼ばれる．プーリング処理を行うことで，特徴マップ内の特徴的な要素を残しつつダウンサンプリングし，計算コストを削減

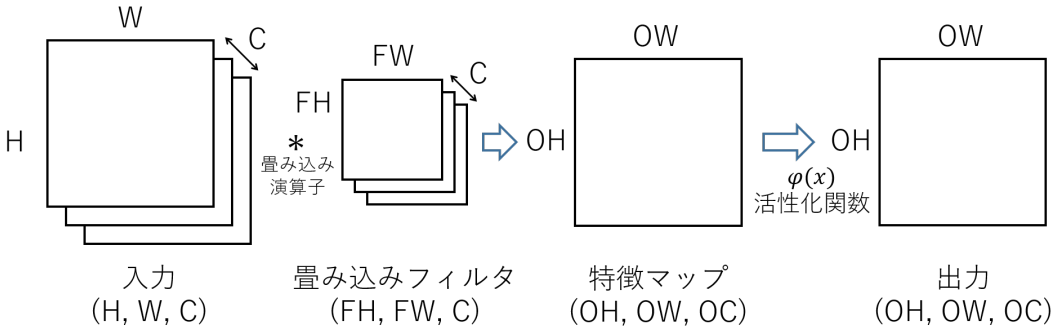


図 2.1　畳み込み層

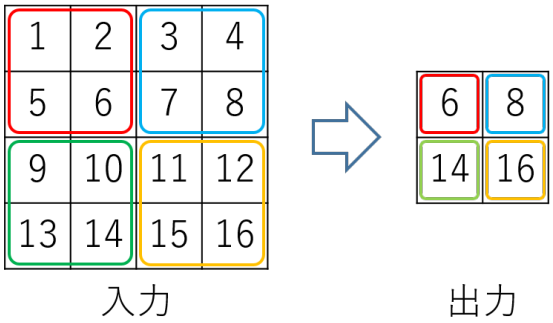


図 2.2　最大値プーリング

できる．また，プーリングには微小な位置変化や回転に対するロバスト性を高める効果もあり，画像認識や物体検出において有用である．

第 3 章

従来法

3.1 手法

フレーム補間に関する研究として，フレーム間のフロー推定からフレーム合成を学習する機械学習手法があり，Deep Voxel Flow (DVF) [8] と呼ばれている．ここでは，この手法について説明する．

図 3.1 に DVF のモデル構造を示す．DVF では，CNN により入力フレーム間の空間成分と時間成分で構成されるボクセルフローを学習し，後に続くボリュームサンプリング層によって中間フレームの合成を行う．まず，任意の動画内での連続する 3 フレーム (I_1, I_2, I_3) を学習データとし，中間フレーム I_2 を除いた (I_1, I_3) をネットワークへの入力 $\mathbf{X}$ とする．ここで，ネットワークへの入力は 256×256 にリサイズされ，[-1, 1] の範囲に正規化される．ネットワークは，中間フレームを直接学習するのではなく，フレーム間のボクセルフロー $\mathbf{F} = (\Delta x, \Delta y, \Delta t)$ を学習する．ここで，$(\Delta x, \Delta y)$ は入力フレーム間のオプティカルフローである空間成分，Δt は時間成分を表している．ネットワークに続くボリュームサンプリング層では，まずボクセルフローの空間成分を用いて各入力フレームから対応位置 $\mathbf{L}^0$, $\mathbf{L}^1$ を推定し，周囲 4 ピクセルをサンプリングする．次に，時間成分を入力フレーム間の重みとして使用し，ピクセルを合成することで最終的な出力である合成フレーム $\hat{\mathbf{Y}}$ を生成する．

ネットワークの構成は，3 つの畳み込み層とプーリング層によるエンコーダ部，3 つの畳み込み層とアップサンプリング層によるデコーダ部，1 つのボトルネック層となっている．プーリング層では最大値プーリングにより特徴マップをダウンサンプリングし，アップサンプリング層ではバイリニア補間により特徴マップをアップサンプリングする．また，入力フレーム間の空間情報をよりよく保持するために，エンコーダ-デコーダ間の対応する畳み込み層の特徴マップを結合する skip connection が追加されている．

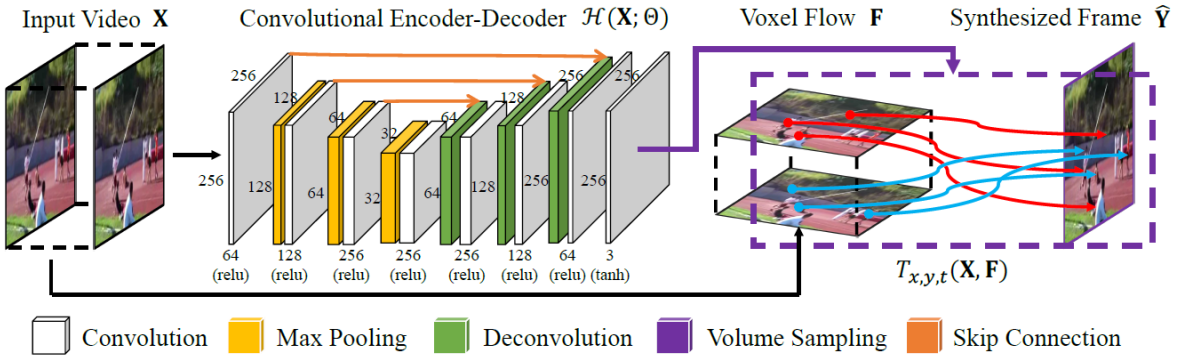


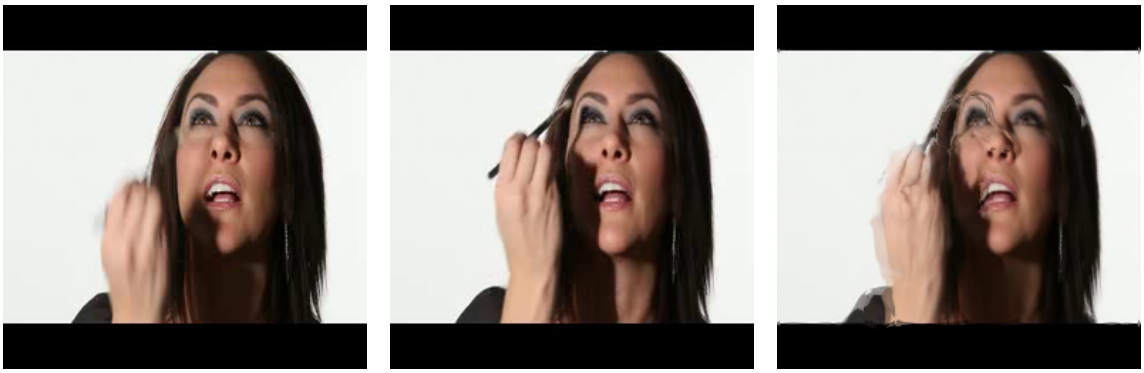
図 3.1　DVF のモデル構造（文献 [8] より引用）

3.2　問題点

従来法では，ネットワークの畳み込み層に通常の畳み込みを使用しているため，図 3.2 のようにフレーム間の動き領域が小さい場合はうまく機能するが，図 3.3 のようなフレーム間の動き領域が大きい場合ではオプティカルフローの推定精度が低下し，乱れた画像が生成されてしまう．また，ネットワークがエンコーダ-デコーダ構造であり，プーリング層を使用しているためフレーム間の位置情報等の情報欠落が発生する．

(a) 入力フレーム I_1　(b) 入力フレーム I_3　(c) 生成画像 $\hat{\mathbf{Y}}$

図 3.2　DVF の生成画像例 1



(a) 入力フレーム I_1　(b) 入力フレーム I_3　(c) 生成画像 $\hat{\mathbf{Y}}$

図 3.3　DVF の生成画像例 2

第 4 章

提案法

4.1　提案モデル

図 4.2 に提案法のモデル構造を示す．提案法では，フレーム間の大きな動きに対応することを目的として，ネットワークに拡張畳み込みを使用する．提案法は DVF に基づいており，ネットワークで入力フレーム間の空間成分と時間成分を学習し，ボリュームサンプリング層によって中間フレームを合成する．ここで，ネットワークへの入力は 256×256 にリサイズされ，[-1, 1] の範囲に正規化される．

従来法のようなエンコーダ-デコーダ構造のネットワークでは，プーリング層で特徴マップをダウンサンプリング，アップサンプリング層でアップサンプリングすることにより，高レベルな特徴を学習しつつ計算コストを削減することができる．しかし，この構造では学習の過程で特徴マップの解像度が低下するため，空間情報が失われ推定精度が低下する．提案法では，通常の畳み込み層とプーリング層の代わりに拡張畳み込みを使用することで，特徴マップの解像度を保持しつつネットワークの受容野を指数関数的に増加させる構造とする．

4.2　拡張畳み込み (Dilated convolution)

拡張畳み込み [9] は，通常の畳み込みに対しパラメータ数の増加なしでより広い受容野での学習を可能にする畳み込み手法である．受容野とは，畳み込み層の出力において 1 つの要素が算出される際に畳み込まれる入力の要素領域であり，受容野が大きいほど入力に対して広い領域の特徴を学習可能となる．拡張畳み込みの例を図 4.1 に示す．図 4.1 において，色付き箇所が畳み込みフィルタ係数を表しており，マス目は入力の各要素を表している．拡張畳み込みでは，例に示すように畳み込みフィルタ係数が一定間隔離れて配置されており，入力の広い領域の要素を畳み込むことができる．ここで，畳み込み

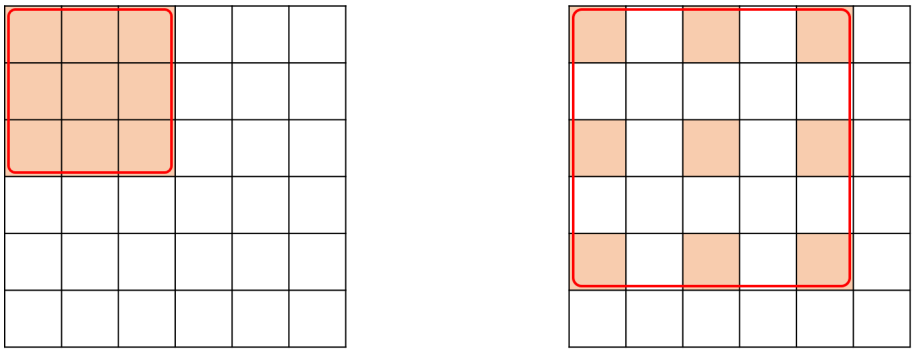


図 4.1　通常の畳み込みと拡張畳み込みの違い

フィルタ係数の間隔は rate(または Dilation) と呼ばれ，rate=2 ではフィルタ係数の間隔が 1，rate=4 では間隔が 3 となる．本論文では，以降 rate を拡張パラメータと呼ぶ．

通常の畳み込みでは受容野のサイズは層数に対して線形であるのに対し，拡張畳み込みでは拡張パラメータを指数関数的に増加させることで受容野のサイズを層数に対して指数関数的に増加させることが可能となる．

4.3　ネットワーク構成

ネットワークの基本的な構成は文献 [9] のコンテキストモジュールを参考にしている．コンテキストモジュールとは，セマンティックセグメンテーション分野において性能向上を目的として提案された追加モジュールであり，マルチスケールのコンテキスト情報を集約する設計となっている．コンテキストモジュールは，異なる拡張パラメータの拡張畳み込み層を複数用いた構成となっており，受容野を指数関数的に増加させることができる．提案法では，このコンテキストモジュールのネットワーク構成をフレーム補間に適用する．具体的には，提案法のネットワークは図 4.2 に示すように，1 つの畳み込み層と 5 つの拡張畳み込み層，そして 2 つの畳み込み層の計 8 層で構成される．ネットワークの各層の詳細なパラメータを表 4.1 に示す．提案法の受容野は表 4.1 に示すように，拡張畳み込みによって指数関数的に増加する．さらに，提案法のネットワークには skip connection を追加する [10]．skip connection では，対応する層の特徴マップを結合することによりネットワークの学習情報を保持する役割を持っている．提案法では，skip connection を用いて各層の特徴マップをネットワーク後方で結合することにより，複数の受容野の特徴マップを保持し，広い受容野のみに対する学習を抑える．

提案法の畳み込み層と拡張畳み込み層の活性化関数には $\alpha = 0.2$ の Leaky ReLU を適用し，最後の畳み込み層である出力層では Tanh 関数により出力値を [-1, 1] に正規化する．

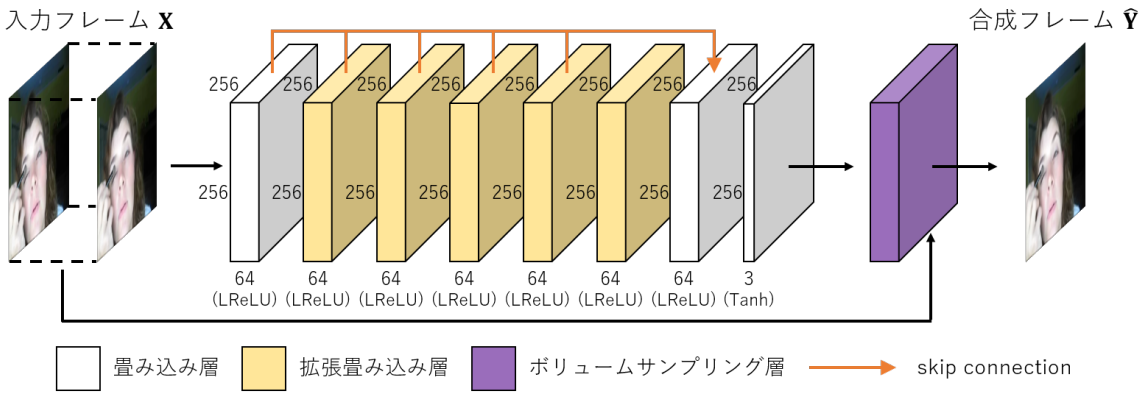


図 4.2　提案法のモデル構造

表 4.1　提案モデルのネットワークパラメータ

タイプ	フィルタサイズ	rate	受容野	出力特徴マップサイズ
畳み込み層	5 × 5	-	5 × 5	256 × 256 × 64
拡張畳み込み層	3 × 3	2	9 × 9	256 × 256 × 64
拡張畳み込み層	3 × 3	4	17 × 17	256 × 256 × 64
拡張畳み込み層	3 × 3	8	33 × 33	256 × 256 × 64
拡張畳み込み層	3 × 3	16	65 × 65	256 × 256 × 64
拡張畳み込み層	3 × 3	32	129 × 129	256 × 256 × 64
畳み込み層	5 × 5	-	133 × 133	256 × 256 × 64
畳み込み層	5 × 5	-	137 × 137	256 × 256 × 3

4.4　モデル学習

提案モデルでは，平均絶対誤差 (Mean Absolute Error : MAE) と Total Variation 正則化を損失関数として使用する．損失関数は以下の式で定義される．

$$L = \|\mathbf{Y}_{\mathrm{gt}} - \hat{\mathbf{Y}}\| + \lambda_1 \|\nabla \mathbf{F}_{\mathrm{motion}}\| + \lambda_2 \|\nabla \mathbf{F}_{\mathrm{mask}}\| \tag{4.1}$$

MAE はモデルの出力である合成フレーム $\hat{\mathbf{Y}}$ と正解画像である中間フレーム $\mathbf{Y}_{\mathrm{gt}}$ を用いて計算され，$\|\nabla \mathbf{F}_{\mathrm{motion}}\|$ はボクセルフローの空間成分 $(\Delta x, \Delta y)$ の正則化，$\|\nabla \mathbf{F}_{\mathrm{motion}}\|$ は時間成分 Δt の正則化である．ネットワークは，最適化手法として学習率 0.00001，$\beta_1 = 0.9$，$\beta_2 = 0.999$ の Adam を用いて学習される．また，ネットワークに Batch Normalization を適用することで，各層の入力がバッチごとに正規化され，学習の安定性が向上し収束が早まる．

第 5 章

実験

5.1 実験条件

提案法によるフレーム補間と従来法である DVF [8] によるフレーム補間の比較実験を行う．提案モデルではバッチサイズ 8，DVF ではバッチサイズ 16 で学習を行う．また，損失関数の正則化係数はそれぞれ，$\lambda_1 = 0.01, \lambda_2 = 0.005$ とする．

5.1.1 データセット

本論文では，データセットとして UCF101 [11] を用いる．このデータセットには様々な人間の行動のビデオが含まれており，101 のカテゴリに属する 13320 のビデオがある．学習データとテストデータには，データセット内の動画を連続する 3 フレームで 1 セットとして分割し，256×256 にリサイズ，[-1, 1] に正規化したものを使用する．また，学習データ数を 99083 セット，テストデータ数を 3337 セットとした．

5.1.2 評価指標

評価指標には，ピーク信号対雑音比 (Peak signal-to-noise ratio: PSNR)，構造的類似性 (Structural Similarity: SSIM) [12] を用いる．

PSNR は，信号の最大パワーと元画像に対する評価画像のノイズとの比率を示す指標であり，式（5.1）で求められる．ここで，MAX は画像の最大画素値であり，MSE (Mean Squared Error) は元画像と評価画像から算出される．評価目安として，値が大きいほど良好であり 40[dB] 以上で目視による見分けがつかない程度の再現度，30[dB] 以下で明らかな劣化が見られる再現度とされている [13].

$$\mathrm{PSNR} = 10 \cdot \log_{10} \frac{\mathrm{MAX}^2}{\mathrm{MSE}} \tag{5.1}$$

SSIM は，元画像と評価画像間での構造的類似度を示す指標であり，画像内の局所領域ごとに類似度を計算し，その平均値を出力値とする．SSIM は式（5.2）で求められ，本論文ではウィンドウサイズを 7×7 としている．ここで，μ_x, μ_y は各領域の画素値の平均値，μ_x, μ_y は標準偏差，σ_{xy} は共分散，C_1, C_2 は発散防止定数である．評価目安として，値が 1 に近いほど良好であり 0.98 以上で目視による見分けがつかない程度の再現度，0.90 以下で明らかな劣化が見られる再現度とされている [13].

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5.2}$$

5.2　実験結果

すべてのテストデータに対する PSNR，SSIM の平均値を表 5.1 に示す．提案法では，DVF に対し PSNR で 2.46 [dB]，SSIM で 0.027 上回る結果が得られた．

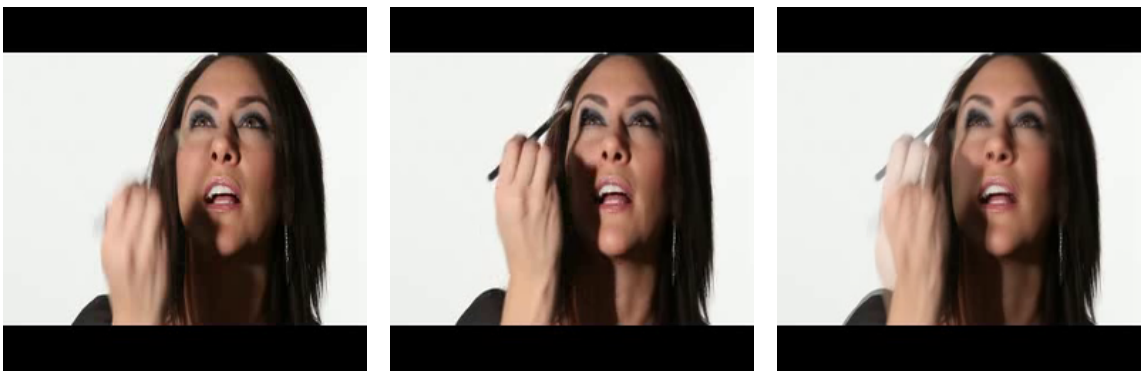
次に，実験条件に対する DVF（従来法）と提案法の生成結果例を図 5.1〜5.5 に示す．図 5.1〜5.5 において，(a) および (b) はネットワークへの入力フレーム，(c) はフレーム間の動き領域を簡易的に示すための入力フレーム合成画像，(d) は正解画像である中間フレーム，(e) は DVF による生成結果，(f) は提案法による生成結果である．

図 5.1 は入力フレーム間の動きが大きい画像に対する生成結果例を示している．DVF では動きの大きい右手部分に加えて顔部分まで生成画像が大きく乱れているのに対し，提案法ではある程度の外形を生成できていることがわかる．図 5.2 は入力フレーム間の動きが中程度の画像に対する生成結果例を示している．生成画像では DVF，提案法ともに右手周辺にぼやけが生じており，PSNR と SSIM も同程度の結果となっている．図 5.3 は入力フレーム間の動きが小さい画像に対する生成結果例を示している．図 5.3(c) に示すように，フレーム間の動きがかなり小さいため，DVF と提案法の生成結果はともに高い数値となっている．また，提案法では弓の外形や人の輪郭部分が DVF よりぼやけが少なく生成できており，PSNR と SSIM も DVF より高い結果となっている．図 5.4 は入力フレーム間の動きのある対象物が大きい画像に対する生成結果例を示している．図 5.4(c) に示すように，フレーム間では人全体に動きがあるため，DVF の生成結果も人部分が大きく乱れている．提案法では，ある程度のぼやけが生じているものの大きな乱れが少ない生成結果となっている．図 5.5 は入力フレーム間全体に動きのある画像に対する生成結果例を示している．図 5.5(c) に示すように，フレーム間全体に動きがあるため，DVF の生成結果も画像全体が乱れている．提案法では，画像全体の乱れが抑えられており，生成結果が大きく改善されていることがわかる．

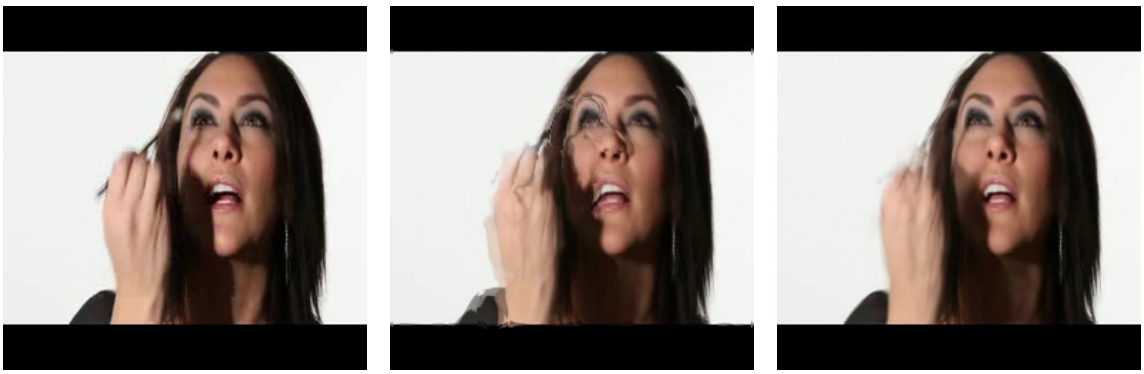
以上の結果から，提案法は，従来法に対し同等以上の結果を達成したと言える．

表 5.1　すべてのテストデータに対する平均値

手法	PSNR	SSIM
DVF	30.61 [dB]	0.900
提案法	33.07 [dB]	0.927

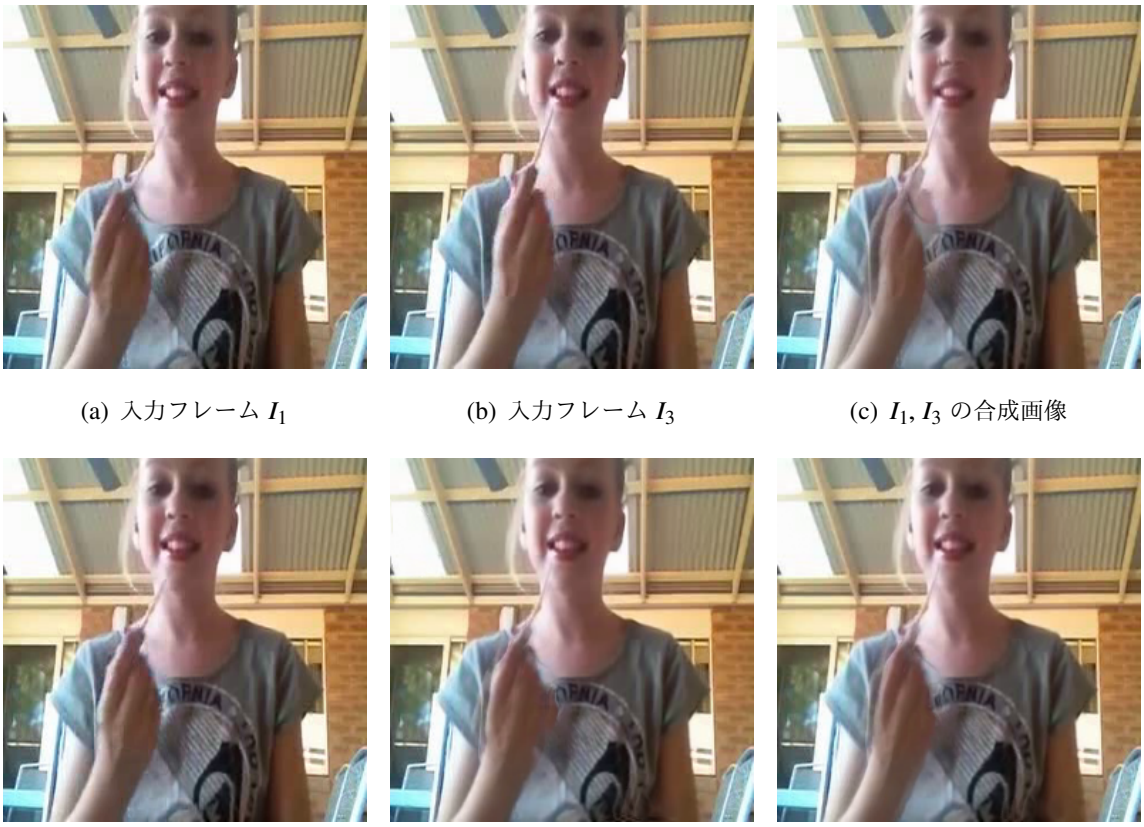


(a) 入力フレーム I_1　(b) 入力フレーム I_3　(c) I_1, I_3 の合成画像



(d) 正解画像 $\mathbf{Y}_{\mathrm{gt}}$　(e) DVF (PSNR=24.30, SSIM=0.873)　(f) 提案法 (PSNR=26.03, SSIM=0.908)

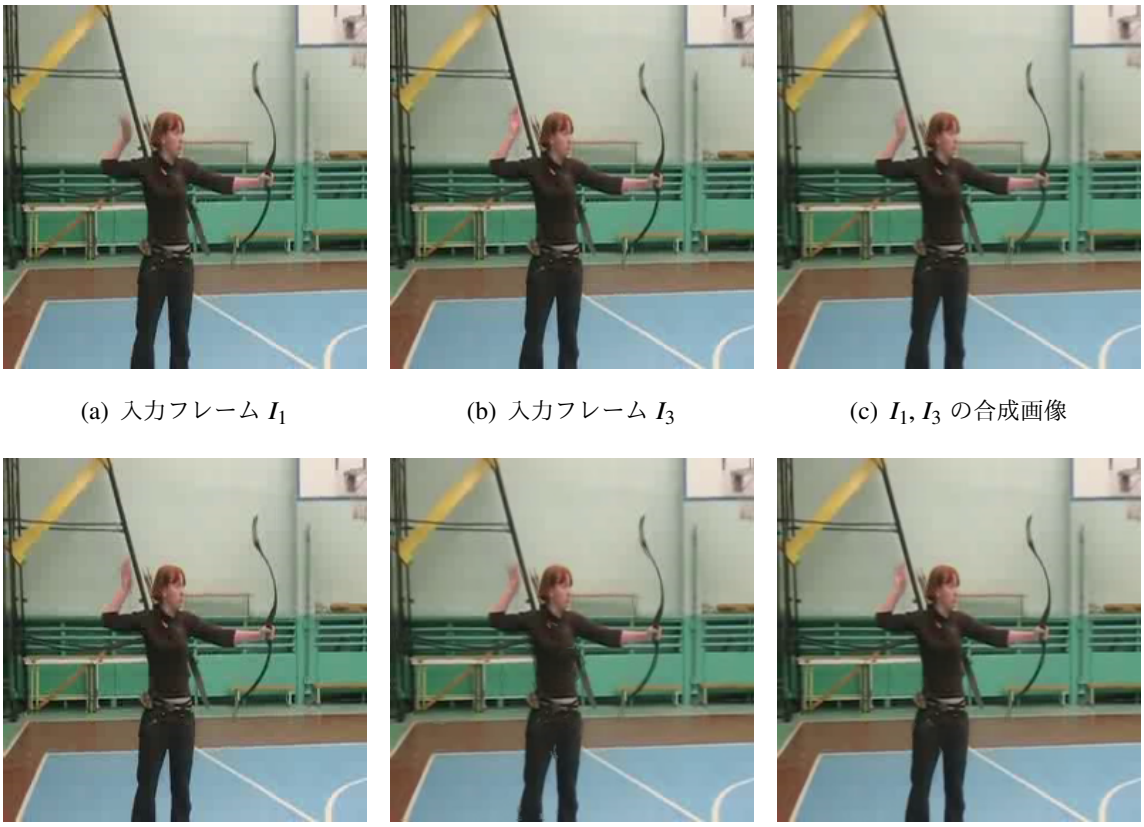
図 5.1　生成画像例 1



(a) 入力フレーム I_1　(b) 入力フレーム I_3　(c) I_1, I_3 の合成画像

(d) 正解画像 $\mathbf{Y}_{\mathrm{gt}}$　(e) DVF (PSNR=29.47, SSIM=0.935)　(f) 提案法 (PSNR=29.33, SSIM=0.935)

図 5.2　生成画像例 2



(a) 入力フレーム I_1　(b) 入力フレーム I_3　(c) I_1, I_3 の合成画像

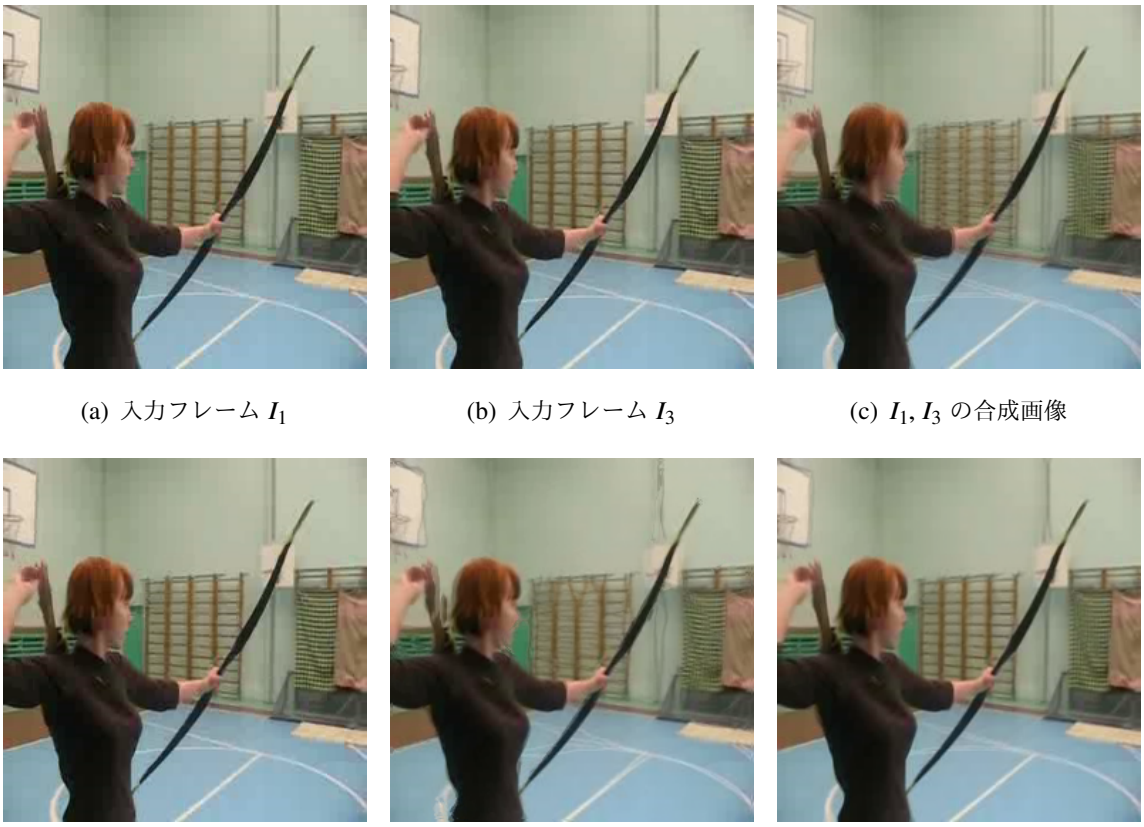
(d) 正解画像 $\mathbf{Y}_{\mathrm{gt}}$　(e) DVF (PSNR=32.64, SSIM=0.955)　(f) 提案法 (PSNR=34.13, SSIM=0.967)

図 5.3　生成画像例 3

(a) 入力フレーム I_1　(b) 入力フレーム I_3　(c) I_1, I_3 の合成画像

(d) 正解画像 $\mathbf{Y}_{\mathrm{gt}}$　(e) DVF (PSNR=24.44, SSIM=0.826)　(f) 提案法 (PSNR=29.51, SSIM=0.929)

図 5.4　生成画像例 4



(a) 入力フレーム I_1　(b) 入力フレーム I_3　(c) I_1, I_3 の合成画像

(d) 正解画像 $\mathbf{Y}_{\mathrm{gt}}$　(e) DVF (PSNR=27.99, SSIM=0.832)　(f) 提案法 (PSNR=32.12, SSIM=0.928)

図 5.5　生成画像例 5

第 6 章

おわりに

6.1　まとめ

本論文では，フレーム間の大きな動きに対応させることを目的に，拡張畳み込みを用いたフレーム補間手法を提案した．第 1 章では，フレーム補間に関する最近の手法と問題点，本論文の研究目的について述べた．第 2 章では，本研究で用いる畳み込みニューラルネットワーク（CNN）の基礎理論について述べた．第 3 章では，従来法として Deep Voxel Flow (DVF) を取り上げ，その手法と問題点について述べた．第 4 章では，本論文の提案法である拡張畳み込みを用いたフレーム補間手法について述べた．また，提案法に用いる拡張畳み込みについて説明した．第 5 章では，提案法の有用性を検証するために，同一の学習データを用いて DVF と提案法の評価実験を行った．DVF と提案法の生成画像を評価指標である PSNR と SSIM で比較したところ，様々な入力画像例において DVF と同等かそれ以上の結果が得られ，提案法の有用性が確認できた．また，すべてのテストデータに対する PSNR，SSIM の平均値を比較したところ，提案法では PSNR で 2.46 [dB]，SSIM で 0.027 上回る結果が得られた．以上から，ネットワークに拡張畳み込みを用いることで大きな動きに対するフレーム補間の精度向上が実現できたと言える．

6.2　今後の課題

本論文では，従来法である DVF に基づき拡張畳み込みを用いた手法を提案した．しかし，現状では依然として生成結果にぼやけや乱れが生じているため，生成精度が低い．したがって，生成画像の精度を上げるために，深度情報などのオクルージョンを考慮した手法 [14] を組み込むことが今後の課題である．また，提案法では現状単一のフレーム補間にのみ対応しているため，マルチフレーム補間に対応させることも今後の課題と言える．

謝辞

本研究を進めるにあたり，ご指導・ご助言を頂いた杉田泰則准教授に厚く御礼申し上げます．また，論文の審査において多くのご指示を頂きました，本学電気系岩橋政宏教授ならびに圓道知博准教授に厚く御礼申し上げます．最後に，本研究に関して多くの指摘をくださいました信号処理応用研究室の皆様に深く感謝いたします．

令和 2 年 2 月

参考文献

[1] Dhruv Mahajan, Fu-Chung Huang, Wojciech Matusik, Ravi Ramamoorthi, Peter Belhumeur, "Moving Gradients: A Path-Based Method for Plausible Image Interpolation", ACM Transactions on Graphics, 28(3), Aug. 2009.

[2] Simone Meyer, Oliver Wang, Henning Zimmer, Max Grosse, Alexander Sorkine-Hornung, "Phase-Based Frame Interpolation for Video", IEEE Conference on Computer Vision and Pattern Recognition, pp. 1410-1418, June 2015.

[3] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, Thomas Brox, "FlowNet: Learning Optical Flow with Convolutional Networks", IEEE International Conference on Computer Vision, pp. 2758-2766, Dec. 2015.

[4] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, Thomas Brox, "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks", IEEE Conference on Computer Vision and Pattern Recognition, pp. 1647-1655, July 2017.

[5] Michael Mathieu, Camille Couprie, Yann LeCun, "Deep multi-scale video prediction beyond mean square error", International Conference on Learning Representations, 2016.

[6] Simon Niklaus, Long Mai, Feng Liu, "Video Frame Interpolation via Adaptive Convolution", IEEE Conference on Computer Vision and Pattern Recognition, pp. 2270-2279, July 2017.

[7] Simon Niklaus, Long Mai, Feng Liu, "Video Frame Interpolation via Adaptive Separable Convolution", IEEE International Conference on Computer Vision, pp. 261-270, Oct. 2017.

[8] Ziwei Liu, Raymond Yeh, Xiaoou Tang, Yiming Liu, and Assem Agarwala, "Video frame synthesis using deep voxel flow", IEEE International Conference on Computer Vision, pp. 4473-4481, Oct. 2017.

[9] Fisher Yu, Vladlen Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions",

International Conference on Learning Representations, 2016.

[10] Yi Zhu, Shawn Newsam, "Learning Optical Flow via Dilated Networks and Occlusion Reasoning", IEEE International Conference on Image Processing, pp. 3333-3337, Oct. 2018.

[11] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild", CRCV-TR-12-01, Nov. 2012.

[12] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, Eero P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", IEEE Transactions on Image Processing, vol. 13, No. 4, pp. 600-612, Apr. 2004.

[13] 小箱 雅彦, "月刊 IM", Vol.50, No. 5, pp. 21-24, May 2011.

[14] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, Ming-Hsuan Yang, "Depth-Aware Video Frame Interpolation", IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3698-3707, June 2019.

付録

A　提案法の skip connection なしモデル

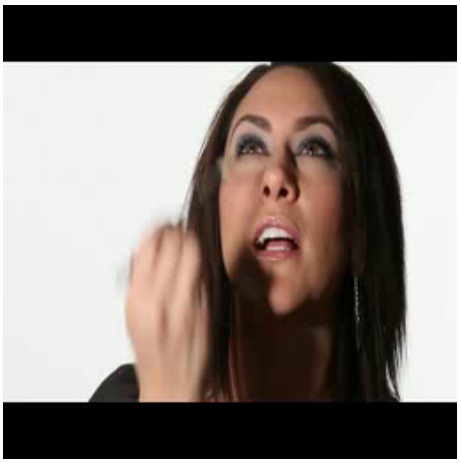
4 章及び 5 章の補足として，提案法の skip connection なしモデルでの実験結果を示す．ここで，skip connection なしモデルのバッチサイズは 16 である．

A.1　生成結果の比較

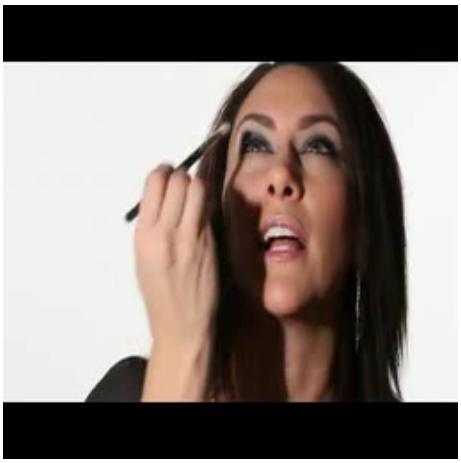
すべてのテストデータに対する PSNR，SSIM の平均値を表 A.1 に示す．次に，実験条件に対する DVF（従来法）と提案法の生成結果例を図 A.1〜A.5 に示す．図 A.1〜A.5 において，(a) および (b) はネットワークへの入力フレーム，(c) は正解画像である中間フレーム，(d) は DVF による生成結果，(e) は提案法 (skip connection なし) による生成結果，(f) は提案法 (skip connection あり) による生成結果である．

表 A.1　すべてのテストデータに対する平均値

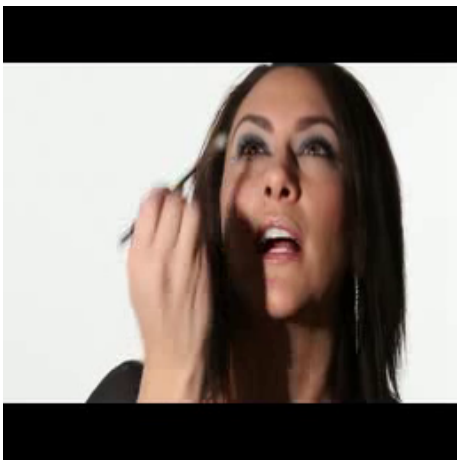
手法	PSNR	SSIM
DVF	30.61 [dB]	0.900
提案法 (skip connection なし)	31.98 [dB]	0.912
提案法 (skip connection あり)	33.07 [dB]	0.927



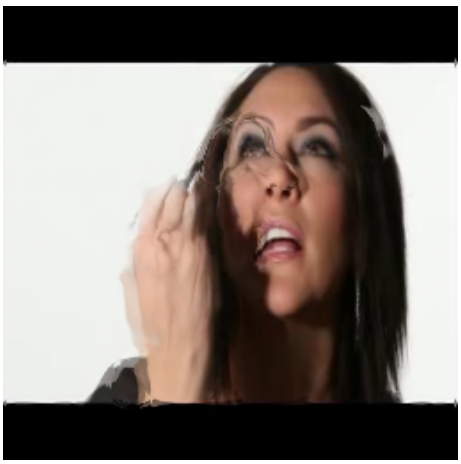
(a) 入力フレーム I_1



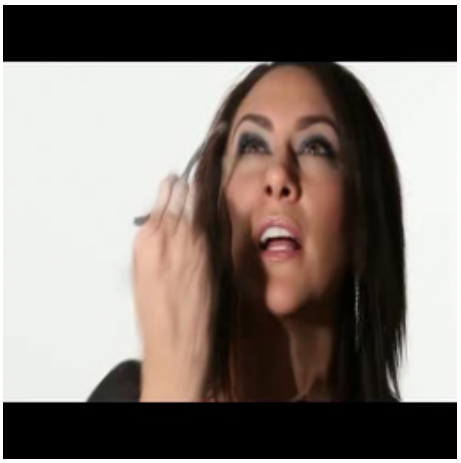
(b) 入力フレーム I_3



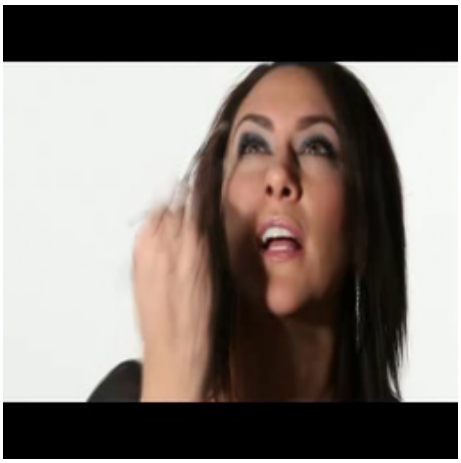
(c) 正解画像 $\mathbf{Y}_{\mathrm{gt}}$



(d) DVF
(PSNR=24.30, SSIM=0.873)

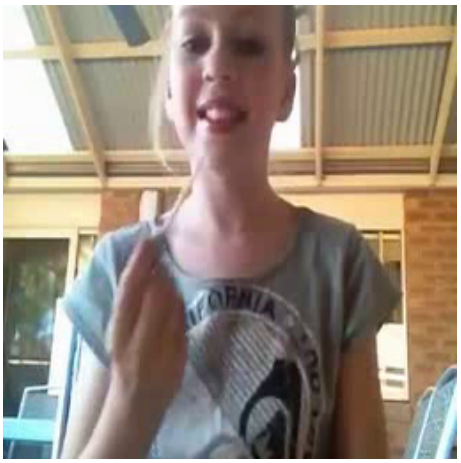


(e) 提案法 (skip connection なし)
(PSNR=25.64, SSIM=0.900)



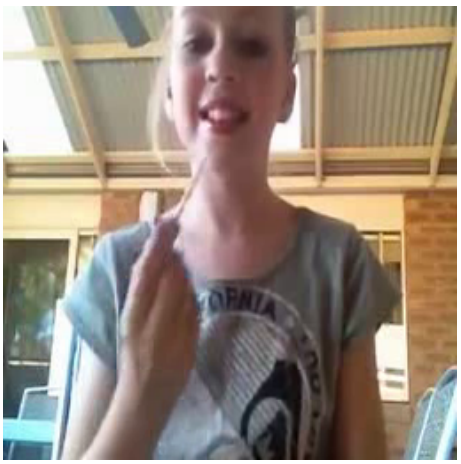
(f) 提案法 (skip connection あり)
(PSNR=26.03, SSIM=0.908)

図 A.1　生成画像例 1



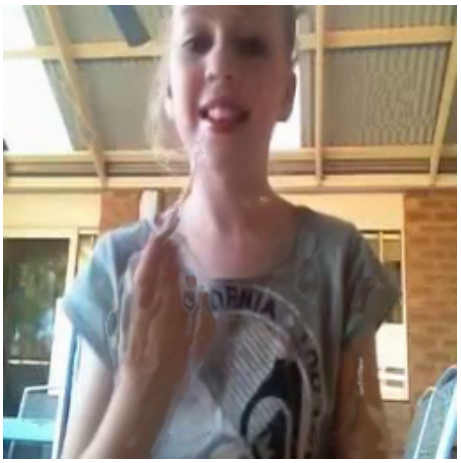
(a) 入力フレーム I_1

(b) 入力フレーム I_3



(c) 正解画像 $\mathbf{Y}_{\mathrm{gt}}$

(d) DVF (PSNR=29.47, SSIM=0.935)



(e) 提案法 (skip connection なし) (PSNR=28.23, SSIM=0.890)

(f) 提案法 (skip connection あり) (PSNR=29.33, SSIM=0.935)

図 A.2　生成画像例 2

(a) 入力フレーム I_1

(b) 入力フレーム I_3

(c) 正解画像 $\mathbf{Y}_{\mathrm{gt}}$

(d) DVF (PSNR=32.64, SSIM=0.955)

(e) 提案法 (skip connection なし) (PSNR=31.34, SSIM=0.940)

(f) 提案法 (skip connection あり) (PSNR=34.13, SSIM=0.967)

図 A.3　生成画像例 3

(a) 入力フレーム I_1

(b) 入力フレーム I_3

(c) 正解画像 $\mathbf{Y}_{\mathrm{gt}}$

(d) DVF (PSNR=24.44, SSIM=0.826)

(e) 提案法 (skip connection なし) (PSNR=27.54, SSIM=0.899)

(f) 提案法 (skip connection あり) (PSNR=29.51, SSIM=0.929)

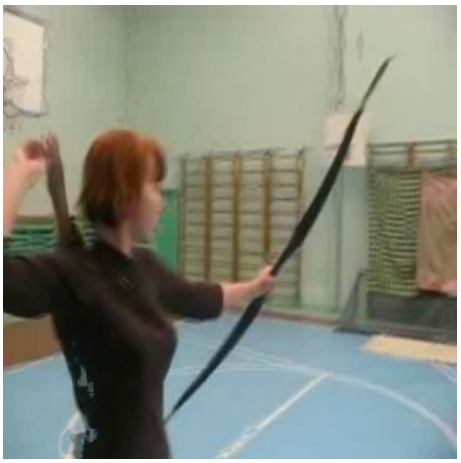
図 A.4　生成画像例 4

(a) 入力フレーム I_1

(b) 入力フレーム I_3

(c) 正解画像 $\mathbf{Y}_{\mathrm{gt}}$

(d) DVF
(PSNR=27.99, SSIM=0.832)



(e) 提案法 (skip connection なし)
(PSNR=28.78, SSIM=0.862)

(f) 提案法 (skip connection あり)
(PSNR=32.12, SSIM=0.928)

図 A.5　生成画像例 5