Vol.2020-CG-177 No.5
2020/3/17

物体情報の活用による画像の超解像

笹俣匡平 [1)]　　岡部誠 [1)]

概要：低解像度の画像を高解像度の画像に変換（超解像）する技術は過去にも多く提案されているが，幾何学的な構造やテクスチャが復元できず，物体特有の見た目が失われがちであった．我々は，物体認識器から得られる物体情報（人，自動車，建物，道路，樹木などのラベル）と超解像器を組み合わせて用い，精度の向上を図る．具体的には，画像中の各物体毎に，その物体カテゴリ専用に用意された超解像器を適用する，というのが基本的なアイデアである．実験の結果，既存手法と比較して，1/90 程度のデータセットしか学習に用いない場合でも，提案手法の結果は既存手法に比べ，匹敵するか，もしくは，上回る画質を達成できる可能性が確認できた．

キーワード：超解像，ディープラーニング，物体認識，ユーザインタフェース

1. はじめに

低解像度の画像を高解像度の画像に変換（超解像）する技術は過去にも多くの提案がある [1]．しかし，常に高品質な結果が得られるとは限らない．特に，既存技術では幾何学的な構造やテクスチャを復元することが難しく，物体特有の見た目が失われがちである．そこで，我々は物体認識器から得られるような物体情報（人，自動車，建物，道路，樹木などのラベル）と超解像器を組み合わせて用いることで，画質の向上が得られないかという点に着目し，研究を行っている．具体的には，入力画像に物体認識器を適用し，画像中の何処にどのカテゴリの物体が存在しているかを特定した上で，各物体の存在箇所に対し，その物体カテゴリ専用に準備した超解像器を適用する，というのが基本的なアイデアである．

一方，物体認識器も完璧ではない．全自動で計算され出力される結果を観察すると，おおまかには各物体の存在箇所が特定できているものの，ピクセル単位で正確な情報を得ることが難しかった．そこで，今回は物体認識器は用いず，ユーザが手作業で物体情報を入力することで実験を行った．超解像器は物体カテゴリの数だけ存在する．各超解像器の対応できる物体の範囲を絞り込んだため，各超解像器の学習に必要なデータセットの数が大幅に削減できるということが分かった．具体的には，既存手法と比較して，1/90 程度のデータセットしか学習に用いなかったにも関わらず，提案手法の結果は既存手法の結果に比べ，匹敵するか，もしくは，上回る画質を達成できる可能性が見えた．

2. 既存手法

SRGAN は 2016 年に Ledig らによって発表された超解像器である [2]．8192 枚の画像を学習した超解像器は高画質な結果を生成でき，その後の多くの文献で引用されている．我々が着目している SRGAN の問題点は，幾何学的な構造やテクスチャを復元することが難しく，物体特有の見た目が失われてしまう，という点である．例えば，図 1 -中央の SRGAN の結果を見ると，髪飾りやネックレスの部分などの構造が崩れてしまっている．

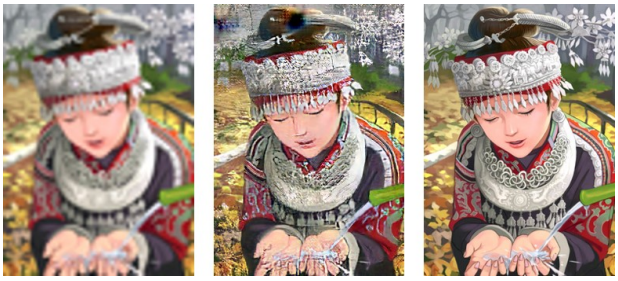

図 1: SRGAN による画像の超解像結果 [2]．左から入力画像（125×120 ピクセル），出力画像（500×480 ピクセル），正解画像．

1) 静岡大学 大学院総合科学技術研究科 工学専攻

ⓒ 2020 Information Processing Society of Japan

我々が調査した限りでは，SRGAN 以降に発表された論文で，SRGAN よりも優れた結果を出力できているものは，ある分野に特化した超解像器が多いという印象である．例えば，人の顔に特化した超解像器は高画質な超解像結果を出力できる [3].また，画像検索器を介し，データセットの中から入力画像に類似した画像を検索し，その上で超解像に用いる手法 [4]がある．ただし，この手法は検索された類似画像を超解像のためのサンプリングの対象として用いる手法であり，そのため，超解像された出力画像の見た目が入力画像から大きく変化してしまう可能性がある．また，動画の超解像については時間的に補間し合える情報を利用できるため，驚くほど高画質な超解像結果を得ることが可能となる [5]．しかし，この手法を静止画像の超解像に適用することはできない．

Blau らは 2018 年に発表した論文で超解像器の評価を行っている [6]．超解像器は通常，入力された低解像度の画像からの歪みがなるべく小さくなるように，高解像度な画像を生成しようとする．この時，結果の品質と歪みの間に図 2 のような関係があることが確かめられている．図 2 の縦軸は超解像画像の知覚品質であり，ユーザテストによって主観的に評価される定性的な値である．図 2 の横軸は歪みである．これは定量的な値であり，低解像度画像と超解像画像の間で計算される MS-SSIM 誤差である．図 2 より，結果の品質を高く保ちつつ，歪みを小さくすることは同時には達成できず，即ち，品質の高い超解像画像を生成したいならば，歪みは大きくなるべきであることが示されている．我々は知覚品質の向上を第一の目標と捉え，図 2 のなるべく右下に位置するようなアルゴリズムを開発したい．

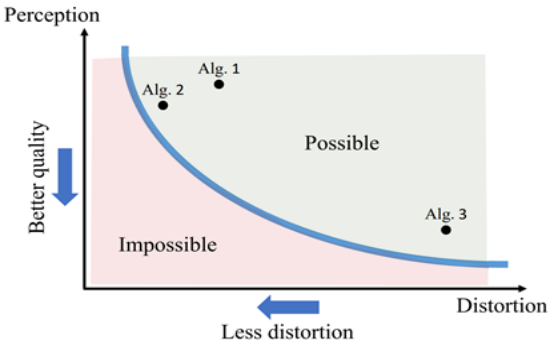

図 2: 縦軸が超解像画像の知覚品質，横軸が低解像度画像と超解像画像の間の歪み [6]

3. SRGAN

SRGAN は，generative adversarial network (GAN)に基づいた超解像器である [2]．GAN は generator と discriminator という 2 つのネットワークから成る．SRGAN の generator と discriminator を図 3 に示す．

まず，generator は低解像度画像と高解像度画像のペアからなるデータセット（8192 ペア）を用い，低解像度画像を入力すると対応する高解像度画像が出力されるように学習しておく．この時，次の MSE 誤差に基づく loss 関数を最小化することを目標に学習を行う：

$$l^{SR}_{VGG/i,j} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}(\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2$$

ただし，I^{LR}は低解像度画像，I^{HR}は高解像度画像，$\phi_{i,j}(I)$は画像 I を学習済み VGG19 モデルに入力した時に出力される特徴量のベクトルとする．

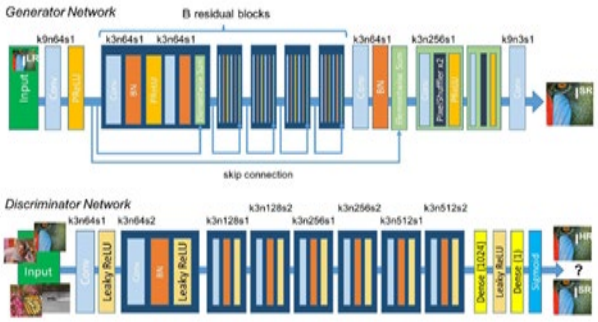

図 3: SRGAN のネットワーク構造 [2]

上記の学習で得られる generator を出発点とし，GAN の学習を行う．まず，discriminator の学習を行う．Discriminator の仕事は入力された高解像度画像が generator によって生成されたものか否かを正しく判定することである．Generator によって生成された超解像画像が入力された際は 0 を，オリジナルの高解像度画像を入力された際には 1 を出力するように discriminator の学習を行う．次に，generator の学習を行う．GAN モデルにおける generator の仕事は，discriminator が 1 を出力してしまうような画像を生成する，即ち，discriminator が超解像画像をオリジナルの高解像度画像と見間違えてしまうような画像を生成することである． この generator の学習の際は，discriminator ネットワーク内のパラメータは更新しないようにしつつ，discriminator が 1 を出力するように generator ネットワーク内のパラメータのみを更新する．以上の discriminator と generator の学習を交互に行うことによって両者の精度を上げていく．

GAN における generator の目的は discriminator をだます(1 を出力させる）ことであり，言い方を変えれば，だませるのであれば generator の出力は何でも構わず，従って，generator の出力が入力した低解像度画像の超解像画像になる保証がない．そこで，MSE 誤差に基づく loss 関数を最適化したモデルを GAN の学習の出発点とし，それの性能を上げていくような学習を行っている．

ⓒ 2020 Information Processing Society of Japan

4. 提案手法

本手法では，物体カテゴリ毎に超解像器を用意した．超解像器のモデルは SRGAN と同じである．入力画像にそれらを適用することで，物体カテゴリと同数の超解像画像が得られる．物体毎に適切な超解像画像を使いたいため，ユーザは手作業で物体の存在箇所を指定し，合成を行う．既存手法では，8192 枚のデータセットを長時間掛けて学習し，単一のモデルを作っていた（図 2）．一方で，提案手法では各超解像器毎に学習すべきデータセットは，その物体カテゴリの画像のみとなるためデータセットの量が少なく済み，その結果，短時間で学習が済むと考えた．

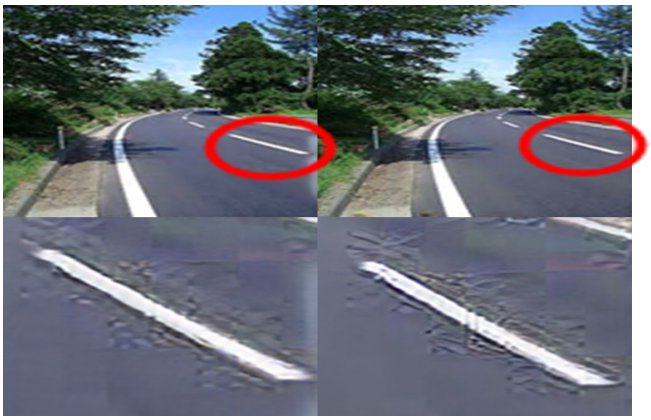

図 4: 左は提案手法の「道路」カテゴリの超解像器（32 枚の画像を学習）の結果．右は既存手法（8192 枚の画像を学習）の結果．

図 4 に提案手法の「道路」カテゴリの超解像画像と，既存手法による超解像画像を掲載した．両者を比較すると，提案手法は圧倒的に少ないデータセットしか学習していないにも関わらず，既存手法に匹敵した結果が出力できている．データセットの量は既存手法に比べ 1/256 も少ないが，データセットを道路のみに絞ることで，道路の物体構造をより素早く学習できたからではないかと考えている．一方で，データセットが削減できたことに伴い，学習時間も大幅に削減することが可能になることを期待したが，学習時間は 1/5 程度は必要となり，それ以上に学習時間を削減すると，満足のいく結果が出せないという結果になった．

	既存手法	提案手法
データセット	8192枚	32枚
学習時間	約50時間	約10時間

図 5: 図 4 の結果を出力するためのデータセットの量と学習時間の比較．

5. 結果と考察

図 6 に，3 つの物体カテゴリ「建物」「道路」「樹木」について学習させた超解像器を用いた結果を掲載する．図 6-上は 3 つの超解像器から出力された結果であり，図 6-下はそれらを合成した結果である．各色の円で囲んだ領域が合成されている．この領域の指定と，3 つの画像の合成には Photoshop のスタンプツールを利用した．合成を手作業で行う理由は，全自動の物体認識器の出力が完璧でないためである．図 7 に示すように，物体に対し完璧なマスクをかけることが難しい．そこで，高画質な合成結果を得るために，提案手法ではユーザが手作業により合成を行うことで，この問題を解決することとした．

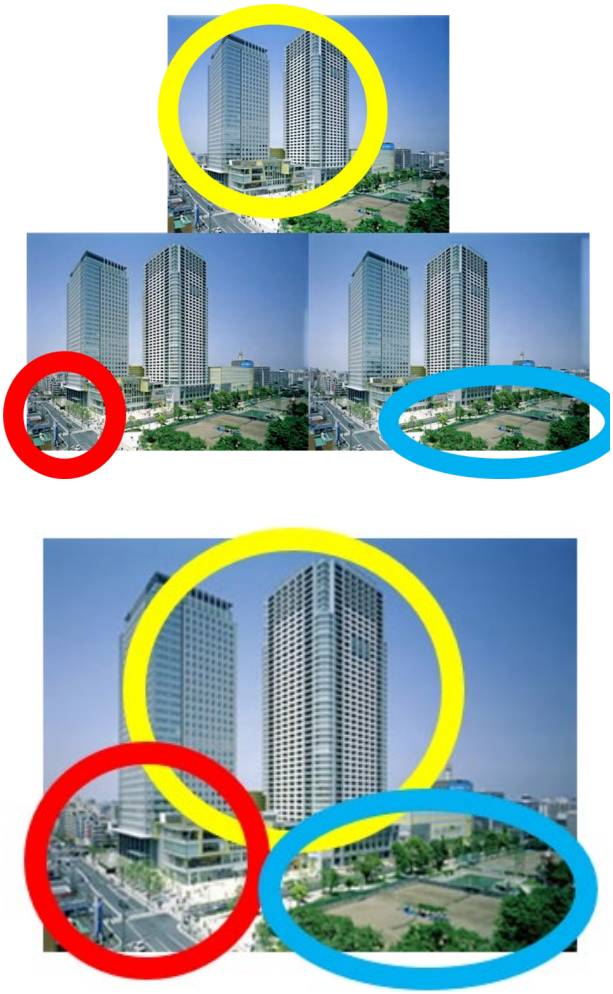

図 6: 建物，道路，樹木の超解像とその合成結果

図 7：既存手法による物体認識の結果 [7]

ⓒ 2020 Information Processing Society of Japan

次に提案手法で作成された超解像画像（図 6）と既存手法により作成された超解像画像との部分的な比較結果を図 8 と図 9 に示す.

図 8：建物部分の比較結果（左が既存手法,右が提案手法）

図 9：樹木部分の比較結果（上が既存手法,下が提案手法）

図 8 について，既存手法の結果を見ると，建物の窓部分に不要な細かい模様が出現してしまっているのに対し，提案手法の結果にはそういった模様がなく，建物の構造をより正しく再現できているのは提案手法の方ではないかと思われる．一方で，図 9 については，既存手法の結果を見ると，樹木に必要な細かい模様が再現出来ていないのに対し，提案手法の結果にはそれらが見られ，樹木の質感をより再現できている．

図 10 の結果を作成するにあたっては，「雪」に特化した超解像器と，図 6 を作成する際にも使用した「樹木」に特化した超解像器を用い，それぞれの結果を合成した．部分毎に拡大した画像を比較すると，雪に関しては既存手法に見られる不要な細かい模様を提案手法では削減できており，また，樹木に関しては既存手法では再現できなかった樹木の細かい模様の質感が提案手法によって再現されている．

6. 今後の方針

今後も物体カテゴリに特化した超解像器の開発を進めたい．現在は超解像器のモデルに既存手法である SRGAN をそのまま用いているが，モデルのネットワーク構造の変更や，損失関数の与え方などを変化させることで，より物体カテゴリに特化した超解像器を作成したい．また，現段階では手作業で行っている合成作業も，物体認識器の出力を上手く利用することで，ユーザの負担を減らしたい．

図 10：雪と樹木の合成結果 （左が既存手法，右が提案手法）

参考文献

[1] Zhihao Wang, Jian Chen, Steven C. H. Hoi. Deep Learning for Image Super-resolution: A Survey. arXiv:1902.06068 [cs.CV].

[2] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, Wenzhe Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. CVPR 2016.

[3] Deokyun Kim, Minseon Kim, Gihyun Kwon, Dae-Shik Kim,Progressive Face Super-Resolution via Attention to Facial Landmark, BMVC 2019

[4] Chieh-Chi Kao, Yuxiang Wang, Jonathan Waltman, Pradeep Sen. Patch-Based Image Hallucination for Super Resolution with Detail Reconstruction from Similar Sample Images. CVPR 2018.

[5] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, Nils Thuerey. Learning Temporal Coherence via Self-Supervision for GAN-based Video Generation. CVPR 2019.

[6] Yochai Blau, Tomer Michaeli. The Perception-Distortion Tradeoff. CVPR 2018.

[7] Jonathan Long, Evan Shelhamer, Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. CVPR 2015.

ⓒ 2020 Information Processing Society of Japan