卒 業 論 文 概 要 書

Summary of Bachelor's Thesis

Date of submission: 1/31/2023 (MM/DD/YYYY)

学科名 Department	情報通信	氏名 Name	進藤嵩紘	指 導 教 員 Advisor	渡辺 裕 ㊞
研究指導名 Research guidance	オーディオビジュアル 情報処理研究	学籍番号 Student ID number	1W192192-3		
研究題目 Title	VVC 符号化映像における YOLO-v7 の特徴量を用いた物体検出精度改善 Accuracy Improvement of Object Detection in VVC Coded Video Using YOLO-v7 Features				

1. まえがき

近年，画像認識のための動画像符号化技術に関する研究が注目を集めている．このような符号化技術の標準化作業は，Video Coding for Machines (VCM)と呼ばれる．視聴用の符号化技術である Versatile Video Coding (VVC)[1]による符号化映像を画像認識に適用すると，符号化雑音により画像認識精度の低下を招く場合がある．本研究では，CNN を用いて，VVC による符号化映像を処理することで，画像認識精度を改善する手法を提案する．提案手法は，Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN)[2]の構造をもとに作成する CNN であり，YOLO-v7[3]から抽出する特徴量を用いて学習する．評価手法には YOLO-v7 の学習済みモデルによる物体検出精度を用いる．提案手法により，符号化映像の物体検出精度を改善できることを実験により示す．

2. 従来手法

通常の映像符号化では符号化雑音が加わることによる映像品質の低下が起こる．Enhancing VVC Through CNN-Based Post-Processing[4]は，CNN を用いた符号化雑音除去処理により，VVC による符号化映像の品質を改善する手法である． SRGAN[5]の生成器をもとに構成されるモデルにより画像処理が行われる．原画像と出力画像の絶対誤差を用いて学習がなされる．PSNR を用いた評価により，VVC 符号化映像の品質が改善されている．

3. 提案手法

3.1 方針

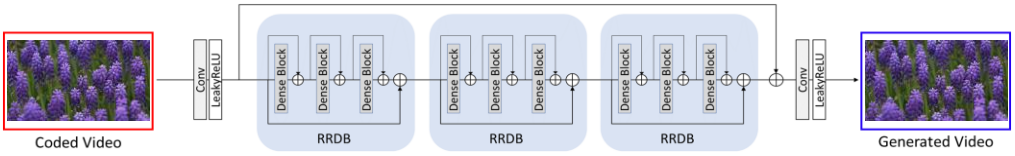
従来の CNN を用いた符号化映像の映像品質改善技術は，符号化雑音を低減することを目的とする．しかし，低減できる雑音量は小さく，画像認識精度の改善には結びつかない．そこで，符号化雑音を処理するモデルの学習に，物体検出モデルから抽出する特徴量を取り入れることで，符号化映像の画像認識精度の改善を目指す．

3.2 モデル構造

ESRGAN の構造を参考に，SRGAN をもとに構成される従来のモデル構造を改変する．ESRGAN は SRGAN と同様に，超解像処理に適したモデルである．SRGAN の残差ブロック[6]を Residual-in-Residual Dense Block (RRDB)に置き換えることで，より層の深いモデルを構成し，画像の絵柄の細かい部分の再現を可能にする．本研究の提案手法では，従来モデルの残差ブロックを RRDB に置き換える．提案するモデル構造を図 1 に示す．

3.3 損失関数

提案手法では物体検出モデルから得られる特徴量を用いた学習を行う．正解画像と出力画像の絶対誤差を損失計算に用いる従来の手法とは異なり，YOLO-v7 の学習済みモデルに入力したときに得られる特徴量の平均二乗誤差(MSE)を損失計算に用いる．提案手法で用いる損失を式(1)に示す．



*Pictures in this figure are images, "HoneyBee" [7]

図 1　提案する符号化映像処理モデルの構造

$$loss = MSE\left(YOLO(I_{output}), YOLO(I_{gt})\right) \quad (1)$$

ここに，$YOLO$ は YOLO-v7 の特徴抽出器，I_{output} は CNN の出力画像，I_{gt} は正解画像を表す．

4. 評価実験と結果

符号化雑音環境下において物体検出精度が改善されることを実験により示す．学習用データセットには，SJTU データセット[7]，UVG データセット[8]，MCML-4K-UHD データセット[9]を用いる．これらのデータセットはすべて，画像サイズが 4K の原画像データを含む．この中から 30 シーケンスを選択し，VTM10.0 [10]により符号化する．量子化係数(QP)を 27，32，37，42，47 とし，参照構造をランダムアクセスとする．モデルの入力には VVC による符号化映像を用い，YOLO-v7 の学習済みモデルから抽出する特徴量で学習を行う．この際，符号化前の原画像を正解画像とする．

テスト用データセットには，SFU-HW-Objects-v1 データセット[11]のクラス A を用いる．これは原画像と，物体検出用のアノテーションからなるデータセットであり，シーケンスの画像サイズによりクラス分けされている．クラス A の画像サイズは 2560×1600 画素であり，最も画像サイズの大きいシーケンスを含む．テスト用データも学習用データと同一の手法で符号化する．

評価手法は YOLO-v7 による物体検出精度を用いる．検出精度は mean Average Precision (mAP)を用いて測る．VVC による符号化映像と，提案するモデルの出力映像の二つを YOLO-v7 の学習済みモデルに入力し，Intersection over Union を 0.5 に設定したときの mAP を計測する．符号化映像の bitrate と mAP の関係を図 2 に示す．

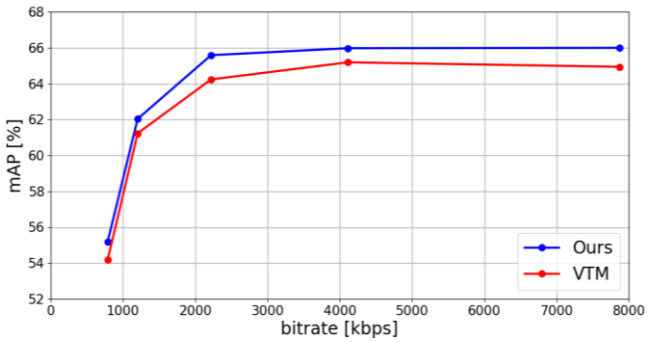


図 2　テストシーケンスの bitrate と mAP の関係

この図より，実験で使用したすべての QP において，YOLO-v7 による物体検出精度が改善されている．以上より，YOLO-v7 から抽出する映像の特徴量を用いた, VVC による符号化映像の符号化雑音処理は，物体検出精度の改善に有効であることが分かる．

5. 結論

本研究では，VVC 符号化映像に対して CNN を用いた画像処理を行うことで，画像認識精度を改善する手法を提案した．実験により，YOLO-v7 による物体検出精度が改善されることを確認した．今後，より高効率な画像認識用の映像符号化手法を検討する必要がある．

参考文献

[1] Versatile Video Coding, Standard ISO/IEC 23090-3, ISO/IEC JTC 1, Jul. 2020.

[2] X. Wang, *et al.*, “ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks,” in IEEE ECCV, pp. 63-79, Sep. 2018.

[3] C. Y. Wang, *et al.*, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for realtime object detectors.” arXiv preprint arXiv:2207.02696, 2022.

[4] F. Zhang, *et al.*, “Enhancing VVC Through Cnn-Based Post-Processing,” in IEEE ICME, pp. 1-6, Jul. 2020.

[5] C. Ledig, *et al.*, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network,” in IEEE CVPR, pp. 4681-4690, Jul. 2017.

[6] K. He, *et al.*, “Deep Residual Learning for Image Recognition,” in IEEE CVPR, pp. 770-778, Jun. 2016.

[7] L. Song, *et al.*, “The SJTU 4K Video Sequence Dataset,” in International Conference on Quality of Multimedia Experience, pp. 34-35, Jul. 2013.

[8] A. Mercat, *et al.*, “UVG dataset: 50/120fps 4K Sequences for Video Codec Analysis and Development,” in ACM Multimedia Systems Conference, pp. 297-302, Jun. 2020.

[9] M. Cheon, *et al.*, “Subjective and Objective Quality Assessment of Compressed 4K UHD Videos for Immersive Experience,” in IEEE TCSVT, vol. 28, pp. 1467-1480, Jul. 2018.

[10] S. K. J. Chen, Y. Ye, Algorithm description for Versatile Video Coding and Test Model 10 (VTM 10). JVETS2002, 2020.

[11] H. Choi, *et al.*, “A dataset of labelled objects on raw video sequences.” Data in Brief, 34:106701, 2021.

2022 年度　卒業論文

VVC 符号化映像における YOLO-v7 の特徴量を用いた物体検出精度改善

Accuracy Improvement of Object Detection in VVC Coded Video Using YOLO-v7 Features

提出日　2023 年 1 月 31 日

指導教員　渡辺 裕 教授

早稲田大学基幹理工学部 情報通信学科

1W192192-3

進藤　嵩紘

目次

第１章　序論

１．１　研究背景

動画像符号化技術は，映像データの冗長な部分を省き，データ量を削減する技術である．限られた通信資源のもとで，効率よく映像データを送受信するために必要とされている．近年, 4K・8K の高解像度の映像の需要が増加しているだけでなく，動画配信サービスの利用者は伸び続けており, 映像データの送受信が通信トラヒックに占める割合が激増している．この需要に応えるため，より圧縮率の高い映像符号化技術が求められており，Versatile Video Coding (VVC) [1]を代表とする動画像符号化技術の研究開発が続けられている．

一方，視聴用の映像を符号化する技術のみならず，Video Coding for Machines (VCM) と呼ばれる，画像認識のための映像を符号化する技術の研究開発も進められている．Moving Picture Experts Group (MPEG) が VCM に関する技術の評価方法やデータ[2]を公表しており，標準化に向けた動きも加速している．MPEG は，画像認識モデルの特徴量圧縮と，画像認識用の映像圧縮の二つの観点から，VCM の標準化活動の探索を行っており，そのなかでも後者をベースとする標準化活動では，主に二つのアプローチから VCM に関する技術が提案されている．一つ目はニューラルネットワークを用いた映像符号化モデルを使用するアプローチであり．二つ目は VVC などの既存の符号化方式とニューラルネットワークを組み合わせるアプローチである．VCM ではこれらのアプローチにより，VVC を超えるより高い圧縮率と，画像認識精度が期待されている．

１．２　関連研究と問題点，および研究目的

VVC は，次世代映像符号化方式として 2020 年 7 月に標準化が完了した最新の動画像符号化技術である．視聴用の映像を符号化する手法であり，専門家の知見を集めたアルゴリズムに基づいて映像を圧縮する．VVC は 4K・8K の映像が普及し始めている現在，High Efficiency Video Coding (HEVC) [3]を超える圧縮性能をもつため，注目されている符号化方式である．しかし，動画像の符号化では，符号化雑音が加わることによる映像品質の低下が起こる．そこで，ニューラルネットワークを用いて符号化映像を処理することで，映像品質を改善させる手法が提案されている．Enhancing VVC Through Cnn-Based Post-Processing [4]では，Convolutional Neural Network (CNN)を用いて符号化雑音を除去する手法を提案している．Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network [5]で提案されている，SRGAN の生成器を参考にしたニューラルネットワークを用いて, 符号化映像を符号化前の映像に近づけることを可能にしている．また，近年 Deep Neural Network (DNN) を用いた視聴用映像の符号化技術の研

究開発も進んでいる．DVC: An End-to-end Deep Video Compression Framework [6]や，Deep Contextual Video Compression (DCVC) [7]は，CNN を用いて視聴用の映像を符号化する手法を提案している．データのエンコード，デコード，動き補償をニューラルネットワークを用いて行う手法である．

画像認識のための映像符号化手法もいくつか提案されてきている．一般的に，画像認識に必要な映像の情報は，人間が視聴の時に必要とする映像の情報よりも少ないと考えられている [8]．High Efficiency Compression for Object Detection [9]は YOLO9000 [10]による物体検出結果を用いて符号化モデルを学習させることで，静止画像の圧縮率と物体検出精度で HEVC を上回ることを示している．End-to-end Compression Towards Machine Vision; Network Architecture Design and Optimization [11]は，CNN を用いた画像符号化モデルを，Faster-RCNN [12]による物体検出結果により学習させることで，静止画像の圧縮率と物体検出精度で VVC を上回ることを示している．これらのモデルによる映像符号化は視聴用の映像符号化方式を圧縮率で上回るが，使用用途は物体検出のみに限られる．この理由は，他の画像認識タスクや，人間の視聴に必要な情報を含まないためである．一方，既存の視聴用符号化方式である VVC をそのまま用いる映像符号化手法も検討されている．この手法では，VVC の映像圧縮効率を超えることは出来ないが，ニューラルネットワークによる映像処理と組み合わせることにより，画像認識精度を改善することが出来る．また，一つの符号化手法により，視聴用と画像認識用の二つの映像を作成することが可能である．

しかし，これらの画像認識用の映像符号化手法はすべて，比較的構造のシンプルな画像認識モデルを想定しており，最先端の画像認識モデルを使用していない．これは VCM の技術を研究するうえで，問題設定を簡単にするためであると考えられる．そこで本研究では，最先端の物体検出モデルである YOLO-v7 [13]のための映像符号化手法について検討する．YOLO-v7 は有名な物体検出モデルの一つである YOLO [14]の最新モデルである．符号化手法には，最新の動画像符号化方式である VVC を用いることで高い圧縮性能を達成する．VVC の符号化映像をニューラルネットワークにより処理することで，YOLO-v7 の物体検出に有効な映像に変換することを目的とする．YOLO-v7 の学習済みモデルから抽出できる特徴量を用いて，符号化映像を処理するニューラルネットワークの学習を行う．提案する符号化映像処理手法により，YOLO-v7 による物体検出精度が改善できることを実験により示す．

１．３　本論文の構成

以下に本論文の構成を示す．

第１章　研究の背景および目的について述べる．

第２章　関連研究について述べる．

第３章　提案手法について述べる．

第４章　評価実験と，その結果について述べる．

第５章　結論と今後の課題について述べる．

第２章　　関連研究

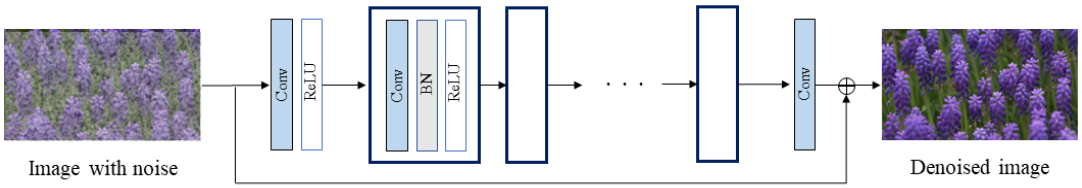
２．１　　まえがき

本章では，関連研究について述べる．まず CNN を用いた雑音除去技術の概要と，その技術を符号化雑音に適用したモデルについて説明する．次に物体検出モデルの概要と，その代表的なモデルである YOLO について説明し，最後に画像認識のための映像符号化モデルについて述べる．

２．２　　CNN を用いた雑音除去処理

２．２．１　　CNN を用いた雑音除去技術の概要

CNN を用いた画像処理技術の進歩は著しく，画像の雑音除去処理や，画像生成，超解像など様々な用途で用いられている．ここでは画像の雑音処理技術について説明する．Denoising Convolutional Neural Network [15]は，初めて CNN を雑音除去手法に用いた手法である．その後，Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising [16]では，DnCNN と呼ぶネットワークを，雑音除去モデルに提案した．DnCNN の構造を図 1 に示す．入力画像を出力画像の前で加算することにより，画像に加わった雑音のみを学習させる．学習に用いる損失は，正解画像と出力画像の Mean Squared Error (MSE)を用いて計算される．



*Pictures in this figure are images, “HoneyBee” [21]

図 1　DnCNN のネットワーク構造

２．２．２　符号化雑音の除去処理

Enhancing VVC Through Cnn-Based Post-Processing では，CNN を用いて VVC による符号化雑音を除去する手法を提案している．符号化雑音が加わった映像の雑音を除去することで，品質を改善することを目的としている．ネットワークの構造は SRGAN の生成器の構造をもとに構成される．SRGAN は超解像を行うモデルであり，residual block [17]を用いて入力画像の特徴を解析する．Enhancing VVC Through Cnn-Based Post-Processing が提案するネットワークモデルも residual block を利用しており，その構造を図 2 に示す．DnCNN と同様に，出力画像の前で入力画像を加算することにより，符号化雑音のみを学習する．符号化前の画像を正解画像とし，正解画像と出力画像の絶対誤差を損失としてニューラルネットワークの学習を行う．

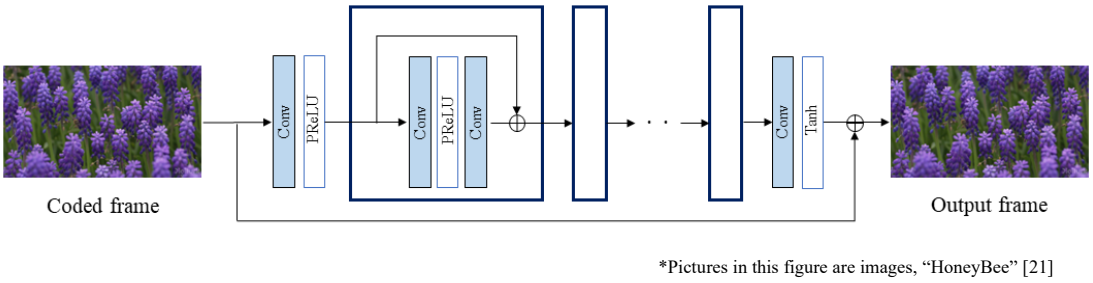


図 2　Enhancing VVC Through Cnn-Based Post-Processing のモデル構造

２．３　物体検出

２．３．１　物体検出技術の概要

物体検出とは，数ある画像認識タスクの中でも特に有名なタスクであり，点群や画像などのデータから，ある特定の物体の位置，大きさ，種類を特定するタスクである．現在ではニューラルネットワークを用いて実装されることが多く，代表的なモデルには，R-CNN[18]，YOLO，SSD[19]などがある．一般的には検出した物体をバウンディングボックスで囲うことで，視覚で物体の位置を確認することが出来る．本論文で用いる技術は画像を用いた物体検出技術である．

２．３．２　YOLO

YOLO は広く利用されている物体検出モデルの一つである．リアルタイムで動作可能な検出時間の短さと，検出精度の高さを兼ね備えており，バージョン 7 まで公開さ

れている．YOLO は 2015 年に You Only Look Once: Unified, Real-Time Object Detection という論文で提案されたモデルであり，当時初めて物体の位置，大きさ，種類を一度に推論可能にしたモデルである．それまでの物体検出モデルでは，まず物体の位置を推測し，次に大きさの推測と，種類の推測を行う，段階処理を必要とした．この検出手法の変更により，検出速度を高速化し，リアルタイムでの物体検出を可能にした．YOLO は最新バージョンの YOLO-v7 に至るまで，検出精度を改善し続けている．

YOLO-v7 は最先端の物体検出モデルであり，検出精度，検出速度ともに優れた性能を持つ．モデル構造は，Backbone と Head から構成されている．Backbone は入力画像から特徴を抽出する役割を持ち，Head はその特徴から物体の位置と種類を推測する役割を持つ．YOLO-v7 のモデル構造を次の図 3 に示す．

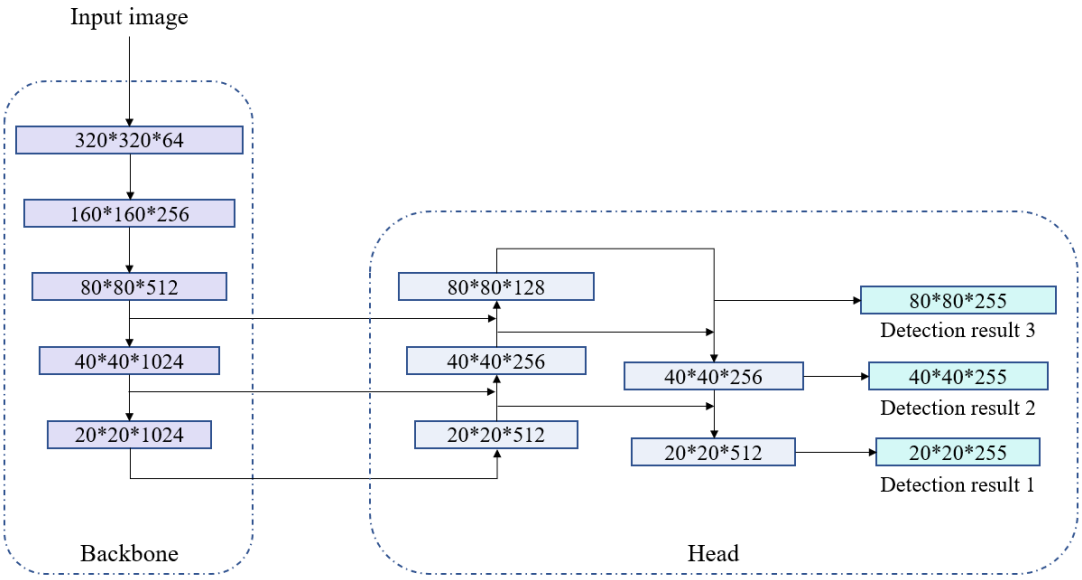


図 3　YOLO-v7 のモデル構造

２．４　画像認識用の画像圧縮手法

画像認識用の画像圧縮を行うモデルの学習には，一般的にはそのタスクを行うモデルの出力結果や特徴量を用いる．End-to-end Compression Towards Machine Vision; Network Architecture Design and Optimization は，GRAO (generalized rate-accuracy optimization)と呼ばれる，視聴用と Faster-RCNN のための画像符号化手法を提案している．CNN を用いた画像符号化モデルを，Faster-RCNN による物体検出結果により学習させることで，静止画像の圧縮率と物体検出精度で VVC を上回る．また，再構成画像と正解画像の平均二乗誤差でも学習を行うことで，鮮明な画像の再構成を可能とし，画像品質の点でもほ

とんど VVC に劣らない符号化モデルを提案する．符号化モデルの学習処理を図 4 に示す．

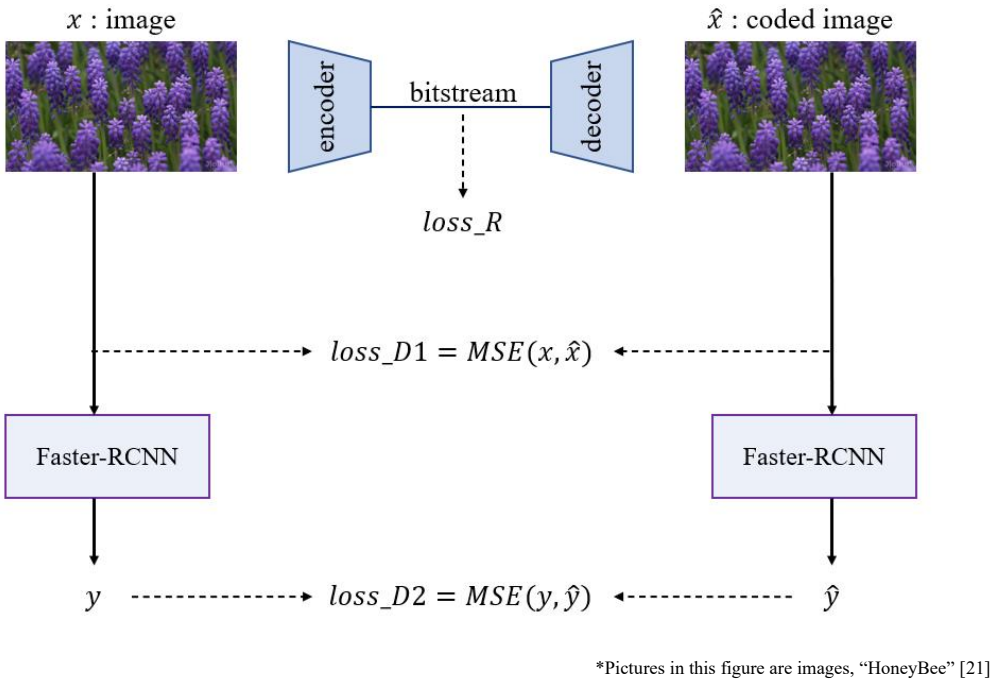


図 4　GRAO の学習方法

図 4 において，x は入力画像，$\hat{x}$ はx の符号化画像，y はx を学習済み Faster-RCNN に入力して得られた物体検出結果，$\hat{y}$ は$\hat{x}$ を学習済み Faster-RCNN に入力して得られた物体検出結果を表す．また，ニューラルネットワークにより構成される符号化モデルの学習で用いる損失関数は式(1)で表される．

$$LOSS = loss\_R + \lambda_1(loss\_D1 + \lambda_2 loss\_D2) \tag{1}$$

ここに，$loss\_R$ はエンコーダーが出力する bit 量を表し、損失計算で用いることで，符号化器の圧縮効率を高める役割を持つ．$loss\_D1$ は符号化前の画像と符号化後の再構成画像との平均二乗誤差であり，デコーダーの画像の再構成を助ける役割を持つ．$loss\_D2$ は符号化前の画像と符号化後の再構成画像を学習済み Faster-RCNN に入力したときの検出結果の誤差であり，再構成画像の物体検出精度を高める役割を持つ．λ_1とλ_2はともに定数であり，λ_1を変更することで符号化モデルの圧縮率を変えることができる．λ_1を大きく設定することで，圧縮率を下げ鮮明な画像を最構成する．λ_2を変えることで，再構成画像の使用用途にあわせた画像作成を可能にする．例えば，λ_2を大きく設定することで，より物体検出に有利な画像を再構成する．

２．５　むすび

本章では，CNN を用いた雑音除去手法，物体検出技術，画像認識のための画像符号化手法について説明した．雑音除去手法である DnCNN と Enhancing VVC Through Cnn-Based Post-Processing について述べ，物体検出手法である YOLO について説明した．また, 物体検出のための符号化モデルについて, おもに学習方法について詳しく説明した.

第３章　提案手法

３．１　まえがき

本章では，提案手法について述べる．

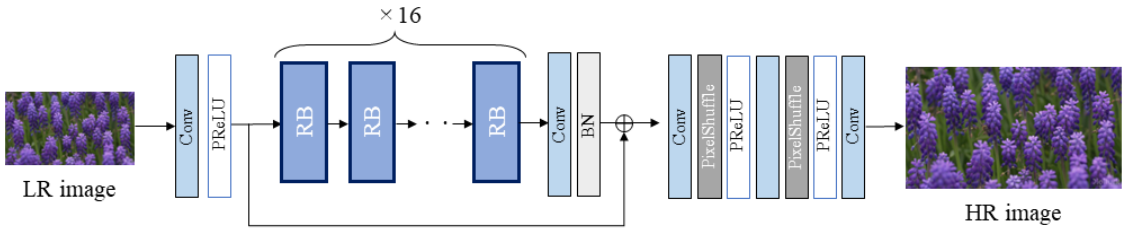
３．２　提案手法

３．２．１　方針

YOLO-v7 のための映像符号化手法を提案する．VVC を用いて符号化した映像を，CNN を用いて処理することで，YOLO-v7 に有効な映像を作成する．VVC は最新の映像符号化手法であり，視聴用映像の符号化手法である．符号化された映像は，映像品質のみならず，画像認識精度も低下するため，CNN を用いて画像認識精度の改善を試みる．前章で紹介した Enhancing VVC Through Cnn-Based Post-Processing による符号化映像処理では，映像の品質改善を目的とし，符号化雑音を除去するが，除去できる雑音の大きさは小さく，画像認識精度の改善には結びつかない．そこで，符号化雑音を処理するモデルの学習に，物体検出モデルから抽出する特徴量を取り入れることで，符号化映像の画像認識精度の改善を図る．学習に用いるモデル構造と損失関数を以下に示す．

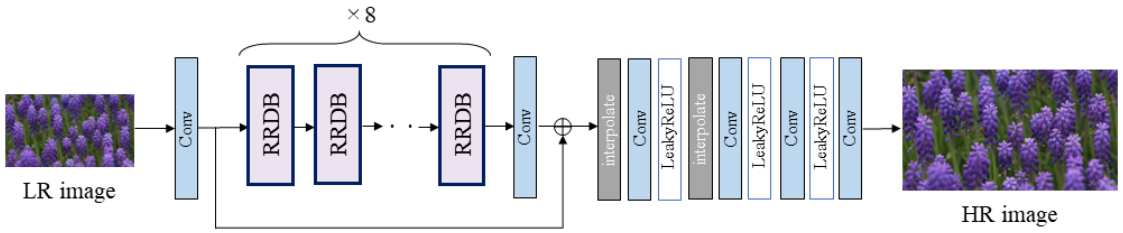
３．２．２　モデル構造

従来手法の雑音除去を行うニューラルネットワークの構成は，SRGSN の生成器をもとに構成される．提案するモデルでは，ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks [20]が提案している ESRGAN と呼ばれるモデル構造を参考にモデルを作成する．ESRGAN は SRGAN と同じく，画像の超解像を行うモデルである．SRGAN の residual block (RB)を Residual-in-Residual Dense Block (RRDB)に置き換えることで，より層の深いニューラルネットワークを構成し，画像の絵柄の細かい部分の再現を可能にしている．SRGAN と ESRGAN の生成器の構造をそれぞれ図 5 と図 6 に示す．また，RB の構造を図 7 に，RRDB の構造と，RRDB を構成する residual dense block (RDB)の構造を図 8 と図 9 に示す．本論文の提案するモデルでも，従来モデルの residual block を RRDB に置き換える．提案するモデル構造を図 10 に示す．



*Pictures in this figure are images, “HoneyBee” [21]

図 5　SRGAN の生成器の構造



*Pictures in this figure are images, “HoneyBee” [21]

図 6　ESRGAN の生成器の構造

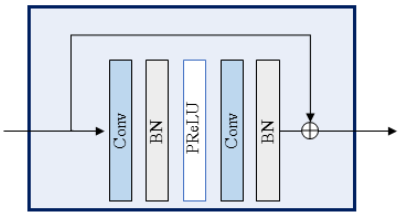


図 7　RB の構造

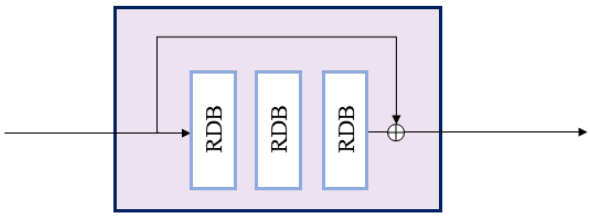


図 8　RRDB の構造

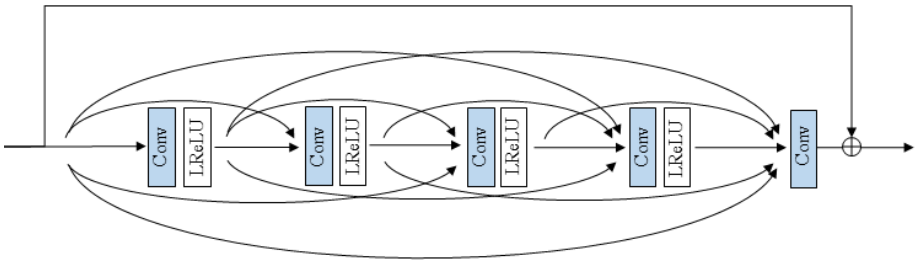
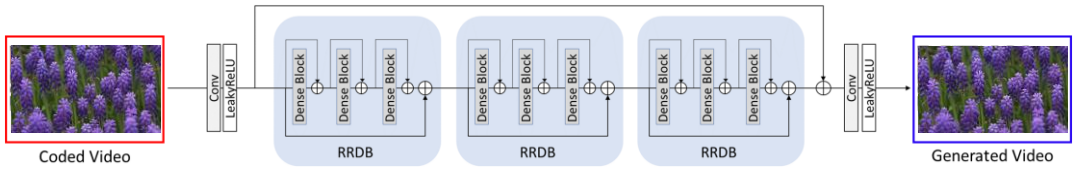


図 9　RDB の構造



*Pictures in this figure are images, “HoneyBee” [21]

図 10　提案するモデル構造

３．２．３　損失関数

Enhancing VVC Through Cnn-Based Post-Processing は符号化映像を処理することで，映像品質の改善を目指すため，生成画像と正解画像の絶対誤差(L1)を用いてネットワークの学習を行う．一方，画像認識のための映像符号化モデルの学習には，式(1)のように画像認識モデルの出力や，そのモデルの特徴量が用いられる．本論文では，YOLO-v7 の検出精度の改善を目指すため，YOLO-v7 の学習済みモデルを用いた学習を行う．生成画像と正解画像を YOLO-v7 に入力することで得られる映像の特徴量の差を損失関数とする．特徴量は YOLO-v7 の backbone の三か所から抽出する．YOLO-v7 の backbone では，入力画像は 5 個のサイズの特徴量にリサイズされるため，そのうち 3 つを使用する．用いる損失関数を式(2)に示す．また，特徴抽出の様子を図 11 に示す．

$$LOSS = \ MSE\left(yolo\left(I_{output}\right), yolo\left(I_{gt}\right)\right) \tag{2}$$

ここに，MSE は平均二乗誤差，$yolo$ は YOLO-v7 の特徴抽出器，I_{output} はニューラルネットワークの出力画像，I_{gt} は正解画像を表す．

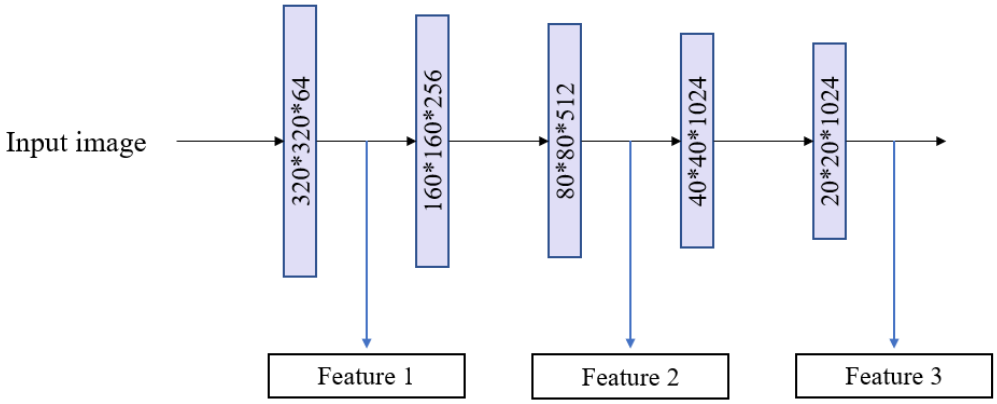


図 11　YOLO-v7 による特徴抽出

３．３　むすび

本章では，提案手法について説明した．符号化映像を処理するためのニューラルネットワークの構造と，その学習に用いる損失関数について述べた．

第４章　実験と結果

４．１　まえがき

本章では，提案手法の有効性を確認するための実験内容と，実験結果について述べ，性能評価を行う．

４．２　実験

４．２．１　学習方法

提案手法により，符号化雑音環境下において物体検出精度が改善することを実験により示す．学習用データセットには，SJTU データセット[21]，UVG データセット[22]，MCML-4K-UHD データセット[23]を用いる．これらのデータセットはすべて，画像サイズが 4K の符号化されていない映像データを含む．この中から 30 シーケンスを選択し，VTM10.0[24]により符号化する．この際，量子化係数(QP)を 27，32，37，42，47 とし，参照構造をランダムアクセスとする．モデルの入力にはこれらの符号化映像を用い，YOLO-v7 の学習済みモデルから抽出する特徴量で学習を行う．

４．２．２　評価方法

テストデータセットには，SFU-HW-Objects-v1 データセット[25]を用いる．このデータセットは VCM の Common Test Condition [26]にも指定されており，符号化されていない 18 個の映像と，それらの映像に対応した物体検出用のアノテーションを含んだデータセットである．データセット内のシーケンスは画像サイズにより A~E の 5 つのクラスに分けられている．学習用のシーケンスはすべて画像サイズが 4K であるため，テストでもできる限り画像サイズの大きいものを使用する．テストに使用するシーケンスの情報を表 1 に示す．これらの 9 つのシーケンスを VTM10.0 により符号化する．この際，量子化係数を 27，32，37，42，47 とし，フレームの参照構造はすべてランダムアクセスとする．

これらの符号化映像に対して，提案手法を適用することで，物体検出精度の改善を図る．手法の適用前と後での物体検出精度を比較することで，提案手法の有効性を確かめる．物体検出手法には YOLO-v7 の学習済みモデルを使用し，検出精度の測定には mean Average Precision (mAP)を用いる．mAP 測定時の Intersection over Union (IoU) は 0.5 とする．

表 1　テストシーケンスの詳細

class	sequence name	size	frame number
A	PeopleOnStreet	2560x1600	150
A	Traffic	2560x1600	150
B	BQTerrace	1920x1080	600
B	BasketballDrive	1920x1080	500
B	ParkScene	1920x1080	240
C	BQMall	832x480	600
C	BasketballDrill	832x480	500
C	PartyScene	832x480	500
C	RaceHorsesC	832x480	300

４．３　実験結果

まず，クラス A の PeopleOnStreet シーケンスを用いた実験結果について述べる．このシーケンスに含まれる物体の 97%が「person」であるため，mAP とともに，人物の検出精度も測定する．Bitrate と mAP の関係を図 12 に示し，bitrate と人物の検出精度を表す AP の関係を図 13 に示す．提案する符号化映像の処理手法により，mAP がすべての量子化係数に関して改善している．また，人物の検出精度についても，すべての量子化係数に関して，AP を 5%程度改善できることが分かる．

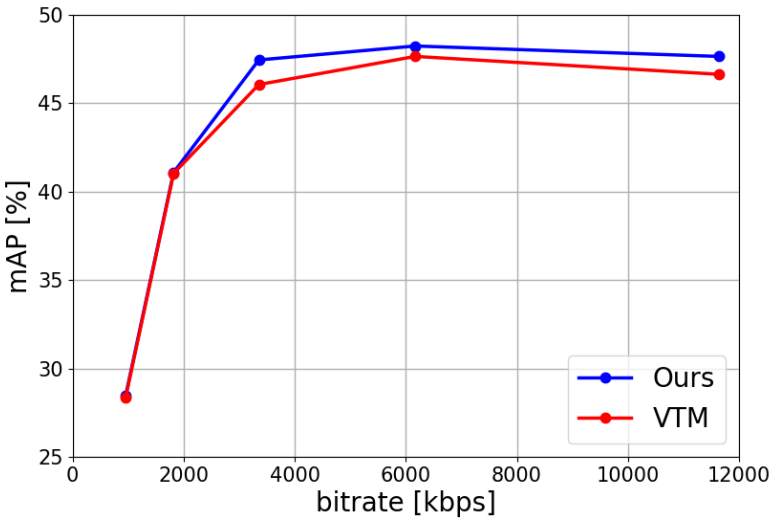


図 12　PeopleOnStreet シーケンスの mAP と bitrate の関係.

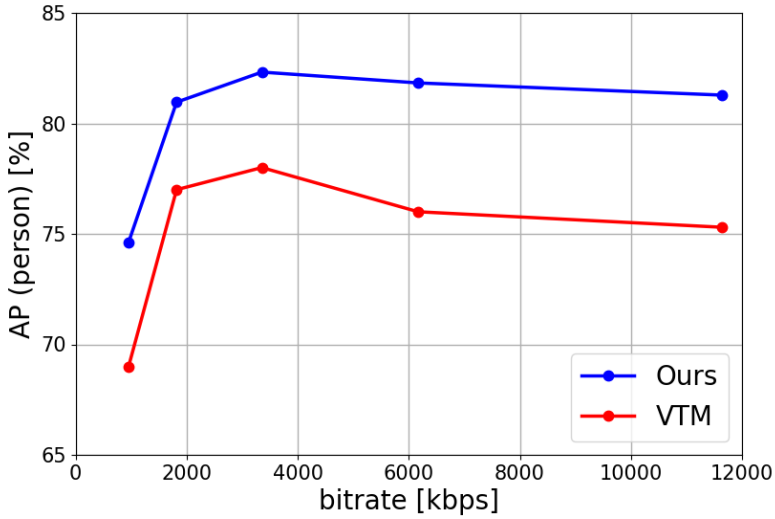


図 13　PeopleOnStreet シーケンスの AP(person)と bitrate の関係.

次に，クラス A の Traffic シーケンスを用いた実験結果について述べる．このシーケンスに含まれる物体の 99%が「car」であるため，mAP とともに，車の検出精度も測定する．Bitrate と mAP の関係を図 14 に示し，bitrate と車の検出精度を表す AP の関係を図 15 に示す．図 14 より，提案手法を適用することで，mAP はすべての量子化係数に関して 2%程度改善できることが分かる．また図 15 より，車の検出精度を表す AP は 3%程度改善できることが分かる．

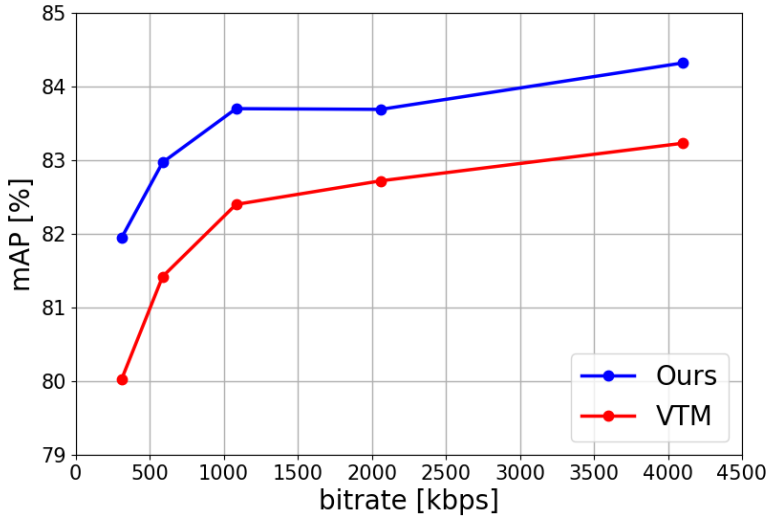


図 14　Traffic シーケンスの mAP と bitrate の関係.

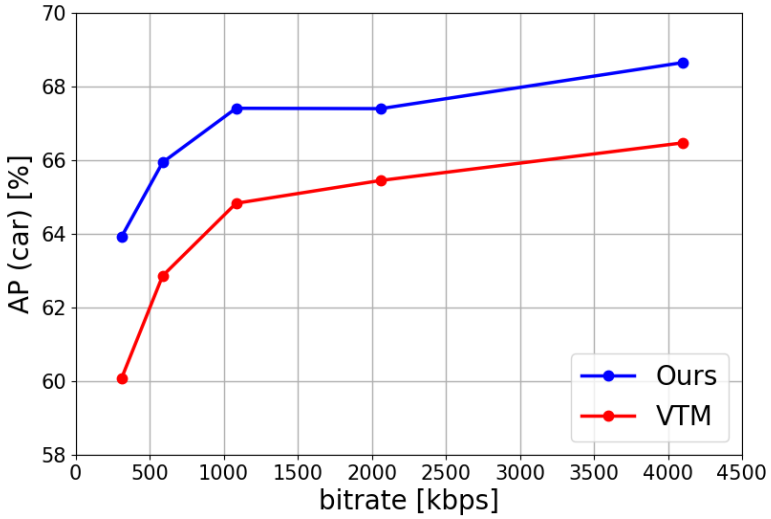


図 15　Traffic シーケンスの AP(car)と bitrate の関係.

クラス B，C のシーケンスを用いた実験結果について述べる．これらのシーケンスについては mAP のみ測定する．クラス B と C のシーケンスにおける，YOLO-v7 による物体検出結果と bitrate の関係を，それぞれ図 16，17 に示す．クラス B，C のシーケンスについても，提案手法による物体検出精度の改善が確認できる．

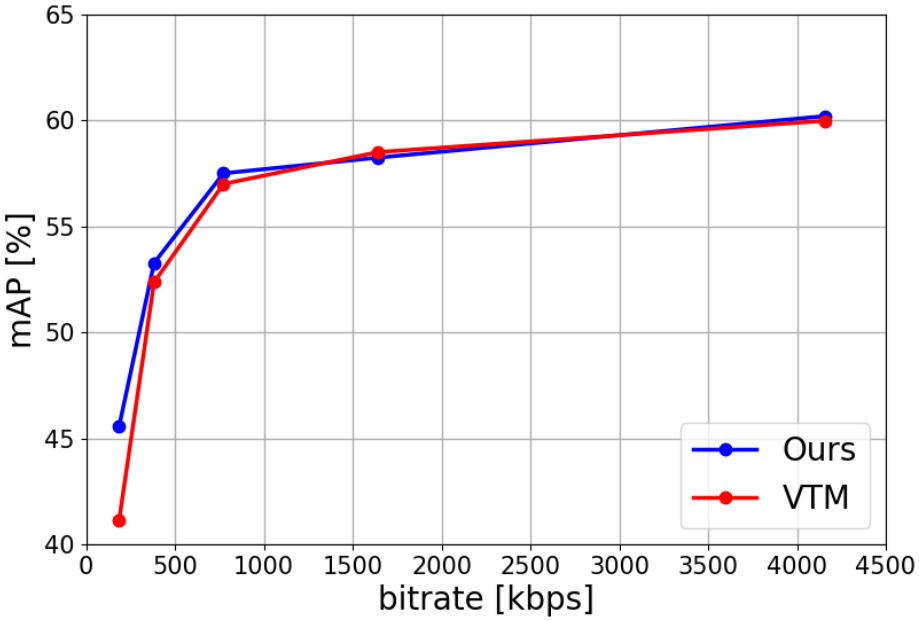


図 16　クラス B シーケンスの bitrate と mAP の関係.

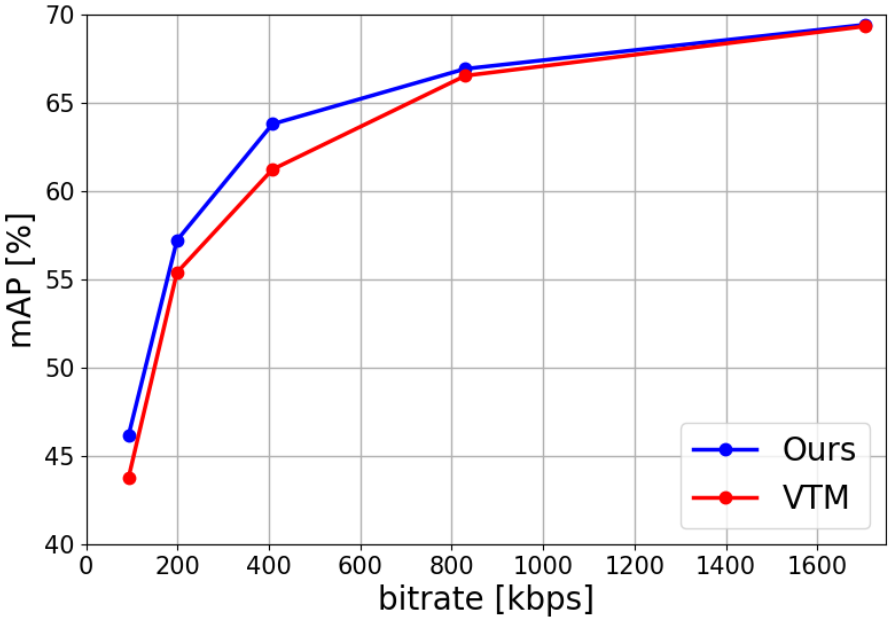


図 17　クラス C シーケンスの bitrate と mAP の関係

最後に，これらの実験結果を表 2 にまとめる．表 2 において，「gap」は提案手法の適用前と後の mAP の値の差を表す．この表より，符号化映像に提案手法を適用することで，物体検出精度が改善できることが分かる．

表 2　提案手法適用前と後の物体検出精度(mAP)の比較

class	method	QP				
		27	32	37	42	47
A	VTM	64.94	65.19	64.23	61.23	54.19
	Ours	65.99	65.97	65.58	62.08	55.20
	gap	+1.05	+0.78	+1.35	+0.85	+1.01
B	VTM	59.99	58.51	57.01	52.37	41.12
	Ours	60.21	58.25	57.51	53.28	45.55
	Gap	+0.22	-0.26	+0.50	+0.91	+4.43
C	VTM	69.34	66.54	61.23	55.39	43.77
	Ours	69.43	66.93	63.80	57.21	46.16
	gap	+0.09	+0.39	+2.57	+1.82	+2.39

４．４　考察

図 12, 14 より，人物検出精度や車の検出精度の改善幅よりも，mAP の改善幅が小さいことが分かる．原因の 1 つに考えられるのは，学習用データセット不足である．学習用のシーケンスには，人物や車が多く含まれている．しかし，「Bus」や「Handbag」などの物体は，テストシーケンスに含まれているにも関わらず，学習用のシーケンスにはあまり含まれていない．これらの物体の検出精度が上がらないことにより，mAP の改善幅が小さくなっていると考えられるため，データの増強が必要である．

４．５　むすび

本項では，提案手法の有効性を確認するための実験と，その結果について述べた．

第５章　結論

５．１　結論

YOLO-v7 の特徴量を用いて学習した CNN により，符号化映像を処理することで，物体検出精度を改善する手法を提案した．映像符号化手法には，最新の動画像符号化方式である VVC を使用し，物体検出モデルには，YOLO の最新バージョンである YOLO-v7 を使用した．YOLO-v7 により映像から抽出する特徴量を用いて，CNN を学習させることで，YOLO-v7 の物体検出に有効な映像の作成を可能にした．また，提案手法の有効性を実験により確認した.

５．２　今後の課題

本論文の実験では、学習データセット不足により，mAP の改善幅が小さくなってしまった可能性があるため, データセットを増量し, 同様の実験を行う必要がある. また，本論文では，VVC 符号化映像の処理手法を検討したため，VVC よりも高い圧縮率を得ることができない．今後は，さらなる画像認識精度の改善と，圧縮効率の改善を目指す必要がある.

謝辞

本研究に際して，丁寧なご指導をしてくださり，実験環境および快適な研究環境を与えてくださった渡辺教授に心より感謝いたします.

日頃から貴重な意見をくださり，研究室における温かい環境を提供してくださった渡辺研究室の皆様に感謝いたします.

最後に，私をここまで育ててくださり，常に心を支えてくださり，生活を支えてくださっている家族に感謝いたします.

参考文献

[1] Versatile Video Coding, Standard ISO/IEC 23090-3, ISO/IEC JTC 1, Jul. 2020.

[2] Call for Proposals on Video Coding for Machines, ISO/IEC JTC 1/SC 29/WG 2, N002 20, Jul. 2022.

[3] High Efficiency Video Coding, Standard ISO/IEC 23008-2, ISO/IEC JTC 1, Apr. 2013.

[4] F. Zhang, C. Feng, and D. R. Bull, “Enhancing VVC Through Cnn-Based Post-Processi ng,” in IEEE ICME, 2020, pp. 1-6.

[5] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-Realistic Single Image Super-Resolution U sing a Generative Adversarial Network,” in IEEE CVPR, 2017, pp. 4681-4690.

[6] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, “DVC: An End-to-end Dee p Video Compression Framework,” in IEEE CVPR, 2019, pp. 11006-11015.

[7] J. Li, B. Li, and Y. Lu, “Deep contextual video compression,” Advances in Neural Info rmation Processing Systems, vol. 34, 2021, pp. 18114-18125.

[8] H. Choi, and I. V. Bajic, “Scalable Image Coding for Humans and Machines,” IEEE Tr ansaction on Image Processing, vol. 31, 2022, pp. 2739-2754.

[9] H. Choi, and I. V. Bajic, “High Efficiency Compression for Object Detection,” in IEEE ICASSP, 2018, pp. 1792-1796.

[10] J. Redmon, and A. Farhadi, “YOLO9000: better, faster, stronger,” in IEEE CVPR, 2017, pp. 6517-6525.

[11] S. Wang, Z. Wang, S. Wang, and Y. Ye, “End-to-end Compression Towards Machine V ision: Network Architecture Design and Optimization,” IEEE Open Journal of Circuits a nd Systems, vol. 2, 2021, pp. 675-685.

[12] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object D etection with Region Proposal Networks,” Advances in Neural Information Processing S ystems, vol. 28, 2015, pp. 91-99.

[13] C. Y. Wang, A. Bochkovskiy, and M. Liao, “YOLOv7: Trainable bag-of-freebies sets ne w state-of-the-art for realtime object detectors.” arXiv preprint arXiv:2207.02696, 2022.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, rea l-time object detection.” arXiv preprint arXiv:1506.02640, 2015.

[15] Q. Xu, C. Zhang, and L. Zhang, “Denoising Convolutional Neural Network,” in IEEE I CIA, 2015, pp. 1184-1187.

[16] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a Gaussian Denoiser: R esidual Learning of Deep CNN for Image Denoising,” IEEE Transactions on Image Pro cessing, vol. 26, 2017, pp. 3142-3155.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in IEEE CVPR, 2016, pp. 770-778.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich Feature Hierarchies for Accura te Object Detection and Semantic Segmentation,” in IEEE CVPR, 2014, pp. 580-587.

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, “SS

D: Single shot multibox detector," In IEEE ECCV, 2016, pp. 21–37.

[20] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," in IEEE ECCV, 2018, p p. 63-79.

[21] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The SJTU 4K Video Sequence Da taset," in International Conference on Quality of Multimedia Experience, 2013, pp. 34-3 5.

[22] A. Mercat, M. Viitanen, and J. Vanne, "UVG dataset: 50/120fps 4K Sequences for Vide o Codec Analysis and Development," in ACM Multimedia Systems Conference, 2020, p p. 297-302.

[23] M. Cheon, and J. S. Lee, "Sbjective and Objective Quality Assessment of Compressed 4K UHD Videos for Immersive Experience," in IEEE Transactions on Circuits and Syst ems for Video Technology, vol. 28, 2018, pp. 1467-1480.

[24] S. K. J. Chen, Y. Ye, Algorithm description for Versatile Video Coding and Test Model 10 (VTM 10). JVETS2002, 2020.

[25] H. Choi, E. Hosseini, S. R. Alvar, R. A. Cohen, and I. V. Bajic. "A dataset of labelle d objects on raw video sequences." Data in Brief, 34:106701, 2021.

[26] S. Liu, and H. Zhang, "Common test conditions for video coding for machines," ISO/I EC JTC 1/SC 29/WG 04, Nov. 2022.

図一覧

表一覧

研究業績

[1] T. Shindo, T. Watanabe, R. Yano, M. Arimoto, M. Takahashi, and H. Watanabe, “Super Resolution for QR code Images,” in IEEE Global Conference on Consumer Electronics (GCCE), pp. 281-284, Oct. 2022.

[2] 進藤嵩紘, 渡部泰樹, 渡辺裕 : “符号化雑音環境下における物体検出精度の改善手法 (A Method for Improving Object Detection Accuracy in Coding Noise Environment)”, 2022 年度画像符号化シンポジウム・2022 年度映像メディア処理シンポジウム (PCSJ 2022 / IMPS 2022), P1-01, Nov. 2022.

[3] T. Watanabe, T. Shindo, H. Watanabe, “Novel CNN approach for video prediction based on FitVid,” in International Workshop on Advanced Image Technology (IWAIT 2023), Jan. 2023.

[4] 進藤嵩紘, 渡部泰樹, 渡辺裕 : “VVC と CNN を組み合わせた YOLO-v7 のための映像符号化手法 (Video Coding Scheme for YOLO-v7 Combining VVCand CNN)”, 2023 年 電子情報通信学会総合大会, Mar. 2023. (発表予定)

[5] 渡部泰樹, 進藤嵩紘, 渡辺裕 : “YOLOV を用いた物体予測検出の一検討 (A Study on Future Object Detection Using YOLOV)”, 2023 年 電子情報通信学会総合大会, Mar. 2023. (発表予定)

Super Resolution for QR Code Images

Takahiro Shindo
School of FSE
Waseda University
Tokyo, Japan
taka_s0265@ruri.waseda.jp

Taiju Watanabe
School of FSE
Waseda University
Tokyo, Japan
lvpurin@fuji.waseda.jp

Remina Yano
Graduate School of FSE
Waseda University
Tokyo, Japan
yano.remina@toki.waseda.jp

Marika Arimoto
Graduate School of FSE
Waseda University
Tokyo, Japan
m.arimoto@akane.waseda.jp

Miho Takahashi
Graduate School of FSE
Waseda University
Tokyo, Japan
miho.takahashi@akane.waseda.jp

Hiroshi Watanabe
Graduate School of FSE
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—In this paper, we propose an image denoising and a super resolution method for converting unreadable low resolution QR code images into readable high resolution ones. We propose an image denoising and a super resolution method using a simple CNN-based model. Image denoising using CNN-based image generation models can be applied to various noises by training model with different levels of noise. On the other hand, denoising using the conventional image processing method can only be applied to the specific type of noise. Therefore, image denoising using CNN is considered to be superior to image processing methods in terms of generalization performance. We further implement super resolution method to convert into high resolution images. We propose QRCNN and QRGAN, simple image denoising and super resolution models for QR code images. QRCNN and QRGAN are based on SRResNet and SRGAN, respectively. However, they both have simpler structure like SRCNN. Given QR code images for the dataset, we can reduce the computational complexity and memory usage by modifying model structure.**

***Keywords*—QRCNN, QRGAN, SRResNet, SRGAN, SRCNN**

I. Introduction

Nowadays, there are a lot of opportunities for people to scan QR codes due to the rise of QR code payments and site guidance. However, we sometimes encounter scanned QR codes images that are printed blurry or taken from a distance. Our proposed methods, QRCNN and QRGAN can convert unreadable low resolution QR code images into readable high resolution ones. QRGAN is a GAN-based [1] model based on QRCNN. In order to reduce computational complexity, these models have simple model structures and only allow grayscale images for the dataset. QRCNN and QRGAN are simpler model of SRResNet and SRGAN [2], respectively. In contrast to SRResNet and SRGAN, our proposed methods can reduce memory usage by not using pretrained VGG [3] for loss calculation. Super resolution model, SRCNN [4] is widely known to have very simple model structure whose number of layers and parameters are similar to our proposed methods. We compare the performance of our methods with SRCNN and show superiority to SRCNN.

II. Related Works

A. SRCNN

The most common super resolution method based on CNN is SRCNN. This model consists of only three convolution layers and two activation functions (ReLU [5]). In SRCNN, low resolution input images are upscaled to the desired image size using bicubic interpolation [6] before input to the network. Therefore, the size of the image does not change in the neural network. The loss function of SRCNN is given by

$$l^{SR} = l^{SR}_{MSE}. \tag{1}$$

Only mean squared error (MSE) of the ground truth and the generated image is used for loss function.

B. SRGAN

The most common super resolution method based on GANs with generators and discriminators is SRGAN. The generator attempts to produce an image that is close to the ground truth. The discriminator attempts to distinguish between the ground truth and the image produced by the generator. SRGAN utilizes this principle to improve super resolution tasks. Contrary to SRCNN, SRGAN enables super resolution by upscaling the size of the image in the middle of the neural network. Therefore, the input image is smaller than the output image. The generator consists of deep layers with 16 residual blocks and skip connections. Each residual block contains two convolutional networks. The discriminator consists of 8 convolutional networks. The loss function of SRGAN is given by

$$l^{SR} = \underbrace{l^{SR}_{X}}_{\text{content loss}} + \underbrace{10^{-3} l^{SR}}_{\text{adversarial loss}}, \tag{2}$$

where the function consists of a content loss and an adversarial loss. Content loss is calculated by MSE or obtained from feature maps derived from VGG. Adversarial loss, on the other hand, results from the generator and the discriminator competing each other.

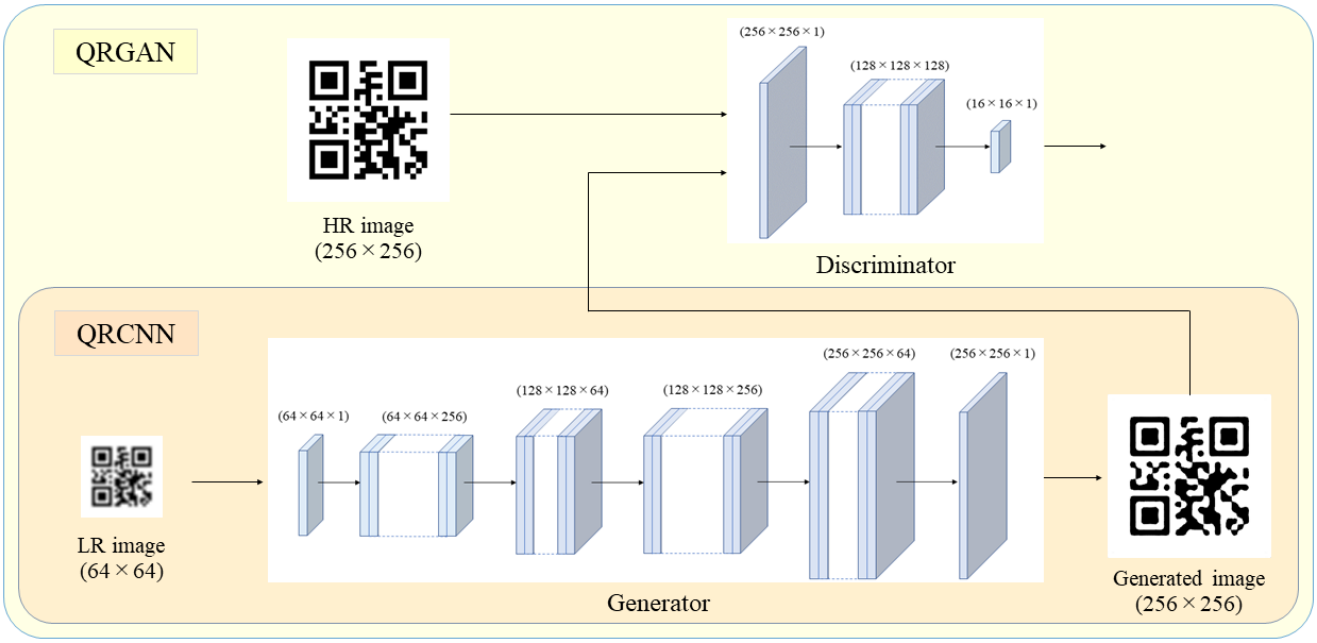


Fig. 1. Model structure of QRCNN and QRGAN

C. SRResNet

This model has similar deep structure to SRGAN, but does not have the discriminator. It is a model which we deprive the adversarial components of SRGAN. Similar to SRGAN, the generator attempts to produce an image that is close to the ground truth. Due to the lack of a discriminator, the loss function of SRResNet is given by

$$l^{SR} = \underbrace{l_X^{SR}}_{\text{content loss}}, \tag{3}$$

where the adversarial loss is not included. Same as SRGAN, content loss is calculated by MSE or obtained from feature maps derived from VGG.

III. Proposed Method

A. Model Structure

The model structure of both QRCNN and QRGAN are simple and the number of parameters of the generator used in these models is 176,449. Fig. 1 shows the model structure of QRCNN and QRGAN. Similar to SRResNet and SRGAN, our proposed methods enable super resolution by upscaling the input image in the middle of the network. The generator has three convolutional layers, two pixel shuffle layers and two activations (LeakyReLU). The discriminator of QRGAN has two convolutional layers, a batch normalization layer and an activation function (LeakyReLU). Contrary to SRResNet and SRGAN, proposed models are applied to grayscale images and does not use residual blocks or skip connections.

B. Loss Function

Like SRResNet, loss function of QRCNN has only content loss. Similar to SRGAN, loss function of QRGAN consists of content loss and adversarial loss. SRResNet and SRGAN use feature maps of VGG for content loss. However, our proposed methods use mean squared difference of pixels between the generated image and the ground truth image for the content loss. The loss function of QRCNN is given by

$$l^{SR} = \underbrace{l_{MSE}^{SR}}_{\text{content loss}}. \tag{4}$$

The loss function of QRGAN is given by

$$l^{SR} = \underbrace{l_{MSE}^{SR}}_{\text{content loss}} + \underbrace{10^{-3} l^{SR}}_{\text{adversarial loss}}. \tag{5}$$

IV. Experiment

A. Dataset

We use 1100 images from QR code image dataset [**?**], which is available on Kaggle. These QR code images contain linked numbers. When we scan these QR codes correctly, we can obtain these numbers. We use 1000 images for training and 100 images for evaluation. Since the size of the original image varies, we normalized them to a single size 256×256 [pixels] using bicubic method.

B. Training Details

As for training, we use 1000 QR code images. To simulate small QR code images, we convert training images to 64×64 [pixels] using bicubic method and apply Gaussian blur or add Gaussian noise. We use these resized images for the input of the generator of QRCNN and QRGAN. As for the input of the discriminator of QRGAN, the output image (256×256 [pixels]) generated by the generator of QRGAN and the ground truth image (256×256 [pixels]) are used. The number

of epochs is just 3 and 10 for Gaussian blur and Gaussian noise, respectively, so the learning process is very short.

C. Evaluation Method

We use 100 images for evaluation. Similar to the training process, we convert testing images to 64×64 [pixels] using bicubic and apply Gaussian blur or add Gaussian noise. We use these as the input images for the pretrained generator of QRCNN and QRGAN and simulate whether the output images are readable. In other words, we simulate whether the linked number is obtained by scanning the output image. For this simulation, we use python module, pyzbar. Pyzbar can scan QR code images and determine what is linked with these QR codes.

D. Comparison

We compare the performance of our proposed methods and SRCNN. To ensure the fairness of this comparison, the number of input channels of SRCNN is fixed to one. SRCNN has three convolutional layers. The number of channels and the kernel size of these convolutional layers inherit the values recommended in the paper of SRCNN. The number of output channels of these convolutional layers are 128, 64 and 1, while the number of the kernel size of these layers correspond to 9, 5 and 5. The number of parameters of SRCNN is 216,961, which is larger than our proposed method. We train SRCNN similar to QRCNN and QRGAN. In SRCNN, low resolution input images are upscaled to the desired image size using bicubic interpolation before input to the network.

E. Evaluation

1) Gaussian blur: To evaluate the performance of QRCNN and QRGAN, we first prepare blurred low resolution QR code images. We use Gaussian blur for this purpose. To control blurriness of the input images, we change the parameters of Gaussian blur used in training and evaluation. The parameters are standard deviation and kernel size. We use these prepared QR codes images for the input of SRCNN, QRCNN and QRGAN. The result of the training process is shown in Fig. 2. (a) refers to the input image that Gaussian blur is applied, (b) is the output image of SRCNN, (c) is the output image of QRCNN, (d) is the output image of QRGAN and (e) is the ground truth. All output images are readable. Table I-V shows the result of the readability (%) corresponding to different parameter sets. The blur values refer to the number of QR code images which are already readable even before applying QRCNN or QRGAN. Other values refer to the number of readable QR code images which are generated by SRCNN, QRCNN and QRGAN. For example, when the standard deviation is 1.10 and the kernel size is 3, blurred images will be completely readable by applying QRCNN or QRGAN. From this simulation, some unreadable QR code images with Gaussian blur can be converted into readable ones by SRCNN, QRCNN and QRGAN. Moreover, our proposed methods show better performance than SRCNN even though the number of parameters is less.

TABLE I
RESULTS OF THE READABILITY OF GAUSSIAN BLUR (KERNEL SIZE 3)

Method	standard deviation					
	1.1	1.2	1.3	1.4	1.5	1.6
Blur	26	18	22	19	15	18
SRCNN	100	100	100	100	98	97
QRCNN	100	100	100	100	100	100
QRGAN	100	100	100	100	100	100

TABLE II
RESULTS OF THE READABILITY OF GAUSSIAN BLUR (KERNEL SIZE 5)

Method	standard deviation					
	1.1	1.2	1.3	1.4	1.5	1.6
Blur	4	0	0	0	0	0
SRCNN	100	86	66	35	24	6
QRCNN	100	100	75	56	53	55
QRGAN	100	100	74	59	50	55

TABLE III
RESULTS OF THE READABILITY OF GAUSSIAN BLUR (KERNEL SIZE 7)

Method	standard deviation					
	1.1	1.2	1.3	1.4	1.5	1.6
Blur	3	0	0	0	0	0
SRCNN	100	100	99	76	57	27
QRCNN	100	100	100	100	85	49
QRGAN	100	100	100	100	88	50

TABLE IV
RESULTS OF THE READABILITY OF GAUSSIAN BLUR (KERNEL SIZE 9)

Method	standard deviation					
	1.1	1.2	1.3	1.4	1.5	1.6
Blur	4	0	0	0	0	0
SRCNN	100	100	96	75	65	52
QRCNN	100	100	100	100	82	74
QRGAN	100	100	100	100	82	74

TABLE V
RESULTS OF THE READABILITY OF GAUSSIAN BLUR (KERNEL SIZE 11)

Method	standard deviation					
	1.1	1.2	1.3	1.4	1.5	1.6
Blur	4	0	0	0	0	0
SRCNN	100	100	95	75	65	51
QRCNN	100	100	100	100	76	74
QRGAN	100	100	100	100	77	74

TABLE VI
RESULTS OF THE READABILITY OF GAUSSIAN NOISE

Method	standard deviation					
	0.7	0.8	0.9	1.0	1.1	1.2
Noise	55	39	12	0	0	0
SRCNN	100	100	94	80	65	48
QRCNN	100	100	100	95	81	66
QRGAN	100	100	100	100	80	73

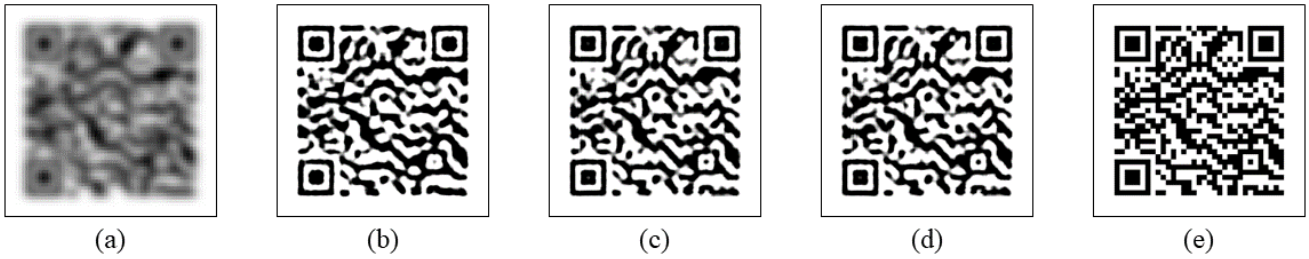


Fig. 2. Results of training for blurred images : (a) Input with kernel size 7 and standard deviation 1.4 (height and width are expanded by 4 to align with others); (b) Output of SRCNN; (c) Output of QRCNN; (d) Output of QRGAN; (e) Ground truth.

Fig. 3. Results of training for noise images : (a) Input with standard deviation 1.0 (height and width are expanded by 4 to align with others); (b) Output of SRCNN; (c) Output of QRCNN; (d) Output of QRGAN; (e) Ground truth.

2) Gaussian noise: Next, we prepare low resolution QR code images with noise. We use Gaussian noise for this purpose. To control the levels of noise, we change the parameter of Gaussian noise used in training and evaluation. The parameter we change is only standard deviation. We use these prepared QR code images as an input of SRCNN, QRCNN and QRGAN. The result of the training process is shown in Fig. 3. (a) refers to the input image with Gaussian noise, (b) is the output image of SRCNN, (c) is the output image of QRCNN, (d) is the output image of QRGAN and (e) is the ground truth. All output images are readable. Table VI shows the result of the readability (%) corresponding to different parameters. The noise values refer to the number of QR code images which are already readable even before applying QRCNN or QRGAN. Other values refer to the number of readable QR code images which are generated by SQCNN, QRCNN and QRGAN. For example, when the standard deviation is 0.7, images with noise will be completely readable by applying QRCNN or QRGAN. Similar to the first simulation, the result of this simulation also show the superiority of our proposed methods. QRCNN and QRGAN can be applied to images not only with blur but with noise.

V. Conclusion

In this paper, we propose QRCNN and QRGAN. QRCNN and QRGAN can convert unreadable QR code images into readable ones. From the experiments, our proposed methods are effective for QR code images with blur or noise. Moreover, our methods are superior to a typical CNN-based model, SRCNN, in terms of QR code image super resolution. We simplify the model structure by only allowing QR code images for the target of image denoising and super resolution. As a result, both QRCNN and QRGAN are able to reduce the number of parameters and computational complexity. Our model does not incorpolate pretrained model, so the memory usage is also small. Therefore, QRCNN and QRGAN can be implemented without relying on high performance GPUs.

Acknowledgment

The results of this research are based on the “Research and Development of Ultra-Coverage Beyond 5G Wireless Communications and Video Coding Standardization Technology through International Collaboration among Japan, the United States, and Australia” of the “Beyond 5G Research and Development Promotion Project (General Type)” commissioned by the National Institute of Information and Communications Technology (NICT), a research and development project for innovative information and communications technology (adopted in FY2022).

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio,“Generative adversarial nets”, NIPS, 2014.
[2] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., “Photo-realistic single image superresolution using a generative adversarial network”, CVPR, 2017.
[3] K. Simonyan, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, ICLR, 2015.
[4] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution”, ECCV, 2014.
[5] V. Nair and G.E. Hinton, “Rectified linear units improve restricted boltzmann machines”, ICML, 2010.
[6] R. G. Keys, “Cubic convolution interpolation for digital image processing”, IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-29, pp. 1153-1160, Dec. 1981.
[7] Cole Dieckhaus, “QR Codes”, Kaggle, 18 Feb. 2020.

符号化雑音環境下における物体検出精度の改善手法
A Method for Improving Object Detection Accuracy in Coding Noise Environment

進藤嵩紘 † 渡部泰樹 † 渡辺裕 †‡
Takahiro Shindo† Taiju Watanabe† Hiroshi Watanabe†‡

† 早稲田大学基幹理工学部
†School of Fundamental Science and Engineering, Waseda University

‡ 早稲田大学大学院基幹理工学研究科
‡Graduate School of Fundamental Science and Engineering, Waseda University

Abstract: Research and standardization activities for Video Coding for Machine (VCM) has been intensified. In this paper, we propose a method to improve the accuracy of image recognition by processing the coding noise in VVC encoded video. The proposed method is based on ESRGAN which is a Convolutional Neural Network (CNN). The evaluation method is the accuracy of object detection by YOLOv7. Experimental results show that the proposed coding noise processing improves object detection accuracy.

1 はじめに

近年，画像認識のための動画像符号化技術に関する研究が行われている．Versatile Video Coding(VVC)[1] による符号化映像には符号化雑音が加わるため，画像認識精度の低下を招く．本稿では，ニューラルネットワークを用いて符号化映像の符号化雑音を処理することにより，画像認識の精度を改善する手法について提案する．提案手法は，Enhanced Super-Resolution Generative Adversarial Networks(ESRGAN)[2] の生成器の構造をもとに作成する Convolutional Neural Network (CNN) である．評価手法には YOLOv7[3] の学習済みモデルによる物体検出精度を用いる．提案する符号化雑音処理により，物体検出精度が改善できることを実験により示す．

2 従来手法

Enhancing VVC Through CNN-Based Post-Processing [4] では，CNN を用いた雑音除去手法により，VVC による符号化映像の品質向上手法を提案する．ネットワーク構造は SRGAN[5] の生成器を参考に構成された CNN である．ネットワークは VVC による符号化映像のフレームを入力とし，出力画像と符号化前の画像との絶対誤差を用いて学習する．評価手法では PSNR(Peak Signal to Noise Ratio) を用い，CNN を用いた符号化雑音処理により，VVC による符号化雑音が低減することを示す．

3 提案手法

従来手法は，VVC による符号化映像の符号化雑音を低減し，符号化前の映像に近づけることを目的とする．しかし，低減できる符号化雑音の大きさは小さく，画像認識精度の改善には結びつかない．そこで，VGG[6] により得られる特徴量を損失計算に用いることで，VVC による符号化映像の画像認識精度の改善を目指す．また，ESRGAN の生成器の構造を参考に，SRGAN を参考に構成される従来手法のネットワーク構造を改変する．

3.1 モデル構造

ESRGAN は SRGAN と同じく，画像の超解像を行うモデルである．このモデルは，SRGAN の residual block[7] を Residual-in-Residual Dense Block(RRDB) に置き換えた構造を持ち，より深いニューラルネットワークを構成することで，画像の超解像において細かな絵柄の再現を可能にする．本稿の提案手法でも，従来手法の residual block を RRDB に置き換たモデルを用いる．提案するネットワーク構造を図 1 に示す．

3.2 損失関数

VGG は画像認識精度の高さから，特徴抽出手法として広く利用されるモデルの一つである．出力画像と符号化前の画像の絶対誤差を損失計算で用いる従来手法とは異なり，提案手法ではそれらの平均二乗誤差と，それらから VGG により抽出される特徴量の平均二乗誤差を損失計算に用いる．従来手法で用いる損失を式 (1) に，提案手法で

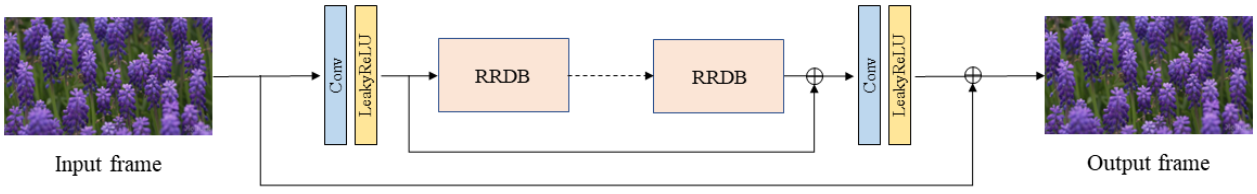


図 1: 提案手法のモデル構造

図 2: Marathon シーケンス

(a)正解フレーム (b)入力フレーム (c)出力フレーム

図 3: 検出結果

用いる損失を式 (2) に示す.

$$loss = l_{L1} \tag{1}$$

$$loss = l_{MSE} + l_{VGG} \tag{2}$$

4 実験及び結果

提案手法により，符号化雑音環境下において画像認識精度が改善することを実験により示す．学習用データセットには，SJTU データセット (Marathon シーケンスは除く)[8]，UVG データセット [9]，MCL-JCV データセット [10] を用いた．これらのデータセットを VTM10.0[11] を用いて符号化する．参照構造はすべてランダムアクセス，量子化係数は 37 である．符号化後の映像をモデルの入力とし，VGG の学習済みモデルにより抽出される特徴量を用いてモデルの学習を行う．テスト用データセットには，SJTU データセット内の Marathon シーケンスを用いる．1 秒間のフレーム数は 25 であり，総フレーム数は 60 である．Marathon シーケンスのフレーム画像を図 2 に示す．テストデータの符号化方法は学習に用いるデータの符号化手法と同一である．

評価手法は入力動画像と出力動画像と正解動画像のそれぞれに対して，YOLOv7 の学習済みモデルにより人物を検出した結果を用いる．検出時に用いる信頼度の閾値を 0.25 から 0.95 まで 0.05 刻みで変化させたときに，検出される人数をそれぞれ計測する．

検出結果の一例を図 3 に示す．検出人数と信頼度の閾値の関係を図 4 に示す．計測に用いたすべての信頼度の閾値において，提案手法による画像認識精度の改善が確認できる．つまり，VGG により抽出される特徴量を用いた符号化雑音処理は，符号化映像の画像認識精度の改善に有効である．

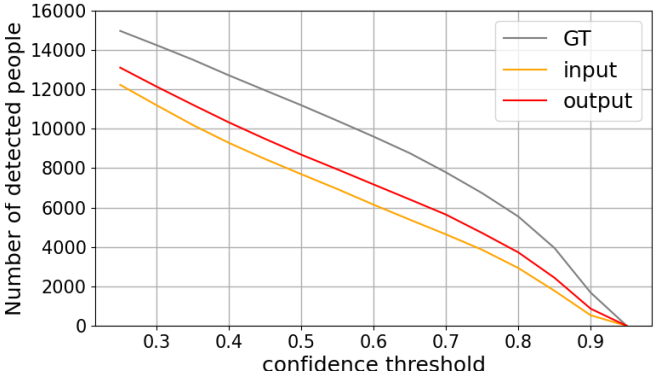


図 4: 検出人数と信頼度の閾値の関係

5 まとめ

VGG による特徴抽出を用いた符号化雑音処理により，符号化映像の画像認識精度を改善できることを示した．VVC による符号化映像を提案手法により処理することで，YOLOv7 による人物検出精度が改善することを実験により確認した．今後，提案手法の汎用性を確かめるために，テスト用のシーケンスを増やす必要がある．

謝辞

本研究成果は，国立研究開発法人情報通信研究機構の委託研究 (05101) により得られたものである．

参考文献

[1] S. L. B. Bross, J. Chen, Versatile Video Coding (Draft 10). JVET-S2001, 2020.

[2] X. Wang *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks", ECCV Workshop, 2018.

[3] Chien-Yao Wang *et al.*, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for realtime object detectors" arXiv preprint arXiv:2207.02696 (2022).

[4] F. Zhang *et al.*, "Enhancing VVC Through CNN-Based Post-Processing", ICME, 2020.

[5] C. Ledig *et al.*, "Photo-realistic single image superresolution using a generative adversarial network", CVPR, 2017.

[6] K. Simonyan *et al.*, "Very deep convolutional networks for large-scale image recognition", ICLR, 2015.

[7] K. He *et al.*, "Deep residual learning for image recognition", CVPR, 2016.

[8] L. Song *et al.*, "The SJTU 4K video sequence dataset", QoMEX, 2013.

[9] A. Mercat *et al.*, "UVG dataset: 50/120fps 4K sequences for video codec analysis and development", ACM Multimedia Systems Conference, 2020.

[10] H. Wang *et al.*, "MCL-JCV: A JND-based H.264/AVC video quality assessment dataset", ICIP, 2016.

[11] S. K. J. Chen, Y. Ye, Algorithm description for Versatile Video Coding and Test Model 10 (VTM 10). JVET-S2002, 2020.

早稲田大学基幹理工学部情報通信学科
〒*162-0072* 東京都新宿区大久保 *3-14-9* 早大 *66-401*
Phone: 03-5286-2509, E-mail: taka_s0265@ruri.waseda.jp

Novel CNN approach for video prediction based on FitVid

Taiju Watanabe, Shindo Takahiro, Hiroshi Watanabe
School of Fundamental Science and Engineering, Waseda University
Shillman Hall 401, 3-14-9 Okubo, Shinjuku, Tokyo, 169-0072 Japan

ABSTRACT

Video prediction is a task in computer vision that predicts future frames from the past few frames of video. In video prediction, a simple CNN-based approach called SimVP has marked remarkable performance without using RNN or vison transformer (ViT). In this paper, we propose a model structure to improve performance of video prediction based on FitVid. FitVid is a regression-based method of predicting future videos using not only video but also motion. We focus on video prediction only conditioned on videos. To this goal, we introduce network modules used in SimVP to FitVid. Experimental results show that the proposed structure shows better prediction accuracy compared to SimVP.

Keywords: Video prediction, SimVP, FitVid, RNN, ViT

1. INTRODUCTION

Video prediction is a task in computer vision that predicts future frames from the past few frames of video. Video prediction is applied to scene understanding and high efficiency video coding. Use of both convolutional and recurrent neural networks is a recent trend of video prediction models. SimVP marks state-of-the-art accuracy on several video prediction datasets, despite using only a simple structure of convolutional layers. FitVid is a regression-based method of predicting future videos using not only video but also motion. The prediction model uses combinations of convolutional and recurrent neural networks. It is the first model to show overfitting on several video prediction tasks. However, the training process is inefficient because of its huge number of parameters. In this paper, we focus on video prediction only conditioned on past videos. We propose to further improve video prediction accuracy of SimVP by introducing a similar model structure as FitVid except for motion information. We compare our model with SimVP on two datasets. The model performed better when applied to difficult video prediction datasets and obtained competitive results when applied to easier video prediction datasets without much increase in computational complexity.

2. RELATED WORKS

2.1 SimVP

SimVP [1] is a video prediction model with simple architectures based on CNNs. It does not incorporate any complex strategies such as adversarial training, teacher-student distilling, and optical flow. SimVP consists of three parts, an encoder, a translator, and a decoder. These modules are all built with CNNs. The encoder is used to obtain spatial features, the translator learns time evolution, and decoder uses spatial and temporal information to predict future frames. With these simple architectures, SimVP has marked the state-of-the-art performance on several prediction datasets. Since SimVP has a very simple configuration, there is room to improve prediction efficiency by adding a more complex structure.

2.2 FitVid

FitVid [2] is a regression-based method of predicting future videos using not only video but also motion. It produces future frames from past frames and actions with autoregressive manner. FitVid consists of an encoder, a dynamics model, and a decoder. The encoder and decoder use similar residual encoding and decoding cells as presented in NVAE [3]. NVAE is a deep hierarchical variational autoencoder which uses spectral regularization and residual parameterization for improving KL optimization. It also uses depth wise convolution in the generator to acquire wide range of information in the image. As for the dynamics model, two layers of LSTMs are incorporated to predict future latent variables from encoded frames. However, FitVid may not be effective, when used for prediction tasks without using actions. Moreover, the training process is inefficient due to its large number of parameters.

3. PROPOSED METHOD

3.1 Encoder and decoder

Our model consists of three parts, an encoder, a translator, and a decoder. The encoder and decoder are used to learn spatial features of the input, and the translator is used to learn temporal information of the input.

As for the encoder and the decoder, we follow the structure to FitVid. The number of hidden channels does not change in the encoder and decoder of SimVP. However, we expand the hidden channels by two when the image size is scaled to half. Each cell consists of two convolutional layers with Group normalization and swish activation function. We also incorporate skip connection [4] to achieve efficient learning. In the convolution layer, 10% of the output of the previous layer is added to the current layer. This enables the model to learn information of previous features. Similar to FitVid, we add Squeeze-and-Excitation blocks (SE blocks) [5] to further enhance the effectiveness of skip connection. By use of SE blocks, channels are adaptively recalibrated by intensifying the correspondence between channels. SimVP uses transposed convolution for up sampling in the decoder.

3.2 Translator

We use same architecture for a translator as SimVP. In the translator, inception modules are used to learn temporal evolution.

We created four models, a baseline model, the baseline with skip connections, the baseline with skip connections and Squeeze-and-Excitation blocks (SE blocks), the baseline with skip connections, SE blocks and nearest neighbor interpolation. All these models are based on CNNs. Model structure of our final model is shown in Fig. 1. We only use mean squared error (MSE) of the output and the ground truth for the loss function.

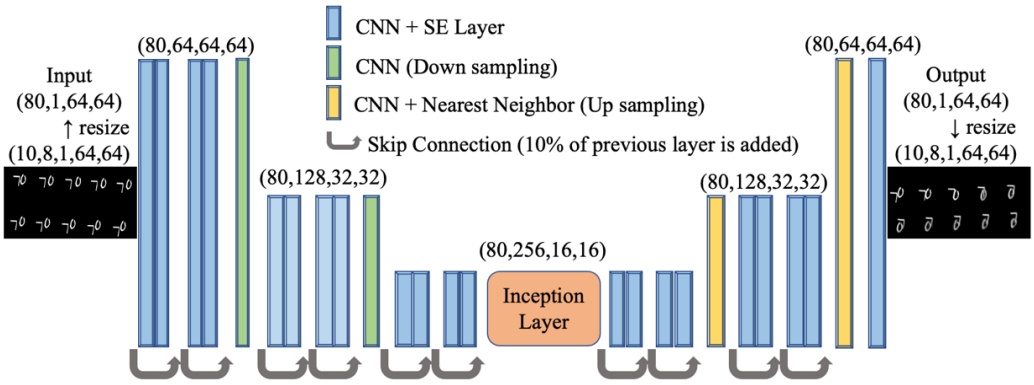


Fig. 1. Model structure of our final model.

4. EXPERIMENTS

4.1 Datasets

To evaluate the performance of our proposed model, we use two video prediction datasets, Moving MNIST [7] and TrafficBJ. Moving MNIST contains videos of two handwritten numbers moving around the scene. It contains 10000 videos for training and 10000 videos for testing. Each video sequence has a resolution of 64x64 pixels. Our task is to predict future 10 frames given 10 previous frames. TrafficBJ contains the trajectory data in Beijing collected from taxicab GPS with two channels, inflow or outflow defined in [8]. Similar to SimVP, we normalized the data into [0,1]. It contains 19627 videos for training and 1334 videos for testing. Each video sequence has a resolution of 32x32 pixels. Our task is to predict future 4 frames given 4 previous frames. The statistics are summarized in Table. 1.

Table. 1. The statistics of datasets. N_{train} is the number of training samples and N_{test} is the number of testing samples. Each sample contains T previous frames and T' future frames with the size of (C, H, W).

Dataset	N_{train}	N_{test}	(C, H, W)	T	T'
Moving MNIST	10000	10000	(1, 64, 64)	10	10
TrafficBJ	19627	1334	(2, 32, 32)	4	4

4.2 Metrics

We use mean squared error (MSE), mean absolute error (MAE), Structural Similarity Index Measure (SSIM) to evaluate the performance of our models.

4.3 Results

We compare our models with SimVP on two datasets with three metrics. The result is shown in Table. 2. From the result, our final model, baseline with skip connection, Squeeze-and-Excitation block, and nearest neighbor interpolation shows better video prediction performance than SimVP. Especially, in difficult video prediction setting where Moving MNIST is used for the dataset, our final model shows significant performance. Our models are trained only for 1200 epochs where SimVP is trained for 2000 epochs. This indicates that our models can be trained more efficiently. Our models show competitive performance with TrafficBJ dataset. Moreover, our model improves by applying several techniques such as skip connection and nearest neighbor interpolation. The visualization results of SimVP and our final model are shown in Fig. 2., and Fig. 3.

Table. 2. Comparison of our models and SimVP. The optimal results are marked by bold.
(SC : skip connection, SE : Squeeze-and-Excitation block, NN : nearest neighbor interpolation)

Method	Moving MNIST			TrafficBJ		
	MSE	MAE	SSIM	MSE × 100	MAE	SSIM
SimVP	23.8	68.9	0.948	41.4	**16.2**	**0.982**
Baseline	23.7	71.0	0.948	41.4	**16.2**	**0.982**
Baseline + SC	23.3	73.6	**0.949**	**39.1**	16.3	0.981
Baseline + SC + SE	22.9	70.7	0.946	44.0	16.4	**0.982**
Baseline + SC + SE + NN	**22.5**	**68.3**	**0.949**	39.4	16.6	0.981

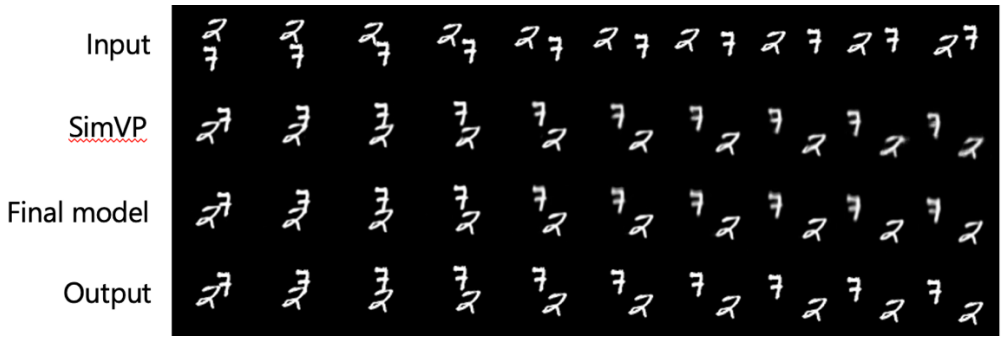


Fig. 2. Result of Moving MNIST.

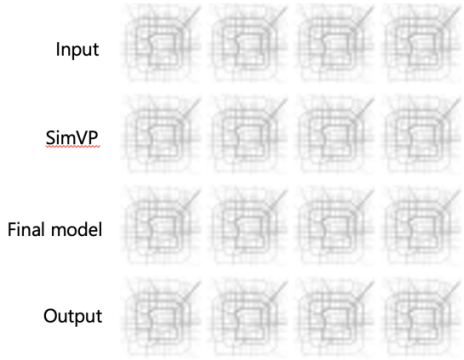


Fig. 3. Result of TrafficBJ.

4.4 Consideration

From Table. 2., our model shows better performance than SimVP when dealing with difficult video prediction tasks. However, our model did not show superiority when it comes to simpler settings. When training TrafficBJ with our final model, the performance did not change from 40 epochs to 80 epochs. This indicates that our final model is overfitting with TrafficBJ like it is shown in FitVid. From Fig. 2., it is more difficult to predict distant future frames.

5. CONCLUSION

In this paper, we propose a model structure to improve performance of video prediction based on FitVid. From the experiment, our model shows better performance than SimVP, the state-of-the-art model for video prediction tasks based on CNNs. Our model is also based on CNN, but by applying additional features inspired by FitVid, it shows better performance than SimVP. Our future work is to seek better model structure and apply to different settings such as human activity estimation.

6. ACKNOWLEDGEMENT

These research results were obtained from the commissioned research (No.05101) by National Institute of Information and Communications Technology (NICT), Japan.

REFERENCES

[1] Z. Gao, C. Tan, L. Wu and S. Z. Li, “SimVP: Simpler yet Better Video Prediction”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3160-3180, Jun. 2022.
[2] M. Babaeizadeh, M. T. Saffar, S. Nair, S. Levine, C. Finn and D. Erhan, “FitVid: Overfitting in Pixel-Level Video Prediction”, arXiv:2106.13195, 1-25, Jun. 2021.
[3] A. Vahdat and J. Kautz, “NVAE: A Deep Hierarchical Variational Autoencoder”, Conference on Neural Information Processing Systems (NeurIPS), 1-13, Dec. 2020.
[4] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition”, arXiv:1512.03385, 1-12, Dec. 2015.
[5] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, “Squeeze-and-Excitation Networks”, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, 2011-2023, Aug. 2020.
[6] O. Rukundo and H. Cao, “Nearest Neighbor Value Interpolation”, International Journal of Advanced Computer Science and Applications (IJACSA), vol. 3, no. 4, 1-6, Mar. 2012.
[7] N. Srivastava, E. Mansimov and R. Salakhudinov, “Unsupervised Learning of Video Representations using LSTMs”, International Conference on International Conference on Machine Learning (ICML), vol. 37, 843-852, Jul. 2015.
[8] J. Zhang, Y. Zheng and D. Qi, “Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction”, AAAI Conference on Artificial Intelligence, 1655-1661, Feb. 2017.

VVC と CNN を組み合わせた YOLO-v7 のための映像符号化手法

Video Coding Scheme for YOLO-v7 Combining VVC and CNN

進藤嵩紘　　渡部泰樹　　渡辺裕
Takahiro Shindo　　Taiju Watanabe　　Hiroshi Watanabe

早稲田大学基幹理工学部
School of Fundamental Science and Engineering, Waseda University

1. まえがき

近年，画像認識技術の発達により，AI を用いた映像解析が急速に拡大している．そこで 2019 年，Moving Picture Experts Group (MPEG) では，Video Coding for Machines (VCM) を画像認識のための映像符号化と位置づけ，標準化作業を開始している．VCM では，より高い映像の圧縮率と画像認識精度が求められる．本稿では，CNN と Versatile Video Coding (VVC) を組み合わせることにより，YOLO-v7 による物体検出精度が高くなる映像符号化手法を提案する．YOLO-v7 の学習済みモデルにより，映像から抽出する特徴量を用いて CNN を学習することで，物体検出に有効な映像を作成する．

2. 提案手法

VVC[1]は最新の動画像符号化方式であり，高い圧縮率と映像品質を達成している．しかし，視聴用符号化方式として設計されているため，CNN と VVC を組み合わせることにより，YOLO-v7[2]による物体検出のための映像符号化手法を提案する．まず CNN により映像の画像サイズを半分に縮小し，その映像を VVC により圧縮することで高い圧縮率を達成する．VVC デコーダの出力映像を，CNN により元のサイズに戻すことで，物体検出に必要な映像情報の復元を目指す．VVC の符号化処理には VTM10.0[3]を使用し，参照構造は「lowdelay_P」とする．提案する映像符号化手法を図 1 に示す．

CNN の学習には YOLO-v7 の特徴量を用いる．映像の特徴抽出には学習済みモデルの backbone を使用し，生成映像と正解映像の特徴量の平均二乗誤差 (MSE) を損失計算に用いる．学習に用いる損失関数を次の式(1)に示す．

$$LOSS = MSE\big(yolo(I_{coded}), yolo(I_{raw})\big) \qquad (1)$$

ここに，$yolo$ は YOLO-v7 の backbone を用いた特徴抽出器を表し，I_{coded} は生成画像，I_{raw} は正解画像を表す．

3. 実験と結果

学習には, SJTU[4]，UVG[5]，MCML-4K-UHD[6]の三つのデータセットを用いる．テストでは VCM の Common Test Condition (CTC)[7] で用いられる，SFU-HW-Objects-v1[8]を使用する．学習に用いるシーケンスはすべて画像サイズが 4K であるため，テストでは最も画像サイズが大きい class A の Traffic シーケンスを用いる．

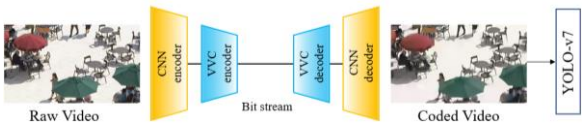


図 1　提案する映像符号化手法

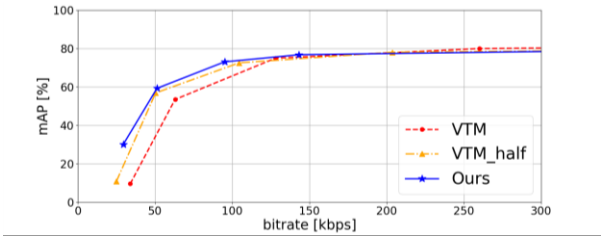


図 2　各符号化手法における bitrate と mAP の関係

提案手法による符号化映像と VTM10.0 による符号化映像の物体検出精度を比較する．また，画像サイズを半分に縮小し，VTM10.0 により符号化した場合の物体検出精度とも比較する．物体検出モデルは YOLO-v7 とし，比較する VTM の参照構造は「randomaccess」とする．検出精度は mean Average Precision (mAP) を用いて計測し，計測時に用いる Intersection over Union の閾値は 0.5 とする．符号化映像の bitrate と mAP の関係を図 2 に示す．図 2 より，提案手法は，映像の圧縮効率と物体検出精度で，VVC を上回ることが分かる．さらに，単純に画像サイズを縮小してから符号化する場合よりも，物体検出精度が高いことが分かる．

4. むすび

本稿では，YOLO-v7 のための映像符号化手法として，CNN と VVC を組み合わせた手法を提案した．CNN を YOLO-v7 の学習済みモデルを用いて学習させることで，物体検出に有効な映像を作成できることを実験により示した．

謝辞

本研究成果は，国立研究開発法人情報通信研究機構の委託研究 (05101) により得られたものである．

参考文献

[1] Versatile Video Coding, Standard ISO/IEC 23090-3, ISO/IEC JTC 1, Jul. 2020.
[2] C. Y. Wang, *et al.*, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for realtime object detectors.” arXiv preprint arXiv:2207.02696, 2022.
[3] S. K. J. Chen, *et al.*, Algorithm description for Versatile Video Coding and Test Model 10 (VTM 10). JVETS2002, 2020.
[4] L. Song, *et al.*, “The SJTU 4K Video Sequence Dataset,” in International Conference on Quality of Multimedia Experience, 2013.
[5] A. Mercat, *et al.*, “UVG dataset: 50/120fps 4K Sequences for Video Codec Analysis and Development,” in ACM Multimedia, 2020.
[6] M. Cheon, *et al.*, “Subjective and Objective Quality Assessment of Compressed 4K UHD Videos for Immersive Experience,” in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, 2018.
[7] Common test conditions for video coding for machines ISO/IEC JTC 1/SC 29/WG 04, Nov. 2022.
[8] H. Choi, *et al.* “A dataset of labelled objects on raw video sequences.” Data in Brief, 34:106701, 2021.

YOLOV を用いた物体予測検出の一検討

A Study on Future Object Detection Using YOLOV

渡部泰樹　進藤嵩紘　渡辺裕
Taiju Watanabe　Shindo Takahiro　Hiroshi Watanabe

早稲田大学基幹理工学部
School of Fundamental Science and Engineering, Waseda University

１．まえがき

物体検出とは，画像の中から定められた物体の種類，位置を正確に特定するタスクである．本稿では，動画を用いた物体予測検出手法を提案する．物体予測検出 (future object detection) とは，過去のフレームから将来の物体の種類，位置を特定するタスクである[1]．物体予測検出では，将来の物体の動向を把握することができるため，危険予測などに応用できる．提案手法では動画物体検出手法の YOLOV[2]を予測用に修正し，動画予測モデルである SimVP[3]の構造を付加することにより，時系列情報の取得を可能としている．

２．提案手法

物体予測検出とは，$\mathrm{t}=1$から$\mathrm{t}=\mathrm{T}$までの過去のフレーム$\{\mathbf{I}^{\mathrm{t}}\}_{\mathrm{t}=1}^{\mathrm{T}}$を物体検出モデル $\mathbf{F}$ の入力として，将来の物体の位置，信頼度，クラス予測確率を含んだベクトル$\mathbf{O}^{\mathrm{T}+\tau}$を出力するタスクである．この物体検出モデル $\mathbf{F}$ は，時刻$\mathrm{t}=\mathrm{T}$から$\mathrm{t}=\tau$までの物体予測検出モデルであり，式(1)のように定式化できる．

$$\mathbf{O}^{\mathrm{T}+\tau}=\mathbf{F}(\mathrm{I}^{1},\mathrm{I}^{2},\mathrm{I}^{3},\cdots,\mathrm{I}^{\mathrm{T}}) \tag{1}$$

本稿では，簡単のために$\mathrm{T}=3$, $\tau=3$として，過去 3 フレームから将来の 3 フレームに存在する物体情報を推定する．物体予測検出手法として，動画物体検出手法の YOLOV を利用する．YOLOV は事前学習済みの YOLOX [4]を利用したモデルであり，動画物体検出タスクにおいて高い性能を示している．そこで，YOLOV を物体予測検出用に修正する．具体的には過去 3 フレームを入力として，正解を将来の 3 フレームの物体情報として学習させる．しかしながら，この YOLOV は時系列情報を保持できないという問題点がある．そこで，入力のフレーム間情報を取得するために動画予測モデルである SimVP の構造を取り入れる．SimVP では中間特徴量に対して，複数の Inception モジュールを適応させることにより入力フレームの時系列情報を保持する．提案する検出モデルの構造を図１に示す．

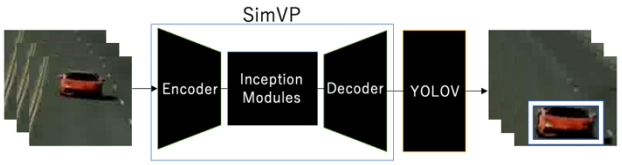


図１．提案する検出モデルの構造

３．実験

提案手法の有効性を検証するために実験を行った．データセットは ImageNet VID [5]を用いる．ImageNet VID は動画の物体検出用のデータセットであり，訓練用として 3862 個の動画，検証用として 555 個の動画が用意されている．また，クラス数は 30 である．YOLOV では，パラメタ数によって，三つのモデル(YOLOV-S, YOLOV-L, YOLOV-X)が用意されており，それぞれについて SimVP を用いて時系列情報を含めた場合との比較を行う．いずれのモデルについても 10 エポック学習とする．評価指標として，IOU の閾値を 50 とした AP50 を各クラスについて計算し，その mAP を用いた．表１に物体予測検出結果を示す．表１から SimVP によって時系列情報を保持したモデルの方が，物体検出精度が向上していることがわかる．

表１．物体予測検出結果

モデル	mAP
YOLOV-S	68.2
YOLOV-S + SimVP	**68.5**
YOLOV-L	74.2
YOLOV-L + SimVP	**74.4**
YOLOV-X	73.8
YOLOV-X + SimVP	**73.9**

４．まとめ

本稿では，YOLOV を用いた物体予測検出手法を提案した．SimVP を用いて時系列情報を保持させることで，物体検出精度が向上することを実験により確認した．

謝辞

本研究成果は，国立研究開発法人情報通信研究機構の委託研究 (No. 05101) により得られたものである．

参考文献

[1] A. Tonderski, et al. “Future Object Detection with Spatiotemporal Transformers”, arXiv: 2204.10321, 1-22, Apr. 2022.

[2] Y. Shi, et al. “YOLOV: Making Still Image Object Detectors Great at Video Object Detection”, arXiv:2208.09686, 1-11, Aug. 2022.

[3] Z. Gao, et al. “SimVP: Simpler yet Better Video Prediction”, CVPR, 3160-3180, Jun. 2022.

[4] Z. Ge, et al. “YOLOX: Exceeding YOLO Series in 2021”, arXiv:2107.08430, 1-7, Jul. 2021.

[5] O. Russakovsky, et al. “ImageNet Large Scale Visual Recognition Challenge”, IJCV, 211-252, Apr. 2015.