****

**本 科 生 毕 业 论 文 正 文**

**（ 2021 届）**

|  |  |  |  |
| --- | --- | --- | --- |
| 论文题目 | 基于多帧的视频质量增强技术与实现 | | |
|  |  | | |
|  | | | |
| 学生姓名 | 褚灵强 | 学 号 | 2017212212122 |
| 专 业 | 计算机科学与技术 | 班 级 | 计算机174 |
| 指导教师 | 丁丹丹 | 职 称 | 讲师 |

**杭州国际服务工程学院教务科**

基于多帧的视频质量增强技术与实现

摘 要

在当今社会，随着腾讯视频、抖音等视频应用的快速发展，以及监控摄像头等视频采集设备增加，视频逐渐成为互联网的主要流量，对高分辨率视频的压缩、低分辨率视频的增强需求日益扩大。目前主流采用的AV1，HEVC等编码框架，均为有损压缩，对视频画质产生了较大的影响。一般来说，我们会采用PSNR，SSIM对图像进行客观质量评价，近几年主观质量评价也逐渐兴起。如何提高视频的主客观质量是研究人员关注的热点。过去，大家大都采用传统手段对视频进行各种方式的增强，如超分辨率、插帧、HDR等。近几年随着机器学习的快速发展，大量基于CNN,RNN等神经网络的解决方案被提出，这些方案的效果相比之下高于传统方案，吸引了越来越多的研究人员使用机器学习来解决视频增强这一问题。但是大多数情况下，人们只是把视频增强简单的理解为多次图像增强，没有考虑到视频帧之间的时空相关性。目前，最新的研究发现，合理运用视频帧之间的相关性，采用多帧融合的方式，可以在一定程度上提升视频增强的效果。

关键词：超分辨率，EDVR, 金字塔结构，多帧融合

**Multi-frame-based video quality enhancement technology and implementation**

**ABSTRACT**

In today's society, with the rapid development of video applications such as Tencent Video and Tictok, and the increase of video capture devices such as surveillance cameras, video has gradually become the main traffic of the Internet. There is a demand for high-resolution video compression and low-resolution video enhancement. Expanding day by day. At present, the mainstream AV1, HEVC and other coding frameworks are all lossy compression, which has a greater impact on video quality. Generally speaking, we will use PSNR and SSIM to evaluate the image quality objectively. In recent years, subjective quality evaluation has gradually emerged. How to improve the subjective and objective quality of video is a hot topic for researchers. In the past, most people used traditional methods to enhance video in various ways, such as super-resolution, frame insertion, and HDR. In recent years, with the rapid development of machine learning, a large number of solutions based on CNN, RNN and other neural networks have been proposed. The effects of these solutions are higher than traditional solutions, attracting more and more researchers to use machine learning. To solve the problem of video enhancement. But in most cases, people simply interpret video enhancement as multiple image enhancements, without considering the spatio-temporal correlation between video frames. At present, the latest research has found that the rational use of the correlation between video frames and the use of multi-frame fusion can improve the effect of video enhancement to a certain extent.

**Keywords:** Deep Learning, Neural Network Structure, Video Processing Demonstration, EDVR, Integration

目录

[第一章 绪论 5](#_Toc68704150)

[1.1 引言 5](#_Toc68704151)

[1.2 基于深度学习的超分辨率处理现状研究 5](#_Toc68704152)

[1.3 研究意义 6](#_Toc68704153)

[第二章 相关技术概述 7](#_Toc68704154)

[2.1 开发环境 7](#_Toc68704155)

[2.1.1 开发工具简介 7](#_Toc68704156)

[2.1.2 环境配置 7](#_Toc68704157)

[2.2 相关技术 7](#_Toc68704158)

[2.2.1 pytorch相关技术特点 7](#_Toc68704159)

[2.2.2 多帧对齐相关技术特点 7](#_Toc68704160)

[2.2.3 Resnet相关技术特点 8](#_Toc68704161)

[2.3 图像评价标准 8](#_Toc68704162)

[2.3.1主观评价方法 8](#_Toc68704163)

[2.3.2客观评价方式 8](#_Toc68704164)

[第三章 针对AV1编码的超分辨率算法 9](#_Toc68704165)

[3.1 EDVR算法介绍 9](#_Toc68704166)

[3.1.1 PCD对齐模块 9](#_Toc68704167)

[3.1.2 TSA融合模块 9](#_Toc68704168)

[3.2 训练集准备 9](#_Toc68704169)

[3.3 训练改进 10](#_Toc68704170)

# 第一章 绪论

1.1 引言

视频超分辨率用于是解决因硬件因素或压缩后导致的视频高频细节丢失问题，从低分辨率视频中还原其原有的高频细节，提升其主客观质量。随着科技和社会的发展，人们对分辨率的要求越来越高，从过去的480p到现在的4k，超高分辨率的需求与日俱增，不仅仅是日常视频观看，高分辨率对安防、医疗、遥感等方面也具有重要作用。虽然目前硬件快速发展，传感器所能采集的像素值越来越大，成像技术快速发展，但在实际运行过程中，需要考虑系统成本和稳定运行的要求，往往在一定程度上牺牲了分辨率，所以超分辨率技术仍然十分重要，具有广阔的应用前景。通过超分辨率技术，我们无需购买昂贵的最新传感器，可以在已有传感器的基础上，超越传感器低分辨率的限制，获得主客观质量更高的高分辨率视频。

现阶段，研究人员主要通过插值、重建、学习三种方式对视频进行超分辨率处理，其中又可分为基于频域、空域、时域的方法。最初，大多数研究人员采用基于频域的方式，主要利用傅里叶变换算法，分离视频低频分量中被隐藏的高频信息，回复视频部分细节。其主要基于数学计算，基础原理清晰，理论推导方便，计算复杂度较低，具有高效率的特点，但其对于误差较为敏感，对有噪声和运动的视频，处理结果差强人意。不久，人们逐渐将目光转向了基于空域的算法，采用插值等方式，增加像素值，算法较为灵活，种类繁多，但其计算一般较为复杂。

1.2 基于深度学习的超分辨率处理现状研究

2016年，Dong等人提出了SRCNN(Super-Resolution Convolutional Neural Network)网络，首次将卷积神经网络应用于单张图像超分辨率重建中，开创了深度学习在图像超分辨率问题中的应用。SRCNN将深度学习与传统稀疏编码之间的关系作为依据，将3层网络划分为图像块提取(Patch extraction and representation)、非线性映射(Non-linear mapping)以及最终的重建(Reconstruction)，其重建效果远远优于其他传统算法。2017年，Li D, Wang Z提出了一种新颖的视频超分辨率算法，使用运动补偿与残差网络（MCResNet）。他们将光流算法用于运动估计与运动补偿以进行预处理，然后使用深度残差卷积神经网络，使用多个运动补偿观测值来预测高分辨率图像。他们所提出的CNN模型可以保留低频内容且有助于恢复高频细节，并自适应地处理大型和复杂的运动。同年，Caballero J等人引进了一种时空亚像素卷积网络，能够对视频进行超分辨率重建并且做到实时速度，实现了将动作补偿和视频超分辨联合起来的算法，并且可以端到端训练。

目前的改进思路大概有以下几种，采用更深的网络、使用残差连接、图像进行补0操作等。例如在2017年的CVPR上提出的EDSR网络。一般来说越深的网络对训练要求更高，会面临梯度消失的问题，也更难收敛，EDSR中采用了增强的resnet引入残差学习，使网络只学习原始图像与目标图像之间从残差，使得收敛速度快。同时，研究人员发现在超分辨率任务中，过拟合并不常见。对于SR来说给定输入低分辨率图像，输出高分辨率图像，中间过程学习的是细节信息。其输入输出的分布基本一致，加入BN后，BN白化中间的特征的方式完全破坏了原始空间的表征，使得输入和输出的数据分布发生了变化，而在重建时，BN还要拿出一部分参数做这部分的恢复。有研究人员在训练中尝试加入BN，发现加入BN的网络loss的收敛速度明显比没有BN的要慢很多且loss的波动会非常大。因此，EDSR中去掉了BN，大约节省了40% 内存，大大加快了训练速度。

1.3 研究意义

目前，主流的视频压缩编码标准有AV1、HEVC等，利用视频图像在时域、空域上的冗余对视频图像进行有损压缩。经过有损压缩的视频图像必然存在失真，影响主客观质量，降低了用户的体验。因此，在保证视频、图片质量的前提下，针对视频传输速率和带宽需求，对通过后处理的手段提高视频图像质量提出了更高的要求和挑战。

主流编码标准之一的AV1采用的视频图像后处理方式是环路滤波，包括了Loop Filter、CDEF、Loop Restoration、帧超分辨率等。但由于其处理过程中的参数、映射关系是人为确定的，不一定最大限度的还原了原始图像。近几年来，深度学习逐渐运用于图像处理之中，展现出了巨大的发展潜力，不再依靠人工去挖掘图像特征，而利用神经网络基于大数据的学习能力，从大量数据中获得相关的非线性映射关系。目前这些内容已经应用到视频压缩、处理的方方面面。超分辨率处理技术，既可以解决成像过程中产生的噪声、模糊等，又可以提高视频图像的分辨率，具有相当的灵活性，适应性强，可用于医学影像上的超分辨率重建，低分辨率摄像头人脸或者外形的恢复等。

第二章 相关技术概述

2.1 开发环境

本节主要介绍本软件系统的开发环境以及运用的编程工具等。

2.1.1 开发工具简介

本系统采用Windows 10操作系统，包含Windows API的支持，使用面向对象的编程工具JetBrains PyCharm Community Edition 2019.2.5 x64和Notepad++ 7.4.2文本编辑器进行项目开发。前者作为 Python 编辑器，具有跨平台性，常用于深度学习开发；后者是一种代码编辑器，也是Windows 操作系统中的程序，用于文本编辑，其支持JavaScript、ES6等语言的编写。

2.1.2 环境配置

在运行过程中配置cuda、cudnn、pytorch等。主要环境如下：cupy-cuda90 6.2.0、numpy 1.16.4、torch 1.1.0、torchvision 0.3.0等。

2.2 相关技术

本节主要对于在模型开发过程中使用的相关技术进行阐述。

2.2.1 pytorch相关技术特点

本作品使用pytorch作为模型训练框架，Pytorch是torch的python版本，专门针对GPU加速的神经网络编程框架，为python用户提供了经典机器学习库 Torch 的端口。Pytorch在设计时，并没有像tensorflow一样提出了大量的新概念，而是使用了非常简洁的设计，它始终坚持高维数组、变量、神经网络模块三个由低到高的抽象层次，简洁的设计使其源码远远少于tensorflow，大约只有后者的十分之一，更少的新概念和更直观的设计使得pytorch的代码更容易编写与阅读。

2.2.2 多帧对齐相关技术特点

视频超分辨率处理首先要解决的就是多帧对齐问题，传统上我们使用光流算法来计算物体的移动。光流是空间运动物体在观察成像平面上的像素运动的瞬时速度，也就是指相邻帧之间物体的位移。一般我们假设运动前后亮度不变，对亮度表达式进行泰勒展开，可得到像素值在x轴和y轴方向上位移的一个约束方程，即其中亮度在x轴和y轴以及时间上的偏导可以由图像序列得到。由于我们有两个未知变量，故仍然需要一个约束方程，从不同的条件引入约束方程产生了各种不同类型的光流算法。此外，实践计算过程中，我们一般会采用一种金字塔结构，将帧从小到大，进行计算，每一层都可以使用下一层的计算结果来得知当前层大概的运动方向。

由此，在深度学习当中，产生了类似原理的用于多帧对齐融合的金字塔结构的模型。通过与Resnet性质相结合，采用金字塔模型对不同尺寸的feature map进行对齐，并逐层传递计算残差，最后得到对齐融合的特征残差。

2.2.3 Resnet相关技术特点

传统网络对输入输出进行映射，输出即为所求结果，网络对结果进行了全部的运算，计算机量较大。但在实际情况中，输出与输入存在部分相同的信息，如对图像进行超分处理，输入图像中有一个人，那么输出图像中一定有一个类似的人，这部分信息是重合的。如果采用传统方法，那么网络也需要对这个人的信息进行全部计算，复杂度大大增加。Resnet网络考虑到了客观存在的这一特性，提出了残差的概念，来降低复杂度。一般我们认为原始的低分辨率图片与高分辨率图片之间，低频分量几乎一样，缺失的是高频分量，即图片细节。所以我们只需要网络给出高分辨率图像与低分辨率图像相差的残差情况，而不必要直接给出结果。那么训练的时候，只需要针对高频分量进行训练就行了，还能加快训练速度。

2.3 图像评价标准

本节主要介绍图像的主观和客观两种方式下的评价方法

2.3.1主观评价方法

主观评价方法就是指用人的肉眼来观察并且评价图像质量，主观评价和人的感觉相一致，但是这种评价方式容易受到周围环境、人性格心境等因素的影响，而且采用这种方式来进行大量评价也不现实。近年来也有一些用于主观质量评价的模型出现，例如Netflix的VMAF视频质量评估工具。它是一种视频质量指标，将人类视觉建模与机器学习相结合，是一种模拟人类主观评价的客观评价方式。

目前，主观评价方式也在越来越受到研究人员的关注。因为提升图像质量的目的，最终的提升观看者的主观体验。人眼对图像的感知与客观上像素值对比有所不同，对某些质量的敏感度较高，如人在观察一张图像时对人脸细节的关注度会自然的高于其衣服纹理的细节。

2.3.2客观评价方式

我们主要采用的客观评价方式是PSNR，即Peak Signal to Noise Ratio的缩写，是一个表示信号最大可能功率和影响它的表示精度的破坏性噪声功率的比值的工程术语,其核心思想是对比原图像与处理后图像像素值的均方误差，通过判断像素值的差距来确定图像质量。另外，SSIM也是常用的图像客观评价方式之一，与PSNR不同它主要关注的是图像的结构相似度，人类的视觉系统在观看图像时习惯于关注其结构信息，相比之下更符合人眼对图像的质量评价。

第三章 针对AV1编码的超分辨率算法

本章主要介绍基于EDVR网络结构，减少网络参数，训练针对AV1编码后图像超分的模型。

3.1 EDVR算法介绍

EDVR算法主要贡献是创造性的提出了PCD对齐模块和TSA融合模块。

3.1.1 PCD对齐模块

PCD对齐模型主要是受到了TDAN算法的启发，TDAN使用可形变卷积来对齐相邻帧的特征图。但与TDAN不同，PCD对齐模型采用了金字塔结构，通过以粗到细的方式执行对齐以处理大而复杂的运动。他们首先将较低比例的特征与粗略估计对齐，然后将偏移和对齐特征传播到更高的比例，以便于精确的运动补偿，类似于光流估计中采用的概念。

PCD对齐模块采用从粗到细的处理提高了像素对齐的精确性，同时降低了时间复杂度。作者在消融实验部分阐述了PCD对齐模型的有效性，另外，PCD对齐模块不需要和光流法一样的额外的监督或者在其他任务上的预训练，而是随着整个网络进行端到端的训练的

3.1.2 TSA融合模块

因为相邻帧特征的重要度不同，参考价值不同，所以不能简单的将相邻帧的特征不加以区分统统用于当前帧中。EDVR算法使用了叫做TSA融合模块的算法，帮助融合多个对齐特征的信息，采用了一种称为相关系数的参数对每个位置处的每个相邻特征进行赋权，指示重建参考图像的信息量，将所有帧中被赋权的特征会卷积、融合到一起。

TSA融合模块主要解决两大问题，一个是由于运动、遮挡、模糊等因素导致不同帧中的相关信息可靠程度不同，不同帧能够提供的信息也各不相同，通过赋予不同相关系数的形式可以在一定程度上解决该问题。另一个是减少了PCD对齐模块的误差导致的错误，前一阶段的对齐并不一定是完全正确的，必然存在一定误差，TSA融合模块可以弥补这种对齐误差向下传递产生的错误。

3.2 训练集准备

训练集采用了REDS（REalistic and Dynamic Scenes），是由NTIRE19发布的全新数据集，提供了视频去模糊和超分辨率的逼真的动态的场景数据集，它为模型提出了两大难题：1.如何在有大幅度运动的情况下对齐多个帧2.如何有效的融合带有不同动作和模糊的不同帧。本次训练使用了其中约1万张图片进行模型训练。

训练集处理流程如下，首先使用FFmpeg及批处理命令将其提供的png格式图片转化为yuv格式，供AV1编码器使用，随后修改AV1工程代码，在其中插入保存图片的代码，否则其只输出编码后的码流文件。对于该训练集，使用了QP37的参数对这一万张图片进行编码，最后将其再转回png文件，得到最终的训练集。

3.3 训练改进

由于DCN训练的时候无法稳定的收敛，导致整个网络不好收敛，这是作者认识到的一个问题，也是这个模型存在的训练上的缺陷。作者自己实验时采用的方法是当出现难以收敛或者DCN偏移过大时，回退训练版本，重新训练。作者也尝试过从较小的模型训练大型模型，并对DCN使用较小的学习率,但仍然会出现问题。该模型的训练是一个较大的缺陷。

考虑到实验过程中硬件、训练时间等实际情况，训练中将其模型卷积核由128改为64，减少网络复杂度，并且减少DCN学习率。