利用 YOLO-v7 特征提升 VVC 编码视频中的目标检测准确率

1st TakahiroShindo早稻田大学基础科学与工学部日本东京 taka s0265@ruri.waseda.jp

2nd TaijuWatanabe早稻田大学基础科学与工学部日本东京 lvpurin @fuji.waseda.jp

3rd Kein Yamada 4th HiroshiWatanabe早稻田大学基础科学与工学部、早稻田大学基础科学与工学研究生院日本东京、日本东京 stslm738.ymd@toki.waseda.jp hiroshi.watanabe@waseda.jp

arXiv:2304.00689v1 [cs.CV] 3 Apr 2023

摘要—**随着基于深度学习的图像识别技术不断发展，人工智能自动视频分析正日益普及。随着用于图像识别的视频用量增加，亟需能够高效压缩此类视频数据的方法。通常，图像编码导致画质下降时，图像识别准确率也会降低。因此，本文提出一种基于神经网络的方法，通过对编码视频进行后处理来提高图像识别准确率，尤其是目标检测准确率。本文采用通用视频编码（VVC）作为视频压缩方法，因为它是目前最新、编码性能最佳的视频编码方法。该神经网络使用最新目标检测模型 YOLO-v7 的特征进行训练。通过采用 VVC 作为视频编码方法，并使用 YOLO-v7 作为检测模型，即使在低比特率下也能实现较高的目标检测准确率。实验结果表明，在目标检测准确率方面，所提方法与 VVC 结合后的编码性能优于常规 VVC。**

索引词—**VCM、视频编码、VVC、YOLO-v7、目标检测、后处理**

I. 引言

近年来，用于图像识别的视频数量迅速增加。消费级摄像机拍摄的大多数视频都会经过编码，用于人类视觉观看和图像识别。因此，人们需要为此用途开发新的视频编码方法并推动其标准化。动态图像专家组（MPEG）将这一研究领域命名为“面向机器的视频编码（VCM）”。MPEG 正从两个角度尝试推动 VCM 标准化：特征编码和视频编码。其中，基于视频编码进行标准化主要有两种方法。第一种是压缩用于图像识别的视频。图像识别所需的视频信息少于观看视频所需的信息 [1]。因此，专为图像识别设计编码方案时，可以实现更高的压缩率。不过，用于图像识别的压缩视频不包含重建可观看视频所需的足够信息，因此其应用范围有限。第二种方法是先压缩供观看的视频，再将其转换为适用于机器处理的形式。对于供观看的视频压缩，可以采用高效视频编码（HEVC）[2] 和 VVC [3]。使用这些视频编码方法后，再通过神经网络进行后处理。这种方法无法超越 VVC 等现有编码方法的压缩率。不过，通过一次视频编码即可同时实现供观看和图像识别的视频编码。本文旨在探讨如何通过后处理提高 VVC编码视频的图像识别准确率。

近年来，图像识别技术取得了显著进展。尤其是，目标检测的准确率大幅提升。然而，目前尚未研究采用最新目标检测模型的视频编码与多任务方法。因此，我们提出一种针对最新目标检测模型处理编码视频的方法。用于目标检测的模型是 YOLO-v7 [4]。YOLO [5] 是最受欢迎的目标检测模型之一，而 YOLO-v7 兼具较高的检测准确率和较快的检测速度。用于视频编码的模型是 VVC，它是最新的视频编码标准。我们使用卷积神经网络（CNN）处理 VVC 编码视频，以提高目标检测准确率。用于处理编码视频的所提神经网络如图 1 所示。在视频编码与多任务中，需要了解执行图像识别任务的神经网络的特性，并据此进行相应的视频压缩和处理。在我们的方法中，我们提取 YOLO-v7 的特征来训练用于处理编码视频的神经网络。该方法能够生成保留 YOLO-v7 目标检测所需信息的视频。实验表明，所提出的

这些研究成果来自日本国立信息通信技术研究所（NICT）委托的研究项目（编号：05101）。

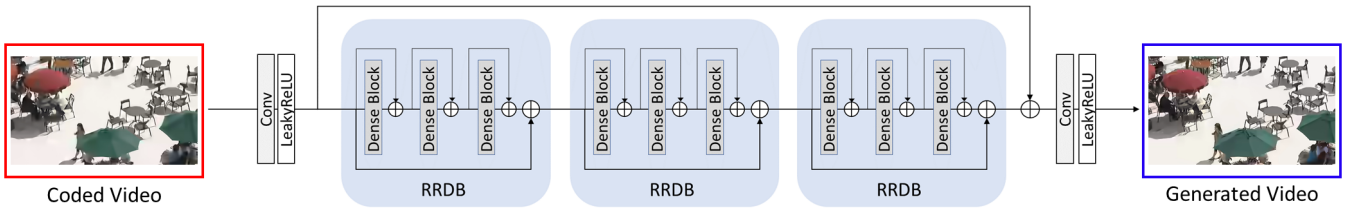

图 1. VVC 编码视频后处理的模型结构。

视频处理能够提高编码视频中的目标检测准确率。

本文其余部分安排如下：第 2 节和第 3 节分别介绍相关工作和所提出的方法。第 4 节讨论实验和结果，最后一节给出结论。

II. 相关工作

A. 面向人类视觉的视频编码

视频压缩技术对于在有限的通信资源下发送和接收视频至关重要。针对人类视觉的视频编码技术研发已开展多年，并制定了许多视频编码标准。其中，VVC 是最新的视频编码标准，其第一个版本于 2020 [3] 年 7 月完成。与以往的视频编码方法类似，它结合了帧内预测和帧间预测，并基于手工设计的算法运行。

另一方面，近年来，利用神经网络进行视频编码的研究也十分活跃。例如，已有使用循环神经网络的视频压缩模型 [6]、使用生成对抗网络的图像压缩模型 [7]，，以及帧间预测模型 [8]，[9]。此外，也有研究提出了用于帧内预测和帧间预测的基于卷积神经网络的视频编码方法[10]。而且，最新模型展现出的编码性能可与 VVC 相媲美 [11]。

许多研究致力于通过使用神经网络进行后处理 [12]–[15] 来提升编码视频的质量。在 VVC 和 HEVC 中，会执行基于块的运动补偿和变换。这种方法会使编码图像产生块噪声，从而降低图像质量。为减少块噪声，会使用去块滤波器。然而，并非总能去除噪声。因此，研究人员已对使用神经网络去除编码噪声的模型展开研究。

B. 面向机器的视频编码

随着图像识别准确率的提高，机器进行视频分析的机会也日益增多[16]。因此，面向图像识别的编码技术正受到关注，相关标准化进程也在加快。机器所需的视频信息被认为不同于人类观看图像时所需的信息 [1]。考虑到这一差异，捕捉图像识别所需的视频信息特征至关重要。因此，视频信息应根据视频的用途进行提取。

许多研究致力于提取目标检测所需的视频信息；目标检测是常见的图像识别任务之一。一种视频编码方法提取 YOLO9000 所需的信息 [17]。在该模型中，视频在编码前输入 YOLO9000，获得的特征会用于视频编码 [18]。这些特征代表 YOLO9000 重点关注的图像部分。编码模型会为图像的这部分分配更多比特，以便解码出有利于目标检测的图像。另一种方法则将视频编码模型与 R-CNN 集成 [19]。在这种编码模型中，使用 R-CNN 的检测结果来训练卷积神经网络[20]。与现有面向人类视觉的视频编码方法相比，这两种方法的目标检测准确率更高，视频压缩比也更高。能够压缩多任务所需视频信息的模型也不断涌现。MSFC [21] 是一种用于目标检测和分割的图像编码模型。用于图像压缩的神经网络通过目标检测和分割结果计算得到的损失进行训练。虽然其图像识别准确率不如 JPEG [22]，但编码效率显著高于 JPEG。

此外，目前也在研究将用于人类视觉的视频编码方法（如 VVC）应用于图像识别编码方法。在一些实验中，研究人员使用神经网络从视频中提取特征，并用 VVC 对其进行编码；这些实验表明，VVC 也能有效应用于视频编码与多任务（VCM）。

III. 提出的方法

A. 方法

我们的方法是通过神经网络处理编码视频来提高目标检测准确率。通过后处理，将面向人类视觉的编码视频转换为面向目标检测的视频。我们采用 VVC 作为视频编码方法，并使用 YOLO-v7 作为目标检测模型。所提出的视频处理方法基于神经网络，其结构遵循 ESRGAN 的生成器 [23]。所提出的模型使用从预训练 YOLO-v7 中提取的视频特征进行训练。使用所提出的方法处理 VVC 编码视频，可以在低比特率下实现较高的目标检测准确率。

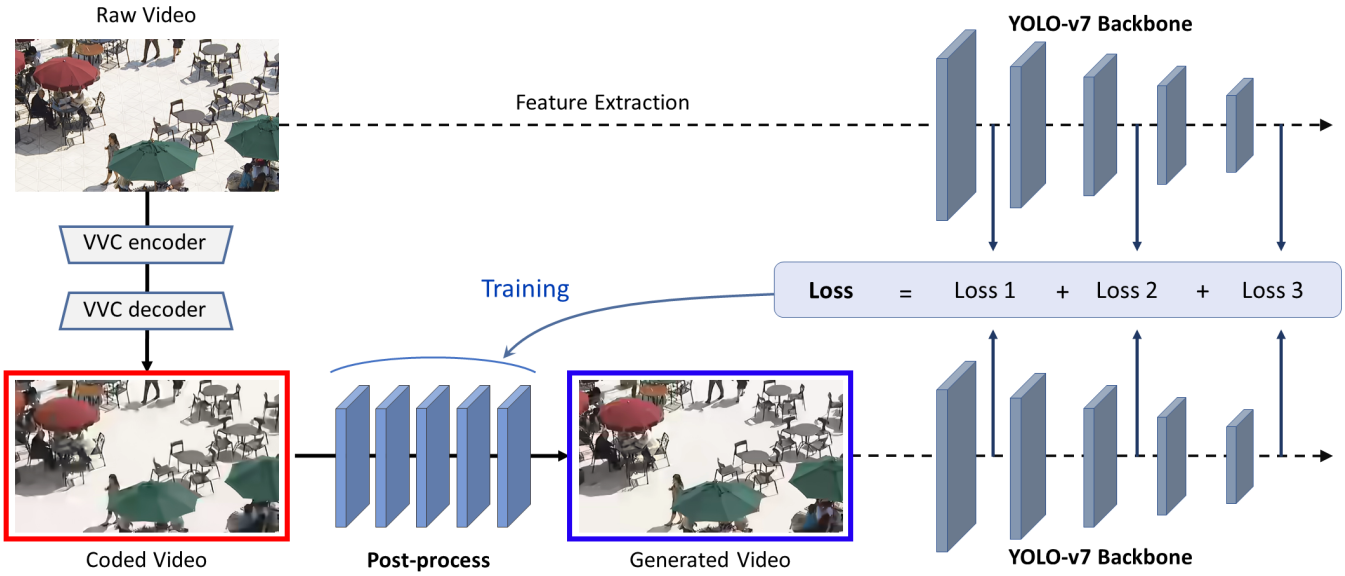

图 2. 所提出神经网络的训练过程。

B. 模型结构

一些用于编码视频后处理的神经网络结构基于 ResNet [24]，，另一些则基于 U-Net [25]。这些模型结构用途广泛，已应用于许多执行图像识别任务的模型中。ResNet 是一种基于残差块的模型，用于 SRGAN [26] 的生成器中。本文提出的网络基于 ESRGAN 生成器中使用的残差中的残差稠密块（RRDB）。与 SRGAN 一样，ESRGAN 是一种图像超分辨率模型。所提出的后处理模型如图 1 所示。与残差块相比，使用 RRDB 能让神经网络在生成图像中再现更精细的纹理。RRDB 的这一特性在编码视频处理中也很有用，因为它有助于重建图像细节。我们的模型由三个 RRDB、两个卷积层和两个激活函数组成。

C. 损失函数

图像识别任务的结果通常用于训练对视频进行压缩以供图像识别的神经网络。例如，在 S. Wang 等人[20] 和 Z. Zhang 等人 [21] 的研究中，使用 R-CNN 的目标检测结果训练图像压缩模型，以提高 R-CNN 的目标检测准确率。在我们的研究中，我们采用 YOLO 的最新版本 YOLO-v7 作为目标检测模型。为了提高 YOLO-v7 在编码视频中的目标检测准确率，使用预训练 YOLOv7 的特征处理这些视频是一种有效的方法。我们使用基于卷积神经网络的神经网络处理编码视频，该网络通过 YOLO-v7特征进行训练。在此次训练中，我们从训练模型的骨干网络中提取 YOLO-v7 的三种特征。计算损失时，我们采用原始视频的特征与所提出模型输出视频的特征之间的均方误差（MSE），训练过程的原始视频特征与所提出模型输出视频的特征。所提出模型的训练过程如图 2 所示。用于训练的损失函数表示为

$$Loss = MSE(yolo(I_{raw}), yolo(I_{output})), \quad (1)$$

其中，$yolo$ 表示 YOLO-v7 的特征提取器。I_{raw}和 I_{output} 分别表示原始视频和输出视频。

IV. 实验

A. 训练细节

本节介绍用于处理编码视频的所提神经网络的训练过程。训练所用的数据集为上海交通大学 [27], UVG [28], 和 MCML-4K- UHD [29]。这些数据集均包含一些 4K（3840x2160）原始视频序列。我们从这些数据集中选取 30 个序列，并使用 VTM10.0 [30] 对其进行编码。帧参考方法的配置为“随机访问”，量化参数（QP）为 27、32、37、42 和 47。以此方式生成的 150 种编码视频用作所提神经网络的输入视频。损失函数如式（1）所示，并使用从 YOLO-v7 训练模型中获得的特征。优化函数为 Adam，学习率为 1e-5。训练神经网络所用的图形处理器为 GeForce GTX 1080ti。

B. 评估方法

我们使用 YOLO-v7 对 VVC 编码视频以及所提出的神经网络的输出视频进行目标检测。为了衡量目标检测准确率，我们需要带有目标标注的视频序列。因此，我们使用 SFU-HW-Objects-v1 数据集 [31] 进行评估。该数据集是少数几个为原始视频提供目标标注的数据集之一。它包含 18 个原始视频序列的目标标注。这些序列根据图像尺寸和

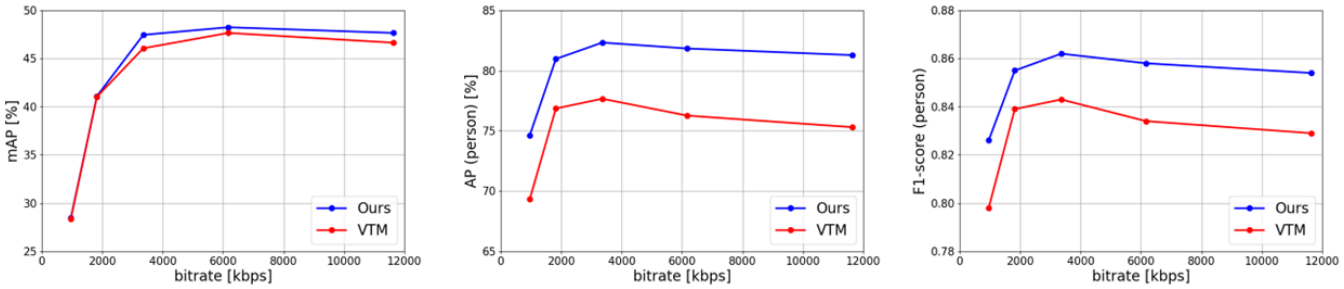

图 3. PeopleOnStreet序列的评估结果：(a) rate-mAP；(b) rate-AP（行人）；(c) rate-F1（行人）。

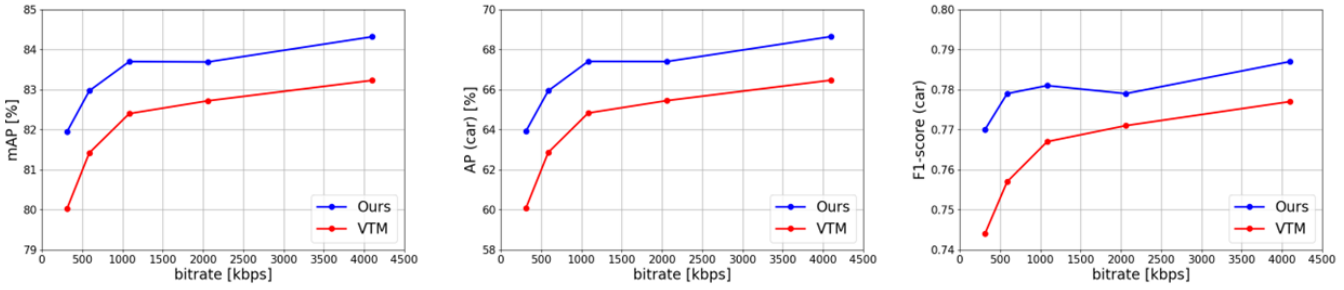

图 4. Traffic序列的评估结果：(a) rate-mAP；(b) rate-AP（汽车）；(c) rate-F1（汽车）。

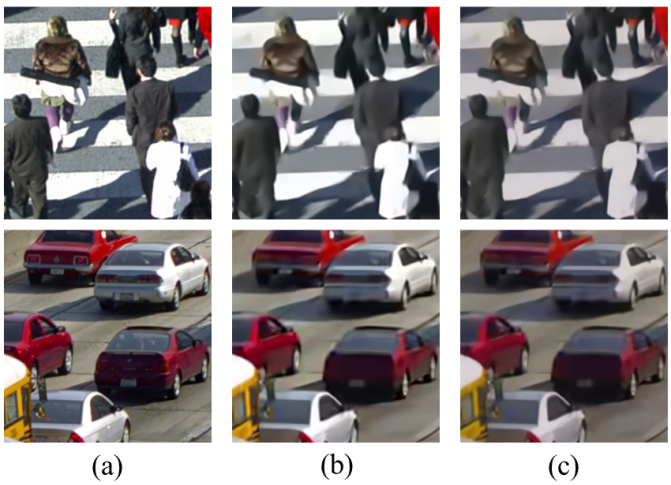

图 5. 帧图像比较（240x270）。(a) 原始帧；(b) 使用 VVC（量化参数：47）编码的输入帧；(c) 所提方法的输出帧。上方图像是PeopleOnStreet序列的一部分。下方图像是 Traffic序列的一部分。

视频特性分为五类。由于训练中使用的视频序列的图像尺寸为 4K，我们选用图像尺寸大于其他类别的 A类、B类和 C类序列。测试序列的详情见表 I。这些序列也使用 VTM 10.0 编码。帧参考方法配置为“随机访问”。两个 A类序列采用五个量化参数（27、32、37、42、47）进行编码，其他序列采用三个量化参数（37、42、47）进行编码。我们采用

表 I 测试序列列表。

类别	序列名称	size	帧编号
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

将所提出的后处理方法应用于编码视频，并比较所提出的处理前后的目标检测准确率。目标检测模型采用预训练的 YOLO-v7，置信度阈值设为 0.25。评估指标采用平均精度（AP）和 F1分数。计算 AP 时，交并比（IoU）阈值始终设为 0.5。

C. 结果

PeopleOnStreet序列中的目标检测结果如图 3 所示。由于“人”占 PeopleOnStreet序列中所有目标的 97%，图 3 也展示了人员检测结果。三幅图均显示了检测准确率与比特率之间的关系。图 (a) 表示平均精度均值（mAP），图 (b) 表示人员检测的 AP，图 (c) 表示人员检测的 F1分数。在该序列中，所提出的方法略微提高了

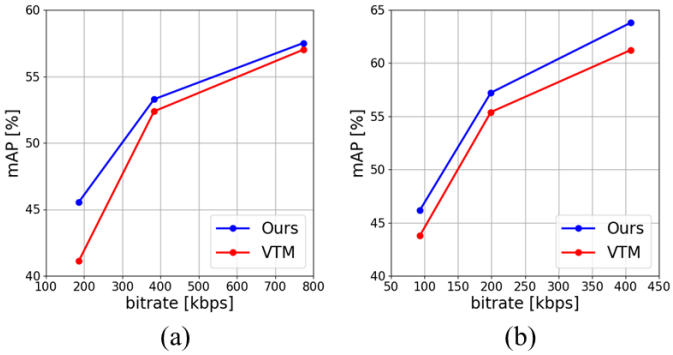

图 6. (a) B类序列的 rate-mAP 评估结果；(b) C类序列的 rate-mAP 评估结果。

mAP 值，并显著提高了人员检测的准确率。无论量化参数如何，人员检测的 AP 值均提高了约 5 个百分点，F1分数提高了约 0.02 至 0.03。所提出的方法对PeopleOnStreet序列帧图像的变换如图 5 所示。图像 (b) 为 VVC 编码视频的帧图像，(c) 为所提出的神经网络的输出视频帧图像。该图显示，所提出的后处理使帧图像的色调略有变化。

Traffic序列的目标检测结果如图 4 所示。由于Traffic序列中“汽车”所占对象比例为 99%，图 4 也展示了“汽车”的检测结果。所有这些图还展示了检测准确率与比特率之间的关系。图 (a) 表示 mAP 值。图 (b) 和 (c) 分别表示汽车检测的 AP 值和 F1分数。在 Traffic序列中，所提出的方法提升了 mAP 值，并使汽车检测的 AP 提高了 2 至 3 个百分点，F1分数提高了约 0.01 至 0.02。图 5 展示了使用所提出的方法对 Traffic序列帧图像进行变换的结果。与 PeopleOnStreet序列的情况类似，在 Traffic序列中，我们可以确认所提出的图像处理导致帧图像的色调发生了变化。

此外，图 6 展示了 B类中的三个序列和 C类中的四个序列的目标检测准确率。图 (a) 显示了 B类三个序列的 mAP 与比特率之间的关系，图 (b) 显示了 C类四个序列的 mAP 与比特率之间的关系。在这两种情况下，所有量化参数下的目标检测准确率均有所提高。

表 II 总结了这些结果。对于表 I 中列出的序列，表中展示了采用 VVC 编码时以及对这些序列进行所提出的后处理时的目标检测准确率（mAP）。表中还以差距的形式展示了 mAP 值的提升幅度。对于两个 A 类序列，与人和汽车的平均精度值相比，所提出的方法对 mAP 值的提升并不明显。训练数据集不足是

表 IImAP 值结果。

类别	方法	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	-	-	57.01	52.37	41.12
	Ours	-	-	57.51	53.28	45.55
	gap	-	-	+0.50	+0.91	+4.43
C	VTM	-	-	61.23	55.39	43.77
	Ours	-	-	63.80	57.12	46.16
	gap	-	-	+2.57	+1.82	+2.39

造成这一结果的原因之一。训练数据集中的序列包含许多人和汽车。然而，尽管雨伞、运动球和椅子等某些物体出现在测试序列中，训练数据集中却不包含这些物体。尽管如此，表 II 显示，所有类别在所有量化参数下的 mAP 均有所提升。mAP 值提升的原因是人、汽车及其他一些物体的平均精度值有所提高。从这些结果可以看出，我们提出的方法能够有效提高 VVC 编码视频的目标检测准确率。

V. 结论

我们提出了一种方法，可提高 YOLO-v7 在 VVC编码视频中的目标检测准确率。实验表明，利用 YOLO-v7特征进行所提出的后处理，可以提高 VVC编码视频中的目标检测准确率。将所提出的方法应用于为人类视觉编码的视频后，可将其转换为适用于 YOLO-v7 的视频。转换后的视频与原始编码视频具有不同的色调，从而有利于 YOLO-v7 进行目标检测。此外，根据方法的使用方式，视频可以适应不同用途，既可用于人类视觉，也可用于目标检测。在这两种情况下，使用 VVC 作为视频压缩方法都能实现较高的压缩率；使用预训练 YOLOv7 作为目标检测模型，则能获得较高的目标检测准确率。未来的工作将探索更有效的方法，以提取图像识别所需的视频信息，进一步提升性能。

参考文献

[1] H. Choi 和 I.V.Bajic，《面向人类和机器的可扩展图像编码》，IEEE图像处理汇刊，第31卷，2022年，2739-2754页。[2] 高效视频编码，标准 ISO/IEC 23008-2，ISO/IEC JTC 1，2013年4月。[3] 通用视频编码，标准 ISO/IEC 23090-3，ISO/IEC JTC 1，2020年7月。[4] C. Y. Wang、A. Bochkovskiy 和 M. Liao，《YOLOv7：可训练的免费赠品组合为实时目标检测器树立新标杆》，arXiv 预印本 arXiv:2207.02696，2022年。[5] J. Redmon、S. Divvala、R. Girshick 和 A. Farhadi，《只需看一次：统一的实时目标检测》，arXiv 预印本 arXiv:1506 02640，2015年。。[6] C. Y. Wu、N. Singhal 和 P. Krahenbuhl，《通过图像插值进行视频压缩》，载于欧洲计算机视觉会议（ECCV）论文集，2018年，425-440页。

[7] O. Ripple 和 L. D. Bourdev，《实时自适应图像压缩》，载于国际机器学习会议（ICML），第70卷，2017年8月，2922-2930页。[8] J.Liu、S. Xia 和 W. Yang，《采用多域层级约束的帧间预测深度参考帧生成》，IEEE 多媒体汇刊，第22卷，2020年10月，2497-2510页。[9] G. Lu、W. Ouyang、D. Xu、X. Zhang、C. Cai 和 Z. Gao，《DVC：端到端深度视频压缩框架》，2019年 IEEE计算机视觉与模式识别会议（CVPR），2019年，10998-11007页。[10] J. Li、B. Li 和 Y. Lu，《深度上下文视频压缩》，载于神经信息处理系统会议（NeurIPS），第34卷，2021年，18114-18125页。[11] J. Li、B. Li 和 Y. Lu，《神经视频压缩的混合时空熵建模》，载于 ACM 多媒体国际会议，2022年10月，1503-1511页。[12] G. Lu、W. Ouyang、D. Xu、X. Zhang、Z. Gao 和 M. T. Sun，《用于视频压缩伪影消除的深度卡尔曼滤波网络》，载于欧洲计算机视觉会议（ECCV）论文集，2018年，568-584页。[13] X. Meng、X. Deng、S. Zhu、S. Liu、C. Wang、C. Chen 和 B. Zeng，《基于联合时空先验的视频编码鲁棒质量增强方法》，IEEE 视频技术电路与系统汇刊，第31卷，第6期，2021年6月，2401-2414页。[14] Q. Xing、Z. Guan、M. Xu、R. Yang、T. Liu 和 Z. Wang，《MFQE 2.0：一种面向压缩视频的多帧质量增强新方法》，IEEE模式分析与机器智能汇刊，第43卷，2021年3月，949-963页。[15] F. Zhang、C. Feng 和 D. R. Bull，《通过基于CNN的后处理增强VVC》，2020年 IEEE多媒体与博览会国际会议（ICME），2020年，1-6页。[16] L. Duan、J. Liu、W. Yang、T. Huang 和 W. Gao，《面向机器的视频编码：协同压缩与智能分析范式》，IEEE图像处理汇刊，第29卷，2020年，8680-8695页。[17] J. Redmon 和 A. Farhadi，《YOLO9000：更好、更快、更强》，2017年 IEEE计算机视觉与模式识别会议（CVPR），2017年，6517-6525页。[18] H. Choi 和 I. V. Bajic，《面向目标检测的高效压缩》，载于 IEEE声学、语音与信号处理国际会议（ICASSP），2018年，1792-1796页。[19] R. Girshick、J. Donahue、T. Darrell 和 J. Malik，《用于准确目标检测和语义分割的丰富特征层次结构》，2014年 IEEE计算机视觉与模式识别会议（CVPR），2014年，580-587页。[20] S. Wang、Z. Wang、S. Wang 和 Y. Ye，《面向机器视觉的端到端压缩：网络架构设计与优化》，IEEE开放电路与系统期刊，第2卷，2021年，675-685页。[21] Z. Zhang、M. Wang、M. Ma、J. Li 和 X. Fan，《MSFC：多任务网络中的深度特征压缩》，2021年 IEEE多媒体与博览会国际会议（ICME），2021年，1-6页。[22] G. K. Wallace，《JPEG静态图像压缩标准》，IEEE消费电子汇刊，第38卷，第1期，xviii-xxxiv页，1992年2月。[23] X. Wang、K. Yu、S. Wu、J. Gu、Y. Liu、C. Dong、Y. Qiao 和 C. C. Loy，《ESRGAN：增强型超分辨率生成对抗网络》，载于欧洲计算机视觉会议研讨会论文集，2018年，0-0页。[24] K. He、X. Zhang、S. Ren 和 J. Sun，《用于图像识别的深度残差学习》，2016年 IEEE计算机视觉与模式识别会议（CVPR），2016年，770-778页。[25] O. Ronneberger、P. Fischer 和 T. Brox，《U-Net：用于生物医学图像分割的卷积网络》，载于医学图像计算与计算机辅助干预国际会议（MICCAI），2015年，234-241页。[26] C. Ledig、L. Theis、F. Huszar、J. Caballero、A. Cunningham、A. Acosta、A. Aitken、A. Tejani、J. Totz、Z. Wang 和 W. Shi，《使用生成对抗网络实现照片级真实感的单图像超分辨率》，2017年 IEEE计算机视觉与模式识别会议（CVPR），2017年，4681-4690页。[27] L. Song、X. Tang、W. Zhang、X. Yang 和 P. Xia，《上海交通大学4K视频序列数据集》，载于多媒体体验质量国际会议，2013年，34-35页。

[28] A. Mercat、M. Viitanen 和 J. Vanne，“UVG 数据集：用于视频编解码器分析与开发的 50/120fps 4K 序列”，载于 ACM 多媒体国际会议，2020年，第 297-302 页。[29] M. Cheon 和 J. S. Lee，“面向沉浸式体验的压缩 4K 超高清视频主观与客观质量评估”，IEEE 视频技术电路与系统汇刊，第 28 卷，2018年，第 1467-1480 页。[30] S. K. J. Chen 和 Y. Ye，《通用视频编码及测试模型 10（VTM 10）的算法描述》。JVETS2002，2020年。[31] H. Choi、E. Hosseini、S. R. Alvar、R. A. Cohen 和 I. V. Bajic，“原始视频序列中标注对象的数据集”。Data in Brief，34，106701，2021年。