

Pruned Lightweight Encoders for Computer Vision

Jakub Žádník, Markku Mäkitalo, Pekka Jääskeläinen
Faculty of Information Technology and Communication Sciences
Tampere University, Finland

{jakub.zadnik, markku.makitalo, pekka.jaaskelainen}@tuni.fi

Abstract—Latency-critical computer vision systems, such as autonomous driving or drone control, require fast image or video compression when offloading neural network inference to a remote computer. To ensure low latency on a near-sensor edge device, we propose the use of lightweight encoders with constant bitrate and pruned encoding configurations, namely, ASTC and JPEG XS. Pruning introduces significant distortion which we show can be recovered by retraining the neural network with compressed data after decompression. Such an approach does not modify the network architecture or require coding format modifications. By retraining with compressed datasets, we reduced the classification accuracy and segmentation mean intersection over union (mIoU) degradation due to ASTC compression to 4.9–5.0 percentage points (pp) and 4.4–4.0 pp, respectively. With the same method, the mIoU lost due to JPEG XS compression at the main profile was restored to 2.7–2.3 pp. In terms of encoding speed, our ASTC encoder implementation is 2.3x faster than JPEG. Even though the JPEG XS reference encoder requires optimizations to reach low latency, we showed that disabling significance flag coding saves 22–23% of encoding time at the cost of 0.4–0.3 mIoU after retraining.

Index Terms—Image Compression, Computer Vision, Texture Compression, Low Latency, JPEG XS, ASTC

I. INTRODUCTION

Automated systems that make fast decisions based on visual input, such as autonomous driving, drone control, or smart factories, rely on a very short response time to prevent damage or injury. Low-latency network transmission enabled by recent development in networking technologies, such as 5G, allows edge devices with low computing power to offload expensive deep neural network (DNN) inference of the vision task to a nearby server. Figure 1 depicts a model scenario of an obstacle suddenly emerging in a trajectory of a self-driving car. Considering the car’s speed of 100 km/h, if the end-to-end latency of the brake control system increased by 40 ms, for example, due to slow compression, the car would travel an additional 1.1 meters, potentially hitting the obstacle instead of stopping in front of it.

Fast compression of source images is necessary to ensure low latencies over a transfer channel, and one way to decrease the latency is to reduce the codec complexity. Removing coding features, however, typically results in decreased vision performance at the same bitrate. Since DNNs have the ability to learn from the input data, it is possible to retrain them on the compressed dataset (after decompressing it) to overcome the coding efficiency lost by pruning the coding features. At the same time, pruning the existing codecs allows to reuse

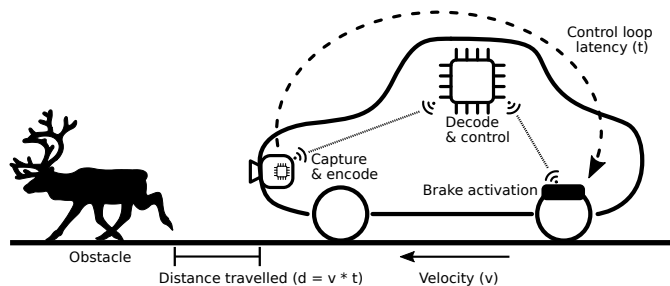


Fig. 1. The reaction time of an autonomous vehicle control system determines whether it can avoid hitting an obstacle.

existing hardware support and retraining does not modify the neural network architecture since only the weights change.

Several types of low-complexity codecs exist. Real-time texture compression can reach very high encoding speeds compared to other methods at the expense of rather low coding efficiency [1]–[3]. A “mezzanine compression” family of codecs is designed specifically to meet ultra-low latency requirements, with JPEG XS [4] as the newest standard in this family. The recently standardized high throughput JPEG 2000 (HTJ2K) [5] simplifies the otherwise complex JPEG 2000 [6] with the goal of 10x throughput improvement. However, it does not offer such a precise rate allocation as JPEG XS. Traditional hybrid video codecs, such as high efficiency video coding (HEVC) [7] or versatile video coding (VVC) [8] offer advanced coding features and great coding efficiency. However, the additional complexity of, for example, the inter or intra prediction and advanced entropy coding can be prohibitive in resource-constrained devices. Joint photographic experts group (JPEG) [9] shares the core transform coding features with hybrid video codecs, but without the additional complexity. At the same time, it can deliver sufficient quality for computer vision applications, as shown in the paper.

In this paper, we explore the idea of pruning the encoding configurations to reduce the encoding time and latency and compensating the lost vision performance by retraining the vision model with the compressed dataset. As two case studies we chose reducing the configuration space of the otherwise very complex adaptive scalable texture compression (ASTC) [10] format and JPEG XS that was designed specifically as a lightweight, low-latency codec. Both codecs operate in a constant bitrate mode which is important for ensuring predictable latency.

We evaluate the effect of ASTC compression artifacts on

the image classification accuracy of ShuffleNet [11] V2 and both ASTC and JPEG XS on the semantic segmentation of LR-ASPP-MobileNetV3 [12]. We compare both to JPEG, and in the case of the segmentation task also to HTJ2K and JPEG 2000.

The contributions of this paper are:

- We propose a lightweight ASTC encoder¹ that is approximately $2.3\times$ faster than JPEG on a Samsung S10 smartphone.
- We study how pruning JPEG XS encoding configurations impacts latency and computer vision performance.
- We demonstrate that the quality vs. latency tradeoff can be alleviated by retraining the classification and segmentation models with the compressed datasets.

II. BACKGROUND AND RELATED WORK

A. Adaptive Scalable Texture Compression

ASTC is the newest and most flexible texture compression format adopted as an OpenGL extension by the Khronos Group. Like other texture compression formats, it quantizes the input block’s color space and represents its pixels as indices pointing at one of the quantized colors. Compared to the older BCn formats, ASTC supports many configuration options: scaling the input block size, partitioning, different color endpoint modes (CEM), endpoint and weight quantizations, dual-plane encoding, and bounded integer sequence encoding (BISE). An important property of texture compression is a fixed compression ratio and random access: The individual pixels are randomly addressable from the compressed representation without decompressing.

Modern graphics processing units (GPUs) have texture fetch units that can perform the decompression online during rendering with a negligible overhead which further enhances the low-latency potential of texture compression.

B. JPEG XS

JPEG XS is a wavelet-based mezzanine codec designed primarily for low complexity, low latency, high bandwidth, and high-quality video delivery. The minimal coding unit of JPEG XS is one precinct whose size can range from less than one pixel line up to several lines of the image.

The JPEG XS rate allocation can predict the bitrate precisely, unlike HTJ2K where a precise rate allocation would require a significant additional complexity [4]. The latest version of the standard also supports direct Bayer data compression [13] which can be used to bypass the traditional image processing pipeline at the sensor side and thus save latency.

To the best of our knowledge, no publicly available JPEG XS encoder currently exists. Therefore, we utilized the reference JPEG XS reference software, version 1.4.0.4 (ISO/IEC 21122-5:2020). In the literature, [14] developed a JPEG XS codec capable of running at 60 frames per second (FPS) at 8K resolution on a 64-core AMD EPYC processor.

C. Compression for Computer Vision

Some previous works optimize the perceptual model of JPEG for computer vision [15], [16], leading to significant quality improvements. [17] optimized the global JPEG XS encoding parameters (gains and priorities) to better capture the characteristics of a computer vision target. Our approach of retraining the vision model with the compressed dataset is complementary to codec parameter optimizations.

[18] used retraining to recover object detection and semantic segmentation performance of BC1 and YCoCg-BC3 [1] texture compression. To the best of our knowledge, no prior work implements a minimal-subset ASTC in the context of computer vision.

[19] proposed a modified loss function of a DNN to achieve more efficient restoration of classification accuracy lost to compression artifacts. They achieved a minor but consistent gain of up to 0.79 percentage points (pp) validation accuracy compared to a simple retraining method used in this work.

The recent exploration of video coding for machines (VCM) by moving picture experts group (MPEG) is an effort to develop a coding scheme with both machine and human perception in mind [20]. The current development is being built on top of VVC which is a more complex format than what we target in this paper. Furthermore, our use case considers only the computer vision performance without the human in the loop. JPEG AI [21] also explores compression for both human and computer vision targets, but focuses on utilizing learning-based coding methods.

Yet another approach to adapting compression for computer vision is “feature compression” which encodes intermediate neural features [22]. Feature compression, however, requires computing some of the convolutional layers on the encoding device which contrasts with our approach of decreasing the encoding complexity.

III. IMPLEMENTATION OF PRUNED CODECS

A. ASTC

Due to the ASTC complexity, exhaustively searching for encoding parameters is not feasible in real time, and such, heuristics must be used to prune the configuration space. In our work, we reduce the configuration space to only one configuration: 5-bit color endpoint and 2-bit and weight quantization with a weight grid of 8×5 . The selected configuration showed the lowest per-pixel distortion measured as peak signal-to-noise ratio (PSNR) on a sample dataset from adjacent configurations without requiring BISE.

Since the only way to scale the ASTC bitrate is to modify the input block size, we implemented both 12×12 and 8×8 input block sizes, implying a compression ratio (CR) of 27:1 and 12:1 ($0.\bar{8}$ and 2.0 bits per pixel (bpp)), respectively.

The encoding of a block starts by selecting the endpoints with a small inset similar to [1]. Then, “ideal weights” are selected by orthogonally projecting the input pixels onto the line defined by the endpoints. Lastly, the “ideal weights” are bilinearly downsampled to the 8×5 grid and quantized into two bits.

¹github.com/cpc/simple-texcomp

B. JPEG XS

The long encoding time of the reference encoder is caused by the rate allocation algorithm exhaustively computing the bit budget for each precinct at all quantization levels and using all possible coding methods. We reduced the number of searched quantizations and coding methods to 13 and 5, respectively, without losing quality as the other combinations were unused in our tests.

To reduce the number of rate allocation passes further, we disabled the significance flag coding. Significance flag coding detects a run of all-zero “significance groups” (groups of 8 adjacent coefficients) that can be encoded with a single flag and requires an additional rate allocation pass. Disabling this method brings the number of utilized coding methods from 5 to 3 and removes the need for “refresh” passes, significantly reducing the encoding time. However, it also reduces the coding efficiency, which we try to recover by retraining with the compressed dataset. We kept the coefficient prediction from a previous line, as disabling it would prevent the encoder meet the target bitrate.

IV. EXPERIMENTAL SETUP

A. Implementation Details

The ASTC encoder uses ARM NEON intrinsics to vectorize the most significant loops using 8-bit fixed-point representation. For a fair runtime comparison with JPEG, we chose the single instruction multiple data (SIMD)-optimized `libjpeg-turbo` library² and developed a wrapper encoder application around the library. The JPEG coding parameters were chosen to match the defaults of the `cjpeg` command-line utility: YCbCr color space with 4:2:0 subsampling and no restart intervals. The quality parameter (Q) 45 was selected so that the bitrate of a random sample of 10000 ImageNet images after encoding is the highest possible at or below the rate of ASTC 12 × 12. Both ASTC and JPEG were evaluated on a single core (A76) of a Samsung S10 smartphone.

The JPEG XS encoder was compiled only for the x86 instruction set and evaluated on a single thread of Intel i7-8650U laptop CPU at a base frequency of 2.1 GHz with disabled frequency scaling. For runtime comparison we chose two open source encoders: `grok`³ for JPEG 2000 and `OpenJPH`⁴ for HTJ2K, both using irreversible discrete wavelet transform (DWT).

B. Vision Tasks

We evaluated the image classification accuracy of ShuffleNet V2⁵ in 0.5× and 1.0× sizes trained on the ImageNet dataset [23] with training hyperparameters derived from [24] and [11]. We also evaluated a semantic segmentation task with LR-ASPP-MobileNetV3⁶ in both large and small versions trained on the Cityscapes dataset [25]. The Cityscapes images

for training were cropped to $\sqrt{2}$ of the original size in each dimension to avoid running out of GPU memory. Both networks were retrained with the dataset compressed with ASTC to recover the vision performance lost by compression artifacts. Unfortunately, the JPEG XS encoder was not able to encode some of the ImageNet images at the bitrate of 0.8 bpp. Therefore, we evaluated only the segmentation task with this codec.

All encoders mentioned in the previous subsection were used for quality evaluations, along with `astcenc`⁷ at the fastest profile for quality evaluations on the Cityscapes dataset.

JPEG-compressed images were used only for retraining the ShuffleNet V2 network. JPEG, JPEG 2000, and HTJ2K reach segmentation mean intersection over union (mIoU) within 1.5% below the uncompressed result, and retraining is expected to bring the results on par with the uncompressed results, therefore, we did not retrain with these codecs.

V. RESULTS

A. Quality

Image Classification: Table I summarizes the highest achieved classification accuracies of ShuffleNet V2 under different conditions. When the compressed data is used as an input to the network trained on uncompressed data (the “orig.” column), the ASTC compression degrades the accuracy by more than 15 pp, while the difference caused by JPEG compression is only 1.3 and 0.6 pp. However, when retrained with the compressed dataset (the “retr.” column), the accuracy decrease for ASTC with 12 × 12 block size is only 4.9 and 5.0 pp for the 0.5× and 1.0× network sizes, respectively. The ASTC 8 × 8 achieves higher quality than 12 × 12: only 2.3–1.8 pp accuracy decrease compared to the uncompressed result. Retraining with the JPEG-compressed dataset brings a 0.2 pp increase in the classification accuracy of the smaller network. The results show that retraining the larger network does not improve the already high accuracy for JPEG.

TABLE I
VALIDATION TOP-1 ACCURACY OF SHUFFLENET V2 ON IMAGENET
VALIDATION SET WITH JPEG AND THE PROPOSED ASTC COMPRESSION
WITH AND WITHOUT RETRAINING.

compression	bpp	0.5x		1.0x	
		orig.	retr.	orig.	retr.
uncompressed	24.0	54.4%		64.3%	
ASTC 12x12	0.89	-16.8	-4.9	-15.1	-5.0
JPEG Q45	~0.89	-1.3	-1.1	-0.6	-0.7
ASTC 8x8	2.00	-6.4	-2.3	-6.6	-1.8

Semantic Segmentation: Figure 2 compares rate-distortion curves of multiple encoders according to three metrics: PSNR, structural similarity index (SSIM), and validation mIoU of LR-ASPP-MobileNetV3 (large vesion) trained on an uncompressed dataset. The small version of the model shows similar relations between the mIoU curves to the large version, therefore, we omitted it for brevity. The plots show that despite

²libjpeg-turbo.org (version 2.1.1)

³github.com/GrokImageCompression/grok (version 9.7.7)

⁴github.com/aous72/OpenJPH (version 0.9.0)

⁵pytorch.org/hub/pytorch_vision_shuffle_net_v2

⁶github.com/ekzhang/fastseg (commit 91238cd)

⁷github.com/ARM-software/astc-encoder (version 3.7)

significant PSNR and SSIM differences between JPEG 2000, JPEG, and HTJ2K, the mIoU difference is relatively small. Disabling significance flag coding of JPEG XS (denoted as “no-sf”) shows a consistent decrease of all metrics in both main and subline profiles. Similarly, when compared to a full-featured `astcenc` encoder at the fastest preset, our pruned implementation achieves significantly lower quality. Both SSIM and mIoU metrics decrease rapidly with JPEG XS at lower bitrates (0.8 and 1.0 bpp), especially the subline profile.

To recover the large mIoU degradation of ASTC and JPEG XS at low bitrates, we retrained the network with the compressed datasets. Table II summarizes the mIoU improvements of retraining LR-ASPP-MobileNetV3 in both small and large variants. For ASTC 12×12 , retraining brought an improvement of 1.1 and 8.6 pp for the small and large networks, respectively. Retraining JPEG XS in the main profile resulted in mIoU around 2.3–2.6 pp lower than the uncompressed result. The subline profile of JPEG XS shows a sharp decline in the vision performance without retraining. Retraining allows recovering most of the quality back. However, the results still do not reach the quality of ASTC 12×12 .

TABLE II
MEAN INTERSECTION OVER UNION (mIoU) OF LR-ASPP-MOBILENETV3 WITH CITYSCAPES VALIDATION SET AND THE JPEG XS COMPRESSION.

compression	bpp	small		large	
		orig.	retr.	orig.	retr.
uncompressed	24.0	61.2%		66.5%	
JPEG XS (main, sf)	0.89	-4.9	-2.3	-6.5	-2.6
JPEG XS (main, no-sf)	0.89	-5.8	-2.7	-7.4	-2.3
JPEG XS (subline, sf)	0.89	-14.4	-5.4	-20.5	-5.3
JPEG XS (subline, no-sf)	0.89	-17.3	-6.7	-25.2	-6.8
ASTC 12×12	0.89	-5.5	-4.4	-12.6	-4.0
ASTC 8×8	2.00	-3.5	-2.8	-6.4	-1.7

Figure 3 shows segmentations of two challenging scenes after the retrained model inference using compressed images to illustrate the effect of compression on the segmentation result. The first image shows shape deformations caused by the ASTC and JPEG XS main profile (“p3”) while the network trained with JPEG XS subline profile (“p5”) fails to detect the people at all. In the second image, all cases detect the person in the foreground but fail to detect some, or all, the people in the distance. It should be emphasized that the MobileNet networks were trained with images containing approximately half of the pixels of the full-resolution images due to GPU memory limitations. Therefore, the examples do not correspond to the best predictions achievable with these networks.

B. Runtime

ASTC Encoder: Table III compares the average encoding time of a Cityscapes image (resolution 2048×1024) of our ASTC encoder with the block size 12×12 and 8×8 to JPEG with the quality parameters Q 85 and 96. The Q parameters were determined by the same procedure as in Subsection IV-A to ensure approximately the same bitrate as ASTC on a random subset of 100 images. For the 12×12 block size, the images

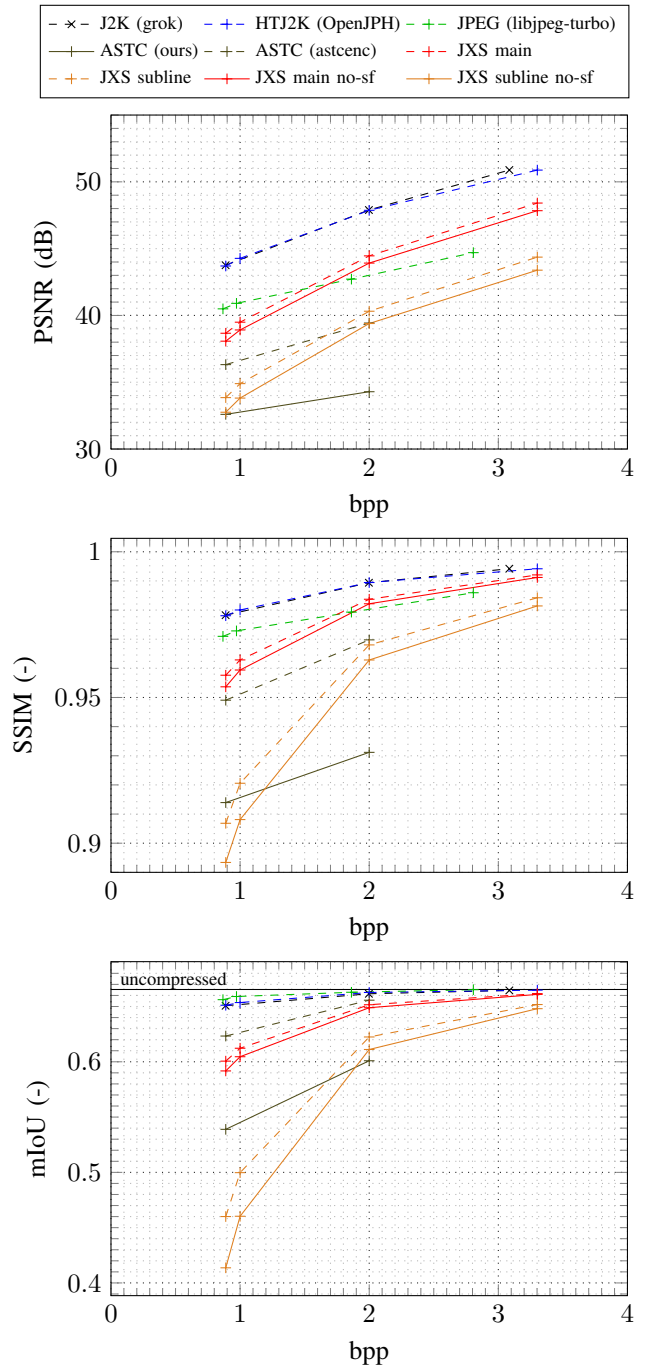


Fig. 2. Mean intersection over union (mIoU) of a FastSeg large network (bottom), SSIM (middle) and PSNR (top) of a Cityscapes validation set compressed with different methods.

were padded to a resolution divisible by a block size of 12 before ASTC encoding.

The results show our simple ASTC 12×12 encoder is approximately $2.3 \times$ faster than the JPEG encoder based on `libjpeg-turbo`. The JPEG decoding was slightly slower than the encoding. While we did not conduct ASTC decoding measurements, in [18], we measured BC1 and YCoCg-BC3 decoding time of an 8K frame as less than 1 ms on a

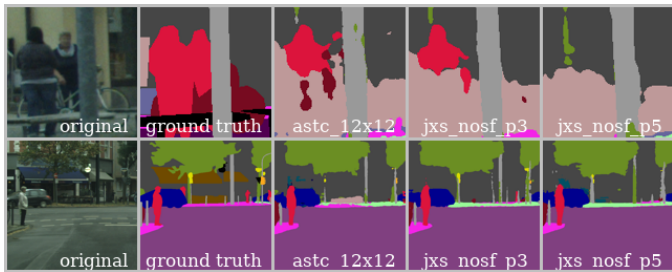


Fig. 3. Visual comparison of LR-ASPP-MobileNetV3 (small version) segmentation of two Cityscapes images. From left to right: Original image (brightened), ground truth, segmentation of the model trained by datasets compressed by the ASTC (12×12), pruned JPEG XS (main profile) and pruned JPEG XS (subline profile) encoders.

desktop GPU. ASTC decoding is more complicated, but the decoding overhead is still expected to be close to negligible in comparison to the encoding.

TABLE III
ENCODING TIME OF THE PROPOSED ASTC ENCODER AND LIBJPEG-TURBO ENCODER AND DECODER (ARM A76 SINGLE-CORE).

	bpp	
	~ 0.89	~ 2.0
ASTC	5.8	7.0
JPEG (enc, libjpeg-turbo)	13.3	16.7
JPEG (dec, libjpeg-turbo)	13.6	21.5

For comparison, we also measured the encoding and decoding time of JPEG at quality parameters 0 and 100 as 11 and 22 ms, and 7 and 32 ms, respectively. These numbers establish the encoding speed bounds of this format.

On a single core of Intel i7-8650U laptop central processing unit (CPU), the AVX2-optimized `astcenc` encoder compressed one Cityscapes image at approximately 44 and 61 ms (block sizes 12×12 and 8×8 , respectively) at the fastest preset, suggesting that both JPEG and our pruned ASTC encoders are faster than a traditional ASTC encoder even with the latter evaluated on a more powerful CPU.

JPEG XS Encoder: Table IV compares JPEG XS with three encoders: JPEG, JPEG 2000, and HTJ2K. The HTJ2K quantization was determined by a similar procedure as in Subsection IV-A. The results show that by disabling the significance flag coding, the encoding time of one Cityscapes frame improved by 22–23%, and is only about 9–20% slower than JPEG 2000. HTJ2K encoding by OpenJPH was 3.8–6.2 \times faster than JPEG XS without the significance coding flags. JPEG by `libjpeg-turbo` brought this difference further by almost another order of magnitude. Kakadu HTJ2K is not publicly available, therefore, we used results published in [26] and extrapolated them to our setup. More specifically, we scaled the result by the number of pixels from 4K to 2048×1024 , multiplied by 4 since the original result was obtained on a 4-core machine, and finally scaled to our frequency of 2.1 GHz from the original 3.4 GHz. Based on this rough estimation, it seems likely the HTJ2K encoder is capable of reaching encoding throughput close to JPEG.

TABLE IV
ENCODING TIMES OF JPEG XS, JPEG, JPEG 2000, AND HTJ2K ENCODERS ON A SINGLE CORE OF I7-8650U CPU

	bpp		
	~ 0.89	~ 2.0	~ 3.3
JPEG XS (main, sf)	625	654	683
JPEG XS (main, no-sf)	477	503	531
JPEG (libjpeg-turbo)	11.3	13.5	16.0
JPEG 2000 (grok)	437	439	441
HTJ2K (OpenJPH)	73.7	105	138
HTJ2K (Kakadu [26])	-	14.4*	-

* extrapolated from the result in the publication

Table V shows JPEG XS encoding time and latency at two different bitrates and three profiles: high, main, and subline. The “precinct” column denotes the number of lines that form one precinct. The “enc” column shows the total frame encoding time and the “latency” column shows the time until the first precinct is done encoding and thus represents the minimum theoretical achievable latency. While the overall frame encoding time does not differ dramatically between presets, the precinct size has a major impact on latency: The precinct of the high profile consisting of three lines shows a latency of 32–33% of the total encoding time, while the latency of a half-line precinct of the subline profile is three times smaller portion of the encoding time. Thus, even without reaching a high throughput, it is possible to achieve latency almost an order of magnitude lower than the encoding time. It should also be noted that the latency includes the wavelet transform over the whole frame and can be further reduced by pipelining it with the rest of the computation.

TABLE V
LATENCY AND THROUGHPUT COMPARISON OF THE PRUNED JPEG XS REFERENCE ENCODER WITHOUT SIGNIFICANCE FLAG CODING.

profile	precinct (lines)	bpp	latency		
			enc (ms)	(ms)	(%)
high	3	0.89	502	167	33%
high	3	2.0	528	166	31%
main	2	0.89	477	136	29%
main	2	2.0	504	137	27%
subline	0.5	0.89	443	53	12%
subline	0.5	2.0	469	53	11%

VI. DISCUSSION

Retraining with the compressed dataset showed the largest improvements when the vision performance without retraining was very low, such as the classification with the ASTC 12×12 and segmentation with the subline JPEG XS encoders. On the segmentation task, the overall quality decrease without retraining was smaller, because the images are less noisy and less prone to compression artifacts. The vision performance could be further enhanced by improving the retraining process and optimizing encoding parameters for computer vision using one of the methods introduced in Subsection II-C.

The lightweight ASTC encoder achieves a higher encoding speed than JPEG, making it the fastest encoder evaluated. On the other hand, the pruned JPEG XS reference encoder

does not achieve sufficient runtime performance. However, its low complexity and results from literature [14] suggest a fast implementation is possible. The rate allocation can be further optimized by, for example, replacing the exhaustive search with a binary search, in combination with other heuristics. The second most expensive operation in the reference encoder is the wavelet transform which we did not modify. In the reference encoder, the wavelet transform is performed over the whole frame before the coding of individual precincts. However, it is possible to interleave the wavelet transform with the precinct coding as the latency of the wavelet transform ranges from a few pixels to 6 lines [4].

To put the results into a practical perspective, let us consider a scenario of compressing a Cityscapes image and sending it over a 500 Mbit/s commercially available 5G network and an embedded transceiver capable of 2 Mbit/s. Assuming a 1 ms latency budget for encoding and network transfer, the ASTC at the bitrate of 0.8 would require a latency of 145 and 3.2 lines, respectively, assuming the encoding speed of 5.8 ms/frame. While the first case allows partitioning the image into larger chunks, in the second case, as shown in Figure 2, lowering the latency of JPEG XS comes at a significant quality loss and thus necessitates the compensation by retraining.

VII. CONCLUSION

We explored decreasing an image encoder complexity to achieve lower latency. Namely, we evaluated a lightweight implementation of an ASTC encoder and a pruned version of a JPEG XS reference encoder.

The ASTC encoder outperforms JPEG in terms of encoding speed by approximately $2.3\times$ at the same bitrate. When retrained with the dataset compressed with ASTC at the lowest bitrate of 0.8 , the classification accuracy was about 5 pp, and the segmentation mIoU 4.4–4.0 pp lower than the output of the networks trained and evaluated without any compression.

The pruned JPEG XS reference encoder is not nearly as fast as ASTC and needs more optimizations to be usable for real-time tasks. Nevertheless, we show that disabling significance flag coding decreases the number of required rate allocation passes, and boosts the encoding speed by 22–23% at the cost of only 0.4–0.3 pp of segmentation mIoU after retraining.

HTJ2K and JPEG outperform both the tested codecs in terms of vision performance. However, ASTC still holds the advantage of the fastest coding speed, while JPEG XS, if sufficiently optimized, is suitable for applications requiring ultra-low latencies. To improve the quality, it is possible to apply computer vision-specific encoding parameter optimizations or improve the retraining process.

ACKNOWLEDGMENT

The work was financially supported by the Tampere University ITC Graduate School. It was also supported by European Union's Horizon 2020 research and innovation program under Grant Agreement No 871738 (CPSoSaware) and in part by the Academy of Finland under Grant 325530.

REFERENCES

- [1] J. Van Waveren and I. Castaño, "Real-time YCoCg-DXT compression," *Tech. Rep., id Software, Inc. and NVIDIA Corp.*, 2007.
- [2] P. Holub, M. Šrom, M. Pulec, J. Matela, and M. Jirman, "GPU-accelerated DXT and JPEG compression schemes for low-latency network transmissions of HD, 2K, and 4K video," *Future Generation Computer Systems*, vol. 29, no. 8, 2013.
- [3] J. Žádník, M. Mäkitalo, J. Vanne, and P. Jääskeläinen, "Image and video coding techniques for ultra-low latency," *ACM Computing Surveys (CSUR)*, 2022.
- [4] A. Descampe, T. Richter, T. Ebrahimi, S. Foessel, J. Keinert, T. Bruylants, P. Pellegrin, C. Buysschaert, and G. Rouvroy, "JPEG XS—a new standard for visually lossless low-latency lightweight image coding," *Proc. of the IEEE*, vol. 109, no. 9, 2021.
- [5] ISO/IEC-15444-15 — ITU-T Rec. T.814, "JPEG 2000 image coding system - part 15: High-throughput JPEG 2000," Standard, Oct. 2019.
- [6] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal processing magazine*, vol. 18, no. 5, 2001.
- [7] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *Trans. on Circuits and Systems for Video Technology*, vol. 22, no. 12, 2012.
- [8] B. Bross, J. Chen, S. Liu, and Y.-K. Wang, "Versatile video coding (draft 8)," JVET-Q2001, 2020.
- [9] G. K. Wallace, "The JPEG still picture compression standard," *Trans. on Consumer Electronics*, vol. 38, no. 1, 1992.
- [10] J. Nystad, S. Lassen, A. Pomianowski, S. Ellis, and T. Olson, "Adaptive scalable texture compression," in *Eurographics / ACM SIGGRAPH Symposium on High Performance Graphics*, 2012.
- [11] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet v2: Practical guidelines for efficient cnn architecture design," in *Proc. of the European Conf. on computer vision (ECCV)*, 2018.
- [12] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan et al., "Searching for MobileNetV3," in *Proc. of the Int. Conf. on Computer Vision*, 2019.
- [13] T. Richter, S. Föbel, A. Descampe, and G. Rouvroy, "Bayer CFA pattern compression with JPEG XS," *Trans. on Image Processing*, vol. 30, 2021.
- [14] K. Itakura, M. Miyazaki, S. Föbel, and M. Van Dorpe, "JPEG-XS codec adapted to 8K and ST 2110," in *SMPTE 2020 Annual Technical Conf. and Exhibition*, 2020.
- [15] X. Xie and K.-H. Kim, "Source compression with bounded DNN perception loss for IoT edge computer vision," in *Proc. of the Int. Conf. on Mobile Computing and Networking (MobiCom)*, 2019.
- [16] Z. Liu, T. Liu, W. Wen, L. Jiang, J. Xu, Y. Wang, and G. Quan, "DeepN-JPEG: A deep neural network favorable JPEG-based image compression framework," in *Proc. of the Design Automation Conf. (DAC)*, 2018.
- [17] B. Brummer and C. de Vleeschouwer, "Adapting JPEG XS gains and priorities to tasks and contents," in *Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
- [18] J. Žádník, M. Mäkitalo, J. Iho, and P. Jääskeläinen, "Performance of texture compression algorithms in low-latency computer vision tasks," in *European Workshop on Visual Information Processing (EUVIP)*, 2021.
- [19] A. Marie, K. Desnos, L. Morin, and L. Zhang, "Expert training: Enhancing ai resilience to image coding artifacts," in *Electronic Imaging, Image Processing: Algorithms and Systems XX*, 2022.
- [20] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," *Trans. on Image Processing*, vol. 29, 2020.
- [21] J. Ascenso and E. Upenik, "White paper on JPEG AI scope and framework v1.0," *ISO/IEC JTC 1/SC 29/WG1 N90049*, 2021.
- [22] J. Shao and J. Zhang, "Bottlenet++: An end-to-end approach for feature compression in device-edge co-inference systems," in *Proc. of Int. Conf. on Communications Workshops (ICC Workshops)*, 2020.
- [23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. of Conf. on Computer Vision and Pattern Recognition*, 2009.
- [24] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [25] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [26] D. Taubman, A. Naman, and R. Mathew, "High throughput block coding in the HTJ2K compression standard," in *Int. Conf. on Image Processing (ICIP)*, 2019.