ResQ: Residual Quantization for Video Perception - Supplementary Material

Davide Abati Haitam Ben Yahia Markus Nagel Amirhossein Habibian

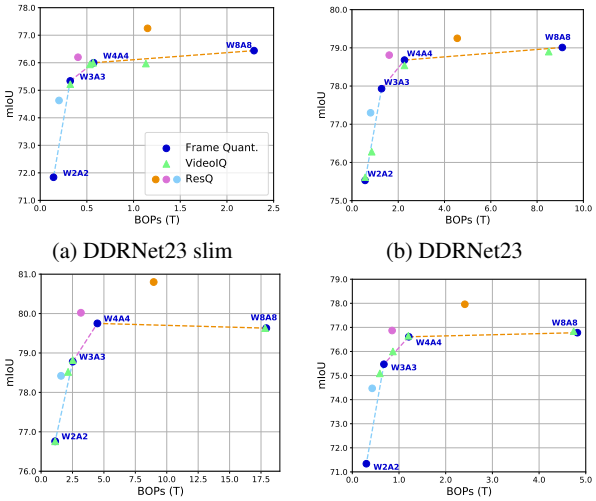
Qualcomm AI Research*

{dabati,hyahia,markusn,ahabibia}@qti.qualcomm.com

1. Additional experiments

Semantic segmentation: QAT. When labeled data is available, Quantization Aware Training (QAT) is a viable option to recover performance lost during Post Training Quantization (PTQ). We hereby report additional results for QAT of segmentation models on Cityscapes. More specifically, we consider the backbones DDRNet23 slim, DDRNet23, DDRNet39 and HRNet-w18-small and employ, for each of them, four different precisions (8, 4, 3 and 2 bits). Moreover, we implement several ResQ models alternating between such precisions in keyframes and residual frames, in cycles of $T = 3$ timesteps. Results for this experiment are reported in Fig. 1, where we represent in blue the baseline results of frame quantization, and in other colors the result from ResQ models. In this respect, the colors representing ResQ results identify the precisions it operates at (w.r.t. the dashed lines between blue points): for example, the orange points represent ResQ models with precision W8A8→W4A4. As the figure shows, even in the presence of QAT our proposal can outperform the baseline of frame quantization. Interestingly, in many cases our final amortized IoU even exceeds the one of the keyframe: we ascribe this gap to the fact that ResQ is a video model, and can learn temporal dynamics during the QAT fine-tuning stage.

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.



(a) DDRNet23 slim (b) DDRNet23

(c) DDRNet39 (d) HRNetw18-small

Figure 1. **QAT results** for different DDRNet and HRNet models with Frame Quantization in blue and ResQ otherwise. Color specifies the operating precision between keyframes and residual frames, following dashed lines. Green triangles represent VideoIQ results.

Bit-width	Quantization	mIoU	Δ mIoU
W8A8→W4A8	PTQ	75.14	-
	QAT	**77.81**	**+2.67**
W8A8→W8A4	PTQ	66.76	-
	QAT	**77.70**	**+10.94**
W8A8→W4A4	PTQ	66.67	-
	QAT	**77.25**	**+10.58**

Table 1. **PTQ vs. QAT.** ResQ results on Cityscapes using DDRNet23-small at various precisions. Note that QAT can recover performance for precisions that are challenging for PTQ.

To further compare the benefit that fine-tuning can bring when using residual quantization, we directly compare ResQ results in the case of PTQ and QAT in Tab. 1. As the table shows, at all tested precisions QAT can grant a significant improvement w.r.t. PTQ. We appreciate the highest gap in the presence of activations quantized to low precision (4 bits), a setting which is particularly harmful for PTQ.

Semantic segmentation: VideoIQ We now aim to empirically compare ResQ to VideoIQ [4] on a frame-prediction task, such as semantic segmentation on Cityscapes. As VideoIQ is tailored for action recognition, it utilizes Temporal Segment Network [5] as a base model. Therefore, we adapt it to use it along with segmentation architectures. Specifically, we consider the QAT Frame Quantization models represented in Fig. 1, quantized at different precisions (8, 4, 3 and 2 bits). We then train, following the objective in [4], a MobileNet v2 [3] policy network that, given a frame, predicts the quantization level to be applied to the segmentation model. By properly weighting

Scheme	Bit-width	mIoU↑	Δ mIoU ↑
Tensor	W8A8→W4A8	73.41	-
Channel	W8A8→W4A8	75.14	+ **1.73**
Tensor	W8A8→W8A4	66.59	-
Channel	W8A8→W8A4	66.76	+ 0.08
Tensor	W8A8→W4A4	66.51	-
Channel	W8A8→W4A4	66.67	+ 0.16
Tensor	W8A4→W4A4	58.43	-
Channel	W8A4→W4A4	60.21	+ **1.78**

Table 2. **Ablation on Tensor vs. Channel Quantization.** Experiments on ResQ at various bit-widths.

the VideoIQ objectives, it is possible to obtain policies that reward BOP reduction and accuracy gains differently. We represent the result of this VideoIQ model in Fig. 1. Importantly, we advantage this model by excluding the cost of the (full precision) policy in the BOP count, and we only measure the strength of the dynamic selection of quantized models. Even so, many times the policy degenerates to non-dynamic decisions, and fixes the predicted precision regardless of the input frame. Moreover, whenever actually predicting bit-widths dynamically, the policy decisions do not outperform random selection (represented by dashed lines between models). These results, consistent across four different backbones, suggest that a VideoIQ-like policy to dynamically select the quantization level based on the frame might be hard to learn for frame-prediction tasks.

Semantic segmentation: channel quantization. We investigate the benefits of performing channel quantization. In Tab. 2, we compare per-tensor and per-channel scale factors for PTQ ResQ on Cityscapes. We observe that channel quantization is on par or exceeds tensor quantization (up to $+1.78$ mIoU) for multiple precisions and is therefore favorable in ResQ .

Semantic segmentation: per class analysis. To further analyze the behaviour of residual quantization, we take a closer look at performances over different classes in Cityscapes. Specifically, in Fig. 2 we measure the mIoU per class obtained with QAT for Frame Quantization and for ResQ, when the frame of interest is processed at 2 bits (for ResQ , we use W3A3→W2A2). We appreciate that ResQ typically performs better than the baseline on classes that have low mIoU (*i.e.*, the challenging classes), such as rider, pole, traffic light and fence. We hypothesize that this behavior is due to the higher precision of the keyframe, serving as a good starting point for the residuals: applying a low precision update might be easier than detecting hard classes with a low precision from scratch.

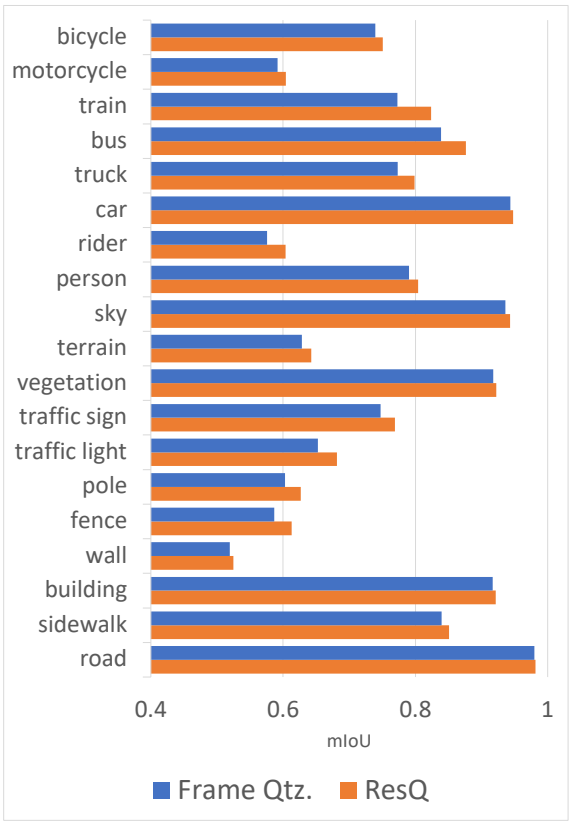


Figure 2. **mIoU per class** on Cityscapes for Frame Quantization and ResQ when the frame of interest is quantized at 2 bits.

Pose estimation: pairwise vs recurrence scheme. To further analyze the temporal aspects of ResQ , we consider an alternative sigma-delta decomposition scheme. Specifically, instead of defining the residual as $\delta^t = x^t - x^k$ (pairwise), we might compute them as instantaneous variations, as $\delta^t = x^t - x^{t-1}$ (recurrent), as done in [1, 2]. In this latter strategy, errors are likely to compound over time, as opposed to the pairwise strategy that only depends on the keyframe. In Fig. 3 we compare both summation strategies in ResQ , when applied to JHMDB (split 1). Although for some precisions no significant difference can be noticed (W8A8→W4A8), we appreciate how pairwise summations prove more stable over long intervals (W8A8→W8A4). We ascribe this finding to the error propagation hampering recurrent summations.

2. BOP and inference time

To investigate whether theoretical BOPs gains would translate to actual runtime improvements that processing at low bit-widths would bring, we rely on a HW simulator for future generations of SnapDragon NPU, a low-power fixed-point accelerator. Tab. 3 shows how bringing weights and activations from 8 to 4 bits reduces the simulated runtime

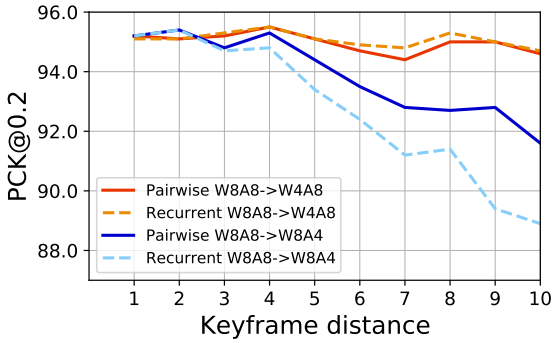


Figure 3. **Ablation on pairwise vs. recurrent** summation on JHMDB for different bit-widths. Note that activation quantization suffers from using recurrent summation and pairwise is better or equal as we increase the distance to the keyframe.

for both pose-estimation and semantic segmentation models.

3. Qualitative results

We report additional qualitative comparison between frame and residual quantization in Fig. 4. We report the comparison at various bit-widths (4, 3, and 2 bits). Similarly to the example reported in Fig. 8 in the main paper, in highlighted regions ResQ takes advantage of the high precision representation granted by the keyframe to better segment the scene at hand with low precision.

References

[1] Amirhossein Habibian, Davide Abati, Taco S Cohen, and Babak Ehteshami Bejnordi. Skip-convolutions for efficient video processing. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2021. 2

[2] Mathias Parger, Chengcheng Tang, Christopher D. Twigg, Cem Keskin, Robert Wang, and Markus Steinberger. Deltacnn: End-to-end cnn inference of sparse frame differences in videos. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2022. 2

[3] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018. 1

[4] Ximeng Sun, Rameswar Panda, Chun-Fu Richard Chen, Aude Oliva, Rogerio Feris, and Kate Saenko. Dynamic network quantization for efficient video inference. In *IEEE International Conference on Computer Vision*, 2021. 1

[5] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision*, 2016. 1

Model	**W8A8**	**W4A8**	**W8A4**	**W4A4**
DDRNet-23s	3.52	2.76	2.43	1.96
HRNet-w32	0.93	0.68	0.72	0.49

Table 3. **Simulated runtimes** in milliseconds for semantic segmentation (DDRNet-23s) and pose estimation (HRNet-w32).

Figure 4. **Qualitative comparison** between frame and residual quantization at low bit-width. All models illustrated here benefit of QAT. In highlighted regions, our proposal significantly outperforms the baseline.