AI and Data Science Journal
Online ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj4042

ROBUSTNESS-AWARE MODEL COMPRESSION FOR MULTIMODAL PERCEPTION UNDER CHALLENGING CONDITIONS

FengYuan Zhang[1*], JieSi Yang[2]
[1]*New York University, New York 10012, NY, USA.*
[2]*The University of Utah, Salt Lake City 84112, UT, USA.*
[*]*Corresponding Author: FengYuan Zhang*

Abstract: Compressed multimodal perception models are usually validated on clean benchmarks, yet they are deployed where visibility is low, sensors are occluded and measurements are noisy. This paper studies RGB-thermal semantic segmentation on the MSRS road-scene benchmark and asks three questions: which predictions break after compression under challenging conditions, how compression, calibration and multimodal reliability cues can repair them, and how accuracy, hard-scenario performance and real latency trade off. We first show that 4-bit quantization and structured pruning remove 2.8-4.9 mIoU points on clean data but up to 7.4 points under unseen corruptions, and that the resulting harmful prediction flips concentrate on low-margin pixels, object boundaries and rare classes. We then propose RAMC, a robustness-aware multimodal compression framework that combines condition-aware Hessian sensitivity and minimax integer programming for mixed-precision post-training quantization with condition-aware Taylor pruning, paired clean-to-degraded distillation and fusion-gate consistency, and we analyze post-compression confidence calibration. At a tight budget of 24 BitOPs per MAC, RAMC improves night mIoU by 1.2 points and mean corrupted mIoU by 0.4 points over HAWQ while reducing harmful flips by 17%. With 75% of the hidden channels removed, it improves mean corrupted mIoU by 1.2 points over distillation; at 50% pruning it exceeds the uncompressed model in clean, night and mean corrupted mIoU while running 1.24-1.47 times faster in ONNX Runtime. Distilled students become underconfident, and a single temperature fitted on held-out degraded data nearly halves their calibration error on unseen corruptions.
Keywords: Model compression; Multimodal perception; RGB-thermal segmentation; Robustness; Mixed-precision quantization; Structured pruning; Calibration

1 INTRODUCTION

Multimodal perception is a core capability of autonomous vehicles and mobile robots. Pairing a visible-light camera with a thermal camera is particularly attractive because long-wave infrared imaging is insensitive to illumination and partially penetrates haze, so the two modalities fail under different conditions [1-2]. RGB-thermal (RGB-T) semantic segmentation networks therefore fuse the two streams at several feature levels [3-4], and recent training strategies explicitly prepare them for missing or degraded inputs [5]. Deployment on embedded platforms, however, requires aggressive model compression, and compression is almost always tuned and reported on clean validation data.

Robustness benchmarks show that perception models degrade substantially under common corruptions such as noise, blur, low light and fog [6-10]. Compression is known to amplify such fragility: pruned and quantized networks preserve top-line accuracy while losing a long tail of difficult examples [11], although carefully designed compression can also preserve or improve out-of-distribution robustness [12]. Whether these observations transfer to multimodal dense prediction, where one sensor may fail while the other remains informative, is largely unexplored. Standard quantization and pruning pipelines measure layer sensitivity and channel importance on clean calibration data [13-18], and knowledge distillation transfers the teacher's behaviour only on the inputs the student sees [19-20]. None of these mechanisms is aware of the operating conditions in which reliability matters most.

This paper addresses robustness-aware compression of RGB-T perception models through three questions. *Q1: which predictions of a compressed model fail under low visibility, occlusion and sensor noise? Q2: how can the compression strategy, confidence calibration and multimodal reliability information be used to repair them? Q3: how do clean accuracy, hard-scenario performance and actual inference latency relate?* We answer these questions on the real day/night MSRS benchmark with a deployment-oriented dual-stream network and six held-out corruption families [2]. The main contributions are as follows.

A condition-aware evaluation protocol for compressed multimodal models that combines a real night split with unseen corruptions of either modality, and a harmful-flip analysis that localises compression errors by prediction margin, object boundaries, classes and modality reliance.

RAMC, a robustness-aware multimodal compression framework comprising condition-aware Hessian sensitivity with minimax bit allocation for mixed-precision quantization and condition-aware Taylor pruning with paired clean-to-degraded distillation and fusion-gate consistency, complemented by an analysis of post-compression confidence calibration.

© By the Author(s) 2024, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

An empirical study of the accuracy-robustness-latency trade-off with real INT8 and pruned models in ONNX Runtime, showing that robustness-aware compression closes most of the gap opened by standard compression and, at moderate pruning, surpasses the uncompressed model.

2 RELATED WORK

2.1 Multimodal Perception under Challenging Conditions

MFNet introduced an RGB-T urban-scene dataset with day and night recordings [1], and MSRS provides registered infrared-visible pairs with segmentation labels derived from it [2]. RTFNet fuses thermal features into an RGB encoder [3], CMX rectifies and fuses cross-modal features at several scales [4], and complementary random masking regularizes fusion networks against modality corruption [5]. Dynamic fusion with quality estimates and confidence regularization when modalities are removed improve multimodal reliability [21-22]. Robustness benchmarks for image classification [6], driving-scene detection and 3D perception apply synthetic corruptions [7-9], and physically based fog synthesis follows the atmospheric scattering model [10]. These works study full-precision models; we study what happens to such reliability mechanisms after compression.

2.2 Compression and Its Side Effects

Integer-only quantization is the standard deployment path on CPUs and accelerators [13]. Mixed-precision methods allocate bit-widths using layer sensitivity, for example the average Hessian trace in HAWQ-V2 [14], while PD-Quant optimizes quantization parameters with a prediction-difference metric [15]. Filter pruning ranks channels by weight magnitude or first-order Taylor importance [16-17], and DepGraph generalizes structural grouping [18]. Distillation and logit-standardized distillation help compact students recover accuracy [19-20]. Hooker *et al.* showed that compression disproportionately affects atypical and corrupted examples [11], whereas Diffenderfer *et al.* found that some compression schemes can improve out-of-distribution robustness [12]. Calibration studies show that modern networks are overconfident and that temperature scaling and sample-adaptive temperatures reduce miscalibration [23-24]. RAMC differs from these methods by injecting operating conditions, real and synthetic, into every compression decision and by exploiting the fusion gates of the multimodal model as a reliability signal.

3 METHODOLOGY

3.1 RGB-T Segmentation Model

Figure 1(a) shows the perception model. Each modality is encoded by a stream built from Universal Inverted Bottleneck (UIB) blocks of MobileNetV4 (a stem convolution [25], one fused inverted bottleneck and seven UIB blocks with ExtraDW, IB and ConvNeXt-like variants), producing features at strides 4, 8, 16 and 32. At every scale a reliability-gated fusion (RGF) module predicts per-pixel modality weights from the concatenated features,

$$\mathbf{w} = \mathrm{softmax}(\varphi[\mathbf{F}_r; \mathbf{F}_t]), \quad \mathbf{F} = \psi(\mathbf{w}_r \odot \mathbf{F}_r + \mathbf{w}_t \odot \mathbf{F}_t) \tag{1}$$

where *phi* is a two-layer 1x1 convolution, *psi* a 1x1 refinement and the softmax is taken over the two modalities. The fused pyramid is decoded by an FPN-style decoder. The network has 0.42 M parameters and requires 0.249 GMACs at 240x320, which places it in the budget of embedded perception. The gates w_t provide an inexpensive, interpretable estimate of how much each pixel relies on the thermal stream; RAMC uses them both as a distillation target and as a calibration signal.

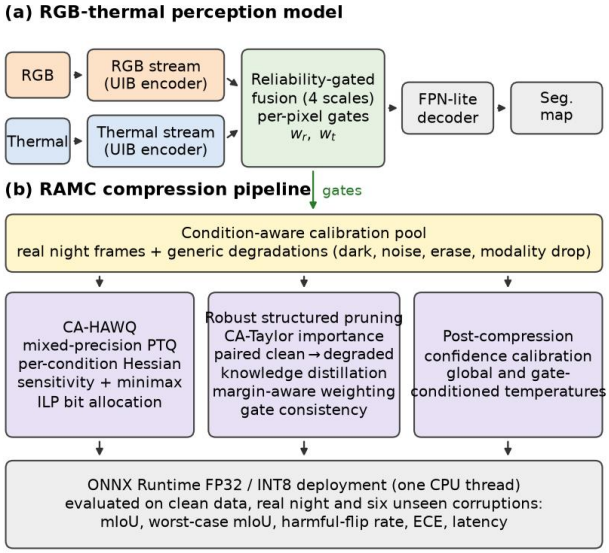


Figure 1 Overview of the RGB-thermal Perception Model (a) and the Robustness-Aware Multimodal Compression Framework RAMC (b)

Note: The condition-aware calibration pool contains real night frames and generic degradations only; the six test corruptions are never seen during training, calibration or compression.

3.2 Challenging-Condition Protocol and Metrics

The real low-visibility condition is the night split of MSRS. In addition, six corruption families are applied to the whole test set at three severities (severity 2 is used unless stated otherwise): *fog*, synthesized with the atmospheric scattering model using a road-scene depth proxy and a weaker attenuation for the thermal band [10]; *low-light*, combining gamma darkening with Poisson-Gaussian sensor noise; *occlusion* of either the RGB or the thermal lens by irregular soft-edged blobs covering 12-32% of the image; *RGB noise*, a Poisson-Gaussian shot and read noise model; and *thermal noise*, Gaussian noise plus column and row fixed-pattern stripes typical of uncooled microbolometers. Training uses a different, generic degradation family (random darkening, Gaussian noise, rectangular erasing and modality dropout), so all six test corruptions are unseen.

Besides mean IoU on clean data (Clean), the night split (Night), the mean over the six corruptions (mCorr) and the minimum over corruptions and night (Worst), we quantify *which* predictions are damaged by compression. Let y_i be the label of pixel i, and let the full-precision (fp) and compressed (c) models predict the labels marked by a hat. The harmful-flip rate

$$\mathcal{H} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\hat{y}_i^{\mathrm{fp}} = y_i \ \wedge \ \hat{y}_i^{\mathrm{c}} \neq y_i\right] \tag{2}$$

counts pixels that the full-precision model classifies correctly and the compressed model breaks. We report it per condition and stratify it by the full-precision prediction margin m_i (top-1 minus top-2 probability), by boundary versus interior pixels, by class and by the full-precision thermal gate. Calibration is measured with the pixel-level expected calibration error (ECE, 15 bins) [23].

3.3 Condition-Aware Mixed-Precision Quantization

Weights are quantized per output channel with symmetric uniform quantizers and activations per tensor with asymmetric quantizers,

$$Q_b(x) = s\big(\mathrm{clip}\big(\lfloor x/s \rceil + z, 0, 2^b - 1\big) - z\big) \tag{3}$$

where the scale s and zero point z follow from a clipping range chosen by minimizing the mean squared quantization error (MSE) on calibration data. Batch normalization is folded into the preceding convolutions, and the first convolution of each stream and the classifier are kept at 8 bits, as is common practice.

Condition-aware calibration pool. Standard pipelines draw calibration images from clean training data. RAMC builds the pool from the same 48 training images but includes the real night frames among them and adds one copy of each image under every generic training degradation, giving $K = 8$ condition groups (day, night, darkening, RGB noise, thermal noise, RGB erasing, thermal erasing and modality drop). No test corruption is used.

Condition-aware sensitivity. HAWQ-V2 measures the sensitivity of layer l by the average Hessian trace of the task loss on clean data [14]. We estimate it separately for every condition group k with Hutchinson's estimator and normalize it within each group,

$$\mathbf{\Omega}_{l,k} = \frac{\mathrm{Tr}\left(\mathbf{H}_l^{(k)}\right)}{n_l} \left\| Q_b(\mathbf{W}_l) - \mathbf{W}_l \right\|_2^2, \quad \widetilde{\mathbf{\Omega}}_{l,k} = \frac{\mathbf{\Omega}_{l,k}}{\sum_j \mathbf{\Omega}_{j,k}} \tag{4}$$

where n_l is the number of weights of layer l and Q_b its low-bit quantizer. Normalization makes groups with very different loss curvature comparable, so that a layer that matters only at night or only under thermal degradation is still visible to the allocator.

Minimax bit allocation. Each quantized layer is either kept at W4A4 or upgraded to W8A8 ($x_l = 1$) at a cost c_l proportional to its MACs. Instead of minimizing the total sensitivity, RAMC minimizes the worst normalized sensitivity over condition groups,

$$\min_{\mathbf{x}\in\{0,1\}^L,\, t} \ t \ \ \text{s.t.} \ \ t \geq \sum_l \widetilde{\mathbf{\Omega}}_{l,k}(1 - x_l) \ \ \forall k, \ \ \sum_l c_l x_l \leq \rho \sum_l c_l \tag{5}$$

where *rho* controls the BitOPs budget. The integer program has fewer than 100 binary variables and is solved exactly in milliseconds with HiGHS through SciPy.

3.4 Robustness-Aware Structured Pruning

Pruning removes hidden channels of every fused and universal inverted bottleneck in both streams, so that the pruned network remains dense and directly accelerates on any runtime. For channel c and condition group k, the first-order Taylor importance is accumulated on the hidden activation a_c feeding the projection convolution and then aggregated across groups with a power mean that emphasizes channels needed by any condition [17],

$$I_{c,k} = \mathbb{E}_{x\in\mathcal{D}_k}\left|\sum_{h,w} a_c \frac{\partial \mathcal{L}}{\partial a_c}\right|, \quad I_c = \left(\frac{1}{K}\sum_{k=1}^{K} \tilde{I}_{c,k}^{\,p}\right)^{\frac{1}{p}} \tag{6}$$

where the tilde denotes normalization by the mean importance of the block within group k and $p = 4$. Channels that are essential only at night or only under thermal noise are thereby protected even if their clean-data importance is small.

The pruned student is fine-tuned with a robust objective that uses *paired* views. For a training sample, x denotes the clean augmented view and x with a tilde the same view after a random generic degradation. The full-precision teacher receives the clean view, which provides privileged soft targets, while the student receives the degraded view:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(f_s(\tilde{x}), y) + \lambda \frac{\sum_i w_i \,\mathrm{KL}(p_t(x)_i \,\|\, p_s(\tilde{x})_i)}{\sum_i w_i} + \frac{\mu}{4}\sum_{\ell=1}^{4} \left\| \mathbf{g}_s^{\ell}(\tilde{x}) - \mathbf{g}_t^{\ell}(\tilde{x}) \right\|_2^2 \quad (7)$$

The pixel weights $w_i = 1 + gamma(1 - m_i)$ increase the distillation weight of low-margin pixels, which Section 4.2 identifies as the main source of harmful flips. The last term aligns the student's thermal gates with the teacher's gates on the degraded input at the four fusion scales, so that compression does not erase the learned modality-reliability behaviour. We use *lambda* = 1, *gamma* = 2 and *mu* = 2.

3.5 Post-Compression Confidence Calibration

Compression and corruption both shift the confidence of the network. Global temperature scaling (TS) applies one temperature to every pixel [23], and adaptive temperature scaling predicts a temperature per sample from its confidence [24]. We additionally evaluate a multimodal variant that conditions the temperature on reliability cues the fused network already computes:

$$T_i = 0.05 + \mathrm{softplus}\left(\theta^{\top} \left[1, g_i, \bar{g}, \sigma_g, \bar{c}\right]\right) \quad (8)$$

where g_i is the thermal gate at pixel i averaged over the four scales, the bar quantities are the image means of the gate and of the uncalibrated confidence, and $sigma_g$ is the spatial standard deviation of the gate. All temperatures are fitted by minimizing the negative log-likelihood on held-out training images, either clean or under the generic degradations. Because the temperature is positive, the predicted labels and the mIoU are unchanged.

4 EXPERIMENTS

4.1 Experimental Setup

Data. MSRS contains 1,083 training and 361 test pairs of registered visible and thermal images of road scenes (547 and 182 of them at night) with nine classes: unlabeled [2], car, person, bike, curve, car stop, guardrail, color cone and bump. Images are resized to 240x320 and, following [1, 3], mIoU is averaged over all nine classes. Every 13th training image (83 images) is held out for fitting calibration temperatures.

Training. The full-precision model is trained from scratch for 40 epochs with AdamW (learning rate $2\text{x}10^{-3}$, weight decay 10^{-4}, cosine schedule), batch size 8, random scaling and 192x256 crops, horizontal flips, photometric jitter, inverse-log-frequency class weights and the generic degradation family with probability 0.5. It reaches 35.34 mIoU on the clean test set (35.43 day, 34.13 night) and 28.71 mCorr.

Compression. Quantization is simulated in PyTorch with 48 calibration images. Mixed precision is evaluated at two budgets, *rho* = 1/3, which equals the BitOPs of uniform W4A8 (33.7 bit-operations per MAC), and *rho* = 0.12 (24 BitOPs per MAC). Pruning removes 50% or 75% of the hidden channels, reducing the cost from 0.249 to 0.151 or 0.109 GMACs, followed by 6 epochs of fine-tuning (AdamW, learning rate 10^{-3}). Deployable models are exported to ONNX and quantized with ONNX Runtime static QDQ INT8 (per-channel weights); latency is the median of 60 runs with batch size 1 on one thread of an Intel Xeon CPU at 2.1 GHz with AVX-512 VNNI.

4.2 Where Do Compressed Models Fail? (Q1)

Tables 1-3 show that the cost of compression grows under challenging conditions. Uniform W4A4 quantization loses 2.80 mIoU on clean data but 3.81 at night and 3.87 in mCorr, with the largest drops under low light (5.93) and RGB noise (4.16). Pruning 75% of the hidden channels with L1 ranking and plain fine-tuning loses 4.94 points on clean data, 5.42 in mCorr and 7.37 under thermal noise. The harmful-flip rate rises accordingly, from 4.34% on clean inputs to 7.68% under corruptions for W4A4 and from 5.38% to 9.32% for the pruned model.

Figure 2 localizes these failures. *Margin:* harmful flips are concentrated on pixels with a small full-precision margin; in the lowest margin decile 16-23% of the pixels flip, against less than 1% in the most confident decile on clean data (Figure 2(a)). Corruptions act in two ways: they enlarge the share of low-margin pixels ($m_i < 0.3$) from 8.8% to 13.8%, and they raise the flip rate even in the most confident decile, from 0.70% to 1.75% for W4A4 and from 0.34% to 1.05% for the distilled 75% pruned model. *Boundaries:* object boundaries cover only 4.1% of the pixels, yet their harmful-flip rate under corruption is 1.39-1.52 times that of interior pixels (Figure 2(b)). *Classes:* thin and rare classes suffer most, e.g., bump (28.9%), curve (21.1%) and car stop (15.4%) for W4A4, compared with 7.0% for the background (Figure 2(c)).

Modality reliance. The fusion gates of the full-precision model respond to degradations in the expected direction: the mean thermal weight rises from 0.514 on clean data to 0.529 under fog and 0.534 under low light and RGB occlusion, and it drops to 0.497 and 0.501 under thermal occlusion and thermal noise; RGB noise is the only degradation that the gates do not compensate (0.495). Compression damages exactly the pixels that rely on the degraded modality: under fog, 19.6% of the RGB-dominant pixels (thermal weight 0.2-0.4) are harmfully flipped by W4A4, against 4.8% of the balanced pixels. Pruning additionally distorts the reliability behaviour itself: under fog the distilled 75% student shifts the thermal weight by -0.031 relative to its teacher, i.e., it partly reverses the teacher's response. In summary,

compressed multimodal models fail on low-margin, boundary and rare-class pixels and on pixels that depend on the degraded sensor, and they lose part of the learned modality-reliability behaviour.

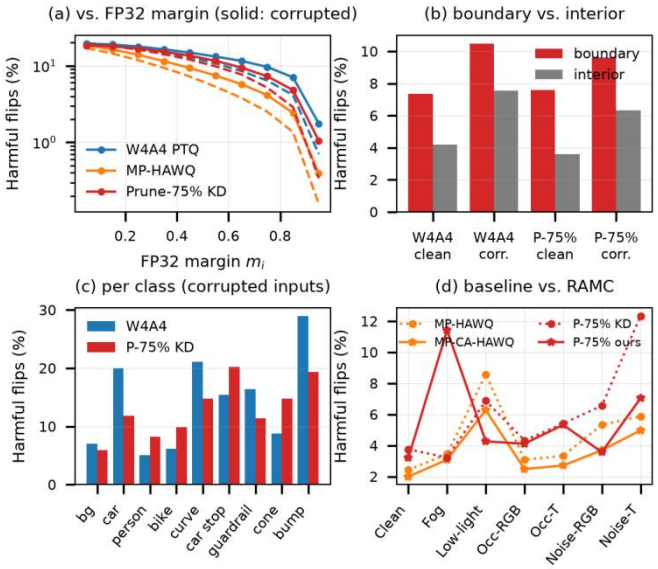


Figure 2 Harmful Prediction Flips After Compression

Note: (a) Flip rate versus the full-precision margin (solid: mean over corruptions, dashed: clean). (b) Boundary versus interior pixels. (c) Per class under corruptions. (d) Per condition for HAWQ versus CA-HAWQ (24 BitOPs per MAC) and distillation versus RAMC (75% pruning).

Table 1 Post-Training Quantization on MSRS (mIoU, %)

Method	BitOPs	Clean	Night	mCorr	Worst	Hc (%)	ECEc (%)
FP32	1024	35.34	34.13	28.71	25.20	-	1.25
W8A8 (clean, MSE)	64.0	35.54	34.54	28.84	25.35	0.56	1.27
W4A8 (clean, MSE)	33.7	34.31	33.12	27.95	25.41	2.89	1.70
W4A4 (clean, MSE)	18.5	32.54	30.33	24.83	21.04	7.68	2.43
W4A4 (CA, MSE)	18.5	33.12	30.39	25.19	21.08	7.82	2.61
MP-HAWQ (clean)	33.7	36.37	34.79	28.87	24.35	3.04	2.02
MP-HAWQ (CA)	33.7	36.25	34.82	28.69	24.22	3.17	2.09
MP-CA-HAWQ (ours)	33.7	36.56	35.22	28.83	24.84	2.68	1.67
MP-HAWQ (clean)	24.0	35.16	33.18	27.57	23.92	4.67	2.23
MP-HAWQ (CA)	24.0	35.63	33.98	27.48	23.67	4.94	2.37
MP-CA-HAWQ (ours)	23.9	36.07	34.36	27.99	24.17	3.88	2.08

Note: BitOPs: average weight-bits x activation-bits per MAC. Rows 6-8: budget equal to uniform W4A8; rows 9-11: tight budget. Night: real night split; mCorr/Worst: mean/minimum mIoU over the six unseen corruptions (Worst also includes night); Hc: harmful-flip rate averaged over corruptions.

4.3 Condition-Aware Mixed-Precision Quantization (Q2)

Table 1 summarizes quantization. W8A8 is lossless (35.54 clean, 28.84 mCorr), uniform 4-bit weights cost about one point, and W4A4 is clearly harmful. Calibrating with the condition-aware pool instead of clean images improves uniform W4A4 by 0.58 points on clean data and 0.36 in mCorr. Mixed precision recovers most of the loss, and the allocation criterion becomes decisive. At the W4A8-equivalent budget CA-HAWQ attains the highest clean and night mIoU (36.56 and 35.22) and the lowest harmful-flip rate (2.68% versus 3.04% and 3.17% for HAWQ), with mCorr on par with HAWQ (28.83 versus 28.87). At the tight budget of 24 BitOPs per MAC, CA-HAWQ improves on HAWQ by 0.91 points on clean data, 1.18 at night and 0.42 in mCorr, and it reduces harmful flips by 17% (3.88% versus 4.67%). The improvement is largest where Q1 located the most dangerous errors: in the most confident decile, the corrupted-input flip rate falls from 0.39% (HAWQ) to 0.17%.

Figure 3 explains the allocation. The normalized sensitivity maps differ markedly across condition groups: thermal erasing and modality drop emphasize the thermal stream and the fusion modules, whereas darkening and RGB noise emphasize early RGB layers. Clean-data HAWQ ignores these patterns; the minimax program of (5) raises the share of 8-bit MACs in the fusion modules from 34.0% to 42.5% and in the thermal stream from 8.0% to 9.4%, and lowers it in the RGB stream from 19.0% to 16.4%. Protecting the layers that implement cross-modal reliability is therefore more valuable for robustness than protecting the layers that dominate the clean loss.

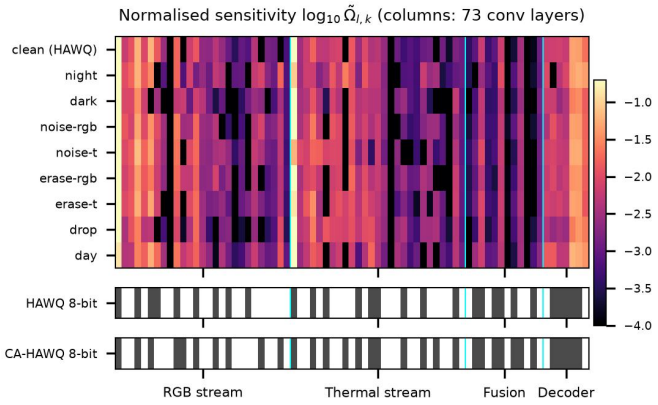


Figure 3 Normalized Layer Sensitivity under Clean Data (HAWQ) and the Eight Condition Groups of the Calibration Pool (top), and the Layers Kept at 8 Bits by HAWQ and CA-HAWQ at 24 BitOPs per MAC (bottom)

Table 2 Structured Pruning on MSRS (mIoU, %)

Method	GMACs	Clean	Night	mCorr	Worst	Hc (%)
FP32 (unpruned)	0.249	35.34	34.13	28.71	25.20	-
L1 + KD	0.151	35.45	34.34	28.17	25.01	3.51
RAMC (ours)	0.151	35.85	35.01	29.13	24.67	2.91
L1 + fine-tune	0.109	30.40	29.20	23.28	17.83	9.32
L1 + KD	0.109	32.72	31.31	25.23	19.45	6.46
RAMC (ours)	0.109	33.92	32.48	26.46	22.27	5.96

Note: Rows 2-3: 50% of hidden channels removed; rows 4-6: 75% removed. All pruned models are fine-tuned for 6 epochs with the same data and augmentation.

4.4 Robustness-Aware Structured Pruning (Q2)

Table 2 compares pruning pipelines at equal cost. At 75% pruning, distillation recovers 1.95 mCorr points over plain fine-tuning, and RAMC adds another 1.20 points on clean data, 1.17 at night and 1.23 in mCorr while lowering harmful flips from 6.46% to 5.96%. At 50% pruning RAMC exceeds the uncompressed model in clean, night and mean corrupted mIoU (35.85, 35.01 and 29.13) with 39% fewer MACs, whereas the distilled baseline remains below it under corruption (28.17 mCorr). Table 3 shows that the largest gains occur under RGB noise (+2.88 at 75%), thermal noise (+2.82) and low light (+1.81), i.e., under the sensor-noise and low-visibility conditions targeted by the paired distillation. Fog is the only corruption on which the 75% RAMC student trails the distilled baseline (23.91 versus 25.12), while at 50% RAMC improves fog by 1.48 points; fog removes contrast smoothly over the whole image, a degradation that is absent from the generic pool and can be added when haze is expected. RAMC also preserves the modality-reliability behaviour: the thermal-gate drift under fog shrinks from -0.031 to -0.009.

Table 3 Per-Corruption mIoU (%) of Compressed Models (severity 2)

Model	Fog	Low-light	Occ-RGB	Occ-T	Noise-RGB	Noise-T
FP32	28.91	26.97	32.78	28.51	29.86	25.20
W4A4 (clean, MSE)	25.39	21.04	29.58	25.69	25.71	21.58
MP-HAWQ, 24 b	28.05	24.24	32.04	28.26	28.60	23.67
MP-CA-HAWQ, 24 b	28.21	24.93	32.43	28.76	29.42	24.17
P-50% L1+KD	26.29	27.04	32.25	28.47	29.94	25.01
P-50% RAMC	27.77	28.75	33.02	28.98	31.61	24.67
P-75% L1+KD	25.12	24.63	29.65	25.95	26.60	19.45
P-75% RAMC	23.91	26.44	30.34	26.36	29.48	22.27

4.5 Confidence Calibration (Q2)

Table 4 reports the pixel-level ECE. The full-precision model is well calibrated (1.25% clean, 1.25% under corruptions), and W4A8 quantization changes its confidence only marginally. Pruning with distillation behaves differently: the 75% student becomes underconfident, with a fitted temperature of 0.82, and its ECE rises to 2.64% on clean data and 3.91% under unseen corruptions. A single global temperature fitted on held-out images reduces these values to 0.97% and 2.09%, and it transfers to the unseen corruptions whether it is fitted on clean or on degraded images. The confidence-adaptive temperature performs similarly on clean data but is less effective under corruption (2.30%) [24]. The gate-conditioned temperature of (8) gives the lowest corrupted ECE for the quantized model (1.53% versus 1.72%) but recovers only part of the pruned model's miscalibration (3.22%): its weights adapt strongly to the gate statistics of the generic degradations, which are not representative of the unseen corruptions. Consequently, RAMC applies global TS as the final post-compression step and refits it after every compression stage, because teacher and student require different temperatures.

Table 4 Expected Calibration Error (%) of Full-Precision and Compressed Models

Method	FP32 clean	FP32 corr.	W4A8-CA clean	W4A8-CA corr.	P-75% clean	P-75% corr.
Uncalibrated	1.25	1.25	0.98	1.72	2.64	3.91
TS (clean val.)	1.55	1.34	1.57	1.68	0.97	2.09
TS (mixed val.)	1.50	1.32	1.63	1.69	1.00	2.09
Adaptive TS (conf.)	1.36	1.34	1.38	1.66	0.88	2.30
Ours (gate-conditioned)	1.45	1.38	1.44	1.53	1.05	3.22

Note: Pixel-level ECE (%) with 15 bins; corr.: mean over the six unseen corruptions. Temperatures never change the arg-max, so mIoU is unaffected.

4.6 Accuracy, Robustness and Latency (Q3)

Table 5 and Figure 4 connect the simulated results to deployed models. Real ONNX Runtime INT8 models (W8A8, per-channel) match or exceed the full-precision accuracy on every split (35.88 clean, 34.88 night), and the calibration data has a negligible effect at 8 bits, consistent with Table 1. Latency, however, does not follow BitOPs. At 240x320 the INT8 graph is slower than FP32 (8.13 ms versus 6.69 ms) because the quantize-dequantize overhead of a 0.25 GMAC network with many depthwise and element-wise operators outweighs the faster integer convolutions; at 480x640 INT8 becomes 1.24 times faster (27.76 ms versus 34.50 ms). Structured pruning yields real speedups at both resolutions: the 50% RAMC model runs in 5.39 ms and 23.47 ms (1.24x and 1.47x), and it is simultaneously more accurate and more robust than the uncompressed model. Removing 75% of the hidden channels brings no further speedup (5.65 ms and 23.39 ms) because memory-bound operators such as depthwise convolutions, gating and upsampling then dominate the run time, while robustness drops by 2.7 mCorr points. For this class of lightweight multimodal networks the best accuracy-robustness-latency trade-off is therefore obtained by moderate, robustness-aware structured pruning; aggressive pruning and low-bit quantization pay off only when the hardware executes them natively and the robustness losses are controlled with the condition-aware criteria of RAMC.

Table 5 Deployment with ONNX Runtime on One CPU Thread

Model	Size (KB)	ms 240x320	ms 480x640	Clean	Night	mCorr
Full FP32	1658	6.69	34.50	35.34	34.13	28.71
Full INT8 (clean cal.)	677	8.13	27.76	35.88	34.88	29.17
Full INT8 (CA cal.)	677	8.86	31.57	35.87	34.84	29.14
P-50% FP32	955	5.39	23.47	35.85	35.01	29.13
P-50% INT8 (CA)	467	6.90	26.33	36.34	35.55	29.52
P-75% FP32	605	5.65	23.39	33.92	32.48	26.46
P-75% INT8 (CA)	362	6.41	22.35	34.31	32.95	26.83

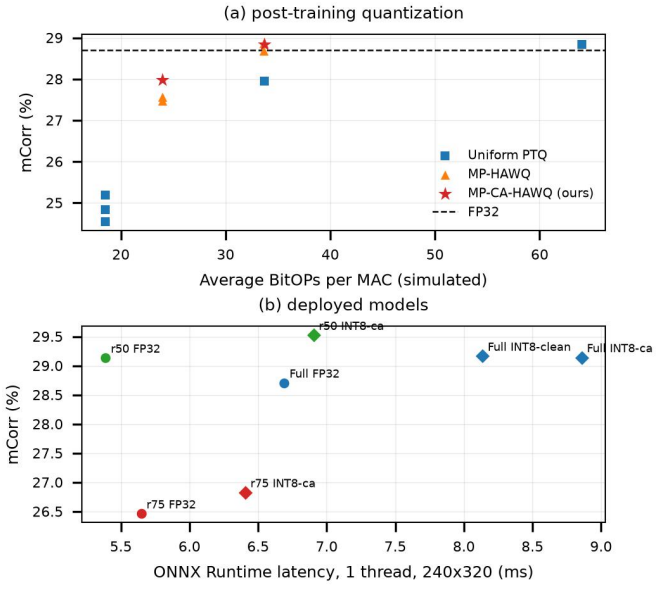


Figure 4 Robustness Versus Cost

Note: (a) Mean corrupted mIoU versus simulated BitOPs for uniform and mixed-precision quantization. (b) Mean corrupted mIoU versus measured ONNX Runtime latency (240x320, one thread) for deployed FP32 and INT8 models.

4.7 Ablation Study

Table 6 isolates the design choices. *Calibration data and ranges:* the condition-aware pool helps, but choosing activation ranges by a worst-case criterion over condition groups hurts (24.54 versus 25.19 mCorr for W4A4), because degenerate groups such as modality drop dominate the maximum and enlarge the clipping ranges; RAMC therefore applies the minimax principle to bit allocation and keeps MSE range estimation. *Aggregation:* minimax and mean

aggregation of the per-condition Hessian sensitivities perform similarly (27.99 versus 27.96 mCorr); the minimax variant gives higher clean and night accuracy and the mean variant fewer flips, and both surpass HAWQ with the same calibration data (27.48 mCorr). *Pruning:* with the L1 criterion, the RAMC fine-tuning objective alone adds 0.41 mCorr points over distillation, and the condition-aware Taylor criterion adds a further 0.82. Removing gate consistency and margin weighting leaves mCorr and the flip rate essentially unchanged (26.49 versus 26.46 and 5.90% versus 5.96%) but lowers clean and night mIoU (33.84 and 32.33 versus 33.92 and 32.48) and increases the gate drift under fog (-0.014 versus -0.009); the two terms therefore improve night performance and preserve the modality-reliability behaviour at no cost in average robustness.

Table 6 Ablation Study (mIoU and Harmful Flips, %)

Variant	Clean	Night	mCorr	Hc (%)
W4A4: clean data + MSE range	32.54	30.33	24.83	7.68
W4A4: CA data + MSE range	33.12	30.39	25.19	7.82
W4A4: CA data + minimax range	32.25	29.98	24.54	8.22
MP-24b, HAWQ	35.63	33.98	27.48	4.94
MP-24b, CA-HAWQ (mean)	35.88	34.21	27.96	3.56
MP-24b, CA-HAWQ (minimax)	36.07	34.36	27.99	3.88
P-75%: L1 + KD	32.72	31.31	25.23	6.46
P-75%: L1 + RAMC fine-tuning	33.09	31.77	25.64	5.97
P-75%: CA-Taylor + RAMC-FT w/o gate & margin	33.84	32.33	26.49	5.90
P-75%: CA-Taylor + RAMC-FT (full)	33.92	32.48	26.46	5.96

Note: Rows 4-6: mixed precision at 24 BitOPs per MAC with CA data and MSE ranges; rows 7-10: 75% pruning.

5 CONCLUSION

This paper examined model compression for RGB-thermal perception under low visibility, occlusion and sensor noise. Compression errors are not uniformly distributed: they concentrate on low-margin, boundary and rare-class pixels and on pixels that rely on the degraded sensor, and compressed models partly lose the modality-reliability behaviour of their teachers. RAMC addresses these failures by introducing operating conditions into every compression decision. Condition-aware Hessian sensitivity with minimax bit allocation improves night mIoU by 1.2 points and reduces harmful flips by 17% over HAWQ at the same BitOPs; condition-aware pruning with paired clean-to-degraded distillation and gate consistency improves mean corrupted mIoU by 1.2 points over distillation and, at 50% pruning, yields a model that is faster, more accurate and more robust on average than the uncompressed one; and distilled students must be recalibrated, for which a single temperature fitted on held-out data nearly halves the calibration error under unseen corruptions. Deployment measurements show that robustness-aware structured pruning gives the most favourable accuracy-robustness-latency trade-off on CPUs, while low-bit quantization requires native hardware support to translate BitOPs into speed. Future work will extend the calibration pool with haze-like degradations and apply RAMC to LiDAR-camera detection.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Ha Q, Watanabe K, Karasawa T, et al. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017.

[2] Tang L, Yuan J, Zhang H, et al. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. Information Fusion, 2022, 83-84: 79-92.

[3] Sun Y, Zuo W, Liu M. RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. IEEE Robotics and Automation Letters, 2019, 4(3): 2576-2583.

[4] Zhang J, Liu H, Yang K, et al. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(12): 14679-14694.

[5] Shin U, Lee K, Kweon I S, et al. Complementary random masking for RGB-thermal semantic segmentation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2024.

[6] Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. In: Proceedings of the International Conference on Learning Representations (ICLR), 2019.

[7] Michaelis C, Mitzkus B, Geirhos R, et al. Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv:1907.07484, 2019.

[8] Dong Y, Kang C, Zhang J, et al. Benchmarking robustness of 3D object detection to common corruptions in autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 1022-1032. DOI: 10.1109/CVPR52729.2023.00105.

[9] Kong L, Liu Y, Li X, et al. Robo3D: Towards robust and reliable 3D perception against corruptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 19994-20006.

[10] Sakaridis C, Dai D, Van Gool L. Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision, 2018, 126(9): 973-992.
[11] Hooker S, Courville A, Clark G, et al. What do compressed deep neural networks forget? arXiv:1911.05248, 2019.
[12] Diffenderfer J, Bartoldson B, Chaganti S, et al. A winning hand: Compressing deep networks can improve out-of-distribution robustness. In: Advances in Neural Information Processing Systems (NeurIPS), 2021, 34.
[13] Jacob B, Kligys S, Chen B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 2704-2713.
[14] Dong Z, Yao Z, Arfeen D, et al. HAWQ-V2: Hessian aware trace-weighted quantization of neural networks. In: Advances in Neural Information Processing Systems (NeurIPS), 2020, 33.
[15] Liu J, Niu L, Yuan Z, et al. PD-Quant: Post-training quantization based on prediction difference metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 24427-24437.
[16] Li H, Kadav A, Durdanovic I, et al. Pruning filters for efficient ConvNets. In: Proceedings of the International Conference on Learning Representations (ICLR), 2017.
[17] Molchanov P, Mallya A, Tyree S, et al. Importance estimation for neural network pruning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 11264-11272.
[18] Fang G, Ma X, Song M, et al. DepGraph: Towards any structural pruning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023: 16091-16101.
[19] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.
[20] Sun S, Ren W, Li J, et al. Logit standardization in knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024: 15731-15740.
[21] Zhang Q, Wu H, Zhang C, et al. Provable dynamic fusion for low-quality multimodal data. In: Proceedings of the International Conference on Machine Learning (ICML), PMLR, 2023, 202: 41753-41769.
[22] Ma H, Zhang Q, Zhang C, et al. Calibrating multimodal learning. In: Proceedings of the International Conference on Machine Learning (ICML), PMLR, 2023, 202: 23429-23450.
[23] Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks. In: Proceedings of the International Conference on Machine Learning (ICML), PMLR, 2017, 70: 1321-1330.
[24] Joy T, Pinto F, Lim S N, et al. Sample-dependent adaptive temperature scaling for improved calibration. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2023, 37(12): 14919-14926.
[25] Qin D, Leichner C, Delakis M, et al. MobileNetV4: Universal models for the mobile ecosystem. In: Proceedings of the European Conference on Computer Vision (ECCV), 2024, 15098: 78-96.