# Selective Focus: Investigating Semantics Sensitivity in Post-training Quantization for Lane Detection

**Yunqian Fan[1,2], Xiuying Wei[2], Ruihao Gong[2,4], Yuqing Ma[3, 4], Xiangguo Zhang[2], Qi Zhang[2], Xianglong Liu[4*]**

[1]School of Information Science and Technology, ShanghaiTech University
[2]SenseTime Research
[3]Institute of Artificial Intelligence, Beihang University, Beijing, China
[4]State Key Laboratory of Complex & Critical Software Environment, Beihang University, Beijing, China
fanyq2022@shanghaitech.edu.cn, weixiuying966@gmail.com,
{gongruihao, zhangxiangguo, zhangqi3}@sensetime.com, {mayuqing, xlliu}@buaa.edu.cn

## Abstract

Lane detection (LD) plays a crucial role in enhancing the L2+ capabilities of autonomous driving, capturing widespread attention. The Post-Processing Quantization (PTQ) could facilitate the practical application of LD models, enabling fast speeds and limited memories without labeled data. However, prior PTQ methods do not consider the complex LD outputs that contain physical semantics, such as offsets, locations, etc., and thus cannot be directly applied to LD models. In this paper, we pioneeringly investigate **semantic sensitivity** to post-processing for lane detection with a novel Lane Distortion Score. Moreover, we identify two main factors impacting the LD performance after quantization, namely intra-head sensitivity and inter-head sensitivity, where a small quantization error in specific semantics can cause significant lane distortion. Thus, we propose a Selective Focus framework deployed with Semantic Guided Focus and Sensitivity Aware Selection modules, to incorporate post-processing information into PTQ reconstruction. Based on the observed intra-head sensitivity, Semantic Guided Focus is introduced to prioritize foreground-related semantics using a practical proxy. For inter-head sensitivity, we present Sensitivity Aware Selection, efficiently recognizing influential prediction heads and refining the optimization objectives at runtime. Extensive experiments have been done on a wide variety of models including keypoint-, anchor-, curve-, and segmentation-based ones. Our method produces quantized models in minutes on a single GPU and can achieve 6.4% F1 Score improvement on the CULane dataset. Code and supplementary statement can at found on https://github.com/PannenetsF/SelectiveFocus.

## Introduction

Deep neural networks have recently sparked great interest in autonomous driving. As a fundamental component in autonomous driving, lane detection (LD) is fundamental for high-level functions such as lane departure warning, lane departure prevention, etc. Lane detection (Qin, Wang, and Li 2020; Wang et al. 2022b; Tabelini et al. 2021) has garnered significant attention and undergone in-depth research, leading to substantial advancements. However, LD models are often required to run on edge devices within limited sizes, necessitating quantization and introducing formidable challenges to detection performance.

There are two prevalent quantization techniques: Quantization Aware Training (QAT) e.g. (Choi et al. 2018; Esser et al. 2019; Bhalgat et al. 2020; Jain et al. 2020) and Post-training Quantization (PTQ) e.g. (Hubara et al. 2020; Wu et al. 2020a; Wei et al. 2022; Wang et al. 2022a). Though QAT can often yield promising performance, it requires a longer training duration and the whole labeled dataset, raising computation costs and safety concerns. In contrast, PTQ methods have attracted wide attention from both industry and academia due to their speed and label-free nature. Recently, some PTQ methods (Nagel et al. 2020; Li et al. 2021; Wei et al. 2022) propose to tune the weight by reconstructing the original outputs, bringing better performance.

LD models typically regress semantic outputs with physical meanings such as offsets, locations, and angles, and employ complex post-processing to handle these outputs. Notably, the sensitivity of these semantic outputs to post-processing varies, with certain elements having the potential to induce significant lane deformation even with minor perturbations. Prior PTQ approaches employing direct reconstruction methods on feature maps treat all outputs uniformly, overlooking post-processing information.

In this paper, we first propose the **semantic sensitivity** in lane detection models and introduce a Lane Distortion Score to measure the quantization distortion between the original LD model and the corresponding quantized counterpart. Subsequently, we investigate these sensitivities from two perspectives, namely the intra-head sensitivity and the inter-head sensitivity. Specifically, the intra-head sensitivity highlights the heightened sensitivity of a limited number of foreground (lane) regions to quantization noise during post-processing, while the inter-head sensitivity indicates the varying degrees of sensitivity to the quantization of different semantic heads over time, as shown in Figure 4.

To address the sensitivity problems above, we propose a Selective Focus framework to alleviate the semantic sensitivity in post-training quantization for LD models, enhancing the performance. The proposed framework is deployed with a Semantic Guided Focus module and a Sensitivity

*Corresponding author.

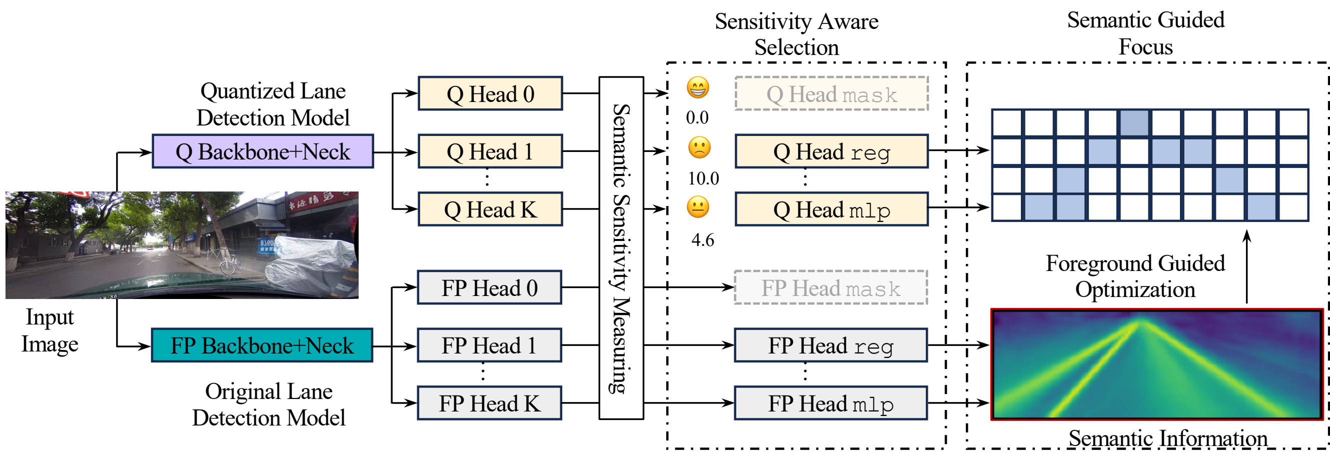


Figure 1: The framework of Selective Focus. Three modules are designed to mine the semantics sensitivity in the post-training quantized lane detection. Semantic Sensitivity Measuring measures the semantic sensitivity quantitatively; Sensitivity Aware Selection adapts the optimization objectives according to dynamic sensitivity. Semantic Guided Focus enables PTQ to focus on the foreground with a practical proxy.

Aware Selection module, respectively targeting the intra-head sensitivity and the inter-head sensitivity. First, the Semantic Guided Focus generates practical proxies of masks from semantics, enhancing the precision of foreground lanes in post-processing. This method guides PTQ to optimize these pivotal areas. Furthermore, the Sensitivity Aware Selection module refines optimization objectives by querying efficiently the real-time sensitivity of each head through our Lane Distortion Score across heads. The proposed framework could tune the models efficiently by introducing the semantic information of the post-process to the optimization implicitly.

To the best of our knowledge, our work is the first to identify the role of semantic sensitivity in PTQ for lane detection models, and we hope it could offer new insight to the community. Extensive experiments on the widely-used CULane dataset and various leading methods validate the effectiveness and efficiency of the proposed Selective Focus framework. In summary, our contributions are listed as follows:

- We introduce the concept of semantics sensitivity in post-processing quantization for lane detection, proposing the Lane Distortion Score metric. Our Selective Focus framework, composed of the Semantic Guided Focus and Sensitivity Aware Selection modules, addresses both intra-head and inter-head sensitivities.
- Considering the intra-head semantics sensitivity, the Semantic Guided Focus module generates practical proxy masks from semantics and thus guides PTQ to optimize these pivotal areas.
- Handling inter-head semantics sensitivity, the Sensitivity Aware Selection module efficiently adjusts optimization objectives based on each head's real-time sensitivity measured by our Lane Distortion Score.
- Our empirical tests across datasets, models, and quantization setups endorse our approach's efficacy. Notably, under the 4-bit setup, performance gains exceed up to 6.4%, on benchmark models with a 6x acceleration.

## Related Works

### Lane Detection Models

The LD task aims to produce lane representations in the given images. Despite the different methods, they all try to set the foreground (lanes) apart from the background. Nowadays the models generally use Convolutional Neural Networks (CNNs) to extract lane features, which can be divided into keypoint-based, anchor-based, segmentation-based, and parameterized-curve-based methods.

**Keypoint-based Methods** predict the mask of lane points and regress them to the real location on the corresponding lanes. CondLaneNet (Liu et al. 2021a) regresses offset between adjacent key points, while GANet (Wang et al. 2022b) regresses offset between each keypoint to the start point of its lane. **Anchor-based Methods** model lanes as pre-defined pairs of start point and angle, and then regress lanes among them. LaneATT (Tabelini et al. 2021) proposes an anchor attention module to aggregate global information for the regression. CLRNet (Zheng et al. 2022) refines the proposals with features at different scales. **Segmentation-based Methods** predict the mask of all the lanes on the image and then cluster them into different lanes. SCNN (Pan et al. 2018) adopts slice-by-slice convolution modules to aggregate surrounding spatial information. RESA (Zheng et al. 2021) further extends the mechanism to aggregate global spatial to every pixel. **Curve-based Methods** model lanes as singular curves, rather than sets of discrete points. For instance, LSTR (Liu et al. 2021b) predicts the parameters for cubic curves and BézierLaneNet (Feng et al. 2022) predicts for Bézier curves.

### Post-traning Quantization

Quantization is widely used in deep learning model deployment to substantially cut down memory and computation re-

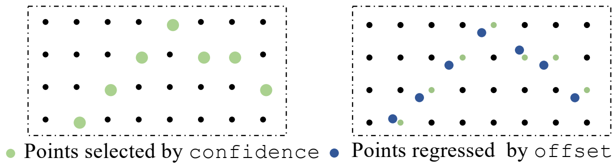


Figure 2: Example of confidence and semantics in post-process. (a) The confidence is used to predict whether a keypoint is at a certain location. (b) The offset is regressed to the shift between the real keypoint and its downscaled grid. Offset is only valid in the indices with positive confidence prediction.

quirements during inference, which is required by the LD models. Compared to QAT (Jacob et al. 2018; Gong et al. 2019; Jain et al. 2020; Esser et al. 2019; Bhalgat et al. 2020) which requires large GPU effort and the whole dataset, PTQ methods have sparked great popularity these days due to their speed and label-free property.

Common PTQ methods like OMSE (Choukroun et al. 2019) and ACIQ (Banner, Nahshan, and Soudry 2019) often identify quantization parameters to minimize the quantized error for tensors, requiring a few batches of forward passes. More recently, some methods have evolved to slightly tune the weights and reconstruct the original outputs. (Wu et al. 2020b) improves the accuracy by setting a well-defined target for the face recognition task. AdaRound (Nagel et al. 2020) initially proposes that adjusting the weight within a small space can be beneficial, and their layer-wise output reconstruction can yield more favorable results with only a marginal increase in optimization time. Building on this, BRECQ (Li et al. 2021) suggests that the outputs of each layer still exhibit some disparity from the final outputs and thus proposes adopting a block-wise reconstruction scheme. Later, QDrop (Wei et al. 2022) investigates the activation quantization under this setting and introduces random activation quantization dropping during tuning, which benefits the performance.

We also opt for model tuning through reconstruction. Nevertheless, prior techniques haven't been applied to lane detection models with multiple heads and complex post-processing functions. We discover that directly reconstructing feature maps for these models overlooks the important post-processing information, ultimately leading to sub-optimal solutions.

# Preliminaries

## Notation

In the context of lane detection models, it's important to note that each head $i$ encompasses two distinct functions: $\mathcal{S}^i(\cdot)$ and $\mathcal{C}^i(\cdot)$. The former function yields outputs with physical significance, which we refer to as **semantics**. These semantics are linked to physical attributes such as distance and angle. The latter function, $\mathcal{C}^i(\cdot)$, produces confidence outputs for each head's semantics. An illustration of how the post-process deals with confidence and semantics is shown in Figure 2.

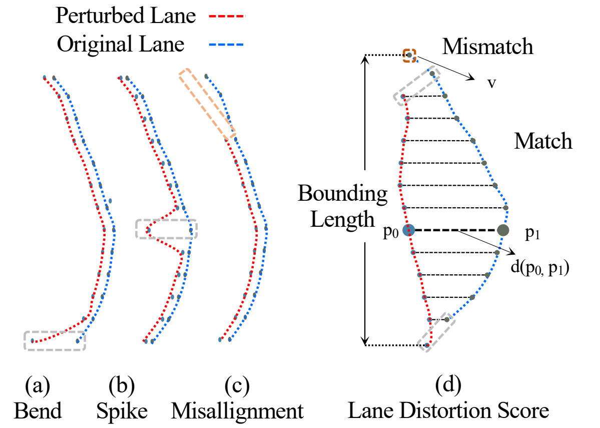


Figure 3: Common types of lane distortion caused by slight perturbation and the proposed Lane Distortion Score. (a) Bend: unexpected shifts at the terminals; (b) Spike: shifts in the middle; (c) Misalignment: missing or extra lane points. Any kind of lane distortion can be represented as a combination of those three types of distortion. (d) We measure the distortion between lanes with two types of point relationships: match and mismatch.

Moreover, let $\mathbf{x}$ represent the vectors of unlabeled data from the calibration dataset $\mathcal{D}$. We use the symbol $\odot$ to indicate element-wise multiplication. Finally, $\hat{\mathcal{S}}$ generates new semantics derived from quantized models.

## Problem Definition

Tuning-based PTQ, as mentioned in the last section, focuses on minimizing the task quantization loss as opposed to minimizing local distance, given by:

$$\min_{\mathbf{w}} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \left( \sum_i \left\| \hat{\mathcal{S}}^i(\mathbf{x}) - \mathcal{S}^i(\mathbf{x}) \right\|_F^2 + \left\| \hat{\mathcal{C}}^i(\mathbf{x}) - \mathcal{C}^i(\mathbf{x}) \right\|_F^2 \right), \tag{1}$$

where the first term corresponds to the reconstruction of the fully pixel-wise semantic outputs, while the second term pertains to the reconstruction of the confidence values. Compression methods (Nagel et al. 2020; Li et al. 2021) optimizes the above equation towards layer-wise and block-wise approximation.

However, such an optimization is not suitable in LD models due to their intricate post-processing steps and semantics rooted in physical interpretation. In the next section, we identify the importance of semantic sensitivity if the post-process. Thus, neglecting the valuable insights offered by post-processing in the optimization objective would ultimately yield less favorable outcomes.

# Method

In this section, we reveal an important factor that significantly impacts PTQ performance in lane detection models: the semantic sensitivity to post-processing, which has been overlooked by other PTQ research before. Then, comprehensive investigations are conducted from both intra-head and inter-head aspects. Building on sensitivity observations within and across heads, we propose a Selective Focus framework including two novel modules, Semantic Guided Focus, and Sensitivity Aware Selection, to allocate appropriate attention to different semantics. Our framework implicitly introduces the post-processing information into quantization optimization. The pipeline is depicted in Figure 1.

## Semantic Sensitivity

**Sensitivity to post-process** Owing to the complexity of optimizing the post-process, prevailing PTQ approaches are confined to adjusting model parameters to align quantized head outputs with their full-precision counterparts (Equation 1) without considering the post-process. Nevertheless, we find the importance of the post-process procedure in ensuring accurate lane generation. Disregarding post-processing information during the quantization optimization phase can result in pronounced distortions, including abrupt bends, spikes, and misalignments, even with a marginal value of (Equation 1), as illustrated in Figure 3.

Motivated by this, we propose to investigate semantic sensitivity, where some outputs of heads can be so important for later post-process that small quantization errors of them can cause severe lane distortion. Incorporating information from the post-process into our model optimization would pave the way for more effective semantic reconstruction.

**Lane Distortion Score** To study the sensitivity, a quantitative evaluation of lane distortion becomes imperative. Given the frequently localized deviations in lanes induced by quantization (as depicted in Figure 3), we abstain from employing the conventional Intersection over Union (IoU) metric (Pan et al. 2018; Feng et al. 2022) which overlooks these local distortions. In response, a direct but effective metric is introduced, which measures the shifts of all points from the perturbed lane to the original one, as shown in Figure 3. The score of matched points is their distance, and the score of mismatched points is a fixed penalty score $v$. Concretely, our devised metric first matches points between perturbed and normal lanes (set $\mathbb{M}$), then calculates the distance for matched points ($d(p_0, p_1)$) and the penalty from mismatched ones:

$$score = \sum_{(p_0,p_1)\in\mathbb{M}} \frac{d(p_0, p_1)}{b(\mathbb{M})} + nv, \tag{2}$$

where the bounding length normalization ($b(\mathbb{M})$) is applied to accommodate lanes of varying lengths and $n$ is the number of mismatched points.

Equipped with the Lane Distortion Score, we can compute the distortion of lanes under perturbation, which strongly supports quantitively analysis of semantic sensitivity to post-process and further method design.

## Semantic Guided Focus

This section explores intra-head semantic sensitivity, where we observe that semantics within each head associated with the foreground region in post-processing play a more significant role and propose a method called Semantic Guided

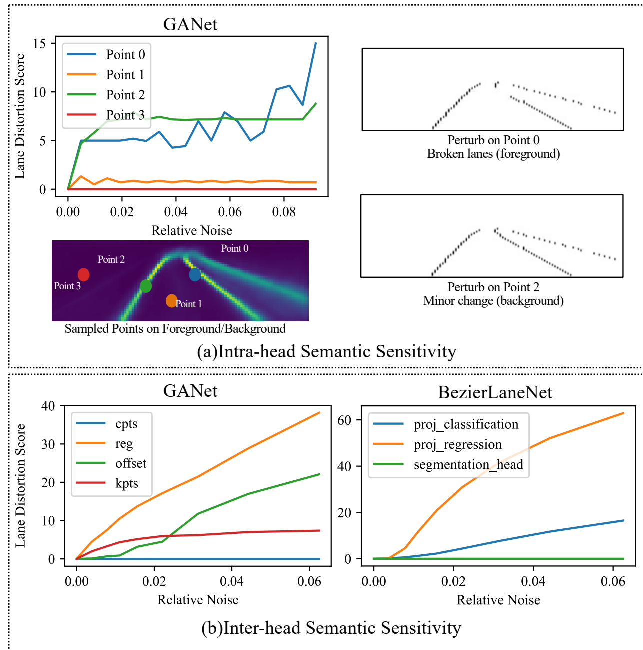


Figure 4: Illustration of semantic sensitivity. The intra-head semantic sensitivity shows that the foreground is sensitive to perturbation while the background is not. The inter-head semantic sensitivity indicates that heads are of different importance for post-processing.

Focus. By leveraging confidence outputs, we can potentially discern between semantics linked to the foreground and background, which enables us to prioritize the former and obtain an improved optimization in a PTQ setting without labels.

**Intra-head Semantic Sensitivity** Considering the relationship between semantics (head outputs) and the post-process, some semantics pertain to the foreground region in the post-process, while others correspond to the background. Also, it's the distortion of the foreground region (lane) that matters. Consequently, we argue that different pixels within each head exhibit distinct sensitivities. To verify, we leverage the Lane Distortion Score of each pixel by adding the same magnitude noise to them. The results showcased in Figure 4 conclusively demonstrate that injecting noise into entries associated with the foreground region can result in more severe lane deformations. Given these findings, allocating equal attention is not reasonable during optimization.

Furthermore, it's worth noting that the number of entries about the foreground is significantly fewer than those related to the background region, which further distracts previous techniques to focus on crucial positions. Therefore, we are motivated to enhance the accurate expression of pixels tied to the foreground region and suppress that of the background.

**Intra-head Sensitivity Focus** Motivated by the findings, the core idea is to distinguish whether each element within each head will be used in the foreground region (the lane) or background.

(1) **Reconstruction on semantic:** We first focus on the semantic term in Equation 1 and introduce a masking function $\mathcal{M}$ to achieve the distingishment. $\mathcal{M}$ removes the error term associated with the background and retains the elements for the foreground region. Then, we incorporate them with an element-wise product, and the optimization objective on semantics becomes:

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \sum_i \left\| \mathcal{M}^i(\mathbf{x}) \cdot (\hat{\mathcal{S}}^i(\mathbf{x}) - \mathcal{S}^i(\mathbf{x})) \right\|_F^2 . \tag{3}$$

However, due to the absence of lane annotations under the PTQ setting, it is unrealistic to identify exact elements tied to the foreground. Fortunately, we find an upper bound of Equation 3 of the models' confidence output. Here we give the theoretical finding of this upper bound:

**Theorem 1.** *Given $\mathcal{M}$ representing a matrix function that discerns elements associated with foreground or background regions, and $\mathcal{C}$ denoting the confidence function of FP models, offering confidence scores for semantics linked with the foreground, the following inequation stands:*

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \sum_i \left\| \mathcal{M}^i(\mathbf{x}) \cdot (\hat{\mathcal{S}}^i(\mathbf{x}) - \mathcal{S}^i(\mathbf{x})) \right\|_F^2 \\ \leq\ &\mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \sum_i \left\| \mathcal{C}^i(\mathbf{x}) \odot (\hat{\mathcal{S}}_i(\mathbf{x}) - \mathcal{S}^i(\mathbf{x})) \right\|_F^2 \end{aligned} \tag{4}$$

Detailed proof of this theorem can be found in (Fan et al. 2024). With this theorem, the semantic-related optimization target can be transformed to:

$$\min_{\mathbf{w}} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \sum_i \left\| \mathcal{C}^i(\mathbf{x}) \odot (\hat{\mathcal{S}}_i(\mathbf{x}) - \mathcal{S}^i(\mathbf{x})) \right\|_F^2 . \tag{5}$$

The theorm means that we can leverage the model output $\mathcal{C}(\mathbf{x})$ as a practical proxy of the annotation mask. In intuition, the knowledge from the well-trained models can be represented in its output, thus the model could generate a mask similar to the annotation. We further assume the mask get from the model follows a binomial distribution paramterized by the confidence output, then the total expectation turns to the result in the theorem. With the upper bound from it, optimization can be more tractable.

(2) **Reconstruction on confidence:** Last, we incorporate the reconstruction loss on confidence values on each head. To prevent elements associated with the background region turns into the foreground, especially under large quantization noise, we propose to enhance the alignment of confidence outputs by adopting a new parameter $\lambda$ with $\lambda > 1$. This heightened penalty on confidence outputs helps retain the fidelity of background-related pixels while enabling a concentrated focus on those linked to the foreground domain via the first term.

$$\begin{aligned} \min_{\mathbf{w}} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \sum_i ( &\left\| \mathcal{C}^i(\mathbf{x}) \odot (\hat{\mathcal{S}}_i(\mathbf{x}) - \mathcal{S}^i(\mathbf{x})) \right\|_F^2 + \\ &\lambda \left\| \hat{\mathcal{C}}^i(\mathbf{x}) - \mathcal{C}^i(\mathbf{x}) \right\|_F^2 ). \end{aligned} \tag{6}$$

Algorithm 1: Sensitivity Aware Selection

**Require:** Semantic head set $\mathbb{H}$, hyper parameter $k$, and calibration dataset $\mathcal{D}$. Semantic function $\mathcal{S}(\cdot)^i$ and confidence functions $\mathcal{C}(\cdot)^i$ on head $i$ on the FP model. Quantized version $\hat{\mathcal{S}}(\cdot)^i$ and $\hat{\mathcal{C}}(\cdot)^i$.
**for** $\mathbf{x} \in \mathcal{D}$ **do**
  **for** $i \in \mathbb{H}$ **do**
    Calculate full precision lanes $L$ from $\{\mathcal{S}(\mathbf{x})^i\}$
    Perturb the model by replacing $\mathcal{S}^i(\mathbf{x})$ with $\hat{\mathcal{S}}_i(\mathbf{x})$.
    Calculate noised lanes $\hat{L}$ from $\hat{\mathcal{S}}(\mathbf{x})^i$.
    Compute the $s$ of $(L, L')$ with Equtaion 2
    Update score of $i$-th head $Score_i \leftarrow Score_i + s$
  **end for**
**end for**
Sort semantic heads by $Score$.
**return** Top-$k$ of semantic head set $\mathbb{H}$

The improved Semantic Guided Focus objective indicates a simple yet elegant principle: a well-trained model inherently possesses the capacity to instruct itself. By utilizing model outputs for both mask estimation and background suppression in PTQ, the optimization process becomes foreground-oriented, leading to more efficient semantic alignment.

## Sensitivity Aware Selection

We also delve into inter-head sensitivity, where we observe that specific heads are more sensitive to post-processing. This insight leads us to introduce the Sensitivity Aware Selection method, which dynamically and efficiently selects the most influential heads during PTQ reconstruction.

**Inter-head Semantic Sensitivity** Considering the diverse roles played by distinct heads in post-processing, we further investigate semantic sensitivity across multiple heads. By injecting noise into each head and calculating our distortion score, the sensitivity curve in Figure 4 is obtained. It can be clearly seen that under the same perturbation magnitude, certain heads like `proj regression` in BezierLane, exhibit notably higher Lane Distortion scores and of course correspond to severely distorted lanes, compared to others. This discrepancy highlights the considerable variation in sensitivity across different heads, which encourages us to discriminate heads and thus implicitly incorporate post-process information during optimization.

**Inter-head Sensitivity Selection** To handle varied semantic sensitivity among heads, we introduce the Sensitivity Aware Selection technique, which efficiently and adaptively selects those sensitive heads during optimization. The algorithm is formulated below and its procedure can be found in the Algorithm 1.

(1) Head selection: The noticeable differences in sensitivity prompt us to focus on the more sensitive heads, optimizing them more effectively. Naturally, given the estimated quantization noise level of each head, we can compute their own Lane Distortion Score. By ranking these scores and

| Bits | Method | Keypoint-based | | | | Curve-based | | | | Achor-based | | Segmentation-based | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Full Precision | Model Baseline | CondLaneNet | | GANet | | LSTR | | BézierLaneNet | | LaneATT | | SCNN | | RESA | |
| | | Small | Mid | Small | Mid | Small | Mid | Small | Mid | Small | Mid | Small | Mid | Small | Mid |
| | | 78.14 | 78.74 | 78.79 | 79.39 | 68.78 | 72.47 | 73.66 | 75.57 | 74.45 | 75.04 | 72.19 | 72.70 | 72.90 | 73.66 |
| W8A8 | ACIQ | 77.95 | 78.58 | 78.58 | 79.21 | 67.50 | 72.14 | 73.43 | 75.36 | 74.34 | 74.56 | 72.03 | 72.55 | 72.76 | **73.64** |
| | QDrop | 78.04 | 78.77 | **78.70** | **79.33** | 68.40 | 72.38 | 73.63 | 75.49 | 74.33 | 74.88 | 72.04 | 72.55 | **72.80** | 73.61 |
| | Ours | **78.10** | **78.90** | 78.53 | 79.30 | **68.58** | **72.40** | **73.63** | **75.50** | **74.38** | **75.01** | **72.33** | **72.69** | 72.53 | 73.49 |
| W8A4 | ACIQ | 58.63 | 37.67 | 5.38 | 20.18 | 47.63 | 12.49 | 23.79 | 4.77 | 54.20 | 0.64 | 62.63 | 49.26 | 54.16 | 49.62 |
| | OMSE | 69.74 | 64.29 | 69.52 | 54.10 | 55.51 | 58.12 | 62.04 | 60.57 | 64.54 | 0.90 | 65.35 | 60.53 | 66.59 | 65.93 |
| | AdaRound | 67.11 | 63.64 | 39.97 | 18.14 | 51.41 | 54.11 | 56.66 | 58.55 | 64.01 | 0.96 | 65.87 | 63.65 | 59.17 | 62.78 |
| | BRECQ | 73.61 | 74.06 | 74.37 | 75.04 | 57.11 | 63.32 | 62.02 | 65.18 | 66.47 | 0.04 | 66.05 | 63.67 | 66.40 | 65.47 |
| | QDrop | 74.76 | 75.49 | 75.77 | 75.56 | 60.34 | 65.25 | 64.48 | 66.91 | 66.58 | 0.06 | 66.85 | 64.83 | 67.27 | 67.54 |
| | Ours | **75.56** | **75.74** | **76.32** | **76.51** | **63.14** | **68.15** | **68.98** | **70.01** | **69.85** | **34.53** | **69.61** | **69.51** | **69.46** | **70.60** |
| W4A4 | ACIQ | 53.96 | 20.84 | 1.54 | 9.02 | 1.47 | 2.37 | 14.10 | 8.82 | 50.65 | 0.32 | 34.47 | 23.70 | 35.56 | 15.45 |
| | OMSE | 63.64 | 55.06 | 49.96 | 37.21 | 1.79 | 17.02 | 52.38 | 46.03 | 62.34 | 0.39 | 51.49 | 49.67 | 57.07 | 50.83 |
| | AdaRound | 20.35 | / | / | / | 20.69 | 7.95 | 50.70 | 48.60 | 34.53 | 0.00 | 6.56 | 0.03 | 68.36 | 64.57 |
| | BRECQ | 74.10 | 75.80 | 75.67 | 75.89 | 30.83 | 50.09 | 66.96 | 70.30 | 68.69 | 28.34 | 54.34 | 52.30 | 67.70 | 69.48 |
| | QDrop | 74.41 | **76.29** | **76.76** | **76.50** | 23.95 | 53.87 | 67.65 | **71.16** | 68.97 | 0.64 | 61.70 | 64.57 | 67.59 | 69.98 |
| | Ours | **74.68** | 75.48 | 76.31 | 76.26 | **34.65** | **60.56** | **68.37** | 70.00 | **69.19** | **37.59** | **68.16** | **68.27** | **69.31** | **70.01** |

Table 1: F1-score performance comparison among different quantization algorithms and models. W8A8 means the weight and activation are all quantized into 8 bits, and so does W4A4 and W8A4. Our method achieves superior performance under most settings.

then selecting the top-$k$ sensitive heads, a new reconstruction loss is constructed. This approach ensures that the optimization process is cognizant of the semantic sensitivity to post-processing, leading to a better-tuned model.

(2) Adaptive selection: Moreover, the optimization process is aimed at minimizing the discrepancy between original and quantized semantics, where we recognize that the quantization loss for individual heads can evolve, leading to dynamic changes in their quantization noise levels and thus sensitivity ranking. Consequently, our selection of top-$k$ sensitive heads must adapt accordingly. In practical terms, we can reassess the Lane Distortion Scores and repeat the aforementioned step at fixed intervals of iterations. However, this repetition could impose a considerable time overhead during tuning.

(3) Efficient adaptive selection: To accelerate it, we propose to apply a pre-processing technique, which first employs the Monte-Carlo method to sample diverse noise levels and derive corresponding Lane Detection scores for each head, then interpolates sampled points from a continuous noise-score curve. This approach empowers us to gauge the semantic sensitivity of different heads based on their respective curves using their current quantization loss as queries, which introduces negligible computational burden during the optimization process.

Based on varied semantic sensitivities across heads, we adeptly and dynamically select the most sensitive heads. This implicit inclusion of post-processing guidance leads to a better-optimized model.

## Experiments

Extensive experiments are conducted to prove the effectiveness of the Selective Focus framework. We first present the experiment setup, and then compare the proposed method with other state-of-the-art PTQ works and the method shows up to 6.4% F1 score improvement. After that, the ablation study of the Selective Focus framework demonstrates the contribution of each component. Finally, we compare the efficiency of the framework with existing PTQ and QAT methods.

### Experiments Setup

We describe the datasets and evaluation protocols, the comparison methods, and the implementation details.

**Datasets and Evaluation** We conduct comprehensive experiments on the CULane dataset and adopt its official evaluation method. CULane contains 88,880 training images and 34,680 test images from multiple scenarios, and the evaluation method provides precision, recall, and F1 score for each scenario. For brevity, we use the F1 score of the whole dataset as the metric. For models, we evaluate LD models in the four major classes: keypoint-, anchor-, segmentation-, and curve-based models, including CondLaneNet (Liu et al. 2021a), GANet (Wang et al. 2022b), LSTR (Liu et al. 2021b), BézierLaneNet (Feng et al. 2022),.LaneATT (Tabelini et al. 2021), SCNN (Pan et al. 2018), and RESA (Zheng et al. 2021).

**Implementation Details** We implement our method based on the PyTorch framework. Weights and activations are both quantized with concrete bits denoted as W/A. Our method is calibrated with 512 unlabeled images on three kinds of quantization bits: W8A8, W8A4, and W4A4. During the optimization, we choose the Adam optimizer with a learning rate set as 0.000025 and adjust weights for 5000 iterations. Because of more computation overhead for layer-wise and block-wise reconstruction, the net-wise reconstruction is adopted here and wins a 6X speedup. Other hyper-parameters including $k$ for Top-$k$ in Sensitivity Aware Selection is kept as 1 for models with two heads and 2 for others, based on our ablation studies.

**Comparison Methods** We implement popular baselines including OMSE (Choukroun et al. 2019), ACIQ (Banner, Nahshan, and Soudry 2019), AdaRound (Nagel et al. 2020), BRECQ (Li et al. 2021), and QDrop (Wei et al. 2022). AdaRound and BRECQ are implemented by leveraging a technical advancement introduced in QDrop, achieving better results for them.

## Main Results

We conduct experiments on CULane datasets. Table 1 shows the results on CULane. With the decreasing activation precision, the proposed method shows advanced performance consistently. For example, the method can achieve more than 3% up to 6.4% F1 score gain under 4-bit activation. As the noise of the semantics used in the post-process would increase significantly and the inter- and intra-head discrepancy would go worse, they may lead to degradation or even failures in methods ignoring it. Also, we note our obvious advantage in the models with specially designed feature aggregation modules, like attention in LSTR, feature flip in BézierLaneNet, and spatial convolution in SCNN and RESA. Those modules usually require information aggregated from the total network, which leads the layer-wise and block-wise methods to a harder situation, while our network-wise framework could take advantage of the cross-layer relationship naturally. Even in hard cases like LaneATT, SCNN, and RESA, the method could outperform others significantly. The proposed Selective Focus framework leverages the post-process information in the PTQ stage and thus tunes the quantized model more efficiently. With advanced performance across different precision configurations and model types, we achieve the new state-of-the-art post-training quantized lane detection and reduce the tuning time by more than 6x.

| Method | Duration (Minutes) | F1 Score |
|---|---|---|
| Ours | 32 | 76.31 |
| w/o Focus | 31 | 73.27 |
| w/o Selection | 29 | 75.30 |
| w/o Focus+Selection | 29 | 72.98 |

Table 2: Abalation study of the proposed framework. Each component contributes to the proposed framework, and the two components are mutually beneficial.

| Network | Method | Duration (Minutes) | F1 Score |
|---|---|---|---|
| CondLaneNet Small | LSQ+ | 1303 | 76.92 |
| | QDrop | 112 | 74.76 |
| | Ours | 33 | 75.56 |
| RESA Small | LSQ+ | 5326 | 69.80 |
| | QDrop | 4378 | 67.59 |
| | Ours | 46 | 69.46 |

Table 3: Comparison between PTQ and QAT methods. The QAT method LSQ+ (gray region) suffers from low computation efficiency, while conventional PTQ methods such as QDrop could save the cost but with an obvious performance drop. In contrast, the proposed method significantly improve the computation efficiency with less performance gap.

## Ablation Study

We first investigate the effect of each component of the proposed framework. Then, we analyze the efficiency of our method, compared to QAT and previous PTQ methods.

**Component Analysis** To elucidate the contributions of individual components in our proposed method, we conducted an ablation study on GANet Small using the W4A4 quantization configuration, as detailed in Table 2. In comparison to the basic network-wise alignment (w/o Focus+Selection), our approach boosts the performance by over 3% in the F1 score. The Semantics Guided Focus emerges as the primary performance driver, underscoring the significance of foreground information and the separate reconstruction benefits for semantics and confidence. While the standalone Sensitivity Aware Selection module enhances the F1 score by a modest 0.3%, its cooperation with Focus amplifies the improvement to 1%, proving the framework's capability to manage semantic sensitivities both within and across heads. Notably, the proposed optimization strategy achieves performance on par with block-wise PTQ, yet maintains a speed akin to network-wise reconstruction.

**Efficiency Analysis** Although QAT comes with cost and privacy concerns, it remains the premier quantization algorithm due to its promising performance. To evaluate the performance and efficiency gap between QAT and PTQ, we performed comparative experiments on CondLaneNet and RESA, utilizing the W8A4 quantization setup. These selected models differ in computational demands, allowing us to thoroughly probe the disparities between QAT and PTQ. Besides, though block-wise PTQ methods, such as QDrop, are much faster than QAT, the storage overhead and processing time for feature maps in intermediate layers are still problems. This is also the reason that we choose the efficient network-wise reconstruction, bringing a 6x speedup.

# Conclusion

This paper sheds light on the post-training quantization in lane detection models leveraging the inherent semantics sensitivity. Our study delves into the essence of semantic sensitivity in the post-process and proposes a novel pipeline for identifying the sensitivity and further leveraging it for optimization. By utilizing the post-processing information, the proposed framework boosts the performance of PTQ for lane detection even with the simplest optimization manner, which could motivate further exploration of the unused information lies in the lane detection models. Future endeavors might encompass efficient embedding of semantics information from post-processing—bypassing intermediate proxies more than the proposed score.

# Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62206010, No.62022009, No. 62306025), and the State Key Laboratory of Software Development Environment (SKLSDE-2022ZX-23).

## References


Banner, R.; Nahshan, Y.; and Soudry, D. 2019. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in Neural Information Processing Systems*, 32.

Bhalgat, Y.; Lee, J.; Nagel, M.; Blankevoort, T.; and Kwak, N. 2020. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 696–697.

Choi, J.; Wang, Z.; Venkataramani, S.; Chuang, P. I.-J.; Srinivasan, V.; and Gopalakrishnan, K. 2018. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*.

Choukroun, Y.; Kravchik, E.; Yang, F.; and Kisilev, P. 2019. Low-bit quantization of neural networks for efficient inference. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 3009–3018. IEEE.

Esser, S. K.; McKinstry, J. L.; Bablani, D.; Appuswamy, R.; and Modha, D. S. 2019. Learned step size quantization. *arXiv preprint arXiv:1902.08153*.

Fan, Y.; Wei, X.; Gong, R.; Ma, Y.; Zhang, X.; Zhang, Q.; and Xianglong, L. 2024. Selective Focus: Investigating Semantics Sensitivity in Post-training Quantization for Lane Detection. https://github.com/PannenetsF/SelectiveFocus.

Feng, Z.; Guo, S.; Tan, X.; Xu, K.; Wang, M.; and Ma, L. 2022. Rethinking efficient lane detection via curve modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17062–17070.

Gong, R.; Liu, X.; Jiang, S.; Li, T.; Hu, P.; Lin, J.; Yu, F.; and Yan, J. 2019. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4852–4861.

Hubara, I.; Nahshan, Y.; Hanani, Y.; Banner, R.; and Soudry, D. 2020. Improving post training neural quantization: Layer-wise calibration and integer programming. *arXiv preprint arXiv:2006.10518*.

Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; and Kalenichenko, D. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2704–2713.

Jain, S.; Gural, A.; Wu, M.; and Dick, C. 2020. Trained quantization thresholds for accurate and efficient fixed-point inference of deep neural networks. *Proceedings of Machine Learning and Systems*, 2: 112–128.

Li, Y.; Gong, R.; Tan, X.; Yang, Y.; Hu, P.; Zhang, Q.; Yu, F.; Wang, W.; and Gu, S. 2021. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*.

Liu, L.; Chen, X.; Zhu, S.; and Tan, P. 2021a. Condlanenet: a top-to-down lane detection framework based on conditional convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3773–3782.

Liu, R.; Yuan, Z.; Liu, T.; and Xiong, Z. 2021b. End-to-end lane shape prediction with transformers. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3694–3702.

Nagel, M.; Amjad, R. A.; Van Baalen, M.; Louizos, C.; and Blankevoort, T. 2020. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, 7197–7206. PMLR.

Pan, X.; Shi, J.; Luo, P.; Wang, X.; and Tang, X. 2018. Spatial as deep: Spatial cnn for traffic scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1.

Qin, Z.; Wang, H.; and Li, X. 2020. Ultra fast structure-aware deep lane detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, 276–291. Springer.

Tabelini, L.; Berriel, R.; Paixao, T. M.; Badue, C.; De Souza, A. F.; and Oliveira-Santos, T. 2021. Keep your eyes on the lane: Real-time attention-guided lane detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 294–302.

Wang, C.; Zheng, D.; Liu, Y.; and Li, L. 2022a. Leveraging Inter-Layer Dependency for Post -Training Quantization. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Wang, J.; Ma, Y.; Huang, S.; Hui, T.; Wang, F.; Qian, C.; and Zhang, T. 2022b. A keypoint-based global association network for lane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1392–1401.

Wei, X.; Gong, R.; Li, Y.; Liu, X.; and Yu, F. 2022. QDrop: randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*.

Wu, D.; Tang, Q.; Zhao, Y.; Zhang, M.; Fu, Y.; and Zhang, D. 2020a. EasyQuant: Post-training quantization via scale optimization. *arXiv preprint arXiv:2006.16669*.

Wu, Y.; Wu, Y.; Gong, R.; Lv, Y.; Chen, K.; Liang, D.; Hu, X.; Liu, X.; and Yan, J. 2020b. Rotation consistent margin loss for efficient low-bit face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6866–6876.

Zheng, T.; Fang, H.; Zhang, Y.; Tang, W.; Yang, Z.; Liu, H.; and Cai, D. 2021. Resa: Recurrent feature-shift aggregator for lane detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4, 3547–3554.

Zheng, T.; Huang, Y.; Liu, Y.; Tang, W.; Yang, Z.; Cai, D.; and He, X. 2022. Clrnet: Cross layer refinement network for lane detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 898–907.