GrooMeD-NMS: Grouped Mathematically Differentiable NMS for Monocular 3D Object Detection

Abhinav Kumar, Garrick Brazil, Xiaoming Liu
Michigan State University, East Lansing, MI, USA
[kumarab6, brazilga, liuxm]@msu.edu
https://github.com/abhi1kumar/groomed_nms

Abstract

Modern 3D object detectors have immensely benefited from the end-to-end learning idea. However, most of them use a post-processing algorithm called Non-Maximal Suppression (NMS) only during inference. While there were attempts to include NMS in the training pipeline for tasks such as 2D object detection, they have been less widely adopted due to a non-mathematical expression of the NMS. In this paper, we present and integrate GrooMeD-NMS – a novel Grouped Mathematically Differentiable NMS for monocular 3D object detection, such that the network is trained end-to-end with a loss on the boxes after NMS. We first formulate NMS as a matrix operation and then group and mask the boxes in an unsupervised manner to obtain a simple closed-form expression of the NMS. GrooMeD-NMS addresses the mismatch between training and inference pipelines and, therefore, forces the network to select the best 3D box in a differentiable manner. As a result, GrooMeD-NMS achieves state-of-the-art monocular 3D object detection results on the KITTI benchmark dataset performing comparably to monocular video-based methods.

1. Introduction

3D object detection is one of the fundamental problems in computer vision, where the task is to infer 3D information of the object. Its applications include augmented reality [2, 68], robotics [43, 74], medical surgery [70], and, more recently path planning and scene understanding in autonomous driving [17, 35, 46, 77]. Most of the 3D object detectors [17, 35, 44, 46, 77] are extensions of the 2D object detector Faster R-CNN [69], which relies on the end-to-end learning idea to achieve State-of-the-Art (SoTA) object detection. Some of these methods have proposed changing architectures [46, 76, 77] or losses [10, 18]. Others have tried incorporating confidence [12, 76, 77] or temporal cues [12].

Almost all of them output a massive number of boxes for each object and, thus, rely on post-processing with a greedy [65] clustering algorithm called Non-Maximal Suppression (NMS) during inference to reduce the number of false positives and increase performance. However, these works have largely overlooked NMS’s inclusion in training leading to an apparent mismatch between training and inference pipelines as the losses are applied on all boxes before NMS but not on final boxes after NMS (see Fig. 1(a)).
We also find that 3D object detection suffers a greater mismatch between classification and 3D localization compared to that of 2D localization, as discussed further in Sec. A3.2 of the supplementary and observed in [12, 35, 76]. Hence, our focus is 3D object detection.

Earlier attempts to include NMS in the training pipeline [31, 32, 65] have been made for 2D object detection where the improvements are less visible. Recent efforts to improve the correlation in 3D object detection involve calculating [77, 79] or predicting [12, 76] the scores via likelihood estimation [40] or enforcing the correlation explicitly [35]. Although this improves the 3D detection performance, improvements are limited as their training pipeline is not end to end in the absence of a differentiable NMS.

To address the mismatch between training and inference pipelines as well as the mismatch between classification and 3D localization, we propose including the NMS in the training pipeline, which gives a useful gradient to the network so that it figures out which boxes are the best-localized in 3D and, therefore, should be ranked higher (see Fig. 1(b)).

An ideal NMS for inclusion in the training pipeline should be not only differentiable but also parallelizable. Unfortunately, the inference-based classical NMS and Soft-NMS [8] are greedy, set-based and, therefore, not parallelizable [65]. To make the NMS parallelizable, we first formulate the classical NMS as matrix operation and then obtain a closed-form mathematical expression using elementary matrix operations such as matrix multiplication, matrix inversion, and clipping. We then replace the threshold pruning in the classical NMS with its softer version [8] to get useful gradients. These two changes make the NMS GPU-friendly, and the gradients are backpropagated. We next group and mask the boxes in an unsupervised manner, which removes the matrix inversion and simplifies our proposed differentiable NMS expression further. We call this NMS as Grouped Mathematically Differentiable Non-Maximal Suppression (GrooMeD-NMS).

In summary, the main contributions of this work include:

- This is the first work to propose and integrate a closed-form mathematically differentiable NMS for object detection, such that the network is trained end-to-end with a loss on the boxes after NMS.
- We propose an unsupervised grouping and masking on the boxes to remove the matrix inversion in the closed-form NMS expression.
- We achieve SoTA monocular 3D object detection performance on the KITTI dataset performing comparably to monocular video-based methods.

2. Related Work

3D Object Detection. Recent success in 2D object detection [26, 27, 48, 67, 69] has inspired people to infer 3D information from a single 2D (monocular) image. However, the monocular problem is ill-posed due to the inherent scale/depth ambiguity [82]. Hence, approaches use additional sensors such as LiDAR [35, 75, 88], stereo [45, 87] or radar [58, 84]. Although LiDAR depth estimations are accurate, LiDAR data is sparse [33] and computationally expensive to process [82]. Moreover, LiDARs are expensive and do not work well in severe weather [82].

Hence, there have been several works on monocular 3D object detection. Earlier approaches [15, 23, 61, 62] use hand-crafted features, while the recent ones are all based on deep learning. Some of these methods have proposed changing architectures [46, 49, 82] or losses [10, 18]. Others have tried incorporating confidence [12, 49, 76, 77], augmentation [80], depth in convolution [10, 22] or temporal cues [12]. Our work proposes to incorporate NMS in the training pipeline of monocular 3D object detection.

Non-Maximal Suppression. NMS has been used to reduce false positives in edge detection [72], feature point detection [29, 53, 57], face detection [85], human detection [11, 13, 20] as well as SoTA 2D [26, 48, 67, 69] and 3D detection [4, 12, 17, 76, 77, 82]. Modifications to NMS in 2D detection [8, 21, 31, 32, 65], 2D pedestrian detection [42, 51, 73], 2D salient object detection [91] and 3D detection [76] can be classified into three categories – inference NMS [8, 76], optimization-based NMS [3, 21, 42, 73, 86, 91] and neural network based NMS [30–32, 51, 65].

The inference NMS [8] changes the way the boxes are pruned in the final set of predictions. [76] uses weighted averaging to update the z-coordinate after NMS. [73] solves quadratic unconstrained binary optimization while [3, 42, 81] and [91] use point processes and MAP based inference respectively. [21] and [86] formulate NMS as a structured prediction task for isolated and all object instances respectively. The neural network NMS use a multi-layer network and message-passing to approximate NMS [31, 32, 65] or to predict the NMS threshold adaptively [51]. [30] approximates the sub-gradients of the network without modelling NMS via a transitive relationship. Our work proposes a grouped closed-form mathematical approximation of the classical NMS and does not require multiple layers or message-passing. We detail these differences in Sec. 4.2.

3. Background

3.1. Notations

Let $B = \{b_i\}_{i=1}^n$ denote the set of boxes or proposals $b_i$ from an image. Let $s = \{s_i\}_{i=1}^n$ and $r = \{r_i\}_{i=1}^n$ denote their scores (before NMS) and rescores (updated scores after NMS) respectively such that $r_i, s_i > 0 \forall i$. $D$ denotes the subset of $B$ after the NMS. Let $O = \{o_{ij}\}$ denote the $n \times n$ matrix with $o_{ij}$ denoting the 2D Intersection over Union (IoU$_{2D}$) of $b_i$ and $b_j$. The pruning function $p$ decides how to rescore a set of boxes $B$ based on IoU$_{2D}$ overlaps.
of its neighbors, sometimes suppressing boxes entirely. In other words, \( p(o_i) = 1 \) denotes the box \( b_i \) is suppressed while \( p(o_i) = 0 \) denotes \( b_i \) is kept in \( D \). The NMS threshold \( N_i \) is the threshold for which two boxes need in order for the non-maximum to be suppressed. The temperature \( \tau \) controls the shape of the exponential and sigmoidal pruning functions \( p \). \( \nu \) thresholds the rescores in GrooMeD and Soft-NMS [9] to decide if the box remains valid after NMS.

\( B \) is partitioned into different groups \( G = \{G_k\} \). \( B_{G_k} \) denotes the subset of \( B \) belonging to group \( k \). Thus, \( B_{G_k} = \{b_i\} \forall b_i \in G_k \) and \( B_{G_k} \cap B_{G_l} = \emptyset \forall k \neq l \). \( G_k \) in the subscript of a variable denotes its subset corresponding to \( B_{G_k} \). Thus, \( s_{G_k} \) and \( r_{G_k} \) denote the scores and the maximum sizes of \( B_{G_k} \) respectively. \( \alpha \) denotes the maximum group size.

\( \lor \) denotes the logical OR while \( |x| \) denotes clipping of \( x \) in the range \([0, 1]\). Formally,

\[
|x| = \begin{cases} 
1, & x > 1 \\
 x, & 0 \leq x \leq 1 \\
 0, & x < 0 
\end{cases}
\]  

(1)

\(|s|\) denotes the number of elements in \( s \). \( \bot \) in the subscript denotes the lower triangular version of the matrix without the principal diagonal. \( \odot \) denotes the element-wise multiplication. \( I \) denotes the identity matrix.

### 3.2. Classical and Soft-NMS

NMS is one of the building blocks in object detection whose high-level goal is to iteratively suppress boxes which have too much IoU with a nearby high-scoring box. We first give an overview of the classical and Soft-NMS [8], which are greedy and used in inference. Classical NMS uses the idea that the score of a box having a high IoU2D overlap with any of the selected boxes should be suppressed to zero. That is, as it uses a hard pruning \( p \) without any temperature \( \tau \). Soft-NMS makes this pruning soft via temperature \( \tau \). Thus,

classical and Soft-NMS only differ in the choice of \( p \). We reproduce them in Alg. 1 using our notations.

### 4. GrooMeD-NMS

Classical NMS (Alg. 1) uses \( \text{argmax} \) and greedily calculates the rescore \( r_i \) of boxes \( B \) and is thus not parallelizable or differentiable [65]. We wish to find its smooth approximation in closed-form for including in the training pipeline.

#### 4.1. Formulation

##### 4.1.1 Sorting

Classical NMS uses the non-differentiable hard \( \text{argmax} \) operation (Line 6 of Alg. 1). We remove the \( \text{argmax} \) by hard sorting the scores \( s \) and \( O \) in decreasing order (lines 2-3 of Alg. 2). We also try making the sorting soft. Note that we require the permutation of \( s \) to sort \( O \). Most soft sorting
methods [6, 7, 60, 63] apply the soft permutation to the same vector. Only two other methods [19, 64] can apply the soft permutation to another vector. Both methods use $O(n^2)$ computations for soft sorting [7]. We implement [64] and find that [64] is overly dependent on temperature $\tau$ to break out the ranks, and its gradients are too unreliable to train our model. Hence, we stick with the hard sorting of $s$ and $O$.

### 4.1.2 NMS as a Matrix Operation

The rescoring process of the classical NMS is greedy set-based [65] and only considers overlaps with unsuppressed boxes. We first generalize this rescoring by accounting for the effect of all (suppressed and unsuppressed) boxes as

$$ r_i \approx \max \left( s_i - \sum_{j=1}^{i-1} p(o_{ij}) r_j, 0 \right) $$

using the relaxation of logical OR $\lor$ operator as $\sum \{ [38, 47]$. See Sec. A1 of the supplementary material for an alternate explanation of (2). The presence of $r_j$ on the RHS of (2) prevents suppressed boxes from influencing other boxes. We next observe that the object detectors output multiple boxes for an object, and a good detector outputs boxes wherever it finds objects in the monocular image. Thus, we cluster the boxes in an image in an unsupervised manner based on IoU$_{2D}$ overlaps to obtain the groups $\mathcal{G}$. Grouping thus mimics the grouping of the classical NMS, but does not recore the boxes. As clustering limits interactions to intra-group interactions among the boxes, we write (6) as

$$ r_{\hat{G}_k} \approx \left( (I + \mathbf{P}_{\hat{G}_k})^{-1} s_{\hat{G}_k} \right). $$

This results in taking smaller matrix inverses in (7) than (6).

We use a simplistic grouping algorithm, i.e., we form a group $\hat{G}_k$ with boxes having high IoU$_{2D}$ overlap with the top-ranked box, given that we sorted the scores. As the group size is limited by $\alpha$, we choose a minimum of $\alpha$ and the number of boxes in $\mathcal{G}_k$. We next delete all the boxes of this group and iterate until we run out of boxes. Also, grouping uses IoU$_{2D}$ since we can achieve meaningful clustering in 2D. We detail this unsupervised grouping in Alg. 3.

### 4.1.4 Masking

Classical NMS considers the IoU$_{2D}$ of the top-scored box with other boxes. This consideration is equivalent to only keeping the column of $\mathbf{O}$ corresponding to the top box while assigning the rest of the columns to be zero. We implement this through masking of $\mathbf{P}_{\hat{G}_k}$. Let $\mathbf{M}_{\hat{G}_k}$ denote the binary mask corresponding to group $\hat{G}_k$. Then, entries in the binary matrix $\mathbf{M}_{\hat{G}_k}$ in the column corresponding to the top-scored box are 1 and the rest are 0. Hence, only one of the columns in $\mathbf{M}_{\hat{G}_k} \odot \mathbf{P}_{\hat{G}_k}$ is non-zero. Now, $\mathbf{I}_{\hat{G}_k} + \mathbf{M}_{\hat{G}_k} \odot \mathbf{P}_{\hat{G}_k}$ is a Frobenius matrix (Gaussian transformation) and we, therefore, invert this matrix by simply subtracting the second term [28]. In other words, $(\mathbf{I}_{\hat{G}_k} + \mathbf{M}_{\hat{G}_k} \odot \mathbf{P}_{\hat{G}_k})^{-1} = \mathbf{I}_{\hat{G}_k} - \mathbf{M}_{\hat{G}_k} \odot \mathbf{P}_{\hat{G}_k}$. Hence, we simplify (7) further to get

$$ r_{\hat{G}_k} \approx \left( (\mathbf{I}_{\hat{G}_k} - \mathbf{M}_{\hat{G}_k} \odot \mathbf{P}_{\hat{G}_k}) s_{\hat{G}_k} \right). $$

Thus, masking allows to bypass the computationally expensive matrix inverse operation altogether.

We call the NMS based on (8) as Grouped Mathematically Differentiable Non-Maximal Suppression or GrooMeD-NMS. We summarize the complete GrooMeD-NMS in Alg. 2 and show its block-diagram in Fig. 1(c).
4.1.5 Pruning Function

As explained in Sec. 3.1, the pruning function \( p \) decides whether to keep the box in the final set of predictions \( D \) or not based on IoU2D overlaps, i.e., \( p(o_i) = 1 \) denotes the box \( b_i \) is suppressed while \( p(o_i) = 0 \) denotes \( b_i \) is kept in \( D \).

Classical NMS uses the threshold as the pruning function, which does not give useful gradients. Therefore, we considered three different functions for \( p \): Linear, a temperature \((\tau)\)-controlled Exponential, and Sigmoidal function.

- **Linear** Linear pruning function [8] is \( p(o) = 0 \).
- **Exponential** Exponential pruning function [8] is \( p(o) = 1 - \exp \left( -\frac{o^2}{\tau^2} \right) \).
- **Sigmoidal** Sigmoidal pruning function is \( p(o) = \sigma \left( \frac{\tau^2 o - N}{} \right) \) with \( \sigma \) denoting the standard sigmoid. Sigmoidal function appears as the binary cross entropy relaxation of the subset selection problem [60].

We show these pruning functions in Fig. 2. The ablation studies (Sec. 5.4) show that choosing for the monocular \(^3\) simplest and the best GrooMeD-NMS.

4.2. Differences from Existing NMS

Although no differentiable NMS has been proposed for the monocular 3D object detection, we compare our GrooMeD-NMS with the NMS proposed for 2D object detection, 2D pedestrian detection, 2D salient object detection, and 3D object detection in Tab. 1. No method described in Tab. 1 has a matrix-based closed-form mathematical expression of the NMS. Classical, Soft [8] and Distance-NMS [76] are used at the inference time, while GrooMeD-NMS is used during both training and inference. Distance-NMS [76] updates the z-coordinate of the box after NMS as the weighted average of the z-coordinates of top-\(K\) boxes. QUBO-NMS [73], Point-NMS [42, 81], and MAP-NMS [91] are not used in end-to-end training. [3] proposes a trainable Point-NMS. The Structured-SVM based NMS [21, 86] rely on structured SVM to obtain the rescores.

![Figure 2: Pruning functions \( p \) of the classical and GrooMeD-NMS. We use the Linear and Exponential pruning of the Soft-NMS [8] while training with the GrooMeD-NMS.](image)

GrooMeD-NMS in Fig. 1(c) provides two gradients - one through \( s \) and other through \( O \).

| Method                  | Train   | Rescore | Prune | #Layers | Par   | #Parameter |
|------------------------|---------|---------|-------|---------|-------|------------|
| Classical              | \( \times \) | \( \times \) | \( \text{Hard} \) |  - | \( \sigma \left( |G| \right) \) |
| Soft-NMS [8]           | \( \times \) | \( \times \) | \( \text{Soft} \) |  - | \( \sigma \left( |G| \right) \) |
| Distance-NMS [76]      | \( \times \) | \( \times \) | \( \text{Hard} \) |  - | \( \sigma \left( |G| \right) \) |
| QUBO-NMS [73]          | \( \times \) | \( \text{Optimization} \) |  - |  - | -          |
| Point-NMS [42, 81]     | \( \times \) | \( \text{Point Process} \) |  - |  - | -          |
| Trainable Point-NMS [3] | \( \times \) | \( \text{Point Process} \) |  - |  - | -          |
| MAP-NMS [91]           | \( \times \) | MAP |  - |  - | -          |
| Structured-NMS [21, 86] | \( \times \) | SSV |  - |  - | -          |
| Adaptive-NMS [51]      | \( \times \) | \( \times \) | \( \text{Hard} \geq 1 \) | \( \sigma \left( |G| \right) \) |
| NN-NMS [31, 32, 65]    | \( \checkmark \) | Neural Network | \( \times \) | \( \geq 1 \) | \( \sigma \left( 1 \right) \) |
| GrooMeD-NMS (Ours)     | \( \checkmark \) | Matrix | Soft | 1 | \( \sigma \left( |G| \right) \) |

Adaptive-NMS [51] uses a separate neural network to predict the classical NMS threshold \(N_t\). The trainable neural network based NMS (NN-NMS) [31, 32, 65] use a separate neural network containing multiple layers and/or message-passing to approximate the NMS and do not use the pruning function. Unlike these methods, GrooMeD-NMS uses a single layer and does not require multiple layers or message passing. Our NMS is parallel up to group (denoted by \( \mathcal{G} \)). However, \( |\mathcal{G}| \) is, in general, << \( |\mathcal{B}| \) in the NMS.

4.3. Target Assignment and Loss Function

**Target Assignment.** Our method consists of M3D-RPN [10] and uses binning and self-balancing confidence [12]. The boxes’ self-balancing confidence are used as scores \( s \), which pass through the GrooMeD-NMS layer to obtain the rescores \( r \). The rescores signal the network if the best box has not been selected for a particular object.

We extend the notion of the best 2D box [65] to 3D. The best box has the highest product of IoU2D and gIoU3D [71] with ground truth \( g_i \). If the product is greater than a certain threshold \( \beta \), it is assigned a positive label. Mathematically,

\[
\text{target}(b_i) = \begin{cases} 
1, & \text{if } \exists g_i \text{ st } i = \text{argmax } q(b_j, g_i) \\
0, & \text{otherwise} 
\end{cases}
\]

with \( q(b_j, g_i) = \text{IoU}_{2D}(b_j, g_i) \left( \frac{1 + \text{gIoU}_{3D}(b_j, g_i)}{2} \right) \). gIoU3D is known to provide signal even for non-intersecting boxes [71], where the usual IoU3D is always zero. Therefore, we use gIoU3D instead of regular IoU3D for figuring out the best box in 3D as many 3D boxes have a zero IoU3D overlap with the ground truth. For calculating gIoU3D, we first calculate the volume \( V \) and hull volume \( V_{hull} \) of the 3D boxes. \( V_{hull} \) is the product of gIoU2D in Birds Eye View (BEV), removing the rotations and hull of the Y dimension. gIoU3D is then given by

\[
\text{gIoU}_{3D}(b_i, b_j) = \frac{V(b_i \cap b_j)}{V(b_i) + V(b_j) - V_{hull}(b_i, b_j) - 1}. \tag{10}
\]

**Loss Function.** Generally the number of best boxes is less than the number of ground truths in an image, as there could
be some ground truth boxes for which no box is predicted. The tiny number of best boxes introduces a far-heavier skew than the foreground-background classification. Thus, we use the modified AP-Loss [14] as our loss after NMS since AP-Loss does not suffer from class imbalance [14].

Vanilla AP-Loss treats boxes of all images in a mini-batch equally, and the gradients are back-propagated through all the boxes. We remove this condition and rank boxes in an image-wise manner. In other words, if the best boxes are correctly ranked in one image and are not in the second, then the gradients only affect the boxes of the second image. We call this modification of AP-Loss as Image-wise AP-Loss. In other words,

$$L_{	ext{Imagewise}} = \frac{1}{N} \sum_{m=1}^{N} \text{AP}(r^{(m)}, \text{target}(B^{(m)})),$$

where $r^{(m)}$ and $B^{(m)}$ denote the scores and the boxes of the $m$th image in a mini-batch respectively. This is different from previous NMS approaches [30–32, 65], which use classification losses. Our ablation studies (Sec. 5.4) show that the Imagewise AP-Loss is better suited to be used after classification losses. Our ablation studies (Sec. 5.4) show that the Imagewise AP-Loss is better suited to be used after NMS than the classification loss.

Our overall loss function is thus given by $L = L_{\text{before}} + \lambda L_{\text{after}}$ where $L_{\text{before}}$ denotes the losses before the NMS including classification, 2D and 3D regression as well as confidence losses, and $L_{\text{after}}$ denotes the loss term after the NMS, which is the Imagewise AP-Loss with $\lambda$ being the weight. See Sec. A2 of the supplementary material for more details of the loss function.

5. Experiments

Our experiments use the most widely used KITTI autonomous driving dataset [25]. We modify the publicly-available PyTorch [59] code of Kinematic-3D [12], [12] uses DenseNet-121 [34] trained on ImageNet as the backbone and $n_b = 1,024$ using 3D-RPN settings of [10]. As [12] is a video-based method while GrooMeD-NMS is an image-based method, we use the best image model of [12] henceforth called Kinematic (Image) as our baseline for a fair comparison. Kinematic (Image) is built on M3D-RPN [10] and uses binning and self-balancing confidence. Data Splits. There are three commonly used data splits of the KITTI dataset; we evaluate our method on all three.

Test Split: Official KITTI 3D benchmark [1] consists of 7,481 training and 7,518 testing images [25].

Val 1 Split: It partitions the 7,481 training images into 3,712 training and 3,769 validation images [12, 16, 77].

Val 2 Split: It partitions the 7,481 training images into 3,682 training and 3,799 validation images [89].

Training. Training is done in two phases - warmup and full [12]. We initialize the model with the confidence prediction branch from warmup weights and finetune using the self-balancing loss [12] and Imagewise AP-Loss [14] after our GrooMeD-NMS. See Sec. A3.1 of the supplementary material for more training details. We keep the weight $\lambda$ at 0.05. Unless otherwise stated, we use $p$ as the Linear function (this does not require $\gamma$) with $\alpha = 100$. $N_e, \gamma$ and $\beta$ are set to 0.4 [10, 12], 0.3 and 0.8 respectively.

Inference. We multiply the class and predicted confidence to get the box’s overall score in inference as in [36, 76, 83]. See Sec. 5.2 for training and inference times.

Evaluation Metrics. KITTI uses AP$_{3D|R_{40}}$ metric to evaluate object detection following [77, 79]. KITTI benchmark evaluates on three object categories: Easy, Moderate and Hard. It assigns each object to a category based on its occlusion, truncation, and height in the image space. The AP$_{3D|R_{40}}$ performance on the Moderate category compares different models in the benchmark [25]. We focus primarily on the Car class following [12].

5.1. KITTI Test 3D Object Detection

Table 2 summarizes the results of 3D object detection and BEV evaluation on KITTI Test Split. The results in Tab. 2 shows that GrooMeD-NMS outperforms the baseline M3D-RPN [10] by a significant margin and several other SoTA methods on both the tasks. GrooMeD-NMS also outperforms augmentation based approach MoVi-3D [80] and depth-convolution based D4LCN [22]. Despite being an image-based method, GrooMeD-NMS performs competitively to the video-based method Kinematic (Video) [12], outperforming it on the most-challenging Hard set.

5.2. KITTI Val 1 3D Object Detection

Results. Tab. 3 summarizes the results of 3D object detection and BEV evaluation on KITTI Val 1 Split at two
IoU thresholds of 0.7 and 0.5 [12, 18]. The results in Tab. 3 show that GrooMeD-NMS outperforms the baseline of M3D-RPN [10] and Kinematic (Image) [12] by a significant margin. Interestingly, GrooMeD-NMS (an image-based method) also outperforms the video-based method Kinematic (Video) [12] on most of the metrics. Thus, GrooMeD-NMS performs best on 6 out of the 12 cases (3 categories × 2 tasks × 2 thresholds) while second-best on all other cases. The performance is especially impressive since the biggest improvements are shown on the Moderate and Hard set, where objects are more distant and occluded.

**AP<sub>3D</sub> at different depths and IoU<sub>3D</sub> thresholds.** We next compare the AP<sub>3D</sub> performance of GrooMeD-NMS and Kinematic (Image) on linear and log scale for objects at different depths of [15, 30, 45, 60] meters and IoU<sub>3D</sub> matching criteria of 0.3 –> 0.7 in Fig. 3 as in [12]. Fig. 3 shows that GrooMeD-NMS outperforms the Kinematic (Image) [12] at all depths and all IoU<sub>3D</sub> thresholds.

**Comparisons with other NMS.** We compare with the classical NMS, Soft-NMS [8] and Distance-NMS [76] in Tab. 4. More detailed results are in Tab. 8 of the supplementary material. The results show that NMS inclusion in the training pipeline benefits the performance, unlike [8], which suggests otherwise. Training with GrooMeD-NMS helps because the network gets an additional signal through the GrooMeD-NMS layer whenever the best-localized box corresponding to an object is not selected. Interestingly, Tab. 4 also suggests that replacing GrooMeD-NMS with the classical NMS in inference does not affect the performance.

**Score-IoU<sub>3D</sub> Plot.** We further correlate the scores with IoU<sub>3D</sub> after NMS of our model with two baselines - M3D-RPN [10] and Kinematic (Image) [12] and also the Kinematic (Video) [12] in Fig. 4. We obtain the best correlation of 0.345 exceeding the correlations of M3D-RPN, Kinematic (Image) and, also Kinematic (Video). This proves that including NMS in the training pipeline is beneficial.

**Training and Inference Times.** We now compare the training and inference times of including GrooMeD-NMS in the pipeline. Warmup training phase takes about 13 hours to train on a single 12 GB GeForce GTX Titan-X GPU. Full training phase of Kinematic (Image) and GrooMeD-NMS takes about 8 and 8.5 hours respectively. The inference time per image using classical and GrooMeD-NMS is 0.12 and 0.15 ms respectively. Tab. 4 suggests that changing the NMS from GrooMeD to classical during inference does not alter the performance. Then, the inference time of our method is the same as 0.12 ms.

**5.3. KITTI Val 2 3D Object Detection**

Tab. 5 summarizes the results of 3D object detection and BEV evaluation on KITTI Val 2 Split at two IoU<sub>3D</sub> thresh-
Table 5: $\text{AP}_{\text{3D}|R_{40}}$ and $\text{AP}_{\text{BEV}|R_{40}}$ comparisons on KITTI Val 2 Cars. [Key: Best, * Released, † Retrained].

| Method                     | $\text{IoU}_{\text{3D}} \geq 0.7$ | $\text{IoU}_{\text{3D}} \geq 0.5$ |
|----------------------------|-----------------------------------|-----------------------------------|
|                            | Easy Mod Hard                     | Easy Mod Hard                     | Easy Mod Hard                     | Easy Mod Hard                     |
|                            | $\text{AP}_{\text{3D}|R_{40}}$     | $\text{AP}_{\text{BEV}|R_{40}}$   | $\text{AP}_{\text{3D}|R_{40}}$     | $\text{AP}_{\text{BEV}|R_{40}}$   |
| M3D-RPN [10] *             | 14.57 10.07 7.51                  | 21.36 15.22 11.28                | 49.14 34.43 26.39                | 53.44 37.79 29.36                |
| Kinematic (Image) [12] †   | 13.54 10.21 7.24                  | 20.60 15.14 11.30                | 51.53 36.55 28.26                | 56.20 40.02 31.25                |
| GrooMed-NMS (Ours)         | 14.72 10.87 7.67                  | 22.03 16.05 11.93                | 51.91 36.78 28.40                | 56.29 40.31 31.39                |

Table 6: Ablation studies of our method on KITTI Val 1 Cars.

| Change from GrooMed-NMS model: | $\text{IoU}_{\text{3D}} \geq 0.7$ | $\text{IoU}_{\text{3D}} \geq 0.5$ |
|--------------------------------|-----------------------------------|-----------------------------------|
|                                | Easy Mod Hard                     | Easy Mod Hard                     | Easy Mod Hard                     | Easy Mod Hard                     |
|                                | $\text{AP}_{\text{3D}|R_{40}}$     | $\text{AP}_{\text{BEV}|R_{40}}$   | $\text{AP}_{\text{3D}|R_{40}}$     | $\text{AP}_{\text{BEV}|R_{40}}$   |
| Training                      |                                   |                                   |                                   |                                   |
| Conf+NMS → No Conf+No NMS     | 16.66 12.10 9.40                  | 23.15 17.43 13.48                | 51.47 38.58 30.98                | 56.48 42.53 34.37                |
| Conf+NMS → Conf+No NMS        | 19.16 13.89 10.96                | 27.01 19.33 14.84                | 57.12 41.07 32.79                | 61.60 44.58 35.97                |
| Conf+NMS → No Conf+No NMS     | 15.02 11.21 8.83                  | 21.07 16.27 12.77                | 48.01 36.18 29.96                | 53.82 40.94 33.35                |
| Initialization               |                                   |                                   |                                   |                                   |
| No Warmup                    | 15.33 11.08 6.78                  | 21.32 16.59 12.93                | 49.15 37.42 30.11                | 54.32 41.44 33.48                |
| Pruning Function             |                                   |                                   |                                   |                                   |
| Linear → Exponential, $\tau = 1$ | 12.81 9.26 7.10                | 17.07 12.17 9.25                | 29.58 20.42 15.88                | 32.06 22.16 17.20                |
| Linear → Exponential, $\tau = 0.5$ | 18.63 13.85 10.98                | 27.52 20.14 15.76                | 56.64 41.01 32.79                | 61.43 44.73 36.02                |
| Linear → Exponential, $\tau = 0.1$ | 18.34 13.79 10.88                | 27.26 19.71 15.90                | 56.98 41.16 32.96                | 62.77 45.23 36.56                |
| Linear → Sigmoidal, $\tau = 0.1$ | 17.40 13.21 9.80                  | 26.77 19.26 14.76                | 55.15 40.77 32.63                | 60.56 44.23 35.74                |
| Group+Mask                   |                                   |                                   |                                   |                                   |
| Group+Mask → No Group        | 18.43 13.91 11.08                  | 26.53 19.46 15.83                | 55.93 40.98 32.78                | 61.02 44.77 36.09                |
| Group+Mask → Group+No Mask   | 18.99 13.74 10.24                  | 26.71 19.21 14.77                | 55.21 40.69 32.55                | 61.74 44.67 36.00                |
| Loss                         |                                   |                                   |                                   |                                   |
| Imagewise AP → Vanilla AP    | 18.23 13.73 10.28                  | 26.42 19.31 14.76                | 54.47 40.35 32.20                | 60.90 44.08 35.47                |
| Imagewise AP → BCE           | 16.34 12.74 9.73                  | 22.40 17.46 13.70                | 52.46 39.40 31.68                | 58.22 43.60 35.27                |
| Inference NMS Scores         |                                   |                                   |                                   |                                   |
| Class*Pred → Class           | 18.26 13.36 10.49                  | 25.39 18.64 15.12                | 52.44 38.99 31.3               | 57.37 42.89 34.68                |
| Class*Pred → Pred            | 17.51 12.84 9.55                  | 24.55 17.85 13.69                | 52.78 37.48 29.37                | 58.30 41.26 32.66                |
| — GrooMed-NMS (best model)   | 19.67 14.32 11.27                  | 27.38 19.75 15.92                | 55.62 41.07 32.89                | 61.83 44.98 36.29                |

5.4. Ablation Studies

Table 6 compares the modifications of our approach on KITTI Val 1 Cars. Unless stated otherwise, we stick with the experimental settings described in Sec. 5. Using a confidence head (Conf+No NMS) proves beneficial compared to the warmup model (No Conf+No NMS), which is consistent with the observations of [12, 76]. Further, GrooMed-NMS on classification scores (denoted by No Conf + NMS) is detrimental as the classification scores are not suited for localization [12, 35]. Training the warmup model and then finetuning also works better than training without warmup as in [12] since the warmup phase allows GrooMed-NMS to carry meaningful grouping of the boxes.

As described in Sec. 4.1.5, in addition to Linear, we compare two other functions for pruning function $p$: Exponential and Sigmoidal. Both of them do not perform as well as the Linear $p$ possibly because they have vanishing gradients close to overlap of zero or one. Grouping and masking both help our model to reach a better minimum. As described in Sec. 4.3, Imagewise AP loss is better than the Vanilla AP loss since it treats boxes of two images differently. Imagewise AP also performs better than the binary cross-entropy (BCE) loss proposed in [30–32, 65]. Using the product of self-balancing confidence and classification scores instead of using them individually as the scores to the NMS in inference is better, consistent with [36, 76, 83]. Class confidence performs worse since it does not have the localization information while the self-balancing confidence (Pred) gives the localization without considering whether the box belongs to foreground or background.

6. Conclusions

In this paper, we present and integrate GrooMed-NMS – a novel Grouped Mathematically Differentiable NMS for monocular 3D object detection, such that the network is trained end-to-end with a loss on the boxes after NMS. We first formulate NMS as a matrix operation and then do unsupervised grouping and masking of the boxes to obtain a simple closed-form expression of the NMS. GrooMed-NMS addresses the mismatch between training and inference pipelines and, therefore, forces the network to select the best 3D box in a differentiable manner. As a result, GrooMed-NMS achieves state-of-the-art monocular 3D object detection results on the KITTI benchmark dataset. Although our implementation demonstrates monocular 3D object detection, GrooMed-NMS is fairly generic for other object detection tasks. Future work includes applying this method to tasks such as LiDAR-based 3D object detection and pedestrian detection.
References

[1] The KITTI Vision Benchmark Suite. http://www.cvlibs.net/datasets/kitti/eval_object.php?objbenchmark=3d. Accessed: 2020-10-11.

[2] Hassan Alhajja, Siva Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. IJCV, 2018.

[3] Samaneh Azadi, Jiashi Feng, and Trevor Darrell. Learning detection with diverse proposals. In CVPR, 2017.

[4] Wentao Bao, Bin Xu, and Zhenzhong Chen. MonoFENet: Monocular 3D object detection with feature enhancement networks. IEEE Transactions on Image Processing, 2019.

[5] Deniz Beker, Hiroharu Kato, Mihai Adrian Morariu, Takahiro Ando, Toru Matsuoka, Wadim Kehl, and Adrien Gaidon. Monocular differentiable rendering for self-supervised 3D object detection. In ECCV, 2020.

[6] Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable perturbed optimizers. In NeurIPS, 2020.

[7] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In ICML, 2020.

[8] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry Davis. Soft-NMS–improving object detection with one line of code. In ICCV, 2017.

[9] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry Davis. Soft-NMS implementation. https://github.com/bharatsingh430/soft-nms/blob/master/lib/nms/cpu_nms.pyx#L98, 2017. Accessed: 2021-01-18.

[10] Garrick Brazil and Xiaoming Liu. M3D-RPN: Monocular 3D region proposal network for object detection. In ICCV, 2019.

[11] Garrick Brazil and Xiaoming Liu. Pedestrian detection with autoregressive network phases. In CVPR, 2019.

[12] Garrick Brazil, Gerard Pons-Moll, Xiaoming Liu, and Bernt Schiele. Kinematic 3D object detection in monocular video. In ECCV, 2020.

[13] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating pedestrians via simultaneous detection & segmentation. In ICCV, 2017.

[14] Kean Chen, Weiyao Lin, Jianguo Li, John See, Ji Wang, and Junni Zou. AP-Loss for accurate one-stage object detection. TPAMI, 2020.

[15] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3D object detection for autonomous driving. In CVPR, 2016.

[16] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew Bergmashawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3D object proposals for accurate object class detection. In NeurIPS, 2015.

[17] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3D object detection network for autonomous driving. In CVPR, 2017.

[18] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. MonoPair: Monocular 3D object detection using pairwise spatial relationships. In CVPR, 2020.

[19] Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable ranks and sorting using optimal transport. In NeurIPS, 2019.

[20] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005.

[21] Chaitanya Desai, Deva Ramanan, and Charless Fowlkes. Discriminative models for multi-class object layout. IJCV, 2011.

[22] Mingyu Ding, Yuqi Hao, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3D object detection. In CVPR Workshops, 2020.

[23] Sanja Fidler, Sven Dickinson, and Raquel Urtasun. 3D object detection and viewpoint estimation with a deformable 3D cuboid model. In NeurIPS, 2012.

[24] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. IJRR, 2013.

[25] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In CVPR, 2012.

[26] Ross Girshick. Fast R-CNN. In ICCV, 2015.

[27] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[28] Gene Golub and Charles Loan. Matrix computations. 2013.

[29] Christopher Harris and Mike Stephens. A combined corner and edge detector. In Alvey vision conference, 1988.

[30] Paul Henderson and Vittorio Ferrari. End-to-end training of object class detectors for mean average precision. In ACCV, 2016.

[31] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. A convnet for non-maximum suppression. In GCPR, 2016.

[32] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In CVPR, 2017.

[33] Peiyun Hu, Jason Ziglar, David Held, and Deva Ramanan. What you see is what you get: Exploiting visibility for 3D object detection. In CVPR, 2020.

[34] Gao Huang, Zhuang Liu, Laurens Maaten, and Kilian Weinberger. Densely connected convolutional networks. In CVPR, 2017.

[35] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. EP-Net: Enhancing point features with image semantics for 3D object detection. In ECCV, 2020.

[36] Kang Kim and Hee Lee. Probabilistic anchor assignment with iou prediction for object detection. In NeurIPS, 2017.

[37] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In ICLR, 2015.

[38] Emile Krieken, Erman Acar, and Frank Harmelen. Analyzing differentiable fuzzy logic operators. arXiv preprint arXiv:2002.06100, 2020.
[39] Jason Ku, Alex Pon, and Steven Waslander. Monocular 3D object detection leveraging accurate proposals and shape reconstruction. In CVPR, 2019.

[40] Abhinav Kumar, Tim Marks, Wenhuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. LUVLi face alignment: Estimating landmarks’ location, uncertainty, and visibility likelihood. In CVPR, 2020.

[41] Animesh Kumar and Vinod Prabhakaran. Estimation of binned signals from the signs of noisy samples. In ICASSP, 2013.

[42] Donghoon Lee, Geonho Cha, Ming-Hsuan Yang, and Songhwai Oh. Individualness and determinantal point processes for pedestrian detection. In ECCV, 2016.

[43] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, Abhinav Kumar, Tim Marks, Wenxuan Mou, Ye Wang, Alykhan Tejani, Sasan Khiamkurth, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In NeurIPS, 2019.

[44] Lijie Liu, Jiwen Lu, Chufan Wu, Jiwen Lu, Lingxi Xie, Jie Zhou, and Tao Li. Augmenting neural networks with first-order logic. In ACL, 2019.

[45] Bojan Pepik, Michael Stark, Bernt Schiele. Learning to filter object detections. In GCPR, 2020.

[46] Sitapa Rujikietgumjorn and Robert Collins. Optimized pedestrian detection for multiple and occluded people. In CVPR, 2013.

[47] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. IJCV, 2004.

[48] Frank Moosmann, Oliver Pink, and Christoph Stiller. Segmentation of 3D LiDAR data in non-flat urban environments using a local convexity criterion. In Intelligent Vehicles Symposium, 2009.

[49] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. ROI-10D: Monocular lifting of 2D detection to 6 pose and metric shape. In CVPR, 2019.

[50] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. Real-time monocular 3D object detection via keypoint estimation. In NeurIPS, 2019.

[51] Zengyi Qin, Jinglu Wang, and Yan Lu. MonoGRNet: A deep fitting degree scoring network for monocular 3D object detection. In CVPR, 2017.

[52] Bojan Pepik, Michael Stark, Peter Gehler, and Bernt Schiele. Multi-view and 3D deformable part models. TPAMI, 2015.

[53] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving. In ICCV, 2019.

[54] Shouqing Ren, Kaiming He, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In CVPR, 2016.

[55] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. Roi-10D: Monocular lifting of 2D detection to 6 pose and metric shape. In CVPR, 2019.

[56] Frank Moosmann, Oliver Pink, and Christoph Stiller. Segmentation of 3D LiDAR data in non-flat urban environments using a local convexity criterion. In Intelligent Vehicles Symposium, 2009.
[74] Ashutosh Saxena, Justin Driemeyer, and Andrew Ng. Robotic grasping of novel objects using vision. *IJRR*, 2008.

[75] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointR-CNN: 3D object proposal generation and detection from point cloud. In *CVPR*, 2019.

[76] Xuepeng Shi, Zhixiang Chen, and Tae-Kyun Kim. Distance-normalized unified representation for monocular 3D object detection. In *ECCV*, 2020.

[77] Andrea Simonelli, Samuel Bulò, Lorenzo Porzi, Manuel Antequera, and Peter Kontschieder. Disentangling monocular 3D object detection: From single to multi-class recognition. *TPAMI*, 2020.

[78] Andrea Simonelli, Samuel Bulò, Lorenzo Porzi, Peter Kontschieder, and Elisa Ricci. Demystifying pseudo-LiDAR for monocular 3D object detection. *arXiv preprint arXiv:2012.05796*, 2020.

[79] Andrea Simonelli, Samuel Bulò, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3D object detection. In *ICCV*, 2019.

[80] Andrea Simonelli, Samuel Bulò, Lorenzo Porzi, Elisa Ricci, and Peter Kontschieder. Towards generalization across depth for monocular 3D object detection. In *ECCV*, 2020.

[81] Samik Some, Mithun Das Gupta, and Vinay Namboodiri. Determinantal point process as an alternative to NMS. In *BMVC*, 2020.

[82] Yunlei Tang, Sebastian Dorn, and Chiragkumar Savani. Center3D: Center-based monocular 3D object detection with joint depth understanding. *arXiv preprint arXiv:2005.13423*, 2020.

[83] Lachlan Tychsen-Smith and Lars Petersson. Improving object localization with fitness NMS and bounded IoU loss. In *CVPR*, 2018.

[84] Alexandru Vasile and Richard Marino. Pose-independent automatic target detection and recognition using 3D laser radar imagery. *Lincoln laboratory journal*, 2005.

[85] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[86] Li Wan, David Eigen, and Rob Fergus. End-to-end integration of a convolution network, deformable parts model and non-maximum suppression. In *CVPR*, 2015.

[87] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Weinberger. Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving. In *CVPR*, 2019.

[88] Pengxiang Wu, Siheng Chen, and Dimitris Metaxas. MotionNet: Joint perception and motion prediction for autonomous driving based on bird’s eye view maps. In *CVPR*, 2020.

[89] Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese. Subcategory-aware convolutional neural networks for object proposals and detection. In *WACV*, 2017.

[90] Xiaqing Ye, Liang Du, Yifeng Shi, Yingying Li, Xiao Tan, Jianfeng Feng, Errui Ding, and Shilei Wen. Monocular 3D object detection via feature domain adaptation. In *ECCV*, 2020.

[91] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Unconstrained salient ob-
GrooMeD-NMS: Grouped Mathematically Differentiable NMS for Monocular 3D Object Detection

Supplementary Material

A1. Detailed Explanation of NMS as a Matrix Operation

The rescoring process of the classical NMS is greedy set-based [65] and calculates the rescore for a box $i$ (Line 10 of Alg. 1) as

$$ r_i = s_i \prod_{j \in d_{<i}} (1 - p(o_{ij})) , \quad (12) $$

where $d_{<i}$ is defined as the box indices sampled from $d$ having higher scores than box $i$. For example, let us consider that $d = \{1, 5, 7, 9\}$. Then, for $i = 7$, $d_{<i} = \{1, 5\}$ while for $i = 1$, $d_{<i} = \emptyset$ with $\emptyset$ denoting the empty set.

This is possible since we had sorted the scores $s$ and $O$ in decreasing order (Lines 2-3 of Alg. 2) to remove the non-differentiable hard argmax operation of the classical NMS (Line 6 of Alg. 1).

Classical NMS only takes the overlap with unsuppressed boxes into account. Therefore, we generalize (12) by accounting for the effect of all (suppressed and unsuppressed) boxes as

$$ r_i = s_i \prod_{j=1}^{i-1} (1 - p(o_{ij}))r_j . \quad (13) $$

The presence of $r_j$ on the RHS of (13) prevents suppressed boxes $r_j \approx 0$ from influencing other boxes hugely. Let us say we have a box $b_2$ with a high overlap with an unsuppressed box $b_1$. The classical NMS with a threshold pruning function assigns $r_2 = 0$ while (13) assigns $r_2$ a small non-zero value with a threshold pruning.

Although (13) keeps $r_i \geq 0$, getting a closed-form recursion in $r$ is not easy because of the product operation. To get a closed-form recursion with addition/subtraction in $r$, we first carry out the polynomial multiplication and then ignore the higher-order terms as

$$ r_i = s_i \left( 1 - \sum_{j=1}^{i-1} p(o_{ij})r_j + \mathcal{O}(n^2) \right) $$

$$ \approx s_i \left( 1 - \sum_{j=1}^{i-1} p(o_{ij})r_j \right) $$

$$ \approx s_i - \sum_{j=1}^{i-1} p(o_{ij})r_j . \quad (14) $$

Dropping the $s_i$ in the second term of (14) helps us get a cleaner form of (19). Moreover, it does not change the nature of the NMS since the subtraction keeps the relation $r_i \leq s_i$ intact as $p(o_{ij})$ and $r_j$ are both between $[0, 1]$.

We can also reach (14) directly as follows. Classical NMS suppresses a box which has a high IoU$_{2D}$ overlap with any of the unsuppressed boxes $(r_j \approx 1)$ to zero. We consider any as a logical non-differentiable OR operation and use logical OR $\lor$ operator’s differentiable relaxation as $\sum [38, 47]$. We next use this relaxation with the other expression $r \leq s$.

When a box shows overlap with more than two unsuppressed boxes, the term $\sum_{j=1}^{i-1} p(o_{ij})r_j > 1$ in (14) or when a box shows high overlap with one unsuppressed box, the term $s_i < p(o_{ij})r_j$. In both of these cases, $r_i < 0$. So, we lower bound (14) with a max operation to ensure that $r_i \geq 0$. Thus,

$$ r_i \approx \max \left( s_i - \sum_{j=1}^{i-1} p(o_{ij})r_j, 0 \right) . \quad (15) $$

We write the rescoring $r$ in a matrix formulation as

$$ \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_n \end{bmatrix} \approx \max \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} , \quad (16) $$

with

$$ \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ \vdots \\ s_n \end{bmatrix} \begin{bmatrix} 0 \\ p(o_{21}) \\ p(o_{31}) \\ \vdots \\ p(o_{n1}) \end{bmatrix} , \quad (17) $$

We next write the above two equations compactly as

$$ r \approx \max(s - Pr, 0) , \quad (18) $$

where $P$, called the Prune Matrix, is obtained by element-wise operation of the pruning function $p$ on $O_L$. Maximum
operation makes (18) non-linear [41] and, thus, difficult to solve.

However, for a differentiable NMS layer, we need to avoid the recursion. Therefore, we first solve (18) assuming the max operation is not present which gives us the solution r \approx (I + P)^{-1}s. In general, this solution is not necessarily bounded between 0 and 1. Hence, we clip it explicitly to obtain the approximation

$$r \approx \left[(I + P)^{-1}s\right],$$

which we use as the solution to (18).

**A2. Loss Functions**

We now detail out the loss functions used for training. The losses on the boxes before NMS, $L_{before}$, is given by [12]

$$L_{before} = L_{class} + L_{2D} + b_{conf} L_{3D} + \lambda_{conf}(1 - b_{conf}),$$

where

$$L_{class} = CE(b_{class}, g_{class}),$$

$$L_{2D} = -\log(IoU(b_{2D}, g_{2D})),$$

$$L_{3D} = \text{Smooth-L1}(b_{3D}, g_{3D}) + \lambda_{a} CE([a_0, a_h], [g_{a_0}, g_{a_h}]),$$

$b_{conf}$ is the predicted self-balancing confidence of each box $b$, while $a_0$ and $a_h$ are its orientation bins [12]. $g$ denotes the ground-truth. $\lambda_{conf}$ is the rolling mean of most recent $L_{3D}$ losses per mini-batch [12], while $\lambda_a$ denotes the weight of the orientation bins loss. $CE$ and Smooth-L1 denote the Cross Entropy and Smooth L1 loss respectively. Note that we apply 2D and 3D regression losses as well as the confidence losses only on the foreground boxes.

As explained in Sec. 4.3, the loss on the boxes after NMS, $L_{after}$, is the Imagewise AP-Loss, which is given by

$$L_{after} = L_{Imagewise} = \frac{1}{N} \sum_{m=1}^{N} AP(r(m), \text{target}(B(m))),$$

Let $\lambda$ be the weight of the $L_{after}$ term. Then, our overall loss function is given by

$$L = L_{before} + \lambda L_{after}$$

We keep $\lambda_a = 0.35$ following [12] and $\lambda = 0.05$. Clearly, all our losses and their weights are identical to [12] except $L_{Imagewise}$.

**A3. Additional Experiments and Results**

We now provide additional details and results evaluating our system’s performance.

**A3.1. Training**

Training images are augmented using random flipping with probability 0.5 [12]. Adam optimizer [37] is used with batch size 2, weight-decay $5 \times 10^{-4}$ and gradient clipping of 1 [10,12]. Warmup starts with a learning rate $4 \times 10^{-3}$ following a poly learning policy with power 0.9 [12]. Warmup and full training phases take 80k and 50k mini-batches respectively for Val 1 and Val 2 splits [12] while take 160k and 100k mini-batches for Test split.

**A3.2. KITTI Val 1 Oracle NMS Experiments**

As discussed in Sec. 1, to understand the effects of an inference-only NMS on 2D and 3D object detection, we conduct a series of oracle experiments. We create an oracle NMS by taking the Val Car boxes of KITTI Val 1 Split from the baseline Kinematic (Image) model before NMS and replace their scores with their true IoU2D or IoU3D with the ground-truth, respectively. Note that this corresponds to the oracle because we do not know the ground-truth boxes during inference. We then pass the boxes with the oracle scores through the classical NMS and report the results in Tab. 7.

The results show that the AP3D increases by a staggering > 60 AP on Mod cars when we use oracle IoU3D as the NMS score. On the other hand, we only see an increase in AP2D by \approx 11 AP on Mod cars when we use oracle IoU2D as the NMS score. Thus, the relative effect of using oracle IoU3D NMS scores on 3D detection is more significant than using oracle IoU2D NMS scores on 2D detection.

### Table 7: Results on using Oracle NMS scores on AP3D/R40, APREV/R40 and AP2D/R40 of KITTI Val 1 Cars. [Key: Best]

| NMS Scores | AP3D/R40(%) Easy Mod Hard | APREV/R40(%) Easy Mod Hard | AP2D/R40(%) Easy Mod Hard |
|------------|--------------------------|---------------------------|---------------------------|
| Kinematic (Image) | 18.29 13.55 10.13 | 25.72 18.82 14.48 | 93.69 84.07 67.14 |
| Oracle IoU2D | 9.36 9.93 6.40 | 12.27 10.43 8.72 | 99.18 95.66 85.77 |
| Oracle IoU3D | 87.93 73.10 60.91 | 93.47 83.61 71.31 | 80.99 78.38 67.66 |
Table 8: AP<sub>3D|R<sub>40</sub> and AP<sub>BEV|R<sub>40</sub> comparisons with other NMS during inference on KITTI Val 1 Cars.

| Inference NMS | IoU<sub>3D</sub⟩ ≥ 0.7 | IoU<sub>3D</sub⟩ ≥ 0.5 |
|---------------|--------------------------|--------------------------|
|               | AP<sub>3D|R<sub>40</sub> (†) | AP<sub>BEV|R<sub>40</sub> (†) | AP<sub>3D|R<sub>40</sub> (†) | AP<sub>BEV|R<sub>40</sub> (†) |
|               | Easy | Mod | Hard | Easy | Mod | Hard | Easy | Mod | Hard | Easy | Mod | Hard |
| Kinematic (Image) [12] | Classical | 18.28 | 13.55 | 10.13 | 25.72 | 18.82 | 14.48 | 54.70 | 39.33 | 31.25 | 60.87 | 44.36 | 34.48 |
| Kinematic (Image) [12] | Soft [8] | 18.29 | 13.55 | 10.13 | 25.71 | 18.81 | 14.48 | 54.70 | 39.33 | 31.26 | 60.87 | 44.36 | 34.48 |
| Kinematic (Image) [12] | Distance [76] | 18.25 | 13.53 | 10.11 | 25.71 | 18.82 | 14.48 | 54.70 | 39.33 | 31.26 | 60.87 | 44.36 | 34.48 |
| Kinematic (Image) [12] | GrooMeD | 18.26 | 13.51 | 10.10 | 25.67 | 18.77 | 14.44 | 54.59 | 39.25 | 31.18 | 60.78 | 44.28 | 34.40 |
| GrooMeD-NMS | Classical | 19.67 | 14.31 | 11.27 | 27.38 | 19.75 | 15.93 | 55.64 | 41.08 | 32.91 | 61.85 | 44.98 | 36.31 |
| GrooMeD-NMS | Soft [8] | 19.67 | 14.31 | 11.27 | 27.38 | 19.75 | 15.93 | 55.64 | 41.08 | 32.91 | 61.85 | 44.98 | 36.31 |
| GrooMeD-NMS | Distance [76] | 19.67 | 14.31 | 11.27 | 27.38 | 19.75 | 15.93 | 55.64 | 41.08 | 32.91 | 61.85 | 44.98 | 36.31 |
| GrooMeD-NMS | GrooMeD | 19.67 | 14.32 | 11.27 | 27.38 | 19.75 | 15.92 | 55.62 | 41.07 | 32.89 | 61.83 | 44.98 | 36.29 |

In other words, the mismatch is greater between classification and 3D localization compared to the mismatch between classification and 2D localization.

A3.3. KITTI Val 1 3D Object Detection

Comparisons with other NMS. We compare our method with the other NMS—classical, Soft [8] and Distance-NMS [76] and report the detailed results in Tab. 8. We use the publicly released Soft-NMS code and Distance-NMS code from the respective authors. The Distance-NMS model uses the class confidence scores divided by the uncertainty in z (the most erroneous dimension in 3D localization [78]) of a box as the Distance-NMS [76] input. Our model does not predict the uncertainty in z of a box but predicts its self-balancing confidence (the 3D localization score). Therefore, we use the class confidence scores multiplied by the self-balancing confidence as the Distance-NMS input.

The results in Tab. 8 show that NMS inclusion in the training pipeline benefits the performance, unlike [8], which suggests otherwise. Training with GrooMeD-NMS helps because the network gets an additional signal through the GrooMeD-NMS layer whenever the best-localized box corresponding to an object is not selected. Moreover, Tab. 8 suggests that we can replace GrooMeD-NMS with the classical NMS in inference as the performance is almost the same even at IoU<sub>3D</sub⟩ = 0.5.

How good is the classical NMS approximation? GrooMeD-NMS uses several approximations to arrive at the matrix solution (19). We now compare how good these approximations are with the classical NMS. Interestingly, Tab. 8 shows that GrooMeD-NMS is an excellent approximation to the classical NMS as the performance does not degrade after changing the NMS in inference.

A3.4. KITTI Val 1 Sensitivity Analysis

There are a few adjustable parameters for the GrooMeD-NMS, such as the NMS threshold N<sub>τ</sub>, valid box threshold v, the maximum group size α, the weight λ for the L<sub>after</sub>, and β. We carry out a sensitivity analysis to understand how these parameters affect performance and speed, and how sensitive the algorithm is to these parameters.

Sensitivity to NMS Threshold. We show the sensitivity to NMS threshold N<sub>τ</sub> in Tab. 9. The results in Tab. 9 show that the optimal N<sub>τ</sub> = 0.3. This is also the N<sub>τ</sub> in [10, 12].

Sensitivity to Valid Box Threshold. We next show the sensitivity to valid box threshold v in Tab. 10. Our choice of v = 0.3 performs close to the optimal choice.

Sensitivity to Maximum Group Size. Grouping has a parameter group size (α). We vary this parameter and report AP<sub>3D|R<sub>40</sub> and AP<sub>BEV|R<sub>40</sub></sub> at two different IoU<sub>3D</sub> thresholds on Moderate Cars of KITTI Val 1 Split in Fig. 5. We note that the best AP<sub>3D|R<sub>40</sub> Performance is obtained at α = 100 and we, therefore, set α = 100 in our experiments.

Sensitivity to Loss Weight. We now show the sensitivity to loss weight λ in Tab. 11. Our choice of λ = 0.05 is
Table 11: $\text{AP}_{\text{3D}}|_{R_{40}}$ and $\text{AP}_{\text{BEV}}|_{R_{40}}$ variation with $\lambda$ on KITTI Val 1 Cars. [Key: Best]

| $\lambda$     | $\text{AP}_{\text{3D}}|_{R_{40}}$ | $\text{AP}_{\text{BEV}}|_{R_{40}}$ |
|---------------|----------------------------------|----------------------------------|
| $\lambda = 0$ | 19.16 13.89 10.96               | 27.01 19.33 14.84               |
| $\lambda = 0.05$ | 19.67 14.32 11.27               | 27.38 19.75 15.92               |
| $\lambda = 0.1$  | 17.74 13.61 10.81               | 25.86 19.18 15.57               |
| $\lambda = 1$   | 10.08 7.26 6.00                 | 14.44 10.55 8.41                |

Table 12: $\text{AP}_{\text{3D}}|_{R_{40}}$ and $\text{AP}_{\text{BEV}}|_{R_{40}}$ variation with $\beta$ on KITTI Val 1 Cars. [Key: Best]

| $\beta$ | $\text{AP}_{\text{3D}}|_{R_{40}}$ | $\text{AP}_{\text{BEV}}|_{R_{40}}$ |
|---------|----------------------------------|----------------------------------|
| $\beta = 0.1$ | 18.09 13.64 10.21               | 26.52 19.50 15.74               |
| $\beta = 0.3$  | 19.67 14.32 11.27               | 27.38 19.75 15.92               |
| $\beta = 0.4$  | 18.91 14.02 11.15               | 27.11 19.64 15.90               |
| $\beta = 0.5$ | 18.49 13.66 10.96               | 27.01 19.47 15.79               |

the optimal value.

**Sensitivity to Best Box Threshold.** We now show the sensitivity to the best box threshold $\beta$ in Tab. 12. Our choice of $\beta = 0.3$ is the optimal value.

**Conclusion.** Our method has minor sensitivity to $N_t$, $\alpha$, $\lambda$ and $\beta$, which is common in object detection. Our method is not as sensitive to $v$ since it only decides a box’s validity. Our parameter choice is either at or close to the optimal. The inference speed is only affected by $\alpha$. Other parameters are used in training or do not affect inference speed.

A3.5. Qualitative Results

We next show some qualitative results of models trained on KITTI Val 1 Split in Fig. 6. We depict the predictions of GrooMeD-NMS in image view on the left and the predictions of GrooMeD-NMS, Kinematic (Image) [12], and ground truth in BEV on the right. In general, GrooMeD-NMS predictions are more closer to the ground truth than Kinematic (Image) [12].

A3.6. Demo Video of GrooMeD-NMS

We next include a short demo video of our GrooMeD-NMS model trained on KITTI Val 1 Split. We run our trained model independently on each frame of the three KITTI raw [24] sequences - 2011_09_29_DRIVE_0026 and 2011_09_26_DRIVE_0009. None of the frames from these three raw sequences appear in the training set of KITTI Val 1 Split. We use the camera matrices available with the raw sequences but do not use any temporal information. Overlaid on each frame of the raw input videos, we plot the projected 3D boxes of the predictions and also plot these 3D boxes in the BEV. We set the frame rate of this demo at 10 fps. The demo is also available in HD at https://www.youtube.com/watch?v=WctKkyWwrno. In the demo video, notice that the orientation of the boxes are stable despite not using any temporal information.

Acknowledgements

This research was partially sponsored by Ford Motor Company and the Army Research Office under Grant Number W911NF-18-1-0330. This document’s views and conclusions are those of the authors and do not represent the official policies, either expressed or implied, of the Army Research Office or the U.S. Government.

We thank Mathieu Blondel and Quentin Berthet from Google Brain, Paris, for several useful discussions on differentiable ranking and sorting. We also discussed the logical operators’ relaxation with Ashim Gupta from the University of Utah. Armin Parchami from Ford Motor Company suggested the learnable NMS paper [32]. Enrique Corona and Marcos Paul Gerardo Castro from Ford Motor Company provided feedback during the development of this work. Shengjie Zhu from the Computer Vision Lab at Michigan State University proof-read our manuscript and suggested several changes. We also thank Xuepeng Shi from University College London for sharing the Distance-NMS [76] code for bench-marking. We finally acknowledge anonymous reviewers for their feedback that helped in shaping the final manuscript.
Figure 6: **Qualitative Results** (Best viewed in color). We depict the predictions of GrooMeD-NMS in image view on the left and the predictions of GrooMeD-NMS, Kinematic (Image) [12], and Ground Truth in BEV on the right. In general, GrooMeD-NMS predictions are more closer to the ground truth than Kinematic (Image) [12].