TOWARDS RELIABLE DETECTION OF EMPTY SPACE: CONDITIONAL MARKED POINT PROCESSES FOR OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks have set the state-of-the-art in computer vision tasks such as bounding box detection and semantic segmentation. Object detectors and segmentation models assign confidence scores to predictions, reflecting the model's uncertainty in object detection or pixel-wise classification. However, these confidence estimates are often miscalibrated, as their architectures and loss functions are tailored to task performance rather than probabilistic foundation. Even with well calibrated predictions, object detectors fail to quantify uncertainty outside detected bounding boxes, i.e., the model does not make a probability assessment of whether an area without detected objects is truly free of obstacles. This poses a safety risk in applications such as automated driving, where uncertainty in empty areas remains unexplored. In this work, we propose an object detection model grounded in spatial statistics. Bounding box data matches realizations of a marked point process, commonly used to describe the probabilistic occurrence of spatial point events identified as bounding box centers, where marks are used to describe the spatial extension of bounding boxes and classes. Our statistical framework enables a likelihood-based training and provides well-defined confidence estimates for whether a region is drivable, i.e., free of objects. We demonstrate the effectiveness of our method through calibration assessments and evaluation of performance.

1 INTRODUCTION

Deep neural networks (DNNs) have demonstrated outstanding performance in computer vision tasks like image classification (Wortsman et al., 2022), object detection (Zong et al., 2023), instance segmentation (Yan et al., 2023) and semantic segmentation (Xu et al., 2023). Object detection describes the task of identifying and localizing objects of a particular class, for example, by predicting bounding boxes or by labeling each pixel that corresponds to a specific instance. Semantic segmentation refers to assigning each pixel in an image to one of a predefined set of semantic classes. These tasks serve as an indispensable tool for scene understanding, providing precise information about the scenario. Computer vision tasks lend themselves to diverse application areas including safety-critical areas like robotics (Cartucho et al., 2021), medical diagnosis (Kang & Gwak, 2019) and automated driving (Xu et al., 2023). Along with high accuracy of the models, *prediction reliability in the form of uncertainty assessment (Maag & Riedlinger, 2024) and calibrated confidence estimates (Mehrtash et al., 2019) is also highly relevant*.

Object detectors provide confidence values ("objectness") about the correctness of each predicted object, while semantic segmentation models output a softmax probability score per pixel indicating the confidence of class affiliation. Both of the former represent notions of probability for correct predictions, which also reflects the prediction uncertainty of the model. For confidence scores to be statistically reliable, the assigned confidence should match the observed prediction accuracy, e.g., out of 100 predictions with a confidence of 70% each, 70 are correct. If this is approximately the case, we call the confidence assignments "well-calibrated". However, DNN confidences are frequently miscalibrated, with confidence values not matching the observed accuracy (Küppers et al., 2022). This problem can often be tackled by confidence re-calibration methods which regularize the training objective or establish calibration by post-processing (Guo et al., 2017). Particularly in the case of

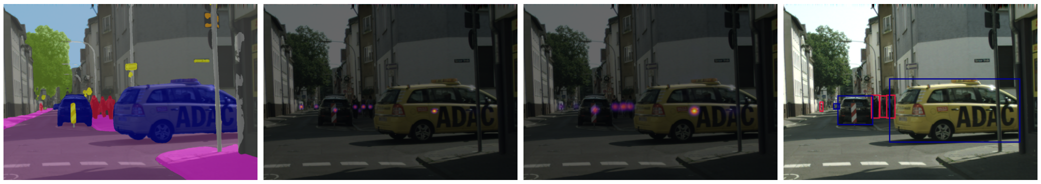

Figure 1: *Left*: Semantic segmentation prediction. *Center left*: Poisson point process intensity. *Center right*: Conditional marked Poisson point process intensity. *Right*: Bounding box prediction.

object detection, the focus is mainly on correctly calibrating objectness for predicted objects, while false negatives are ignored.

The problem of miscalibration often arises from a mis-specification of the architecture and/or the loss function of the model. The latter are often created based on heuristics and practical considerations, e.g., feature pyramid networks (Lin et al., 2017) or anchor box procedures (Redmon et al., 2016), working well empirically but lacking a probabilistic foundation. *The logic of stochastic independence of pixel-wise predictions like in semantic segmentation or objectness are violated in practice.* Moreover, the commonly used cross-entropy loss with one-hot targets does not intrinsically give rise to calibration, since its main goal is to optimize classification accuracy (Liu et al., 2024). The DNN is likely to predict the true class with high probability without calibrating the probabilities of the other classes. Thus, models often predict extremely high or low probabilities tending towards overconfidence.

Aside from the calibration of foreground predictions, object detection exhibits the peculiarity that the areas outside detections are (naturally) not annotated and learned as background during training. *Object detectors do not provide explicit uncertainty assessment for areas outside their own predictions.* That is, the model does not make a probability assessment of whether an area without detected objects is truly free of obstacles since such a mechanism is not built into the loss function. Such a model is then often overconfident in assuming that "no object" means the area is safe. This fact carries significant impact, e.g., in trajectory planning of autonomous cars and robotic agents where it is central to assess whether the planned track actually is collision-free. We consider the absence of adequate methods addressing this question to be a major safety and research gap in the field of safe automated driving. Our approach addresses this flaw and constitutes a step towards safe navigation.

In this work, we propose an object detection model based on spatial statistics, regarding bounding box data as realizations of a marked point process (Møller et al., 2003; Sherman, 2011). Such models have been used to model spatial phenomena, e.g., in astronomy, epidemiology or geostatistics, to describe the probabilistic occurrence of spatial point events. We add marks to center point events describing the spatial extent of objects (width and height) and class affiliation and obtain an object detection model based on the logic of counting processes. When conditioned on the number of events, our model allows for likelihood-based end-to-end training. Out of the box, *our model allows to compute well-defined notions of confidence for the event that a certain region in space is drivable*. We differentiate between two notions of "being drivable" depending on whether we consider to marked or non-marked point process. This gives rise to an object detector which defines confidences that an arbitrary test region in space is truly free or objects. This model conceptually provides aleatoric uncertainty, but can be combined with methods for estimating epistemic uncertainty, such as Bayesian approximation. An exemplary prediction of our model is shown in fig. 1. We summarize our contribution as follows:

- We develop a deep learning framework for modeling the intensity function of spatial point processes based on a negative log-likelihood loss function to obtain probabilistic statements about empty space.
- We propose a mathematically principled model for object detection based on the theory of marked Poisson point processes allowing for end-to-end training. Contrasted with existing object detectors, we have zero training and only a single inference hyperparameter.
- We propose an evaluation protocol for testing calibration of empty space confidences and evaluate our method on street scene data as well as on drone data as practical use cases.
- The proposed model is shown to be well-calibrated on image regions while having comparable performance to standard object detection architectures.

To our knowledge, this is the first time, a spatial statistics approach has been used for deep object recognition in computer vision. Our derivation of the loss function is a principled and mathematically underpinned approach to deep object detection highlighting parallels and differences with typical design choices in object detection. The source code of our method will be made publicly available.

2 Related work

Drivable area prediction. Drivable area or free space detection refers to a task related to autonomous driving that does not focus on objects, but on the classification between drivable and background (everything else). This can be done based on different sensoric data such as LiDAR, radar, camera and aerial imaging (Hortelano et al., 2023) where we focus on camera data. The basis for the drivable area prediction is often a semantic segmentation architecture, since pixel level predictions are desired (Chen et al., 2025; Qiao & Zulkernine, 2021). Different network design choices have been pursued to solve drivable area prediction including fully convolutional models (Yu et al., 2023; Chan et al., 2019), incorporation of LSTM layers (Lyu et al., 2019) or boundary attention units (Sun et al., 2019) to overcome task-specific difficulties. Another approach to improve drivable area detection involves multi-task architectures, e.g., for different segmentation tasks (Lee, 2021), detection of lane lines (Wang et al., 2024) or instance segmentation (Luo et al., 2024). In Qian et al. (2020), a unified neural network is constructed to detect drivable areas, lane lines and traffic objects using sub-task decoders to share designate influence among tasks. In addition to combining tasks, different types of sensor data can also help, such as combining RGB images with depth information (Fan et al., 2020; Jain et al., 2023). An uncertainty-aware symmetric network is presented in Chang et al. (2022) to achieve a favorable speed-accuracy trade-off by fully fusing RGB and depth data.

In contrast to our methodology, drivable area prediction and free space detection are based on design choices and practical considerations and do not exhibit well-defined notions of void confidence. In direct contrast to our approach, which *assigns occupation confidences to arbitrary test regions*, the aim of drivable area prediction is to concretely localize regions that do not contain obstacles.

Object detection models. In computer vision, space occupation by objects is classically treated in the form of object detection. Object detectors typically model the existence of foreground objects by assigning an objectness score to super pixels (Farhadi & Redmon, 2018; Ren et al., 2015; Liu et al., 2016) which is trained via binary cross entropy. Super pixels with sufficient objectness then contribute a bounding box prediction to the final set of detections which is subsequently pruned by non-maximum-suppression to filter double predictions that indicate the same object. Our proposed intensity model shares similarities with objectness in single stage detectors like the YOLO models (Redmon et al., 2016; Redmon & Farhadi, 2017; Farhadi & Redmon, 2018), SSD (Liu et al., 2016), FCOS (Tian et al., 2019) or also CenterNet (Duan et al., 2019) in that a spatial feature map is computed which indicates the existence of foreground objects. The same mechanism is used in region proposal modules of two-stage architectures like the R-CNN family (Girshick et al., 2014; Ren et al., 2015; Girshick, 2015; Cai & Vasconcelos, 2018) or probabilistic two-stage extensions like CenterNetv2 (Zhou et al., 2021). However, there is a central difference in the meaning of the feature maps. Objectness is interpreted on the basis of superpixels and in a binary fashion (foreground/background) where different pixels are stochastically independent as prescribed by the loss function. *Intensity on the other hand is built on a cumulative logic via integration over image regions* and, hence, modeled densely on the same resolution as the input image. Spatial dependence between pixels is incorporated by a free spatial point process loss function presented in section 3 and leads to statistically reliable confidence estimation for space occupation. Due to its focus on modeling center points, anchor-free models like CenterNet (Duan et al., 2019) and FCOS (Tian et al., 2019) share similarities with our model in terms of inference logic. We emphasize, however, that the "centerness" score of CenterNet is conceptually more closely related with objectness than intensity due to the built-in superpixel independence rendering both models misspecified for calibration of space occupation. Additionally, we do not claim that our model is capable of outperforming the above-mentioned object detectors in terms of detection accuracy as our model has not gone through several generations of architectural optimization. Rather, *we present the first ever approach to object detection which is fully probabilistically founded*. Our model is *capable of assigning calibrated void/occupation probabilities to arbitrary regions in space*, a central research gap for autonomous driving and robotic perception. An important evaluation protocol for object detectors in terms of the

PDQ score (Hall et al., 2020) has been introduced viewing object detection in light of the underlying probabilistic modeling assumptions for bounding box regression. Bounding box localization is modeled by Gaussian distributions, however, irrespective of the chosen optimization objective. In contrast, we rigorously derive the correct correspondence between regression loss and probabilistic model for bounding box uncertainty.

3 Method description

In this section, we propose our object detection model architecture and inference logic. To this end, we derive the optimization functional from the theory of spatial point processes and propose an implementation modeling point process intensity coupled with bounding box properties as marks.

3.1 Marked point process models

Bounding box data as marked point configurations. A statistical model has to reflect the structure of the data. In object detection, the data describing an instance is viewed as tuples $z = (\xi, m)$, where $\xi = (\xi_x, \xi_y) \in [0,1]^2$ encodes the location of the instance's center while $m = (h, w, \kappa) \in \mathbb{R}^2 \times \mathcal{C}$ proves height and width of the bounding box as well as the label κ from a discrete set of classes $\mathcal{C}$. We call ξ "(center) point" location and m the "mark" attached to ξ.

The object detection ground truth data for a given image consists of a set $\{z_1, \dots, z_n\}$ of marked points, where $n \in \mathbb{N}_0$ varies between images. The corresponding point configuration $\{\xi_1, \dots, \xi_n\}$ lives in the space of n-point configurations Ξ_n over $[0,1]^2$, where we allow for zero points, i.e., $\Xi_0 = \{\emptyset\}$, containing only the "empty configuration". The center points on a generic image are then specified by an element of $\Gamma = \bigoplus_{n \in \mathbb{N}_0} \Xi_n$, the space of finite point patterns. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, then *a random variable* $X : \Omega \to \Gamma$ *is called a (spatial) point process over* $[0,1]^2$. $N(A) = |X \cap A|$ is the associated counting process counting the number of points $X(\omega) = (\xi_1(\omega), \dots, \xi_{N(\omega)}(\omega))$ inside the measurable set $A \subseteq [0,1]^2$. $N(\omega) = N([0,1]^2)(\omega)$ gives the total count of points for the given random parameter $\omega \in \Omega$. Likewise, *a marked point process* $X_M : \Omega \to \Gamma_M$ *takes values in* $\Gamma_M = \bigoplus_{n \in \mathbb{N}_0} (\Xi_n \times M^n)$*, where the mark space* M *for object detection is given by* $M = \mathbb{R}^2 \times \mathcal{C}$. The projection $\{z_1, \dots, z_n\} \mapsto \{\xi_1, \dots, \xi_n\}$ associates a (non-marked) point process X with X_M and X_M can be considered as a point process over the extended space $[0,1]^2 \times M$.

We are interested in statistical models that are eligible to model the distribution of a marked point process X_M. Here we use the Poisson point process (PPP) as the simplest model and "workhorse" of spatial statistics. It is based on an intensity measure $\Lambda = \Lambda(z)\,\mathrm{d}z$, where $\mathrm{d}z$ is the Lebesgue measure on $[0,1]^2 \times \mathbb{R}^2 \times \mathcal{C}$ where on $\mathcal{C}$ we chose the cardinality measure. For $A_M \subseteq [0,1]^2 \times M$ specify the statistics of $N_M(A_M)$ by the Poisson distribution with intensity $\Lambda(A_M) = \int_{A_M} \Lambda(z)\,\mathrm{d}z$, i.e., the probability of finding $n \in \mathbb{N}_0$ marked point instances in A_M is given by

$$\mathbb{P}\left(N_M(A_M) = n\right) = \tfrac{1}{n!}\Lambda(A_M)^n \cdot \exp\left(-\Lambda(A_M)\right). \tag{1}$$

Choosing $A_M = A \times M$, $A \subseteq [0,1]^2$ measurable, also the associated point process of center points $N(A)$ is a Poisson point process with intensity on $[0,1]^2$ given by

$$\lambda(\xi) = \textstyle\int_M \Lambda(\xi, m)\,\mathrm{d}m. \tag{2}$$

Object detection via conditional marked Poisson point processes. Equation (1) defines a probabilistic model for bounding box data $\{z_1, \dots, z_n\}$ on a fixed image I. *A predictive model can be derived by conditioning the intensity* $\Lambda(z) = \Lambda(z|I)$ *on the input* $I \in [0,1]^2 \times \mathbb{R}^3$ *represented by three values for RGB channels.* We call such a model a conditional marked Poisson point process (CMPPP) model. The input resolution of the pixelized image I_d determines the discretization of the Lebesgue measure over $[0,1]^2$ and, therefore, the area/mass of each pixel. We denote by Π the set of pixel locations in $[0,1]^2$ associated with I_d.

Let us now consider suitable models for object detection from CMPPP. We take a factorizing model

$$\Lambda(z|I) = \Lambda(\xi, m|I) = \widetilde{\lambda}(\xi|I) \cdot p(m|\xi, I), \tag{3}$$

where $\Lambda(z|I) \geq 0$ and $p(m|\xi, I)$ is a Markov kernel that models the probability density of the mark $m = (h, w, \kappa)$ given that there is a bounding box centered in ξ. Since $\int_M p(m|\xi, I)\,\mathrm{d}m = 1$ for any $\xi \in [0,1]^2$, $\widetilde{\lambda}(\xi) = \lambda(\xi)$ follows from eq. (2) and eq. (3). It remains to model the conditional intensity and the conditional probability on mark space as follows

$$\lambda(\xi|I_{\mathrm{d}}) = \exp\left(L_\xi(I_{\mathrm{d}})\right), \qquad p(m|\xi, I_{\mathrm{d}}) = p_{w,h}(B_\xi(I_{\mathrm{d}})) \cdot p_\kappa(C_\xi(I_{\mathrm{d}})|B_\xi(I_{\mathrm{d}})) \tag{4}$$

for a continuous distribution $p_{w,h}$ for the bounding box width and height and some categorical distribution p_κ for the object class. Note, that both, $p_{w,h}$ and p_κ, are technically conditional on I_{d}, which is omitted in the notation for readability. *The functions L, B and C are realized as dense (pixel-wise, i.e., indexed by pixel ξ) output of a neural network* which is fitted based on data $(I_d, z_1, \dots, z_n)$, where $z_i = (\xi_i, w_i, h_i, \kappa_i)$ for $i = 1, \dots, n$ in a maximum-likelihood approach. In our implementation, $p_{w,h}$ is realized as a bivariate (independent) Laplace distribution with location parameter (w, h) and isotropic scale parameter σ which is discussed further in the next section. We model p_κ by a softmax distribution with logits $C_\xi(I_{\mathrm{d}})$, independently of $B_\xi(I_{\mathrm{d}})$.

3.2 Likelihood training approach and inference for CMPPP models

CMPPP loss function. Commonly, one chooses the negative log-density function with respect to the Lebesgue measure $\mathrm{d}x$ as loss function for a parametric model density p_θ, i.e., $\ell(x, \theta) = -\log p_\theta(x)$ where $X \sim \mu_\theta = p_\theta(x)\,\mathrm{d}x$ is a hypothesis on the distribution of the data represented by (independent copies of) the random variable $X : \Omega \to \mathbb{R}^d$. This is no longer feasible in the setting of point processes $X : \Omega \to \Gamma$ as no Lebesgue measure exists on the infinite dimensional space Γ and therefore the notion of a density w.r.t. $\mathrm{d}x$ is ill-defined. The notion of likelihood can, however, be adapted to reference measures other than $\mathrm{d}x$. Let, in the finite dimensional setting, μ be another (probability) measure on $\mathbb{R}^d$ with density $p_\mu(x)$ such that $p_\mu(x) = 0$ implies $p_\theta(x) = 0$. We obtain

$$\ell(x, \theta) = -\log p_\theta(x) = -\log \tfrac{p_\theta(x)}{p_\mu(x)} - \log p_\mu(x) = -\log\left(\tfrac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu}(x)\right) - \log p_\mu(x). \tag{5}$$

Here $\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu}(x) = \frac{p_\theta(x)}{p_\mu(x)}$ is the Radon-Nikodym (RN) derivative of μ_θ with respect to the reference measure μ. As the gradients $\nabla_\theta \ell(x, \theta)$ of eq. (5) do not depend on p_μ, training with the negative log likelihood loss is equivalent to training with the negative log-RN derivative. Furthermore, the gradients also do not depend on the particular choice of μ, provided the RN derivative exists.

The training of stochastic models with an infinite dimensional state space is based on the insight that RN derivatives with respect to adequately chosen reference measures μ still exist, as long as $\mu_\theta = p_\theta\,\mu$ holds, where the relative density $p_\theta(x) = \frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu}(x)$ again is the Radon-Nikodym derivative at $x \in \Gamma$. *As the reference measure, we choose the distribution μ of the homogeneous PPP over $[0,1]^2$ with intensity function $\lambda_\mu \equiv 1$.* The RN derivative of the distribution μ_θ of a process with intensity $\lambda_\theta(\xi)$ w.r.t. μ then is

$$\tfrac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu}(x) = \exp\left(-\textstyle\int_{[0,1]^2}(\lambda_\theta(\xi) - 1)\,\mathrm{d}\xi\right) \cdot \textstyle\prod_{l=1}^n \lambda_\theta(\xi_l), \quad x = \{\xi_1, \dots, \xi_n\} \in \Gamma. \tag{6}$$

This identity is a standard exercise in textbooks on spatial statistics, see, e.g., Daley & Vere-Jones (2006). For the convenience of the reader we provide a proof in appendix A.1. Equation (6) immediately generalizes to marked CMPPP if we replace $x \in \Gamma$ with $x = \{z_1, \dots, z_n\} \in \Gamma_M$, $\lambda_\theta(\xi_i)$ with $\Lambda_\theta(z_i|I)$, ξ with z, $[0,1]^2$ with $[0,1]^2 \times M$ and $\mathrm{d}\xi$ with $\mathrm{d}z$.

Inserting eq. (3) and eq. (4) into the updated version of eq. (6) and taking the negative logarithm, we obtain the CMPPP loss function for $(L^\theta, B^\theta, C^\theta)$:

$$\ell(x, \theta) = \textstyle\int_{[0,1]^2} \mathrm{e}^{L^\theta_\xi(I_{\mathrm{d}})}\,\mathrm{d}\xi - \sum_{i=1}^n L^\theta_{\xi_i}(I_{\mathrm{d}}) - \sum_{i=1}^n \left[\log(p_{w,h}(B^\theta_{\xi_i}(I_{\mathrm{d}})) + \log(p_\kappa(C^\theta_{\xi_i}(I_{\mathrm{d}})\right]. \tag{7}$$

Here, $z = \{z_1, \dots, z_n\} \in \Gamma_M$ is the bounding box configuration observed in the image I. Minimizing the loss function now enables a strictly likelihood-based training.

It is now easy to interpret the first two terms of eq. (7) as the loss for the center point intensity $\lambda_\theta = \exp(L^\theta)$ and hence a loss for a "distributed objectness score". Assuming a Laplace distribution, the $p_{w,h}$-term yields the standard L^1-loss for bounding box regression and the p_κ-term the cross entropy classification loss. The integral in the first term is discretized over $[0,1]^2$ according to the image resolution $H \times W$ as

$$\textstyle\int_{[0,1]^2} \exp\left(L^\theta_\xi(I_{\mathrm{d}})\right)\mathrm{d}\xi \approx \sum_{\xi \in \Pi} \exp\left(L^\theta_\xi(I_{\mathrm{d}})\right) \cdot \tfrac{1}{HW} \tag{8}$$

with each pixel obtaining area $\frac{1}{|\Pi|} = \frac{1}{HW}$. We note that we choose to not fit the scale variable σ of the Laplace distribution by the model. This makes the training of B^θ is independent of the value of $\sigma > 0$ such that training with the L^1-loss can be conducted first resulting in the approximately optimal weights $\widehat{\theta}$. Thereafter, the maximum likelihood equations for σ yield $\widehat{\sigma} = \frac{1}{n}\sum_{i=1}^{n} \left\| \binom{w_i}{h_i} - B^{\widehat{\theta}}_{\xi_i}(I_\mathrm{d}) \right\|_1$, i.e., the mean average deviation of the bounding box regression *allowing for object detection training with zero hyperparameters*.

Probabilistic predictions of empty space. On the basis of a trained model with parameters $(\widehat{\theta}, \widehat{\sigma})$ we now derive a probabilistic conditional prediction that a measurable test region $A \subseteq [0,1]^2$ is "free of objects". We interpret this statement in two different ways.

On the one hand, it may mean "*A is free of object centers*". The random variable X models object centers and $N = \delta_X$ is the associated counting (Dirac-) measure. Then, the event that "A is free" amounts to $X \cap A = \emptyset$ or $N(A) = 0$. Using the Poisson statistic (1) for the associated center point process, i.e., $\lambda_{\widehat{\theta}}(\cdot|I_\mathrm{d}) = \exp(L^{\widehat{\theta}}_{(\cdot)})$ instead of $\Lambda(\cdot)$, we obtain

$$\mathbb{P}_{\widehat{\theta}}(N(A) = 0|I) = \exp\left(-\textstyle\int_A \lambda_{\widehat{\theta}}(\xi|I)\,\mathrm{d}\xi\right) \approx \exp\left(-\frac{1}{HW}\sum_{\xi\in A\cap\Pi} \exp\left(L^{\widehat{\theta}}_{\xi}(I_d)\right)\right). \tag{9}$$

On the other hand, an interpretation is the event that "*A does not intersect any of the bounding boxes*" $[\xi_x - \frac{w}{2}, \xi_x + \frac{w}{2}] \times [\xi_y - \frac{h}{2}, \xi_y + \frac{h}{2}] =: b(z)$ for any $z = (\xi_x, \xi_y, w, h, \kappa) \in X_M$. To this end, consider the critical set $D^\mathrm{c}(\xi) \subseteq [0,1]^2 \times M$ for a point $\xi \in A$, that is, the set of all bounding boxes $z' \in [0,1]^2 \times M$ such that ξ is contained in the corresponding bounding box $b(z')$. It is easily seen that $D^\mathrm{c}(\xi) = \{z' = (\xi', w', h', \kappa') \in [0,1]^2 \times M : |\xi_x - \xi'_x| \le w/2$ and $|\xi_y - \xi'_y| \le h/2\}$. The critical set for the entire region A then is given by $D^\mathrm{c}(A) = \bigcup_{\xi\in A} D^\mathrm{c}(\xi)$ and the probability that no bounding box in a given image I intersects A is

$$\mathbb{P}_{\widehat{\theta}}(N(D^\mathrm{c}(A)) = 0|I) = \exp\left(-\textstyle\int_{D^\mathrm{c}(A)} \Lambda_{\widehat{\theta}}(z|I)\,\mathrm{d}z\right), \tag{10}$$

where $\Lambda(z|I)$ is evaluated using eq. (3) and eq. (4). Let us shortly consider the evaluation of the integral on the right hand side of eq. (10). By our modeling ansatz, the integral over $\Lambda_{\widehat{\theta}}$ separates to

$$\textstyle\int_{[0,1]^2\setminus A} \lambda_{\widehat{\theta}}(\xi|I) \cdot \int_{\{m\in M: b(\xi,m)\cap A\neq\emptyset\}} p(m|\xi, I)\,\mathrm{d}m\,\mathrm{d}\xi. \tag{11}$$

For the special case that the test region A is a rectangle with center point $(\xi^A_x, \xi^A_y) \in [0,1]^2$, width w^A and height h^A, A intersects $b(\xi, w, h, \kappa)$ if both, $|\xi^A_x - \xi_x| \le \frac{1}{2}(w^A + w)$ and $|\xi^A_y - \xi_y| \le \frac{1}{2}(h^A + h)$ hold. The inner integral then factorizes and the integral in eq. (10) becomes

$$\textstyle\sum_{\xi\in\Pi\setminus A} e^{L^{\widehat{\theta}}_{\xi}(I_\mathrm{d})} \cdot \int_{2|\xi^A_x - \xi_x| - w^A}^{\infty} \frac{1}{2\sigma} e^{-\frac{1}{\sigma}|w - B^{\widehat{\theta}}_{\xi,w}(I_\mathrm{d})|}\mathrm{d}w \cdot \int_{2|\xi^A_y - \xi_y| - h^A}^{\infty} \frac{1}{2\sigma} e^{-\frac{1}{\sigma}|h - B^{\widehat{\theta}}_{\xi,h}(I_\mathrm{d})|}\mathrm{d}h \tag{12}$$

which can easily be expressed by the cumulative distribution function of $p_{w,h}$. The evaluation of the void confidence generalizes e.g., to modeling with normal distributions and can easily be algorithmically implemented. While the inner integral over $\{m \in M : b(\xi, m) \cap A \neq \emptyset\}$ may in general be computed by logical querying of pixels and CDFs, more general shapes may also be treated via the inclusion-exclusion principle.

Prediction of foreground objects. For bounding box prediction, instead of the standard non-maximum suppression algorithm, *we exploit the counting statistics of the spatial point process and determine the expected number of center points in* $[0,1]^2$ *by* $\mathbb{E}[N] = \int_{[0,1]^2} \lambda_{\widehat{\theta}}(\xi|I)\mathrm{d}\xi \approx \frac{1}{HW}\sum_{\xi\in\Pi} \lambda_{\widehat{\theta}}(\xi|I_\mathrm{d})$. Afterwards, we extract this number of peaks (see fig. 2) from the intensity function to find the predicted center points. As the intensity function is sharply peaked but still somewhat spread-out, we crop square patches of 32×32 pixels around determined maxima before searching for the next peak. An ablation study on crop square size, *our only hyperparameter*, is provided in appendix A.2. We observe robust behavior over large range of settings. Marks are determined by evaluation of $B^{\widehat{\theta}}(I_\mathrm{d})$ and $C^{\widehat{\theta}}(I_\mathrm{d})$ feature maps at the respective locations.

4 Experiments

4.1 Model design and experimental setting

Network architecture. In order to model the functions L, B and C in eq. (4), we choose deep neural networks that are capable to compute pixel-wise outputs, in particular we utilize architectures used in semantic segmentation with $1 + 2 + |\mathcal{C}|$ output channels modeling the tuple $(L^\theta, B^\theta, C^\theta)$. Given ground truth data $(I_\mathrm{d}, z_1, \ldots, z_n)$, this allows for computing the full CMPPP loss (eq. (7)) and training end-to-end. In appendix A.3, we present additional experiments with a two-stage architecture and additional experiments where we model the residuals as normal distributions instead of Laplace distributions. We implement our CMPPP model in the MMDetection environment (Chen et al., 2019), importing segmentation architectures from MMSegmentation. Our investigations involve a DeepLabv3+ (Chen et al., 2018) model with ResNet-50 backbone, an FCN model with HRNet (Wang et al., 2021) backbone, as well as SegFormer-B5 (Xie et al., 2021) model. Training ran on a Nvidia A100 GPU with 80GBs of memory and standard (pre-set) parameter settings for training on the Cityscapes dataset.

Datasets. In our experiments, we use two datasets: a street scene dataset, where empty spaces are relevant for safety issues such as whether a planned trajectory is actually collision-free, and a aerial dataset e.g. for detecting free landing sites for drones. Regarding the perspective of robotics, we believe that this perspective is covered by the perspective of the ego car and that automated vehicles operate in environments that are more complex than typical environments for mobile robots. The Cityscapes dataset (Cordts et al., 2016) depicts dense urban traffic scenarios in various German cities. This dataset consists of $2{,}975$ training images and 500 validation images of size $1{,}024 \times 2{,}048$ from 18 and 3 different cities, respectively, with labels for semantic and instance segmentation of road users (different vehicles and humans). From the instance labeling, we obtain the center points and bounding box ground truth of the objects. The VisDrone-DET dataset (Zhu et al., 2022) shows aerial images of urban scenarios in different Chinese cities. The dataset consists of $6{,}471$ training and 548 validation images of different resolutions and bounding box annotations for different road-related semantic classes.

4.2 Numerical results

PPP intensity calibration. In this section, we compare our intensity prediction of the (non-marked) PPP with an analogous semantic segmentation prediction regarding the respective calibration of predicting empty spaces. *We aim at answering the question whether a test region A is drivable.* Given an input image I_d and learned weights, a semantic segmentation model computes a probability distribution $p(\cdot|I_\mathrm{d})_\xi \in [0,1]$ over $\mathcal{C}$ for each pixel $\xi \in \Pi$ specifying the probability $p(\kappa|I_\mathrm{d})_\xi \in [0,1]$ for each class $\kappa \in \mathcal{C}$. The probability that a region is drivable is given by $\mathbb{P}_S(A \text{ is drivable}) = \prod_{\xi \in \Pi \cap A} p(\text{“road”}|I_\mathrm{d})_\xi$, as the pixel predictions are assumed to be independent when conditioned on a fixed image I. We consider a region to be drivable if it contains no classes other than “road”. For the PPP model, the probability that the region is free is derived from the case $n = 0$ in eq. (1) under the discretization (8)

$$\mathbb{P}_P(A \text{ is drivable}) = \exp\left(-\textstyle\int_A \lambda(\xi)\,\mathrm{d}\xi\right) \approx \exp\left(-\frac{1}{|\Pi|}\sum_{\xi \in \Pi \cap A} \lambda(\xi)\right). \qquad (13)$$

We consider a region to be drivable if it contains no center point. To determine the calibration of the methods, we sample random boxes (test regions) of a fixed area s (height and width chosen randomly), where the number of boxes per image is fixed (here 50). For each box, we determine whether it is drivable and determine the respective probability. Based on these quantities, we calculate the expected calibration error (ECE) (Naeini et al., 2015) to evaluate calibration.

The results for the Cityscapes dataset are shown in Table 1, depending on the test box size. We observe that calibration for smaller boxes also achieves smaller ECE errors. Furthermore, we find that HRNet is more accurately calibrated than DeepLabv3+ for semantic segmentation. The reason for this could be the difference in model capacity and that HRNet does not use significant downsampling or pyramid architecture, instead relying on high-resolution representations through the whole process. However, both convolutional networks are less well calibrated than the SegFormer. These differences between the models become significantly smaller with our PPP method; only slight trends can still be

Table 1: Calibration values of semantic segmentation model (ECE$_S$) and our PPP method (ECE$_P$) for the Cityscapes dataset and different box sizes s.

	s	250	500	750	1,000	1,500	2,500	5,000	10,000
Deep-Labv3+	ECE$_S$	0.1102	0.1741	0.2033	0.2245	0.2521	0.2667	0.2417	0.1948
	ECE$_P$	0.0012	0.0017	0.0029	0.0029	0.0046	0.0062	0.0109	0.0164
HRNet	ECE$_S$	0.0413	0.0859	0.1142	0.1443	0.1785	0.2206	0.2295	0.1939
	ECE$_P$	0.0008	0.0012	**0.0014**	0.0019	**0.0022**	**0.0041**	**0.0053**	**0.0071**
SegFormer	ECE$_S$	0.0621	0.0585	0.0609	0.0593	0.0588	0.0582	0.0626	0.0713
	ECE$_P$	**0.0006**	**0.0008**	**0.0014**	**0.0018**	0.0027	0.0046	**0.0053**	0.0082

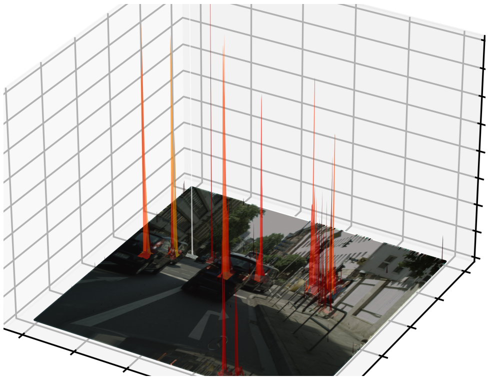

Figure 2: Intensity landscape over an input image from the Cityscapes val dataset. Peaks are mostly sharply localized and indicate foreground detections.

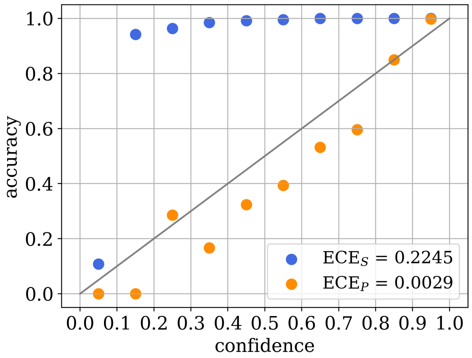

Figure 3: Confidence calibration plots for semantic segmentation (blue) and PPP (orange) with corresponding ECE for the Cityscapes dataset and the DeepLabv3+ detector and $s = 1{,}000$.

identified. It is quite evident that our PPP method is significantly better calibrated than the semantic segmentation prediction. A corresponding confidence calibration plot is shown in fig. 3. The semantic segmentation model is consistently underconfident. If any pixel indicates that the test area may not be drivable, the confidence vanishes as expected from the probability factorization. *In comparison, our method fluctuates close to optimal calibration (gray diagonal in the plot)*. We also observe good calibration on the more complex VisDrone dataset, as shown in table 2 (top row).

Bounding box detection. Using our CMPPP model, instead of the center point of an object, we receive a bounding box prediction with the corresponding class and score value. The latter reflects the probability (12) that a region is drivable, whereby we use the Laplace distribution function as in eq. (12) due to training with the L^1-loss. An example of CMPPP confidence and the bounding box prediction is shown in fig. 1 (see appendix A.4 for further visualizations).We observe the differences in intensity between the objects more clearly than with the PPP because box height and width contribute to the confidence level.

In this object detection application, we consider a test regions to be drivable if the test box and the ground truth bounding box do not overlap. In the same way as before, we sample 50 random test boxes of a fixed size s per image to evaluate the calibration. The results depending on s are given in table 3 for Cityscapes and in table 2 (bottom row) for VisDrone. While the PPP only predicts the center points of objects, CMPPP adds the prediction of the height and width of the bounding box, which is more challenging for confidence calibration. We achieve slightly worse values compared to the PPP predictions, although the calibration errors are still small. The Segformer model consistently exhibits better calibration than its convolutional counterparts. However, per architecture we do not identify a clear trend between error and box size indicating robustness across test region scales.

Object detection is usually evaluated using mean average precision (mAP), which assesses detection capability and accuracy. On the two classes, persons and vehicles, the DeepLabv3+ CMPPP model achieves a mAP of 49.43%, HRNet 55.49% and SegFormer 51.04% on the Cityscapes images. In

Table 2: Calibration values of our DeepLabv3+ PPP model (ECE$_P$) and CMPPP object detector (ECE$_{BB}$) for the VisDrone dataset for different box sizes s.

	250	500	750	1,000	1,500	2,500	5,000
ECE$_P$	0.0111	0.0206	0.0276	0.0352	0.0502	0.0704	0.1114
ECE$_{BB}$	0.0793	0.0968	0.1190	0.1389	0.1829	0.2459	0.3807

Table 3: Calibration values of our CMPPP object detection models (ECE$_{BB}$) for the Cityscapes dataset and different box sizes s.

	250	500	750	1,000	1,500	2,500	5,000	10,000
DeepLabv3+	0.1011	0.0925	0.0869	0.0842	0.0803	0.0697	0.0692	0.0762
HRNet	0.1223	0.1133	0.1089	0.1050	0.0954	0.0781	0.0665	0.0641
SegFormer	0.0905	0.0813	0.0760	0.0737	0.0627	0.0538	0.0476	0.0600

comparison, Faster R-CNN and CenterNet, well-known object detection networks, obtain mAP value of 59.32% and 57.08%, respectively. We do not claim that our model is capable of outperforming these models in terms of object detection performance as our model has not gone through several generations of architectural optimization. Rather, our model is capable of assigning well-calibrated occupation probabilities to arbitrary regions in space. Treating superpixel objectness similarly to softmax confidences, Faster R-CNN and CenterNet both compute highly ill-calibrated void confidences, obtaining ECE values of 0.9915, as expected.

Runtime and scalability. Our models essentially have the complexity of modern semantic segmentation architectures with minimal additional post-processing. Our DeepLabv3+ model (43.6M params, 16.2 FPS) is slightly slower than a comparable Faster R-CNN (41.4M params, 29.4 FPS) while our HRNet model (65.9M params, 15.4FPS) is on par with a Faster R-CNN with ResNeSt50 backbone (65.8M params, 16.4FPS). Overall, we conclude that our model scales well along with existing architectures even without specific tuning for efficiency.

5 LIMITATIONS

In fig. 3, we observe residual miscalibration of the model for center bins which we hypothesize is due to the fact that the model invested significant amounts of capacity to also calibrate $\mathbb{P}(N(A) = n|I)$ for other n. Our model assigns square patch intensity to any found peak during inference which often conflicts with large foreground objects whose intensity is spread over larger areas. This suggests using depth-dependent patch sizes to differentiate between objects from different size scales. Finally, road participants and obstacles occupy physical space while we model "free"/non-interacting point configurations in our model, not incorporating repelling potentials between events.

6 CONCLUSION

In this work, we have introduced a novel object detection architecture and learning objective guided by the question "With what probability is some particular region of the input image devoid of objects, i.e., drivable?". Following a principled approach based on the theory of spatial point processes, we have derived an object detection model which may be trained by a notion of negative log-likelihood to model the object intensity function over the input image. Modeling the point configuration with shape and class distribution markings constitutes an object detection model capable of assigning a meaningful confidence to the event of a test region intersecting any predicted object in the image. We investigate three instances of our model on two application-driven dataset and show in numerical experiments that it is capable of solid object detection performance and is well-calibrated on emptiness. Compared to semantic segmentation and conventional object detectors, we obtain significantly better confidence calibration, and particularly, the first object detection models providing reliable information about object-free areas.

REFERENCES

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving Into High Quality Object Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, Salt Lake City, UT, June 2018. IEEE.

Joao Cartucho, Samyakh Tukra, Yunpeng Li, Daniel S. Elson, and Stamatia Giannarou. Visionblender: a tool to efficiently generate computer vision datasets for robotic surgery. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 9(4):331–338, 2021. doi: 10.1080/21681163.2020.1835546.

Yu-Ching Chan, Yu-Chen Lin, and Peng-Cheng Chen. Lane mark and drivable area detection using a novel instance segmentation scheme. In *2019 IEEE/SICE International Symposium on System Integration (SII)*, pp. 502–506, 2019. doi: 10.1109/SII.2019.8700359.

Yicong Chang, Feng Xue, Fei Sheng, Wenteng Liang, and Anlong Ming. Fast road segmentation via uncertainty-aware symmetric network. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 11124–11130. IEEE Press, 2022. doi: 10.1109/ICRA46639.2022.9812452.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *European Conference on Computer Vision (ECCV)*, pp. 833–851, Cham, 2018. Springer International Publishing. doi: 10.1007/978-3-030-01234-2_49.

Xing Chen, Yujiao Dong, Xinyong Li, Xunjia Zheng, Hui Liu, and Tengyuan Li. Drivable area recognition on unstructured roads for autonomous vehicles using an optimized bilateral neural network. In *Scientific Reports*, 2025. doi: 10.1038/s41598-025-94655-1.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.350.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer Science & Business Media, 2006.

Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. CenterNet: Keypoint Triplets for Object Detection, April 2019. arXiv:1904.08189 [cs].

Rui Fan, Hengli Wang, Peide Cai, and Ming Liu. Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 340–356, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58577-8.

Ali Farhadi and Joseph Redmon. Yolov3: An incremental improvement. In *Computer vision and pattern recognition*, volume 1804, pp. 1–6. Springer Berlin/Heidelberg, Germany, 2018.

Ross Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, December 2015.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1321–1330. JMLR.org, 2017.

David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1031–1040, 2020.

Juan Luis Hortelano, Jorge Villagrá, Jorge Godoy, and Víctor Jiménez. Recent developments on drivable area estimation: A survey and a functional analysis. *Sensors*, 23(17), 2023. ISSN 1424-8220. doi: 10.3390/s23177633.

Mahek Jain, Guruprasad Kamat, Rochan Bachari, Vinayak A. Belludi, Dikshit Hegde, and Ujwala Patil. Affordrive: Detection of drivable area for autonomous vehicles. In Pradipta Maji, Tingwen Huang, Nikhil R. Pal, Santanu Chaudhury, and Rajat K. De (eds.), *Pattern Recognition and Machine Intelligence*, pp. 532–539, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-45170-6.

J. Kang and J. Gwak. Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access*, 7:26440–26447, 2019. doi: 10.1109/ACCESS.2019.2900672.

Fabian Küppers, Anselm Haselhoff, Jan Kronenberger, and Jonas Schneider. *Confidence Calibration for Object Detection and Segmentation*, pp. 225–250. Springer International Publishing, Cham, 2022. ISBN 978-3-031-01233-4. doi: 10.1007/978-3-031-01233-4_8.

Dong-Gyu Lee. Fast drivable areas estimation with multi-task learning for real-time autonomous driving assistant. *Applied Sciences*, 11(22), 2021. ISSN 2076-3417. doi: 10.3390/app112210713.

Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. volume 9905, pp. 21–37. 2016.

Yuchi Liu, Lei Wang, Yuli Zou, James Zou, and Liang Zheng. Optimizing calibration by gaining aware of prediction correctness, 2024.

Tong Luo, Yanyan Chen, Tianyu Luan, Baixiang Cai, Long Chen, and Hai Wang. Ids-model: An efficient multitask model of road scene instance and drivable area segmentation for autonomous driving. *IEEE Transactions on Transportation Electrification*, 10(1):1454–1464, 2024. doi: 10.1109/TTE.2023.3293495.

Yecheng Lyu, Lin Bai, and Xinming Huang. Road segmentation using cnn and distributed lstm. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2019. doi: 10.1109/ISCAS.2019.8702174.

Kira Maag and Tobias Riedlinger. Pixel-wise gradient uncertainty for convolutional neural networks applied to out-of-distribution segmentation. *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, pp. 112–122, 2024. doi: 10.5220/0012353300003660.

Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39:3868–3878, 2019.

Jesper Møller, P. Bickel, Peter J. Diggle, S. Fienberg, K. Krickeberg, I. Olkin, N. Wermuth, and S. Zeger (eds.). *Spatial Statistics and Computational Methods*, volume 173 of *Lecture Notes in Statistics*. Springer, New York, NY, 2003. ISBN 978-0-387-00136-4 978-0-387-21811-3. doi: 10.1007/978-0-387-21811-3.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 2901–2907. AAAI Press, 2015. ISBN 0262511290.

Yeqiang Qian, John M. Dolan, and Ming Yang. Dlt-net: Joint detection of drivable areas, lane lines, and traffic objects. *IEEE Transactions on Intelligent Transportation Systems*, 21(11):4670–4679, 2020. doi: 10.1109/TITS.2019.2943777.

Donghao Qiao and Farhana Zulkernine. Drivable area detection using deep learning models for autonomous driving. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 5233–5238, 2021. doi: 10.1109/BigData52589.2021.9671392.

Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Los Alamitos, CA, USA, June 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.91.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

Michael Sherman. *Spatial statistics and spatio-temporal data: covariance functions and directional properties*. John Wiley & Sons, 2011.

Jee-Young Sun, Seung-Wook Kim, Sang-Won Lee, Ye-Won Kim, and Sung-Jea Ko. Reverse and boundary attention network for road segmentation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 876–885, 2019. doi: 10.1109/ICCVW.2019.00116.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully Convolutional One-Stage Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9626–9635, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-72814-803-8. doi: 10.1109/ICCV.2019.00972.

Huinian Wang, Jingyao Wang, Baoping Xiao, Yizhou Jiao, and Jinghua Guo. Drivable area and lane line detection model based on semantic segmentation. In *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, pp. 1–7, 2024. doi: 10.1109/ICMI60790.2024.10585878.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090. Curran Associates, Inc., 2021.

J. Xu, Z. Xiong, and S. P. Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired by pid controllers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19529–19539, Los Alamitos, CA, USA, 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01871.

B. Yan, Y. Jiang, J. Wu, D. Wang, P. Luo, Z. Yuan, and H. Lu. Universal instance perception as object discovery and retrieval. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15325–15336, Los Alamitos, CA, USA, 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01471.

Yue Yu, Yanhui Lu, Pengyu Wang, Yifei Han, Tao Xu, and Jianhua Li. Drivable area detection in unstructured environments based on lightweight convolutional neural network for autonomous driving car. *Applied Sciences*, 13(17), 2023. ISSN 2076-3417. doi: 10.3390/app13179801.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection, March 2021.

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2022. doi: 10.1109/TPAMI.2021.3119563.

Z. Zong, G. Song, and Y. Liu. Detrs with collaborative hybrid assignments training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6725–6735, Los Alamitos, CA, USA, 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.00621.