Learning point process models for vehicles detection using CNNs in satellite images

Jules Mabon, Mathias Ortner, Josiane Zerubia

Learning point process models for vehicles detection using CNNs in satellite images

1st Jules Mabon
Inria, Université Côte d'Azur
Sophia-Antipolis, France

2nd Mathias Ortner
Airbus Defense and Space
Toulouse, France

3rd Josiane Zerubia
Inria, Université Côte d'Azur
Sophia-Antipolis, France

***Abstract*—We present a method combining marked point processes and convolutional neural networks applied to the detection of small objects in optical satellite images. In such images, objects are densely scattered, and visual information is scarce. The point process framework allows factoring in priors to account for object interactions. Classical point process approaches make use of contrast measures to account for object location. These fail when contrast is low and visual aspect is varied. We replace those with terms build from convolutional neural network outputs. Moreover, we propose a method to learn the parameters of the point process energy model. We show our approach improves results from the straight convolutional neural network outputs. The code will be available at github.com/Ayana-Inria/**

***Index Terms*—object detection, point process, convolutional neural network, energy based model, remote sensing**

I. Introduction

While object detection has been thoroughly studied for the past 20 years [1], small object detection in optical satellite images remains challenging due to the limited spatial resolution, where objects of interest such as vehicles are only few pixels large, and thus lack visual information. Additionally, the dense scattering of objects increases the difficulty in separating instances and introduces interactions between neighboring objects. In our work, we aim at extracting the geometrical configuration of objects as vector information, in images with resolutions around 0.5 m.

Convolutional neural networks (CNNs) have, in recent years, greatly improved the object detection performance on such data [2]. Most of these approaches first extract features through convolutions, then propose a series of boxes (anchors) that are refined by regression afterwards [3]–[5]. Some deal with the inherent scale issues of remote sensing by using feature pyramids [4]. To alleviate both the imbalance issue between objects of interest and background, and limited visual features, others propose to use the relation between the two [6], [7]. Lastly [8] proposes to introduce prior knowledge on objects of interest via the input of text-modal information. However, the majority of these approaches do not consider the interactions between objects more than the non-superposition introduced by the post-processing non-maximum suppression step. We believe the priors over the interactions can complement the limited visual information typical of remotely sensed data.

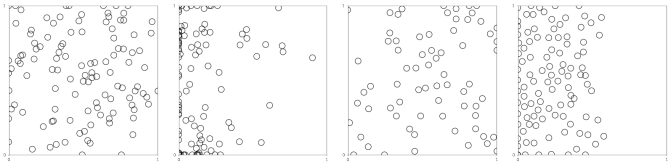
On the other hand, approaches based on Marked Point Processes [9] propose a framework that models the set of objects as a whole instead of independent instances detections. It jointly solves the detection and selection of objects. Such models combine data from the image with priors on the objects and their interactions. These approaches have been applied to detection in microscopy imagery [10], [11], or remote sensing data [11]–[14] among other tasks. However, these previous approaches rely on contrast measures to assert the correspondence of points with the image: while these perform well in highly contrasted images, they start to fail when the scene is more complex (shadows, partial occlusions), and the objects visual aspect is more varied.

With the proposed approach, we leverage the information extraction capabilities of CNNs with the modeling of the MPP: formulating the detection task as an energy minimization problem while introducing priors, building data terms from simple CNN outputs to replace contrast measures, and finally learning the parameters of the energy model on data. We summarize our contributions as follows:

- Introducing a parameter-light interaction model on top of a CNN detection model as a Marked Point Process (MPP).
- Estimating the MPP energy model parameters with contrastive divergence.
- Proposing an efficient sampling of the MPP using diffusion dynamics.
- Demonstrating the geometric regularization capabilities and noise robustness of our method on several datasets.

First we introduce the theoretical base for the MPP, to then define the model combining the CNN data terms and the interaction priors. We present the estimation of the energy model parameters followed by the method to sample the point process. Lastly we compare our models on detection task on benchmark data and data provided by Airbus Defense and Space (ADS).

Thanks to BPI France (LiChiE contract) for funding, and to the OPAL infrastructure from Université Côte d'Azur for providing computational resources and support.



(a) uniform (b) gradient (c) repulsive (d) gradient + repulsive

Fig. 1. Point processes derived from simple energies; (a): $V_a(y) \propto 1$, (b): $V_b(y) \propto y_i$, (c): $V_c(y|\mathcal{N}_y) \propto \min_{y' \in \mathcal{N}_y} d(y, y')$, (d): $V_d(y|\mathcal{N}_y) \propto V_b + V_c$. Horizontal axis corresponds to coordinate i and vertical to j.

II. Point Process for Object Detection

A. Marked Point Process

A configuration of points Y is a finite non-ordered set of elements of $S \times M$, with the image space as $S \subset \mathbb{R}^2$ and M the mark space. A mark can be any random variable from the radius of a circle to a discrete categorization of the object. In our case, a point $y \in Y$ is composed of coordinates y_i, y_j in S, and three marks that describe a rectangle : width y_a, length y_b and angle y_α. The set of all configurations with any number of points is denoted by $\mathcal{Y} = \bigcup_{n=0}^{\infty} (S \times M)^n$.

A configuration of points Y is modeled as the realization of a non-uniform MPP; the density of which is defined by h relative to the uniform point process [9]. The density derives from an energy U, through a non-normalized Gibbs density :

$$h(Y|X) \propto \exp(-U(Y|X)) \tag{1}$$

For object detection, the density is function of the image X; the configuration $\widehat{Y}$ that best fits one image minimizes the energy $U(Y|X)$.

B. Energy model

The total energy of a configuration $U(Y|X)$ is computed for each point as a weighted sum of K energy terms V_k:

$$U(Y|X, \theta) = \sum_{y \in Y} \theta_{w,0} + \sum_{k=1}^{K} \theta_{w,k} V_k(y|X, \mathcal{N}_y, \theta) \tag{2}$$

We distinguish between *prior terms* ($V_k(y|\mathcal{N}_y, \theta)$) that only depend on y and its neighborhood $\mathcal{N}_y$ (set of points in Y within a distance d_{max}), that measure the coherence of the configuration itself, and *data terms* ($V_k(y|X, \theta)$) that are function of y and the image X (also referred as observation in remote sensing): these measure the correspondence of the points with the image.

Here, θ denotes the set of model parameters to learn (*eg.* weights $\theta_{w,1}, \ldots, \theta_{w,K}$).

The sum formulation of (2) allows for good explainability of inferred configurations, along with easy composition of these priors and data terms that each describe object behaviors (see Fig. 1).

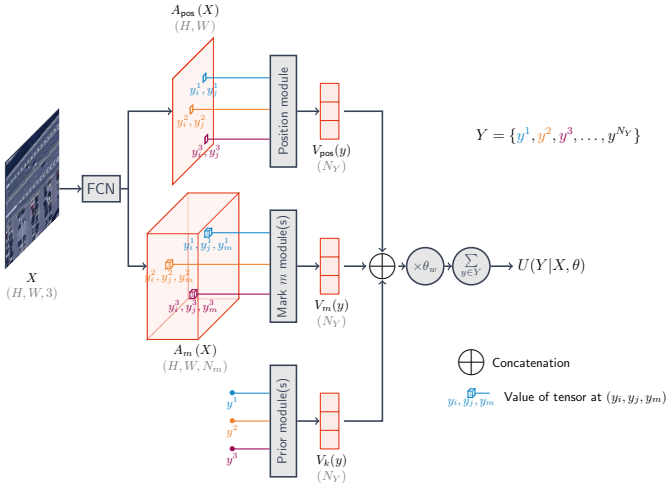


Fig. 2. Simplified model architecture

III. From CNN output to data terms

Previous data terms for MPP [10]–[14] are based on contrast measures, that are crafted to fit each application. Moreover, these measures rely heavily on the high contrast between objects and their background along with limited object and background diversity. On the other hand CNNs have shown to be very efficient at extracting features from images for object detection and classification. In the following section, inspired from [15], we will show how to interpret a CNN based object detector output to get an energy that measures the fitness of a configuration against an image.

The pipeline to compute $U(Y|X, \theta)$ is illustrated in Fig. 2. From an image X, we first extract 4 energy maps (position and three marks). For any configuration $Y \in \mathcal{Y}$, we can compute the energy terms V_k for every point y in configuration Y using the already computed energy maps. Weighting and summing the energy terms gives the energy $U(Y|X, \theta)$ over the whole configuration.

A. Position likelihood term

CNN based object detection method such as [16] make use of a heatmap to find object centers. The object center probability map is obtained as follows: A fully convolutional model such as Unet [17] transforms an image X of size H, W into a tensor $A_{\text{pos}}(X) \in \mathbb{R}^{H \times W}$. The probability-like measure of an object center at coordinates (y_i, y_j) is given by $p(y_i, y_j|X) = \sigma(A_{\text{pos}}(X)[y_i, y_j])$, with $\sigma(\cdot)$ the sigmoid function, and $A_{\text{pos}}(X)[y_i, y_j]$ the value of map $A_{\text{pos}}(X)$ at coordinates (y_i, y_j).

Relating this probability with the Gibbs model, we get the following position energy, with $\theta_{t,\text{pos}}$ a real-valued offset parameter:

$$V_{\text{pos}}(y_i, y_j|X) = \ln(1 + \exp(-A_{\text{pos}}(X)[y_i, y_j] + \theta_{t,\text{pos}})) \tag{3}$$

As the raster map $A_{\text{pos}}(X)$ defines energies over discrete pixel positions, and $(y_i, y_j) \in \mathbb{R}^2$, we perform bilinear interpolation to get energy $V_{\text{pos}}(y_i, y_j|X)$ for every real-valued location in S. In this paper we use the CNN built in [18].

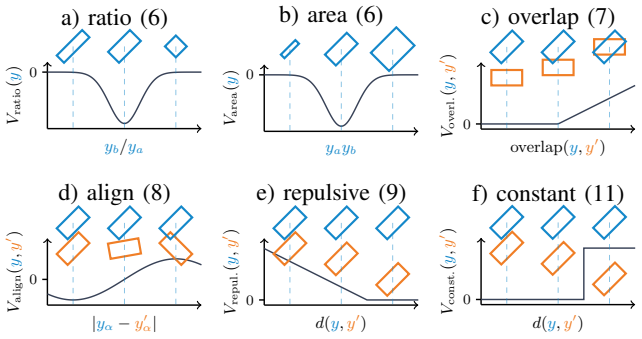


Fig. 3. Illustration of some energy priors.

B. Mark likelihood term

For every mark m (here $m \in \{a, b, \alpha\}$), we define the mark data term $V_m(y_m|y_i, y_j, X)$ from the output of a CNN. We denote (y_i, y_j) a position in image X, and m_k a discrete value of a mark such as $\forall k = 1, \ldots, N_m$, $m_k = m_{\min} + k(m_{\max} - m_{\min})/N_m$. Supposing we have a CNN model trained to output $p\left(m_k|y_i, y_j, X\right)$, this probability of value m_k is given by the softmax function applied to the CNN output vector $A_m\left(X, y_i, y_j\right)$;

$$p\left(m_k|y_i, y_j, X\right) = \frac{\exp\left(A_m\left(X, y_i, y_j\right)[k]\right)}{\sum_{k'=1}^{N_m} \exp\left(A_m\left(X, y_i, y_j\right)[k']\right)}. \quad (4)$$

Thus, we define the energy for the mark m as

$$V_m\left(y_m|y_i, y_j, X\right) = -A_m\left(X, y_i, y_j\right)[k_{y_m}] + \ln\left(\sum_{k=1}^{N_m} \exp\left(A_m\left(X, y_i, y_j\right)[k]\right)\right) \quad (5)$$

with k_{y_m} equal to the mark value y_m mapped to $[0, N_m]$; $k_{y_m} = N_m(y_m - m_{min})/(m_{max} - m_{min})$.

IV. Priors on configurations

The energy U encompasses several priors as energies, those allow to regularize the configuration against the data terms.

A. Object priors

These are functions of the current point y only. For $k \in \{\text{ratio}, \text{area}\}$:

$$V_k(y) = -\exp\left(-0.5\left(f_k(y) - \mu_k\right)^2 \sigma_k^{-2}\right) \quad (6)$$

with $f_{\text{ratio}}(y) = y_b/y_a$ and $f_{\text{area}}(y) = y_a y_b$. Parameters μ_k,σ_k are respectively the average observed value and the standard deviation.

B. Interaction priors

The following priors (illustrated in Fig. 3) depend on the neighborhood of the point y. The term in (7) penalizes overlapping objects (with $\theta_{t,\text{overl.}}$ the overlap tolerated threshold), (8) favors aligned objects, (9) and (10) are respectively repulsive and attractive priors, finally (11) allows to adjust the energy of neighborless points. With $f_{\text{overl.}}(y, y')$ the intersection over the smallest area of y and y', and $\text{Relu}(x) = \max(0, x)$ for $x \in \mathbb{R}$:

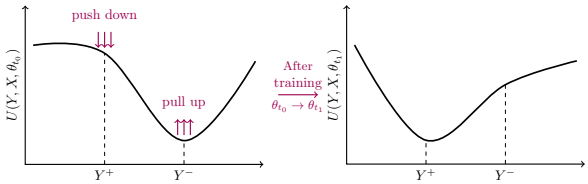


Fig. 4. Effect of training the energy with contrastive samples generated from the current θ_n (figure adapted from [19]).

$$V_{\text{overl.}}\left(y|\mathcal{N}_y, \theta\right) = \max_{y' \in \mathcal{N}_y} \left\{\text{Relu}\left(f_{\text{overl.}}(y, y') - \theta_{t,\text{overl.}}\right)\right\} \quad (7)$$

$$V_{\text{align}}\left(y|\mathcal{N}_y, \theta\right) = \min_{y' \in \mathcal{N}_y} \left\{-\cos\left(|y_\alpha - y'_\alpha|\right)\right\} \quad (8)$$

$$V_{\text{repul.}}\left(y|\mathcal{N}_y, \theta\right) = \max_{y' \in \mathcal{N}_y} \left\{1 - \frac{d\left(y, y'\right)}{d_{max}} - \theta_{t,\text{repul.}}\right\} \quad (9)$$

$$V_{\text{attra.}}\left(y|\mathcal{N}_y, \theta\right) = \min_{y' \in \mathcal{N}_y} \left\{\frac{d\left(y, y'\right)}{d_{max}} - \theta_{t,\text{attra.}}\right\} \quad (10)$$

$$V_{\text{const.}}\left(y|\mathcal{N}_y, \theta\right) = \mathbb{1}_{|\mathcal{N}_y|=0} \quad (11)$$

V. Learning Energy Model parameters

To find the parameters $\widehat{\theta}$ for the energy model, we adapt a method for Energy Based Models [19] that maximizes the likelihood :

$$P(Y_1^+, \ldots, Y_{|\mathcal{S}|}^+|X_1, \ldots, X_{|\mathcal{S}|}, \theta) = \prod_{k=1}^{|\mathcal{S}|} P(Y_k^+|X_k, \theta) \quad (12)$$

Where Y_k^+, X_k is the k^{th} image and ground truth pair of the dataset $\mathcal{S}$. As discussed in [19], minimizing directly the negative log-likelihood, requires computing an intractable integral over $\mathcal{Y}$. To avoid this, Hinton [20], [21] proposes approximating this integral, within a stochastic gradient descent, with a single *contrastive* sample. This sample is drawn with Monte Carlo Markov chain, at step t, from $h(Y|X, \theta_t)$; maximizing the energy of the contrastive sample Y^- while minimizing the energy of the *positive* sample Y^+. Du [22] improves upon this approach by using a replay buffer: using the contrastive sample from last epoch as initialization for drawing the current contrastive sample.

From the above we get the loss

$$\mathcal{L}(\theta_t, Y^+, Y^-, X) = U(Y^+|X, \theta_t) - U(Y^-|X, \theta_t), \quad (13)$$

of which we illustrate the effect in Fig. 4. This loss is minimized with the following procedure:

$\beta \leftarrow \emptyset$, $t \leftarrow 0$
while not converged **do**
  For one X, Y^+ pair in dataset $\mathcal{S}$
  $\widetilde{Y}_0 \sim \beta$ with probability 95% and $\mathcal{U}(\mathcal{Y})$ otherwise
  $Y^- \leftarrow \texttt{Sample}(Y_0, X, \theta_t)$
  $\beta \leftarrow \beta \cup Y^-$
  $\Delta\theta_t \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, Y^+, Y^-, X)$
  Update θ_t to θ_{t+1} based on $\Delta\theta_t$ with SGD [23]
  $t \leftarrow t + 1$
end while

VI. Sampling configurations from an energy

A. RJMCMC

Canonical sampling of Point Processes is performed with Reversible Jump Markov Chain Monte Carlo (RJMCMC) [24]. This method expands upon a Metropolis Hastings algorithm, by allowing to jump across dimensions. In the case of point processes [12]–[14], the jump adds or removes points, as the configurations $Y \in \mathcal{Y}$ have no fixed cardinality. The basic kernels to satisfy the convergence conditions, is the uniform birth and death kernel: it adds or removes points uniformly to the current configuration Y.

The RJMCMC relies on an accept-reject mechanism, driven by the acceptance rate r: kernel Q proposes a move from configuration Y to Y', and is accepted with probability $\min(1, r)$. With temperature T:

$$r = \frac{Q(Y' \to Y)}{Q(Y \to Y')} \exp\left(-\frac{U(Y'|X,\theta) - U(Y|X,\theta)}{T}\right) \quad (14)$$

A uniform birth kernel Q_B proposes a new configuration Y' with an added point y sampled uniformly in $S \times M$. To accelerate convergence, our birth kernel samples the new y in $S \times M$ from a birth density [12] given some pre-computed density derived from the CNN outputs. To sample from this density $d(y)$, we first sample a pixel q in the discretized space[1] P with density $\tilde{d}(q)$ (15), then sample y inside q uniformly (16). Thus, denoting q_y the pixel containing y (with pixel area $|q_y| = 1$):

$$\tilde{d}(q) \propto \exp(-\theta_{w,pos} V_{\text{pos}}(q_i, q_j|X)) \quad (15)$$

$$d(y) = \frac{1}{|q_y|}\tilde{d}(q_y). \quad (16)$$

Previously, computing the birth map could be quite computationally expensive when using contrast measures (as one would need to compute the contrast of a set of differently shaped and oriented objects for every position in the image) or would require some heuristic (e.g. when detecting boats, generating a binary map of water bodies to restrict the search space). Here the birth map is obtained by simple scalar operations (see (15) and (3)) on the CNN output $A_{\text{pos}}(X)$. Finally, this density formulation can be extended to take into account mark terms V_m, in the same way it is done for the position energy.

The birth kernel Q_B from Y to $Y' = Y \cup \{y\}$, with y sampled with density d, is balanced as such (supposing Q_B and Q_D are picked with equal probabilities):

$$Q_B(Y \to Y') = \frac{d(y)}{\lambda} \qquad Q_B(Y' \to Y) = \frac{1}{|Y'|}. \quad (17)$$

The death kernel Q_D, samples a point y in Y to be removed and produce Y'. To ensure detailed balance:

$$Q_D(Y \to Y') = \frac{1}{|Y|} \qquad Q_D(Y' \to Y) = \frac{d(y)}{\lambda}. \quad (18)$$

[1] P is the discretized space $S \times M$, defined as $P = [\![1, H]\!] \times [\![1, W]\!] \times \prod_{m \in \{a,b,\alpha\}} \{m_k,\, k = 1, \ldots, N_m\}$.

B. Jump diffusion

For the RJMCMC Green [24] proposes a transform kernel (i.e. moves, rotates or scales one point randomly) to explore the space of all configurations $\mathcal{Y}$ at a fixed dimension with random perturbations on Y. For more efficiency we leverage the energy gradient over Y with stochastic diffusion (or Langevin dynamics) [25], [26] to update Y. The stochastic diffusion updates Y with step size δ as such[2]:

$$Y' = Y - \delta \nabla_Y U(Y|X,\theta) + dw\sqrt{2T},\ dw \sim \mathcal{N}(0, \delta) \quad (19)$$

We leverage the automatic differentiation of libraries such as PyTorch[3] to automatically compute the gradient $\nabla_Y U$.

C. Parallelization

While the canonical RJMCMC for point processes adds and removes one point at a time in the image, we can leverage the spatial Markovianity (*ie.* points distant enough have independent energies) to run the kernel in parallel over the image. This approach is detailed in [27]. The space S is split into cells c of size $2d_{max} + 2\delta_{\max}$, each cell is assigned to one of four sets s so that no cell is neighboring a cell in the same set. Consequently, for any given set of cells, any perturbation in a cell induces an energy change independent of the other cells. Thus, the perturbations can be applied in any order, or in parallel without change in results. We denote Y^c the subset of points of configuration Y in cell c.

D. Sampling procedure

To ensure proper birth densities over the whole image, we set the following probabilities to select the cells to iterate on, with $d(c)$ the density d summed over cell c:

$$p(s) = \sum_{c \in s} d(c) \qquad p(c|s) = \frac{d(c)}{\sum_{c' \in s} d(c')}. \quad (20)$$

With initial configuration Y_0, image X and parameters θ, the procedure $\texttt{Sample}(Y_0, X, \theta)$, samples a configuration as follows:

for $n = 0$ to N **do**
  kernel $\leftarrow$ *diffusion* with probability 0.8, else *jump*
  pick one set s with probability $p(s)$
  keep each c in s with probability $p(c|s)$ to make $\tilde{s}$
  for all c in $\tilde{s}$ **do**
    if kernel is *diffusion* **then**
      $dw \sim \mathcal{N}(0, \delta)$
      $Y^c_{n+1} \leftarrow Y^c_n - \nabla_{Y^c_n} U(Y^c_n|X,\theta)\delta + dw\sqrt{2T}$
    else
      $Q \leftarrow Q_B$ with probability 0.5 else Q_D
      $Y^{c\prime}_n \sim Q(Y^c_n \to \cdot)$
      $r \leftarrow \frac{Q(Y^{c\prime}_n \to Y^c_n)}{Q(Y^c_n \to Y^{c\prime}_n)} \exp\left(-\frac{U(Y^{c\prime}_n|X,\theta) - U(Y^c_n|X,\theta)}{T}\right)$
      $Y^c_{n+1} \leftarrow Y^{c\prime}_n$ with probability $\min(1, r)$
    end if
  end for
end for
return Y_N

[2] The translation of each point is bounded by $\delta_{\max}$, in order to ensure the parallelization conditions are met.

[3] https://pytorch.org/

Fig. 5. Samples of detection on the test dataset. The score threshold (to not display low score objects) is set to maximize the $F1$ score for each model.

VII. Application and Results

A. Quantitative and qualitative evaluation on benchmark data

Our method is aimed for satellite image of resolution around 0.5 m/pixel. We train all models on a sub-sampled version of DOTA [28] to the desired resolution. We evaluate all methods on the same test split of this dataset.

To obtain the configuration $\widehat{Y}$, we perform the sampling method described above with simulated annealing: $T_{t+1} = \alpha T_t$, $\alpha = 0.997$.

The scoring of individual points (used to evaluate precision-recall curves and Average Precision), is computed as the Papangelou intensity [9]

$$\lambda(y|\widehat{Y} \setminus \{y\}) = \frac{h(\widehat{Y}|X)}{h(\widehat{Y} \setminus \{y\}|X)}, \forall y \in \widehat{Y}. \tag{21}$$

The measure $\lambda(y|Y)dy$ gives the infinitesimal probability to find a point in dy around y for a given configuration Y. Thus, (21) gives a confidence score for y given the image X its context $\widehat{Y} \setminus \{y\}$.

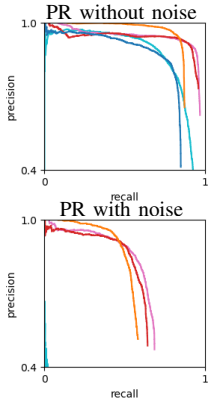
In this paper we show results on three CNN based models, and our two MPP models:

- CNN-local max.: Naive detection from the CNN backbone (used in the MPP models); we find objects through local maxima in the output probability maps (or local minima in the energy maps).
- CNN-MPP$^\blacklozenge$: MPP with minimal inferred parameters: manually set priors (i.e., with manually set parameters $\theta_{t,\text{pos}}$, $\theta_{t,\text{overl.}}$, $\theta_{t,\text{attra.}}, \ldots$), only the weight vector $\widehat{\theta}_w$ is estimated from the training dataset. This is akin to previous MPP parameters estimation [14], [29], where only the weight of the energy terms can be estimated automatically.
- CNN-MPP$^\bigstar$: MPP with parametrized priors. The parameter vector θ (which encompasses the weights θ_w and energy term parameters $\theta_{t,\text{pos}}$, $\theta_{t,\text{overl.}}$, $\theta_{t,\text{attra.}}, \ldots$) is estimated.
- BBA-Vec. and YOLOV5-OBB: Lastly, we compare all of our above-mentioned models with BBA-Vec. from [3] and YOLOV5-OBB from [30], [31].

TABLE I
AVERAGE PRECISION (AP) (LEFT) & PRECISION-RECALL (PR) (RIGHT)

Method	AP_0 [a]	AP_N [b]
BBA-Vec.	0.82	0.19
YOLOV5-OBB	0.86	0.10
CNN-local max.	0.86	0.55
CNN-MPP♦	**0.91**	**0.58**
CNN-MPP★	**0.92**	**0.62**

[a] AP on test data.
[b] AP on test data with noise.



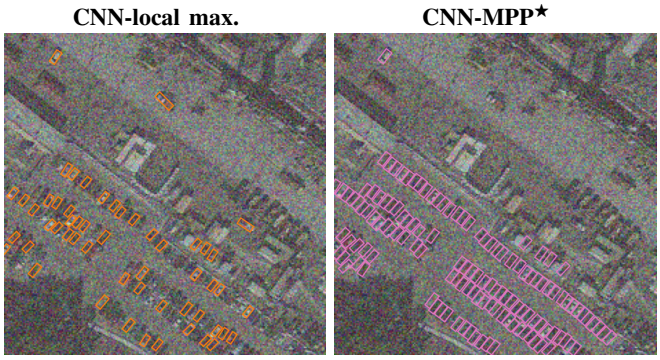


Fig. 6. Sample result on the test dataset with additive noise. The score threshold (to not display low score objects) is set to maximize the $F1$ score for each model. BBA-Vec. and YOLOV5-OBB do not produce any detections on this noisy image.

For every model we compute the Average Precision (AP) in Table I both the test DOTA at resolution of 0.5 m. Some results on sample images are shown in Fig. 5; it shows our CNN and MPP combination allows for regularization of the resulting configurations.

We also test all methods on the same images with additive noise, metrics are displayed in Table I and some results in Fig. 6; While BBA-Vec. and YOLOV5-OBB fail to produce any meaningful results, our MPP methods produce some geometrically regular configurations.

B. Qualitative evaluation on ADS data

We evaluate the methods on data provided by ADS, at a 0.5 m resolution. As this data is not labeled, models are trained only on the benchmark data presented above. Results are presented in Fig. 7; Qualitatively, our MPP model is able to produce regular configurations of vehicles, while missing fewer objects of interest compared to BBA-Vec..

VIII. CONCLUSION

We proposed a novel approach for object detection that leverages interaction between objects while maintaining a low number of parameters within a probabilistic model at the whole image level. It allows adding priors on interactions, while estimating the importance of these priors on some labeled data. We show this allows regularization and robustness on the resulting configuration of points, especially when the information contained in the image is scarce. We believe this could yield interesting results on other applications where the priors have more importance (*eg.* object tracking, where priors on dynamics are strong). However, the inference times of this method has room for improvement, as Markov chain based inference can be quite computationally expensive.

REFERENCES

[1] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023.

[2] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, Jan. 2020.

[3] J. Yi, P. Wu, B. Liu, Q. Huang, H. Qu, and D. Metaxas, "Oriented object detection in aerial images with box boundary-aware vectors," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2150–2159.

[4] Y. Li, Y. Xing, Z. Wang, T. Xiao, Q. Song, W. Li, and J. Wang, "A Framework of Maximum Feature Exploration Oriented Remote Sensing Object Detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[5] Y. Yao, G. Cheng, G. Wang, S. Li, P. Zhou, X. Xie, and J. Han, "On Improving Bounding Box Representations for Oriented Object Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.

[6] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "FarSeg++: Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–16, 2023.

[7] Y. Cao, Y. Bai, R. Pang, B. Liu, and K. Zhang, "Vehicle Detection Algorithm Based on Background Features Assistance in Remote Sensing Image," *Sensors and Materials*, vol. 35, no. 2, p. 607, Feb. 2023.

[8] X. Lu, X. Sun, W. Diao, Y. Mao, J. Li, Y. Zhang, P. Wang, and K. Fu, "Few-Shot Object Detection in Aerial Imagery Guided by Text-Modal Knowledge," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–19, 2023.

[9] M.-C. V. Lieshout, *Markov Point Processes and Their Applications*. London: Imperial College Press, Jul. 2000.

[10] X. Descombes, "Multiple objects detection in biological images using a marked point process framework," *Methods*, vol. 115, pp. 2–8, Feb. 2017.

[11] T. Li, M. Comer, and J. Zerubia, "A Connected-Tube MPP Model for Object Detection with Application to Materials and Remotely-Sensed Images," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. Athens: IEEE, Oct. 2018, pp. 1323–1327.

[12] C. Lacoste, X. Descombes, and J. Zerubia, "Point processes for unsupervised line network extraction in remote sensing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1568–1579, Oct. 2005.

[13] M. Ortner, X. Descombes, and J. Zerubia, "A Marked Point Process of Rectangles and Segments for Automatic Analysis of Digital Elevation Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 1, pp. 105–119, Jan. 2008.

[14] P. Craciun, M. Ortner, and J. Zerubia, "Joint detection and tracking of moving objects using spatio-temporal marked point processes," in *2015 IEEE Winter Conference on Applications of Computer Vision*, 2015, pp. 177–184.

[15] D. Duvenaud, J. Wang, J. Jacobsen, K. Swersky, M. Norouzi, and W. Grathwohl, "Your classifier is secretly an energy based model and you should treat it like one," in *ICLR 2020*, 2020.

[16] Z. Huang, W. Li, X.-G. Xia, and R. Tao, "A General Gaussian Heatmap Label Assignment for Arbitrary-Oriented Object Detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 1895–1910, 2022.

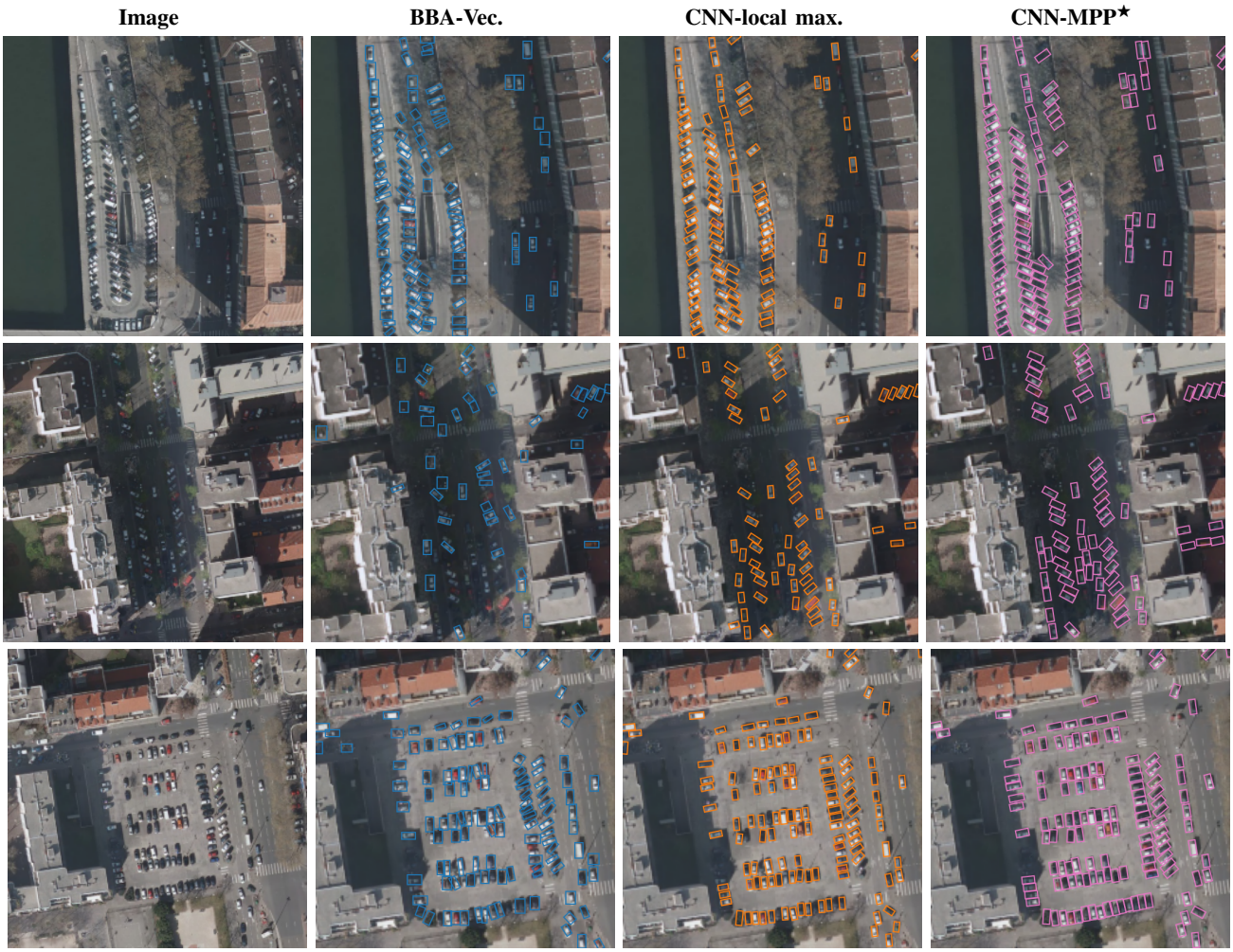


Fig. 7. Samples of detection on the Airbus DS data. The dataset is not annotated.[© Airbus Defense and Space]

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., 2015, pp. 234–241.

[18] J. Mabon, J. Zerubia, and M. Ortner, "Point process and CNN for small object detection in satellite images," in *SPIE, Image and Signal Processing for Remote Sensing XXVIII*, Berlin, Germany, Sep. 2022.

[19] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. J. Huang, "A Tutorial on Energy-Based Learning," *Predicting structured data*, p. 59, 2006.

[20] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[21] Y. W. Teh, M. Welling, S. Osindero, and G. E. Hinton, "Energy-based models for sparse overcomplete representations," *Journal of Machine Learning Research*, vol. 4, no. Dec, pp. 1235–1260, 2003.

[22] Y. Du and I. Mordatch, "Implicit Generation and Modeling with Energy Based Models," *NeurIPS 2019*, p. 11.

[23] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade: Second Edition*, G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 421–436.

[24] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, Dec. 1995.

[25] M. Miller, U. Grenander, J. OSullivan, and D. Snyder, "Automatic target recognition organized via jump-diffusion algorithms," *IEEE Transactions on Image Processing*, vol. 6, no. 1, pp. 157–174, 1997.

[26] F. Lafarge, G. Gimel'farb, and X. Descombes, "Geometric Feature Extraction by a Multimarked Point Process," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1597–1609, Sep. 2010.

[27] Y. Verdié and F. Lafarge, "Efficient Monte Carlo Sampler for Detecting Parametric Objects in Large Scenes," in *Computer Vision – ECCV 2012*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 7574, pp. 539–552.

[28] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A Large-scale Dataset for Object Detection in Aerial Images," *arXiv:1711.10398 [cs]*, May 2019.

[29] Q. Yu and G. Medioni, "Multiple-Target Tracking by Spatiotemporal Monte Carlo Markov Chain Data Association," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2196–2210, Dec. 2009.

[30] X. Yang and J. Yan, "On the Arbitrary-Oriented Object Detection: Classification Based Approaches Revisited," *Int J Comput Vis*, vol. 130, no. 5, pp. 1340–1365, May 2022.

[31] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, K. Michael, TaoXie, J. Fang, imyhxy, Lorna, Z. Yifu, C. Wong, A. V, D. Montes, Z. Wang, C. Fati, J. Nadar, Laughing, UnglvKitDe, V. Sonck, tkianai, yxNONG, P. Skalski, A. Hogan, D. Nair, M. Strobel, and M. Jain, "Ultralytics/yolov5: V7.0 - YOLOv5 SOTA realtime instance segmentation," Zenodo, Nov. 2022.