
Detecting Lesion Bounding Ellipses With Gaussian Proposal Networks

Yi Li

Baidu Research Institute

1195 Bordeaux Dr. Sunnyvale, CA 94089

liyi17@baidu.com

Abstract

Lesions characterized by computed tomography (CT) scans, are arguably often elliptical objects. However, current lesion detection systems are predominantly adopted from the popular Region Proposal Networks (RPNs) [21] that only propose bounding boxes without fully leveraging the elliptical geometry of lesions. In this paper, we present Gaussian Proposal Networks (GPNs), a novel extension to RPNs, to detect lesion bounding ellipses. Instead of directly regressing the rotation angle of the ellipse as the common practice, GPN represents bounding ellipses as 2D Gaussian distributions on the image plain and minimizes the Kullback-Leibler (KL) divergence between the proposed Gaussian and the ground truth Gaussian for object localization. We show the KL divergence loss approximately incarnates the regression loss in the RPN framework when the rotation angle is 0. Experiments on the DeepLesion [27] dataset show that GPN significantly outperforms RPN for lesion bounding ellipse detection thanks to lower localization error. GPN is open sourced at <https://github.com/baidu-research/GPN>.

1 Introduction

Current state-of-the-art object detection systems are predominantly based on deep neural networks that learn to propose object regions which are usually represented by bounding boxes [21, 15, 11, 1, 6, 4]. Region Proposal Networks (RPNs), first introduced in Faster R-CNN [21], simultaneously predicts the objectness and regions bounds at every predefined anchor location on a grid of feature map. When object regions are annotated as bounding boxes, the region bounds in RPN are defined by the two center coordinates, the width and the height of the object region with respect to the corresponding anchor. In this case, regressing the regions bounds is directly optimizing the overlap between the proposed bounding box and the ground truth bounding box.

Lesions characterized by computed tomography (CT) scans are often elliptical and additional geometry information about the lesion regions may be annotated besides bounding boxes. For example, the large-scale medical imaging dataset DeepLesion [27] recently released from NIH is annotated with the response evaluation criteria in solid tumors (RECIST) diameters. Each RECIST-diameter annotation consists of two axies, the first one measures the longest diameter of the lesion and the second one measures the longest diameter perpendicular to the first axis. Therefore, the RECIST diameters closely represent the major and minor axises of a bounding ellipse of the lesion.

Extensions for bounding ellipse detection based on the RPN framework have been introduced to also predict the rotation angle of the object [8, 22, 14, 10, 28, 16, 25, 29, 19]. However, due to the rotation angle, it is not trivial to directly optimizing the overlap between the proposed bounding ellipse and the ground truth bounding ellipse within the RPN framework. Most of the existing methods either directly use an ellipse regressor to minimize the difference between the proposed angle and the ground truth angle, e.g. in the co/tangent domain [8, 14], as an additional term in the regression loss

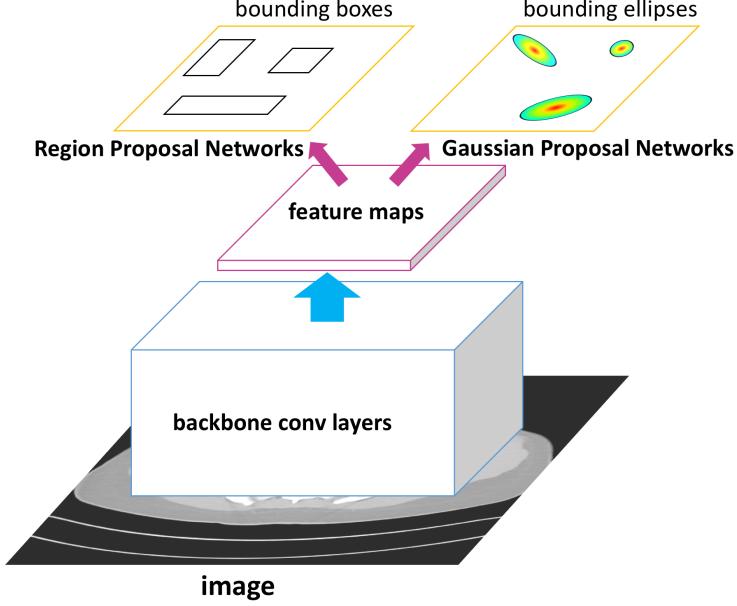


Figure 1: Comparison between Region Proposal Networks and Gaussian Proposal Networks. Instead of proposing bounding boxes, Gaussian Proposal Networks propose bounding ellipses as 2D Gaussian distributions on the image plane and use a single KL divergence loss for object localization.

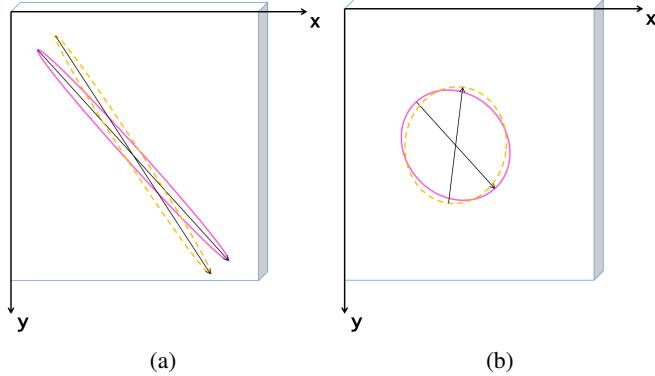


Figure 2: Two illustrative examples of how the rotation angle may affect the overlap between two ellipses differently. The ellipse of solid line is the ground truth and the ellipse of dash line is proposed by the detection model, where both ellipses only differ in the rotation angle. (a) the major axis is significantly longer than the minor axis, (b) the major axis is slightly longer than the minor axis.

for localization. Other methods discretize the rotation angle first, and then predict the angle category as a classification problem [22].

We argue that directly regressing the rotation angle is not optimal for bounding ellipse localization. The rotation angle and the aspect ratio, i.e. the ratio of the major axis to the minor axis, jointly affect the overlap between two ellipses. Figure 2 shows two extreme examples of a proposed bounding ellipse and its ground truth bounding ellipse where they only differ in the rotation angle. In Figure 2a, where the aspect ratio is significantly larger than 1, a small shift in the rotation angle results in a dramatic change of overlap between the two ellipses. However in Figure 2b, where the major axis is almost equal length to the minor axis, the shift of the rotation angle merely affects the overlap between the two ellipses. Therefore, it may not be optimal or sometimes even unnecessary to directly regress the rotation angle, when the essential goal is to optimize the overlap between the proposed ellipse and the ground truth ellipse.

In this work, we present Gaussian Proposal Networks (GPNs) that learn to propose bounding ellipses as 2D Gaussian distributions on the image plane. Figure 1 shows an illustrative comparison between GPN and RPN. Unlike most of the extensions to the RPN framework that introduce an additional term to directly regress the rotation angle for object localization, GPN minimizes the Kullback-Leibler (KL) divergence between the proposed Gaussian distribution and the ground truth Gaussian distribution as one single loss for localization. KL divergence directly measures the overlap of one distribution with respect to a reference distribution. When the two distributions are Gaussian, KL divergence has analytical form and is differentiable. Therefore, GPN can be readily implemented with standard automatic differentiation packages [20] and trained with back-propagation algorithm. We also show that when the rotation angle is 0, the KL divergence loss approximately incarnates the regression loss used in the RPN framework for bounding box localization. We experiment the efficacy of GPN for detecting lesion bounding ellipses on the DeepLesion [27] dataset. GPN outperforms RPN by a significant margin in terms of free-response receiver operating characteristic across different experimental settings. Error analysis shows that GPN achieves significant lower localization error compared to RPN. Further analysis on the distribution of predicted rotation angles from GPN supports our conjecture that it may not be necessary to regress the rotation angle when the ground truth bounding ellipse has similar lengths of major and minor axes.

2 Related work

2.1 Region proposal networks (RPNs)

The design of GPN generally follows the principles of RPN [21]. RPN has a fully convolutional backbone network that processes the input image and generates a feature map grid. The feature vector of each position on the feature map is further processed with a 3×3 convolutional layer and then associated with potentially multiple anchors of varies scales and aspect ratios. The ground truth region of interests (RoIs) are then assigned to anchors that meet certain overlapping criteria, e.g. intersection over union (IoU) greater than 0.7. Finally, the 3×3 convolutional layer is followed by two separate 1×1 convolutional layers, one is predicting the objectness scores of the RoIs and the other is predicting the bounding box offsets of the RoIs with respect to the anchors. RPN is jointly trained with one classification loss and multiple smoothed L1 regression losses.

Compared to RPN, GPN proposes bounding ellipses as 2D Gaussian distributions and minimizes a single KL divergence loss between two Gaussian distributions for localization. GPN could be applied with the same extensions that apply to RPN, such as training with multi-scale feature maps [11, 15, 4], online hard negative mining [23] and focal loss [12]. However, in the Faster R-CNN two-stage detector, a second R-CNN classifier [21] is appended to RPN through a ROI pooling layer [5] and classifies each ROI into specific object categories or background. Performing ROI pooling with bounding ellipses is nontrivial and is beyond the scope of this paper. Thus, GPN currently only applies to one-stage detectors like SSD [15], where object bounds and categories are jointly predicted through one single network.

2.2 Bounding ellipse and KL divergence

Bounding ellipse annotation has been widely used in human faces detection [9]. To generate bounding ellipses, most of the detection methods directly map detected bounding boxes into ellipses [10, 28, 16, 25, 29]. Shi et al. [22] discretize the rotation angle and classify it into different categories. Hu et al. [8] train a separated ellipse regressor to regress the cotangent of the rotation angle using features extracted offline from deep networks. Opitz et al. [19] also point out the problem that different parameters of bounding ellipses, i.e. the center coordinates, the major and minor axes and the rotation angle, impact ellipses overlap differently, thus is in spirit similar to our argument. However, in order to maximize the overlap, Opitz et al. [19] rasterize both the proposed ellipse and the ground truth ellipse and numerically compute the gradient of each ellipse parameter by counting the change of rasterized overlap in pixels. In comparison, GPN uses KL divergence to analytically optimize the overlap between two Gaussian distributions as a surrogate to optimizing ellipses overlap. Najibi et al. [18] also use Gaussian distributions to generate the saliency maps of objects based on their bounding boxes. But they do not consider the rotation angle and do not use KL divergence loss for optimization.

KL divergence has been used to measure the ellipticity and circularity of a given set of points [17]. To the best of our knowledge, the most similar work to our approach is the recent Softer-NMS [7] that also uses KL divergence for bounding box localization. However, the motivation of Softer-NMS is to model the location uncertainty of each corner of proposed bounding boxes through 1D Gaussian distribution. The motivation of GPN is to propose bounding ellipses as 2D Gaussian distributions on the image plain. Therefore, it is not directly comparable to Softer-NMS. We discuss the differences and connections between Softer-NMS and GPN with more details in the Supplementary.

3 Gaussian Proposal Networks (GPNs)

This section describes the mathematical and implementation details of GPN. We first formulate the representation of ellipses as 2D Gaussian distributions in Section 3.1. We then derive the KL divergence between 2D Gaussian distributions using this representation in Section 3.2. Next we draw connections between the KL divergence loss and the regression loss used in the RPN framework in Section 3.3. Finally, we provide some implementation details in Section 3.4.

3.1 Ellipses as 2D Gaussian distributions

The equation of an ellipse in a 2D coordinate system without rotation is given by

$$\frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(y - \mu_y)^2}{\sigma_y^2} = 1, \quad (1)$$

where we denote μ_x, μ_y as the center coordinates of the ellipse, and σ_x, σ_y as the lengths of semi-axes along x and y axes.

The probability density function of a 2D Gaussian distribution is given by

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}))}{2\pi|\boldsymbol{\Sigma}|^{\frac{1}{2}}}, \quad (2)$$

where \mathbf{x} is the vector representation of coordinates (x, y) , $\boldsymbol{\mu}$ is the mean, $\boldsymbol{\Sigma}$ is the covariance matrix and $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix. If we assume the off-diagonal term in $\boldsymbol{\Sigma}$ is 0 and parameterize $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ as

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}, \quad (3)$$

then the ellipse equation in Equation 1 corresponds to the density contour of the 2D Gaussian distribution when

$$(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = 1. \quad (4)$$

When the major axis of the ellipse is rotated of an angle θ with respect to the x axis, we use a rotation matrix $R(\theta)$ to map the coordinates in the original (x, y) system into a new (x', y') system, where the major axis of the ellipse is aligned with the x' axis in the new system as shown in Figure 3

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = R(\theta) \begin{bmatrix} x \\ y \end{bmatrix}, R(\theta) = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}. \quad (5)$$

For example, the coordinates of $(1, 0)$ and $(0, 1)$ in the (x, y) system, are $(\cos \theta, -\sin \theta)$ and $(\sin \theta, \cos \theta)$ in the (x', y') system. If we use σ_l, σ_s to denote the lengths of the semi-major (long) and semi-minor (short) axes of the ellipse in the (x', y') system and assume it is centered at (μ'_x, μ'_y) , then the ellipse equation in the (x', y') system is given by

$$(\mathbf{x}' - \boldsymbol{\mu}')^\top (\boldsymbol{\Sigma}')^{-1}(\mathbf{x}' - \boldsymbol{\mu}') = 1, (\boldsymbol{\Sigma}')^{-1} = \begin{bmatrix} \frac{1}{\sigma_l^2} & 0 \\ 0 & \frac{1}{\sigma_s^2} \end{bmatrix}, \quad (6)$$

and again, \mathbf{x}' and $\boldsymbol{\mu}'$ are the vector representations of (x', y') and (μ'_x, μ'_y) .

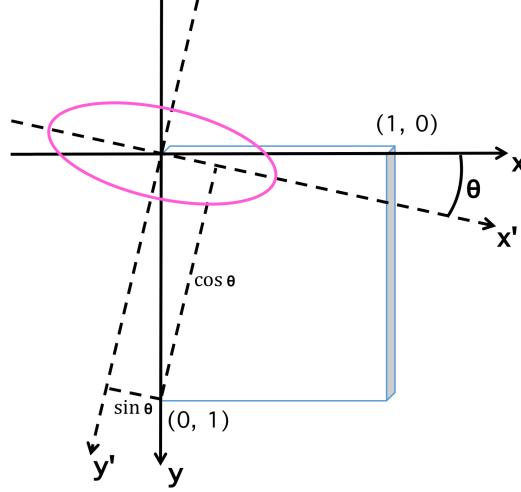


Figure 3: Correspondence between the original coordinate system (x, y) , and the rotated coordinate system (x', y') with a rotation angle θ . The major axis of the ellipse is aligned with the x' axis in the rotated system. The coordinates of $(1, 0)$ and $(0, 1)$ in the (x, y) system, are $(\cos \theta, -\sin \theta)$ and $(\sin \theta, \cos \theta)$ in the (x', y') system.

With the rotation matrix $R(\theta)$, the ellipse equation in the (x, y) system can be derived as

$$\begin{aligned} [R(\theta)(\mathbf{x} - \boldsymbol{\mu})]^\top (\Sigma')^{-1} [R(\theta)(\mathbf{x} - \boldsymbol{\mu})] &= 1, \\ (\mathbf{x} - \boldsymbol{\mu})^\top [R^\top(\theta)(\Sigma')^{-1} R(\theta)](\mathbf{x} - \boldsymbol{\mu}) &= 1, \end{aligned} \quad (7)$$

where $\boldsymbol{\mu}$ is the center coordinates (μ_x, μ_y) of the ellipse in the (x, y) system.

Finally, we can use a 2D Gaussian distribution in the (x, y) system parameterized by

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \Sigma^{-1} = R^\top(\theta) \begin{bmatrix} \frac{1}{\sigma_l^2} & 0 \\ 0 & \frac{1}{\sigma_s^2} \end{bmatrix} R(\theta) \quad (8)$$

to represent the ellipse centered at (μ_x, μ_y) , with semi-major and semi-minor axes of lengths (σ_l, σ_s) , and a rotation angle of θ between its major axis and the x axis. Note that $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

The goal of GPN is to propose bounding ellipses such that their parameters, $(\mu_x, \mu_y, \sigma_l, \sigma_s, \theta)$, match the ground truth ellipses through the criteria of KL divergence.

3.2 KL divergence of 2D Gaussian distributions

The KL divergence between a proposed 2D Gaussian distribution \mathcal{N}_p and a target 2D Gaussian distribution \mathcal{N}_t is given by [2]

$$D_{\text{KL}}(\mathcal{N}_t || \mathcal{N}_p) = \frac{1}{2} \left[\text{tr}(\Sigma_p^{-1} \Sigma_t) + (\boldsymbol{\mu}_p - \boldsymbol{\mu}_t)^\top \Sigma_p^{-1} (\boldsymbol{\mu}_p - \boldsymbol{\mu}_t) + \ln \frac{|\Sigma_p|}{|\Sigma_t|} - 2 \right], \quad (9)$$

where $\text{tr}(\mathbf{X})$ is the trace of matrix \mathbf{X} .

Assuming \mathcal{N}_p and \mathcal{N}_t are parameterized by $(\mu_{x_p}, \mu_{y_p}, \sigma_{l_p}, \sigma_{s_p}, \theta_p)$ and $(\mu_{x_t}, \mu_{y_t}, \sigma_{l_t}, \sigma_{s_t}, \theta_t)$ following Equation 8, we can derive each term in Equation 9 as

$$\text{tr}(\Sigma_p^{-1} \Sigma_t) = \cos^2 \Delta\theta \frac{\sigma_{l_t}^2}{\sigma_{l_p}^2} + \cos^2 \Delta\theta \frac{\sigma_{s_t}^2}{\sigma_{s_p}^2} + \sin^2 \Delta\theta \frac{\sigma_{l_t}^2}{\sigma_{s_p}^2} + \sin^2 \Delta\theta \frac{\sigma_{s_t}^2}{\sigma_{l_p}^2}, \quad (10)$$

$$(\boldsymbol{\mu}_p - \boldsymbol{\mu}_t)^\top \Sigma_p^{-1} (\boldsymbol{\mu}_p - \boldsymbol{\mu}_t) = \frac{(\cos \theta_p \Delta x + \sin \theta_p \Delta y)^2}{\sigma_{l_p}^2} + \frac{(\cos \theta_p \Delta y - \sin \theta_p \Delta x)^2}{\sigma_{s_p}^2}, \quad (11)$$

$$\ln \frac{|\Sigma_p|}{|\Sigma_t|} = \ln \frac{\sigma_{l_p}^2}{\sigma_{l_t}^2} + \ln \frac{\sigma_{s_p}^2}{\sigma_{s_t}^2}, \quad (12)$$

where we define

$$\Delta\theta = \theta_p - \theta_t, \Delta x = \mu_{x_p} - \mu_{x_t}, \Delta y = \mu_{y_p} - \mu_{y_t}. \quad (13)$$

The exact details of deriving each term in Equation 10 through Equation 12 are provided in the Supplementary.

3.3 Connection to the RPN regression loss

The general form of KL divergence derived in the previous section looks rather complex. However, if we omit the rotation angle, i.e. assuming θ_p and θ_t are always 0, then σ_l and σ_s are actually the half width and the half height of the bounding box that tightly surrounds the bounding ellipse. And we obtain a much simpler form of KL divergence as

$$D_{\text{KL}}(\mathcal{N}_t || \mathcal{N}_p) = \frac{1}{2} \left[\frac{w_t^2}{w_p^2} + \frac{h_t^2}{h_p^2} + \frac{\Delta^2 x}{w_p^2} + \frac{\Delta^2 y}{h_p^2} + \ln \frac{w_p^2}{w_t^2} + \ln \frac{h_p^2}{h_t^2} - 2 \right]. \quad (14)$$

where we have replaced $(\sigma_{l_p}, \sigma_{l_t})$ with (w_p, w_t) , and $(\sigma_{s_p}, \sigma_{s_t})$ with (h_p, h_t) for easy comparison with the RPN regression loss.

For bounding box regression, RPN outputs four terms for each proposed bounding box [21]

$$\begin{aligned} t_x^p &= (x_p - x_a)/w_a, & t_y^p &= (y_p - y_a)/h_a, \\ t_w^p &= \ln(w_p/w_a), & t_h^p &= \ln(h_p/h_a), \end{aligned} \quad (15)$$

to match the four targets from the ground truth bounding box

$$\begin{aligned} t_x^t &= (x_t - x_a)/w_a, & t_y^t &= (y_t - y_a)/h_a, \\ t_w^t &= \ln(w_t/w_a), & t_h^t &= \ln(h_t/h_a), \end{aligned} \quad (16)$$

where (x_a, y_a, w_a, h_a) are the center coordinates, width and height of the matching anchor.

RPN uses smoothed L1 loss for regression. When the loss is small, smoothed L1 loss becomes squared loss. Therefore, for center coordinates regression, the squared loss are

$$(t_x^p - t_x^t)^2 = \frac{\Delta^2 x}{w_a^2}, \quad (t_y^p - t_y^t)^2 = \frac{\Delta^2 y}{h_a^2}, \quad (17)$$

which are very similar to the middle two terms in Equation 14, except that the RPN regression loss normalizes $(\Delta x, \Delta y)$ by the anchor width and height (w_a, h_a) , where KL divergence normalizes them by the predicted width and height (w_p, h_p) .

When the loss is large, the L1 losses for width and height regression are

$$|t_w^p - t_w^t| = \left| \ln \frac{w_p}{w_t} \right|, \quad |t_h^p - t_h^t| = \left| \ln \frac{h_p}{h_t} \right|, \quad (18)$$

which are equivalent to the last two terms (ignoring the constant -2) in Equation 14 when $w_p \geq w_t, h_p \geq h_t$. However, since KL divergence is asymmetric, when $w_p < w_t, h_p < h_t$, KL divergence penalizes differently through the first two terms in Equation 14.

The comparison between the KL divergence loss and the smoothed L1 regression loss in RPN, not only provides another perspective towards the efficacy of RPN for object localization, but also suggests an alternative loss for bounding box localization, i.e. Equation 14. However, the efficacy of the KL divergence loss for general bounding box localization needs comprehensive analysis on other benchmark datasets, e.g. PASCAL VOC [3] and MS COCO [13], and is beyond the scope of this paper.

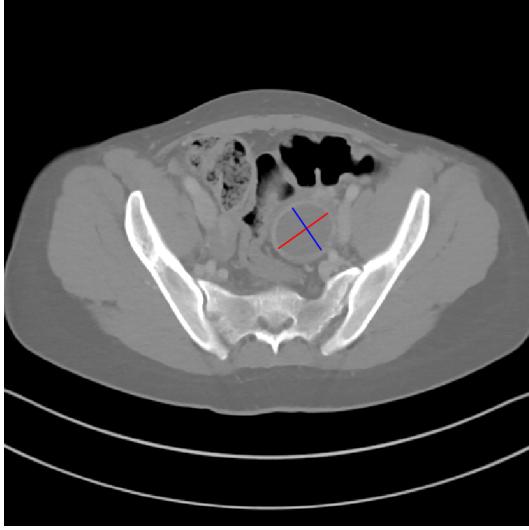


Figure 4: One example of slice image and its RECIST diameters annotation from the DeepLesion dataset [27]. The red line is the major axis that measures the longest diameter of the annotated lesion and the blue line is the minor axis that measures the longest diameter perpendicular to the major axis.

3.4 Implementation details

To implement the KL divergence loss in practice, we follow the design of anchors from RPN and make the four targets the same as Equation 16 based on the matching anchor, plus the tangent of the rotation angle following [8]. We let GPN output four terms the same as Equation 15, plus an additional term for the tangent of the rotation angle. Finally, we plug both the network outputs and the targets into Equation 9 through Equation 12 to compute the KL divergence loss.

The KL divergence loss can be directly added with the classification loss without any balancing factors used in RPN

$$L_{\text{total}} = L_{\text{cls}} + L_{\text{KLD}}. \quad (19)$$

The only important caveat we found to make the KL divergence loss well bounded is to initialize the weights of the 1×1 convolutional layer for bounding ellipse localization within a small range. Specifically, we use a Gaussian distribution with 0 mean and standardization of 0.001 to initialize the weights.

GPN was implemented with PyTorch-0.3.1 [20].

4 Experiments

We present comprehensive evaluation of GPN for detecting lesion bounding ellipses on the DeepLesion dataset [27]. We first introduce some details about the DeepLesion dataset in Section 4.1, and experiments setup in Section 4.2. Next, we show GPN significantly outperforms RPN for bounding ellipse detection across different settings in Section 4.3. Finally, we present a comprehensive error analysis in Section 4.4.

4.1 DeepLesion dataset

DeepLesion is a large-scale medical imaging dataset recently released from NIH [27]. It contains 32,735 lesions in 32,120 CT slice images from 4,427 unique patients. More than 99% of the slice images are 512×512 pixels. Each lesion is annotated with two response evaluation criteria in solid tumors (RECIST) diameters. The first one measures the longest diameter of the lesion and the second one measures the longest diameter perpendicular to the first diameter, so they closely represent the major and minor axes of a bounding ellipse, and we use this notion thereafter. We note this assumption may be inaccurate when the center of the minor axis is not aligned with the center of the

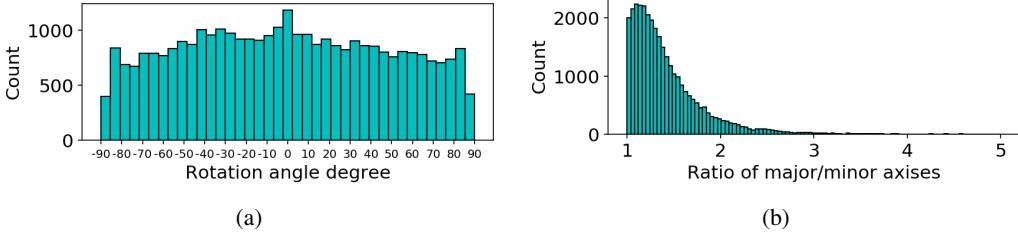


Figure 5: (a) the distribution of the rotation angles between the lesion’s major axis and the x axis in degree, (b) the distribution of lesions’ aspect ratios.

FPs per image	0.5	1	2	4	8	16
GPN-5-anchor	0.36	0.50	0.62	0.72	0.76	0.79
RPN-5-anchor	0.31	0.43	0.53	0.61	0.65	0.68
GPN-1-anchor	0.45	0.56	0.65	0.72	0.75	0.77
RPN-1-anchor	0.41	0.51	0.58	0.63	0.66	0.67
Faster R-CNN	0.57	0.67	0.76	0.82	0.86	0.89

Table 1: Detection sensitivities of GPN and RPN on the test set of DeepLesion at different false positives per image. Ellipse IoU of 0.5 is used as the threshold. The Faster R-CNN results are from [26] where the IoU is computed based on bounding box.

major axis. However, we found that for more than 90% of the RECIST-diameter annotations, the center of the minor axis is within the middle 20% range of the major axis, so we think the assumption approximately holds. Figure 4 shows an example of slice image and its RECIST-diameter annotation. DeepLesion has a wide range of rotation angles and aspect ratios, therefore it is particularly challenge for bounding ellipse detection and localization. Figure 5a shows the distribution of the rotation angles between the lesion’s major axis and the x axis. Figure 5b shows the distribution of lesions’ aspect ratios. For more details about the DeepLesion dataset please refer to [27].

4.2 Experiments setup

We follow the practices from [26] to convert the raw slice images with pixel value in Hounsfield Unit (HU) into 512×512 three channel images with pixel values between 0 and 255. We use the official split from DeepLesion for training (70%), validation (15%), and test (15%). All the networks are trained with 20 epochs. We compute intersection over union (IoU) between ellipses by rasterizing ellipses first and then counting the pixel overlaps. However, this numerical approach is compute intensive, therefore we only use it for performance evaluation. During training, we use the bounding box that tightly surrounds the bounding ellipse to compute IoU for anchor assignment and non-maximum suppression. Complete details about data preprocessing and model training are provided in the Supplementary.

4.3 Performances for bounding ellipse detection

We evaluate the performances of GPN and RPN for bounding ellipse detection using a single-scale feature map of stride 8 and five anchor scales, i.e. (16, 24, 32, 48, 96), following [26]. Compared to the Faster R-CNN two-stage detector used in [26], both GPN and RPN are one-stage detectors that suffer from overwhelming number of background proposals during training [12]. Therefore, we also experiment training with just one anchor scale of 16 to mitigate this issue. Both settings are equally applied to GPN and RPN except that GPN uses the KL divergence loss for localization while RPN uses the default smoothed L1 loss for localization with the additional term to regress the tangent of the rotation angle [8]. We use pretrained VGG-16 [24] as the backbone network following [26] and a single anchor aspect ratio of 1:1.

Figure 6a and Table 1 show the overall performances of GPN and RPN on the test set of DeepLesion across different settings measured by the free-response receiver operating characteristic (FROC). FROC measures the detection sensitivity with different average false positives per image. We consider

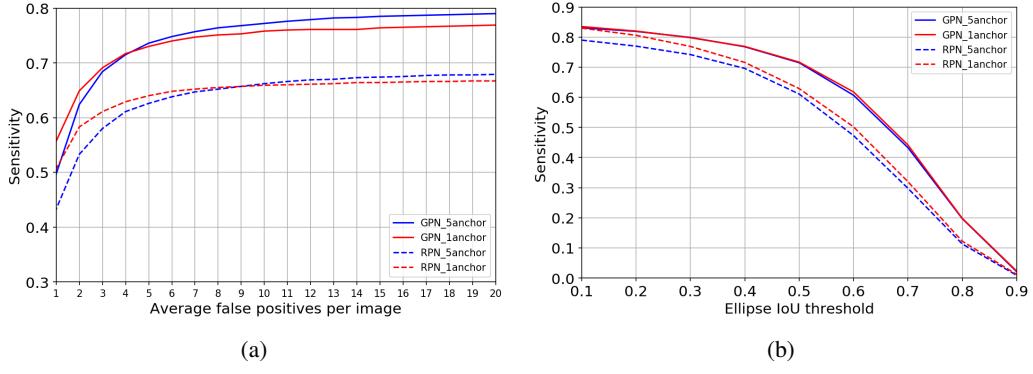


Figure 6: (a) FROC curves of GPN and RPN on the test set of DeepLesion at 0.5 ellipse IoU threshold. (b) Detection sensitivities of GPN and RPN on the test set of DeepLesion with different ellipse IoU thresholds at 4 false positives per image.

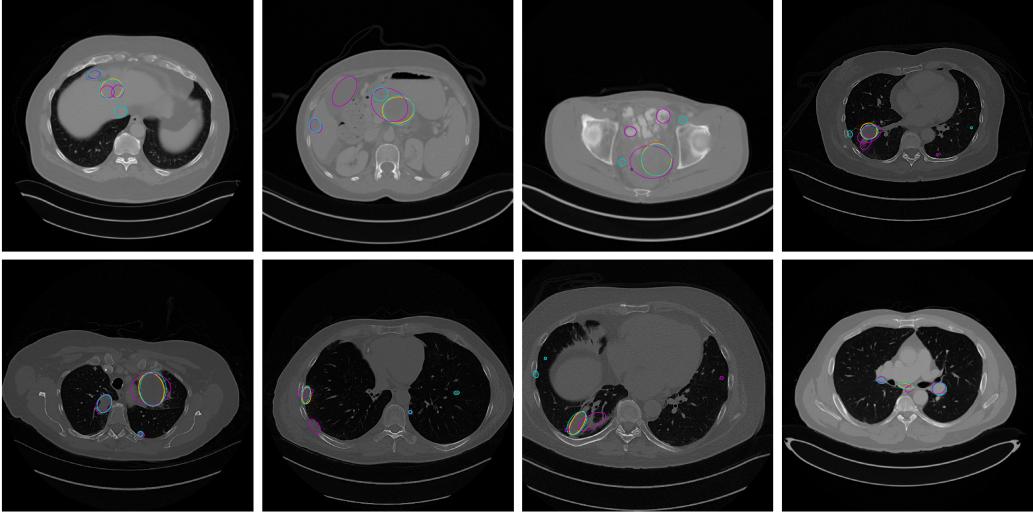


Figure 7: Selected examples of proposed bounding ellipses from GPN-5anchor (cyan) and RPN-5anchor (magenta) compared to the ground truth (orange) on the test set of DeepLesion. Only the top 3 proposed ellipses with the highest classification scores from each model are selected for each image.

a proposed bounding ellipse is correct if its IoU with the ground truth ellipse is greater or equal than 0.5. Note that, IoU between ellipses is a very stringent criteria, especially when the ellipse aspect ratio is significantly larger than one as illustrated by Figure 2a.

GPN consistently outperforms RPN across both settings by a significant margin. We also include the previous state-of-the-art results on the DeepLesion dataset based on Faster R-CNN [26] in Table 1. Yet, the Faster R-CNN results were trained and evaluated on the bounding box that tightly surrounds the bounding ellipse, so it is not directly comparable to our results. One anchor scale training improve the performances of GPN when the average false positives per image is less or equal than 3.

4.4 Error analysis

Figure 7 shows a few examples of proposed bounding ellipses from GPN-5anchor and RPN-5anchor compared to the ground truth on the test set of DeepLesion. To focus comparison on ellipse localization, we only select proposed bounding ellipses that are overlapped with the ground truth by both models. We can see that GPN detects ellipses of various sizes, rotation angles and aspect ratios with more accurate overlaps than RPN. We present detailed error analysis in the next two sections.

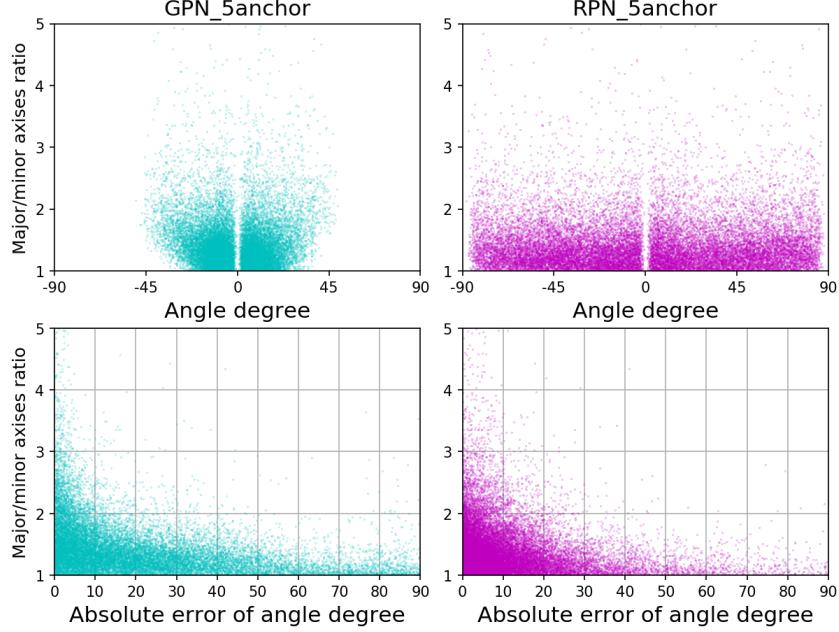


Figure 8: **The upper panels** are the distributions of predicted angles of GPN-5anchor and RPN-5anchor with respect to the ground truth aspect ratio on the training set of DeepLesion. **The lower panels** are the absolute degree errors of predicted angles of GPN-5anchor and RPN-5anchor with respect to the ground truth aspect ratio on the training set of DeepLesion.

4.4.1 Localization error

We first investigate the contribution of localization error to detection performance. Figure 6b shows the detection sensitivities of GPN and RPN with different ellipse IoU thresholds at 4 false positives per image. We can see that, when the IoU threshold is small, both GPN and RPN have comparable detection sensitivities since it is dominated by proposals vs background classification accuracy. As the IoU threshold increases, the performance of RPN decreases significantly faster than GPN, suggesting its localization error is significantly higher than GPN.

4.4.2 Rotation angle error

We also investigate the behaviors of rotation angle prediction of GPN and RPN. In the Introduction Section, we conjecture that it may be unnecessary to directly regress the rotation angle for bounding ellipse localization, especially when the ellipse aspect ratio is close to 1 as illustrated in Figure 2b. The upper panels of Figure 8 show the distributions of predicted angles of GPN-5anchor and RPN-5anchor with respect to the aspect ratio on the training set of DeepLesion. We notice when the aspect ratio is close to 1, GPN mostly predicts the rotation angle around 0 since it merely affects the ellipses overlap. On the other hand, because RPN directly regresses the rotation angle, the distribution of its predicted angles is much wider regardless of the aspect ratio, and similar to the ground truth distribution in Figure 5a. We notice the absolute degrees of predicted angles from GPN rarely exceed 45° . This is because GPN tends to flip the major and minor axes representation when the ground truth rotation angle of the major axis is greater than 45° or less than -45° , since the flipped representation contributes symmetrically to the KL divergence loss.

The lower panels of Figure 8 show the absolute degree errors of predicted angles of GPN-5anchor and RPN-5anchor on the training set of DeepLesion. We use the predicted longer axis as the actual major axis to account for the flipping effect of GPN. We see the angle errors of both GPN and RPN are significantly lower when the ground truth aspect ratio is significantly larger than 1. GPN does have higher angle errors than RPN when the aspect ratio is close to 1 as expected. However, GPN achieves slightly lower angle errors than RPN when the aspect ratio is larger than 2, even GPN is not directly regressing the rotation angle. The general trend that the angle errors are lower when the

aspect ratios are larger still holds on the test set for both GPN and RPN (figures are shown in the Supplementary).

5 Discussion

In this work, we present Gaussian Proposal Networks (GPNs), a new extension to the popular Region Proposal Networks (RPNs) [21], to detect bounding ellipses of lesions on CT scans. Compared to RPN that uses multiple regression losses for bounding box localization, GPN views bounding ellipses as 2D Gaussian distributions on the image plane, and optimizes the KL divergence between the proposed Gaussian and the ground truth Gaussian for localization. We show the KL divergence loss is closely connected to the regression loss used in RPN. On the large-scale medical imaging dataset DeepLesion [27], GPN significantly outperforms RPN for bounding ellipse detection across different experimental settings thanks to much lower localization error through the KL divergence loss. Further error analysis reveals that directly regressing the ellipse rotation angle may be unnecessary when the ellipse aspect ratio is close to 1.

We intend to further investigate the efficacy of GPN on nature image datasets with bounding ellipse annotations, such as the FDDB benchmark [9]. It is also interesting to comprehensively test the KL divergence loss derived in Section 3.3 for general bounding box localization on PASCAL VOC [3] and MS COCO [13] and see if it also outperforms the current regression loss in RPN.

References

- [1] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.
- [2] J. Duchi. Derivations for linear algebra and optimization. *Berkeley, California*, 3, 2007.
- [3] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [4] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.
- [5] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [6] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [7] Y. He, X. Zhang, M. Savvides, and K. Kitani. Softer-nms: Rethinking bounding box regression for accurate object detection. *arXiv preprint arXiv:1809.08545*, 2018.
- [8] P. Hu and D. Ramanan. Finding tiny faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1522–1530, 2017.
- [9] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report.
- [10] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5325–5334, 2015.
- [11] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 936–944, 2017.
- [12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [14] L. Liu, Z. Pan, and B. Lei. Learning a rotation invariant detector with rotatable bounding box. *arXiv preprint arXiv:1711.09405*, 2017.
- [15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [16] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European conference on computer vision*, pages 720–735. Springer, 2014.

- [17] K. Misztal and J. Tabor. Ellipticity and circularity measuring via kullback–leibler divergence. *Journal of Mathematical Imaging and Vision*, 55(1):136–150, 2016.
- [18] M. Najibi, F. Yang, Q. Wang, and R. Piramuthu. Towards the success rate of one: Real-time unconstrained salient object detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1432–1441. IEEE, 2018.
- [19] M. Opitz, G. Waltner, G. Poier, H. Possegger, and H. Bischof. Grid loss: Detecting occluded faces. In *European conference on computer vision*, pages 386–402. Springer, 2016.
- [20] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.
- [21] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [22] X. Shi, S. Shan, M. Kan, S. Wu, and X. Chen. Real-time rotation-invariant face detection with progressive calibration networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2295–2303, 2018.
- [23] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016.
- [24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [25] K. Wang, Y. Dong, H. Bai, Y. Zhao, and K. Hu. Use fast r-cnn and cascade structure for face detection. In *Visual Communications and Image Processing (VCIP), 2016*, pages 1–4. IEEE, 2016.
- [26] K. Yan, M. Bagheri, and R. M. Summers. 3d context enhanced region-based convolutional neural network for end-to-end lesion detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 511–519. Springer, 2018.
- [27] K. Yan, X. Wang, L. Lu, and R. M. Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3):036501, 2018.
- [28] Z. Yang and R. Nevatia. A multi-scale cascade fully convolutional network face detector. In *International Conference on Pattern Recognition*, pages 633–638. IEEE, 2016.
- [29] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S[^] 3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017.