Uncertainty-Aware Gradient Stabilization for Small Object Detection

Supplementary Material

1. Hessian in Norm-based Localization

For the $\mathcal{L}_2$ loss, we analyze the function's gradient and Hessian for center coordinate x:

$$\mathcal{L}_2(T_x, \hat{T}_x) = \left(\frac{x - \hat{x}}{w_a}\right)^2, \tag{1}$$

where w_a denotes the anchor width, x is the ground-truth coordinate, and $\hat{x}$ the predicted coordinate. Note that normalized coordinates are defined as $T_x = \frac{x - x_a}{w_a}$ and $\hat{T}_x = \frac{\hat{x} - x_a}{w_a}$, where x_a is the anchor's center. The gradient can be derived as:

$$\frac{\partial \mathcal{L}_2}{\partial \hat{T}_x} = 2(T_x - \hat{T}_x), \tag{2}$$

and the Hessian is derived as:

$$\mathbf{H}_x = \frac{\partial^2 \mathcal{L}_2}{\partial \hat{T}_x^2} = 2. \tag{3}$$

Mapping back to the original coordinates, the Hessian in terms of $\hat{x}$ is:

$$\mathbf{H}_x = \frac{\partial^2 \mathcal{L}_2}{\partial \hat{x}^2} = \frac{2}{w_a^2} \tag{4}$$

This reveals $K_x \propto 1/w_a^2$, and smaller anchors result in growth in K, leading to steeper loss landscapes.

For size regression using $\mathcal{L}_2$ loss on width w, the gradient with respect to $\hat{w}$ is:

$$\frac{\partial \mathcal{L}_2}{\partial \hat{w}} = 2 \cdot \log\left(\frac{w}{\hat{w}}\right) \cdot \left(-\frac{1}{\hat{w}}\right). \tag{5}$$

The Hessian is derived as:

$$\mathbf{H}_w = \frac{\partial^2 \mathcal{L}_2}{\partial \hat{w}^2} = \underbrace{2 \cdot \frac{1}{\hat{w}^2}}_{\text{Term 1}} + \underbrace{2 \cdot \frac{1}{\hat{w}^2} \log \frac{w}{\hat{w}}}_{\text{Term 2}} \tag{6}$$

Term 2 vanishes as $\hat{w} \to w$, and the Hessian approximates $\mathbf{H}_w = 2/\hat{w}^2$, matching Eq. 4. This reveals that the loss curvature becomes steep when $\hat{w}$ is small, potentially causing instability during optimization.

2. Hessian in IoU-based Localization

Following [1], we consider axis-aligned square boxes with ground truth center x and predicted center $\hat{x}$, the IoU loss is defined as:

$$\mathcal{L}_{\text{IoU}} = -\ln\left(\frac{I}{U}\right), \tag{7}$$

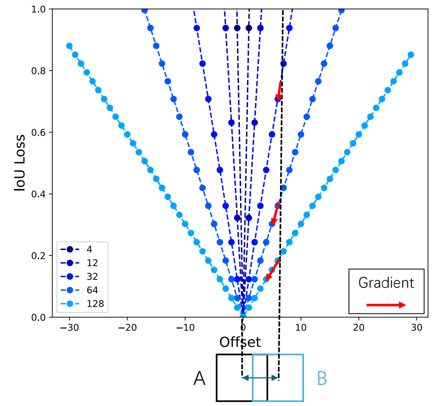

Figure 1. IoU loss and gradient magnitudes for different object sizes. Identical center shifts x produce larger gradients (red arrows) and sharper curvature for smaller boxes.

where $I = \max(0, w - |x - \hat{x}|)$, $U = 2w - I$, and w is the box width. The gradient with respect to $\hat{x}$ derives as:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{IoU}}}{\partial \hat{x}} &= -\frac{1}{U}\frac{\partial I}{\partial \hat{x}} + \frac{I}{U^2}\frac{\partial U}{\partial \hat{x}} \\ &= -\left(\frac{1}{U} + \frac{I}{U^2}\right)\frac{\partial I}{\partial \hat{x}} \quad (\because \partial U/\partial \hat{x} = -\partial I/\partial \hat{x})\end{aligned} \tag{8}$$

$$= \left(\frac{1}{w+d} + \frac{w-d}{(w+d)^2}\right)\text{sign}(x - \hat{x}), \quad d = |x - \hat{x}|. \tag{9}$$

The Hessian for overlapping boxes ($|x - \hat{x}| < w$) becomes:

$$\frac{\partial^2 \mathcal{L}_{\text{IoU}}}{\partial \hat{x}^2} = \frac{4w}{(w^2 - d^2)^2}, \tag{10}$$

For small objects, both gradient and Hessian exhibit inverse scaling:

$$\frac{\partial \mathcal{L}_{\text{IoU}}}{\partial \hat{x}} \propto \frac{1}{w}, \quad \frac{\partial^2 \mathcal{L}_{\text{IoU}}}{\partial \hat{x}^2} \propto \frac{1}{w^3}. \tag{11}$$

This confirms that smaller objects exhibit larger gradients and sharper curvature. Fig. 1 shows that small objects suffer from disproportionately steep gradients despite equal positional errors.

3. Convergence Analysis

Fig. 2 shows mAP progression for UGS versus TPH-YOLOv5-x [2] on VisDrone. Under both (a) 640×640

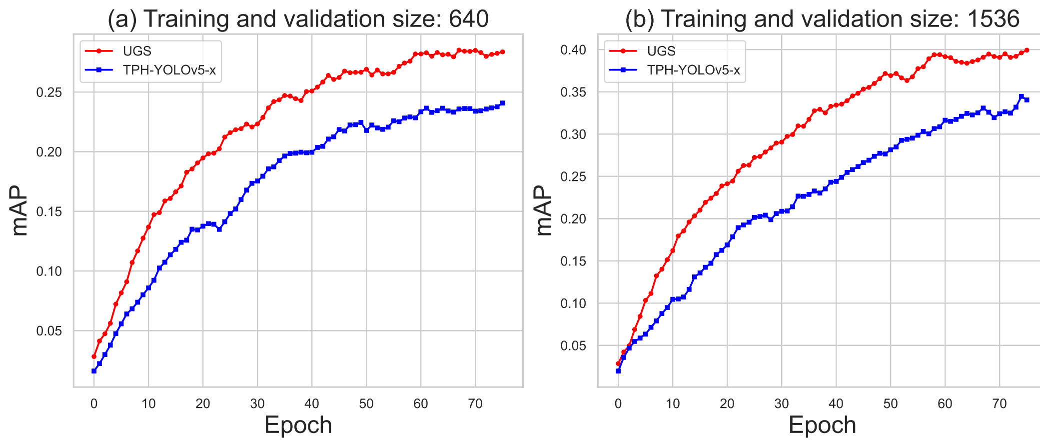

Figure 2. Comparison of training curves on VisDrone under two resolutions: (a) 640×640 and (b) 1536×1536. UGS consistently outperforms TPH-YOLOv5-x [2] across all epochs, achieving faster convergence and higher final mAP.

and (b) 1536×1536 resolutions, UGS achieves higher final mAP with faster convergence. For lower-resolution inputs (Fig. 2a), UGS surpasses the baseline by ∼3% mAP. For higher resolutions (Fig. 2b), it maintains a ∼4% mAP advantage, demonstrating stability across resolution.

References

[1] Chang Xu, Jinwang Wang, Wen Yang, and Lei Yu. Dot distance for tiny object detection in aerial images. In *CVPR*, pages 1192–1201, 2021. 1

[2] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In *ICCV*, pages 2778–2788, 2021. 1, 2