-Appendix-
Revisiting AP Loss for Dense Object Detection: Adaptive Ranking Pair Selection

1. Distant Function Selection

We evaluate the distant function with piece-wise step function $\mathbf{H}(\cdot)$ and sigmoid function $\mathbf{S}(\cdot)$, as shown in Fig. 1 and Fig. 2 respectively. The experimental results based on RetinaNet [1] are given in Table 1. We can observe that the performance gap between piece-wise step function $\mathbf{H}(\cdot)$ and sigmoid function $\mathbf{S}(\cdot)$ is only 0.1% in term of AP (37.4 *v.s.* 37.3). The results demonstrate that these two distance functions have no essential difference. In this paper, we use $\lambda = 8$ for all experiments.

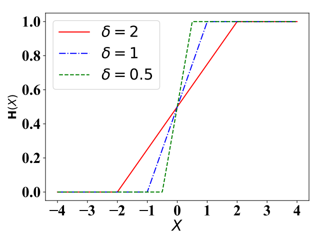

Figure 1. $\mathbf{H}(\cdot)$.

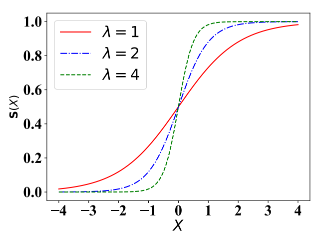

Figure 2. $\mathbf{S}(\cdot)$.

2. The Equivalence between Cross Entropy and Error-Driven Update

Here we find that if pair-wise error loss has the same gradients form as Eq.(7) in the main paper, then Error-Driven Update can be omitted for simplicity. To keep the numerator of pair-wise error gradients as the same as Eq.(7) in the main paper, we follow the common practice on cross entropy loss which adds a logistic function to sigmoid function. To start with, $\mathbf{S}(\cdot)$ is replaced with $\mathbf{CE}(\mathbf{S}(\cdot), 0)/\lambda$, which can be written as:

$$
\begin{aligned}
&\frac{1}{\lambda} CE(S(\hat{P}_v - \hat{P}_u), 0) \\
&= -\frac{1}{\lambda}((1-0) \cdot log(1 - S(\hat{P}_v - \hat{P}_u)) + 0 \cdot (S(\hat{P}_v - \hat{P}_u))) \\
&= -\frac{1}{\lambda} log(1 - S(\hat{P}_v - \hat{P}_u))
\end{aligned} \tag{1}
$$

where the gradients of this distance function w.r.t $S(\hat{P}_v - \hat{P}_u)$ can be calculated as:

$$
\frac{\partial CE(S(\hat{P}_v - \hat{P}_u), 0)}{\lambda \partial S(\hat{P}_v - \hat{P}_u)} = \frac{1}{\lambda(1 - S(\hat{P}_v - \hat{P}_u))} \tag{2}
$$

Table 1. Varying delta and lambda for distance function.

δ	AP	AP_{50}	AP_{75}	λ	AP	AP_{50}	AP_{75}
1	37.0	57.6	39.2	2	36.4	57.1	37.9
0.5	37.4	57.5	39.2	4	36.9	57.5	38.7
0.25	36.8	56.3	38.7	8	37.3	57.4	38.9
0.125	35.1	53.8	36.6	16	36.5	55.9	38.3

Since the gradient of $S(\hat{P}_v - \hat{P}_u)$ w.r.t $\hat{P}_u$ can be written as:

$$
\frac{\partial S(\hat{P}_v - \hat{P}_u)}{\partial \hat{P}_u} = -\lambda S(\hat{P}_v - \hat{P}_u)(1 - S(\hat{P}_v - \hat{P}_u)) \tag{3}
$$

Therefore, we can have the the gradients of distance function w.r.t $\hat{P}_u$:

$$
\begin{aligned}
&\frac{\partial CE(S(\hat{P}_v - \hat{P}_u), 0)}{\lambda \partial \hat{P}_u} = \frac{\partial CE(S(\hat{P}_v - \hat{P}_u), 0)}{\lambda \partial S(\hat{P}_v - \hat{P}_u)} \cdot \frac{\partial S(\hat{P}_v - \hat{P}_u)}{\partial \hat{P}_u} \\
&= \frac{1}{\lambda(1 - S(\hat{P}_v - \hat{P}_u))} \cdot (-\lambda S(\hat{P}_v - \hat{P}_u)(1 - S(\hat{P}_v - \hat{P}_u))) \\
&= -S(\hat{P}_v - \hat{P}_u)
\end{aligned} \tag{4}
$$

Also, to keep the denominator term BC as the same as Eq. (7) in the main paper, we detach it from backpropagation and treat it as a constant. Note that, after employing these two tricks (*i.e.*cross entropy and detaching), we can have the same gradient of our pair-wise error (*i.e.*, $(-\sum_{v \in \mathcal{N}} S(\hat{P}_v - \hat{P}_u))/(rank^+(u) + rank^-(u))$) as AP loss, which theoretically leads to similar performances. The experimental results in Table 1 in the main paper also demonstrate that.

3. Threshold for Selecting Valid Negative Samples

In training processing, the number of negative samples N_{neg} is enormous and might overwhelm the loss. To solve this issue, we utilize a larger margin threshold T to filter out easy negative samples, as shown in Fig. 3. Specifically, we set a valid indicator for each pair-wise error to ignore easy pairs. Here we describe indicator function $\mathbb{1}_{uv}$ as:

$$
\mathbb{1}_{uv} = \begin{cases} 1, & \hat{P}_v - \hat{P}_u > T \\ 0, & else \end{cases} \tag{5}
$$

Table 2. Varying th for N_{neg}.

Balance Constant	T	AP	AP_{50}	AP_{75}
$rank^{+}(u) + rank^{-}(u)$	N/A	37.3	57.4	38.9
N_{neg}	0	36.8	57.1	38.8
N_{neg}	0.2	37.2	57.0	38.9
N_{neg}	0.25	37.3	56.7	39.4
N_{neg}	0.3	36.9	56.3	38.7
N_{neg}	0.5	35.2	53.3	37.1

Table 3. Varying for Q on FCOS [2].

Q	AP	AP_{50}	AP_{75}
10,000	37.6	54.3	40.0
50,000	39.7	57.3	42.3
100,000	40.0	58.1	42.4
200,000	40.0	58.1	42.6

Then N_{neg} is formulated as: $N_{neg} = \sum_{v \in \mathcal{N}} \mathbb{1}_{uv}$. We also study the impact of different thresholds on detection accuracy. As shown in Table 2, when $T = 0.25$, N_{neg} provides the same performance as $rank^{+}(u)+rank^{-}(u)$. This demonstrates the selection of these two balance constants is robust.

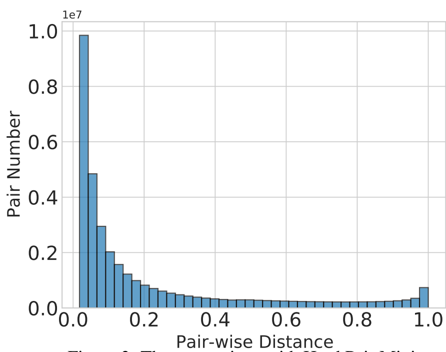

Figure 3. The comparison with Hard Pair Mining

4. Maximum Pair Number

In our experiments, the memory (11GB) of `2080TI` GPU can be ran out because of the extreme large number of pair $\{\hat{P}_v, \hat{P}_u\}$. Thus we adopt a simple yet efficient trick; constricting the input number of pairs.

Here we denote the maximum input number of pairs by Q (*i.e.*the maximum length of $\mathcal{A}_u$ for L_{APE}). Specifically, we manually choose the top Q predictions $\hat{P}_v$ of negative samples in $\mathcal{A}_u$. We conduct experiments varying Q for APE loss on FCOS, and the results are shown in Table. 3. When Q is greater than $100,000$, the performance will no longer be improved. It can be concluded from the results that the promotion from large Q becomes minor as the gradually increasing of Q.

References

[1] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Focal loss for dense object detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 2117–2125, 2017. 1

[2] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 9627–9636, 2019. 2