Supplementary for "Implicit rate-constrained optimization of non-decomposable objectives"

We provide more details, in particular:
Appendix A: Proof of Proposition 1.
Appendix B: *Run-time comparisons*, i.e. progress in terms of the evaluation metric on the validation and test sets as a function of training time.
Appendix C: More experiments on CelebA.
Appendix D: More experiments on BigEarthNet.

A Proof of Proposition 1

The assumption in Proposition 1 holds, for example, when we seek to minimize the FPR subject to FNR $= \beta$. In this case, the FPR objective can be approximated by $\tilde{f}(\theta, \lambda) = \mathbb{E}_{x \sim \mathcal{D}_0}\left[\ell\left(-1, \theta^\top x + \lambda\right)\right]$ and the FNR constraint can be approximated by $\tilde{g}(\theta, \lambda) = \mathbb{E}_{x \sim \mathcal{D}_1}\left[\ell\left(+1, \theta^\top x + \lambda\right)\right] - \beta$, where $\ell(y, z) = \log(1 + e^{-yz})$ is the standard logistic loss, and $\mathcal{D}_0$ and $\mathcal{D}_1$ are respectively the class-conditional distributions over examples with labels 0 and 1. Note that $\tilde{f}(\theta, \lambda)$ is jointly convex in (θ, λ) and is strictly increasing in λ, while $\tilde{g}(\theta, \lambda)$ is jointly convex in (θ, λ) and is strictly decreasing in λ.

Proof of Proposition 1. From the joint convexity of $\tilde{g}$, we have for any (θ, λ) and (θ', λ'):

$$\langle \nabla_\theta \tilde{g}(\theta', \lambda'), \theta - \theta' \rangle + \frac{\partial \tilde{g}(\theta', \lambda')}{\partial \lambda}(\lambda - \lambda') \;\leq\; \tilde{g}(\theta, \lambda) - \tilde{g}(\theta', \lambda').$$

Therefore this also holds for $(\theta, \tilde{h}(\theta))$ and $(\theta', \tilde{h}(\theta'))$:

$$\langle \nabla_\theta \tilde{g}(\theta', \tilde{h}(\theta')), \theta - \theta' \rangle + \frac{\partial \tilde{g}(\theta', \tilde{h}(\theta'))}{\partial \lambda}(\tilde{h}(\theta) - \tilde{h}(\theta')) \;\leq\; \tilde{g}(\theta, \tilde{h}(\theta)) - \tilde{g}(\theta', \tilde{h}(\theta')) \;=\; 0 - 0 = 0.$$

Because g is strictly decreasing in λ, $\frac{\partial \tilde{g}(\theta', \tilde{h}(\theta'))}{\partial \lambda} < 0$, and therefore we can rewrite the above inequality as:

$$\frac{1}{\frac{\partial \tilde{g}(\theta', \tilde{h}(\theta'))}{\partial \lambda}} \langle \nabla_\theta \tilde{g}(\theta', \tilde{h}(\theta')), \theta - \theta' \rangle + \tilde{h}(\theta) - \tilde{h}(\theta') \;\geq\; 0,$$

Using the fact that $\nabla_\theta \tilde{h}(\theta') = -\frac{\nabla_\theta \tilde{g}(\theta', \tilde{h}(\theta'))}{\frac{\partial \tilde{g}(\theta', \tilde{h}(\theta'))}{\partial \lambda}}$ (see (4) in the main text), we have:

$$\langle \nabla_\theta \tilde{h}(\theta'), \theta' - \theta \rangle \;\geq\; \tilde{h}(\theta') - \tilde{h}(\theta),$$

or

$$\tilde{h}(\theta) \;\geq\; \tilde{h}(\theta') + \langle \nabla_\theta \tilde{h}(\theta'), \theta - \theta' \rangle.$$

This shows that $\tilde{h}$ is convex in θ. The convexity of $\tilde{f}(\theta, \tilde{h}(\theta))$ follows from the convexity of $\tilde{f}$ and $\tilde{h}$, and from the fact that $\tilde{f}$ is monotonically increasing in its second argument. □

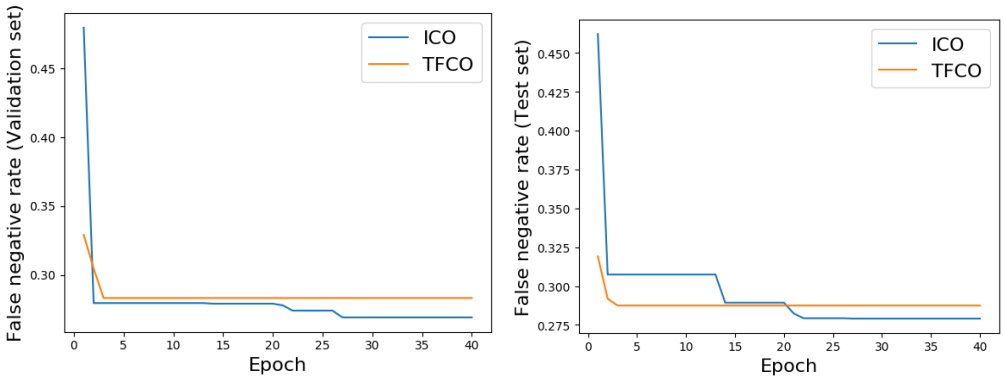

Figure 1: *Minimizing false negative rate (FNR) at a fixed false positive rate (FPR) of 0.05 for CelebA*: FNR as a function of training epochs for TFCO and the proposed ICO.

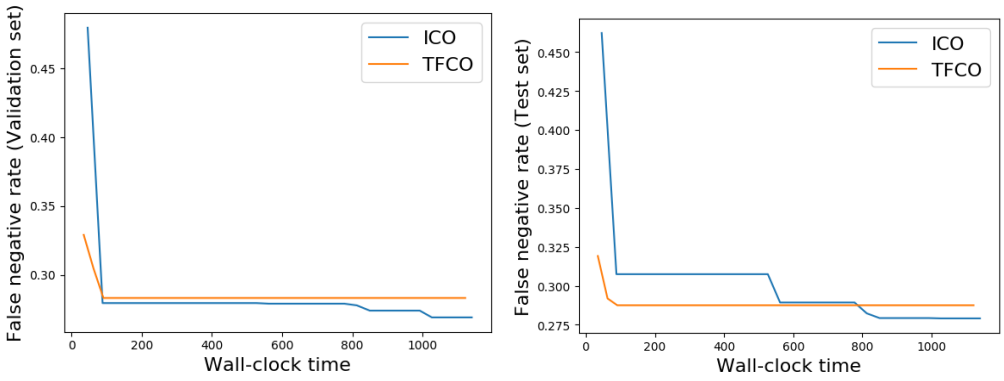

Figure 2: *Minimizing false negative rate (FNR) at a fixed false positive rate (FPR) of 0.05 for CelebA*: FNR as a function of wall-clock time for TFCO and the proposed ICO.

B Timing comparisons

We monitor the performance of TFCO and ICO in terms of value of the evaluation metric as the training proceeds. At the end of every training epoch, we record the best value of metric seen so far on the validation set, and use the same model (that yields the best validation metric) to score the test set.

For the problem of optimizing false negative rate (FNR) at a fixed false positive rate (FPR) on CelebA, Figures 1 and 2 show these FNR values for the attribute *High_Cheekbones* on the validation and test sets as the training proceeds in terms of training epochs and actual wall-clock time, respectively. We observe that while TFCO converges to a much lower FNR at the end of the first epoch (the first data point shown in Figure 1), the proposed ICO eventually achieves a lower FNR on both validation and test sets. Both TFCO and ICO were trained for 40 epochs in these experiments and TFCO was about 1.3x faster in terms of wall-clock time.

We repeat similar experiment for the problem of optimizing the partial area under the ROC curve for FPR in the range of $[0, 0.1]$, again for CelebA. Figure 3 and 4 show the ROC-AUC on validation and test sets for the *High_Cheekbones* attributes as the training proceeds in terms of epochs and wall-clock time, respectively. We observe similar behavior as in the earlier experiment and TFCO converges to a much better ROC-AUC early on in the training at the end of first epoch (first data point in the plots). However, the proposed ICO eventually achieves a better ROC-AUC both on validation and test sets. Both TFCO and ICO were trained for 25 epochs in this experiment and ICO is about 5x faster than TFCO in terms of wall-clock

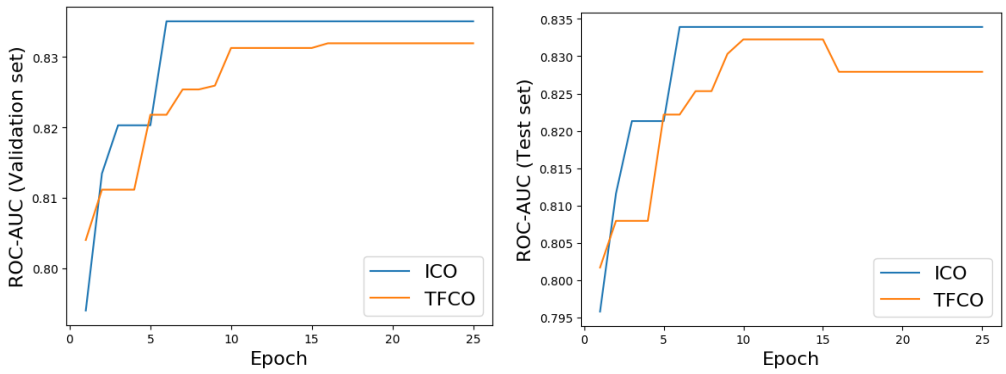

Figure 3: *Maximizing partial area under the ROC curve (for FPR* $\in [0, 0.1]$*) for CelebA*: ROC-AUC as a function of training epochs for TFCO and the proposed ICO.

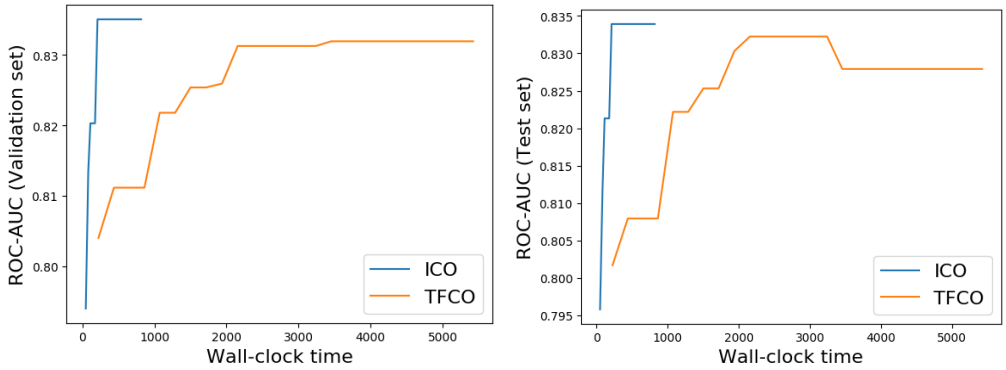

Figure 4: *Maximizing partial area under the ROC curve (for FPR* $\in [0, 0.1]$*) for CelebA*: ROC-AUC as a function of wall-clock time for TFCO and the proposed ICO.

time. This is due to the fact that optimizing ROC-AUC is a problem with multiple constraints (10 in this case) and we do not optimize the thresholds using gradients in ICO and only rely on the threshold correction step after every 100 minibatches. On the other hand, TFCO training time per minibatch slows down due to multiple constraints.

C CelebA results

We report results on more CelebA attributes for the two problems considered in the main text: (i) Minimizing false negative rate (FNR) at a fixed false positive rate (FPR) for FPRs $\in \{1\%, 2\%, 5\%, 10\%\}$ (Table 1), and (ii) Maximizing partial area under the ROC curve (ROC-AUC) for FPR in the range $[0, \beta]$ for $\beta \in \{1\%, 2\%, 5\%, 10\%, 20\%\}$ (Table 2). These results also show the standard deviation over five random trials which were omitted in the main text due to space constraints. For partial AUC in the FPR range $[0, \beta]$, we also compare with a pairwise loss baseline [1] which optimizes the objective $\frac{1}{N^+|S^-|}\sum_{i:y_i=1}\sum_{j\in S^-}\tilde{f}(s^\theta(x_i) - s^\theta(x_j))$, where $s^\theta(x)$ denotes the score (*e.g.*, logits) for example x, N^+ is the number of positive examples in the minibatch, S^- is the subset of negative examples whose scores lie in the top β fraction of all negative examples, and $\tilde{f}$ is the surrogate used for 0-1 loss (either softplus or sigmoid with a temperature hyperparameter as we used for the proposed method).

D BigEarthNet results

We also report results on more BigEarthNet labels for the problem of maximizing partial area under the ROC curve (ROC-AUC) for FPR in the range $[0, \beta]$ for $\beta \in \{5\%, 10\%, 20\%\}$ (Table 3). These results also show the standard deviation over five random trials which were omitted in the main text due to space constraints. We also compare with the pairwise loss baseline for partial AUC as described earlier. The proposed ICO outperforms cross-entropy and pairwise loss baselines in all the cases, and also outperforms TFCO for most cases.

Table 1: Minimizing false negative rate (FNR) at a given false positive rate (FPR) for **CelebA**. The mean FNR (in %) are reported over five random trials, along with std. deviations, for cross-entropy (CE), TFCO and ICO. Proposed ICO outperforms both CE and TFCO by a considerable margin in most cases. *Lower* values are better.

Attributes	FPR	CE	TFCO	ICO
High-cheekbones	1%	53.52 (1.71)	49.09 (1.56)	**46.96 (0.55)**
	2%	44.82 (1.23)	40.89 (0.82)	**39.84 (0.49)**
	5%	32.87 (0.83)	30.11 (0.67)	**28.54 (0.30)**
	10%	22.88 (0.43)	20.37 (0.65)	**19.66 (0.36)**
Heavy-makeup	1%	57.00 (0.84)	52.07 (1.24)	**49.65 (0.81)**
	2%	45.59 (1.24)	41.23 (1.41)	**38.89 (0.80)**
	5%	28.22 (0.80)	25.43 (0.97)	**23.11 (0.70)**
	10%	15.12 (0.44)	13.61 (0.77)	**12.36 (0.19)**
Wearing-lipstick	1%	44.00 (1.28)	42.64 (1.29)	**37.47 (0.97)**
	2%	32.74 (0.88)	30.44 (1.51)	**26.74 (0.50)**
	5%	16.33 (0.43)	14.97 (0.77)	**13.05 (0.25)**
	10%	6.61 (0.27)	5.92 (0.21)	**4.78 (0.12)**
Smiling	1%	37.40 (1.30)	35.93 (1.37)	**33.74 (0.71)**
	2%	29.44 (0.76)	27.80 (1.23)	**26.10 (0.57)**
	5%	18.73 (0.53)	17.04 (0.80)	**16.88 (0.25)**
	10%	11.78 (0.20)	10.74 (0.29)	**10.23 (0.25)**
Black-hair	1%	69.32 (1.78)	64.47 (1.55)	**63.23 (1.19)**
	2%	56.48 (1.48)	52.00 (1.10)	**50.50 (0.67)**
	5%	36.72 (1.61)	**32.41 (0.60)**	32.48 (0.66)
	10%	22.97 (1.87)	19.16 (1.22)	**18.62 (0.43)**
Blond-hair	1%	40.49 (1.18)	38.62 (1.17)	**36.85 (0.58)**
	2%	28.89 (1.17)	25.64 (1.20)	**24.20 (0.72)**
	5%	13.44 (1.01)	11.64 (0.76)	**10.81 (0.24)**
	10%	6.54 (0.46)	4.91 (0.22)	**4.68 (0.20)**
Brown-hair	1%	80.75 (1.82)	77.16 (0.74)	**76.34 (0.78)**
	2%	69.69 (1.77)	66.10 (1.04)	**65.74 (1.28)**
	5%	52.41 (2.55)	**45.83 (0.92)**	46.43 (0.51)
	10%	35.92 (2.71)	**29.94 (0.87)**	30.02 (0.67)
Wavy-hair	1%	85.04 (0.80)	84.42 (0.95)	**83.54 (0.69)**
	2%	78.91 (1.20)	77.02 (1.47)	**76.07 (0.90)**
	5%	65.79 (1.77)	61.81 (0.92)	**60.71 (1.01)**
	10%	50.52 (1.41)	47.49 (0.98)	**45.95 (1.12)**

Table 2: Maximizing area under the ROC curve for **CelebA**, in a given FPR range $[0, \beta]$ for $\beta \in \{1\%, 2\%, 5\%, 10\%, 20\%\}$. The mean ROC-AUC are reported over five random trials, along with std. deviations, for cross-entropy (CE), Pairwise-loss, TFCO and ICO. *Higher* values are better.

Attributes	FPR	CE	Pairwise-loss	TFCO	ICO
High-cheekbones	1%	66.10 (1.14)	68.18 (2.01)	62.96 (10.45)	**69.83 (0.84)**
	2%	70.87 (0.91)	72.85 (0.28)	**74.98 (0.20)**	73.17 (0.84)
	5%	75.89 (1.26)	78.13 (0.36)	73.83 (11.30)	**78.45 (0.53)**
	10%	80.15 (1.02)	81.51 (0.57)	74.07 (12.13)	**82.67 (0.49)**
	20%	84.26 (0.74)	85.70 (0.33)	72.44 (17.48)	**86.82 (0.30)**
Heavy-makeup	1%	65.03 (0.92)	66.33 (1.41)	**68.55 (0.83)**	66.75 (0.31)
	2%	68.82 (0.61)	71.13 (0.80)	**73.43 (0.13)**	71.57 (0.80)
	5%	75.48 (1.44)	77.78 (0.40)	**79.54 (0.09)**	78.32 (0.33)
	10%	81.53 (1.10)	82.85 (0.75)	**85.00 (0.12)**	84.39 (0.47)
	20%	88.22 (0.40)	88.68 (0.36)	89.93 (0.11)	89.81 (0.19)
Wearing-lipstick	1%	70.24 (1.13)	71.77 (0.70)	**74.82 (0.33)**	72.29 (0.77)
	2%	75.42 (0.68)	77.21 (0.39)	**79.28 (0.22)**	78.44 (0.28)
	5%	82.19 (0.37)	83.42 (0.97)	84.68 (0.19)	84.56 (0.28)
	10%	87.70 (0.89)	88.44 (0.25)	89.35 (0.18)	**89.73 (0.27)**
	20%	91.88 (0.44)	93.00 (0.20)	93.19 (0.12)	**93.93 (0.15)**
Smiling	1%	75.39 (0.51)	75.87 (0.63)	**78.03 (0.42)**	75.59 (0.76)
	2%	78.44 (0.41)	79.85 (0.52)	**81.51 (0.20)**	79.80 (0.55)
	5%	83.48 (0.68)	84.50 (0.54)	64.97 (17.24)	**84.81 (0.41)**
	10%	86.76 (0.84)	88.06 (0.22)	73.79 (18.63)	**88.80 (0.30)**
	20%	90.88 (0.52)	91.46 (0.11)	76.61 (18.69)	**92.08 (0.09)**
Black-hair	1%	60.53 (0.86)	57.73 (1.11)	**61.44 (0.44)**	61.24 (0.26)
	2%	64.48 (0.45)	63.93 (1.10)	**66.53 (0.44)**	66.07 (0.61)
	5%	70.87 (1.01)	71.85 (0.25)	73.19 (0.29)	**73.79 (0.45)**
	10%	78.04 (0.71)	78.49 (0.56)	79.88 (0.11)	**80.19 (0.07)**
	20%	84.33 (0.54)	85.45 (0.37)	86.00 (0.09)	86.09 (0.30)
Blond-hair	1%	71.36 (0.62)	70.38 (0.89)	**73.11 (0.46)**	72.11 (0.37)
	2%	76.50 (0.91)	76.25 (0.66)	**79.01 (0.16)**	78.06 (0.54)
	5%	84.74 (0.69)	84.21 (0.37)	**86.18 (0.13)**	85.76 (0.25)
	10%	89.39 (0.59)	89.75 (0.73)	**90.63 (0.15)**	90.49 (0.21)
	20%	93.30 (0.33)	93.56 (0.38)	94.24 (0.10)	94.27 (0.14)
Brown-hair	1%	55.40 (0.54)	52.65 (0.56)	56.09 (0.28)	**56.61 (0.41)**
	2%	58.82 (0.38)	54.62 (1.34)	59.68 (0.21)	**60.10 (0.40)**
	5%	64.87 (0.63)	61.21 (2.25)	66.71 (0.11)	**67.13 (0.40)**
	10%	69.74 (1.14)	70.05 (0.34)	**73.23 (0.27)**	73.01 (0.25)
	20%	77.67 (1.03)	78.25 (0.51)	80.09 (0.16)	80.06 (0.23)
Wavy-hair	1%	54.03 (0.28)	50.91 (0.23)	52.36 (1.31)	**54.33 (0.09)**
	2%	55.85 (0.78)	52.08 (0.50)	52.64 (2.14)	**57.02 (0.31)**
	5%	60.16 (0.73)	54.17 (2.30)	53.70 (2.98)	**61.30 (0.38)**
	10%	64.42 (0.37)	56.70 (2.09)	57.16 (4.32)	**65.34 (0.48)**
	20%	69.26 (1.03)	68.56 (0.49)	66.47 (5.60)	**71.48 (0.32)**

Table 3: Maximizing area under the ROC curve for **BigEarthNet**, in a given FPR range $[0, \beta]$ for $\beta \in \{5\%, 10\%, 20\%\}$. The mean ROC-AUC are reported over five random trials, along with std. deviations, for cross-entropy, Pairwise-loss, TFCO, ICO. *Higher* values are better.

Labels	FPR	CE	Pairwise-loss	TFCO	ICO
Broad-Leaved Forest (BLF)	5%	66.20 (0.59)	52.07 (0.35)	66.43 (1.86)	**69.90 (0.53)**
	10%	71.00 (0.80)	53.78 (1.34)	71.72 (0.78)	**73.91 (0.66)**
	20%	75.42 (0.67)	57.94 (0.81)	76.20 (0.61)	**77.91 (0.77)**
Complex Cultivation patterns (CC)	5%	62.19 (0.48)	52.44 (0.26)	62.71 (2.19)	**63.61 (0.12)**
	10%	67.46 (0.25)	54.71 (0.62)	66.75 (0.82)	**68.35 (0.42)**
	20%	73.81 (0.83)	59.34 (0.64)	**76.01 (0.77)**	74.88 (0.43)
Coniferous Forest (CF)	5%	71.93 (0.66)	59.00 (0.56)	71.49 (3.44)	**74.70 (0.62**)
	10%	78.62 (0.59)	77.66 (2.22)	79.98 (1.26)	**80.76 (0.94)**
	20%	84.82 (0.68)	85.84 (0.31)	**86.62 (0.62)**	86.02 (0.29)
Discontinuous Urban Fabric (DUF)	5%	69.80 (1.45)	55.33 (0.98)	71.76 (1.89)	**73.94 (0.39)**
	10%	75.13 (0.86)	59.15 (2.17)	77.20 (0.87)	**78.03 (0.32)**
	20%	78.86 (1.19)	78.89 (0.38)	81.45 (0.62)	**81.83 (0.57)**
Land principally occupied by Agriculture, with significant areas of Natural Vegetation (ANV)	5%	58.79 (0.49)	51.23 (0.16)	58.80 (0.77)	**60.46 (0.17)**
	10%	62.72 (0.46)	52.65 (0.56)	63.89 (0.98)	**64.38 (0.32)**
	20%	67.77 (0.34)	54.36 (0.63)	69.17 (0.52)	68.97 (0.76)
Mixed Forest (MF)	5%	64.48 (0.76)	54.59 (0.60)	65.50 (0.30)	**65.93 (0.60)**
	10%	71.05 (0.47)	60.41 (0.28)	72.06 (0.33)	71.89 (0.48)
	20%	77.56 (0.69)	76.44 (0.26)	78.83 (0.14)	**79.20 (0.26)**
Non-Irrigated Arable Land (NIAL	5%	70.07 (0.27)	55.10 (0.28)	**72.67 (0.19)**	71.45 (0.48)
	10%	75.29 (0.37)	59.62 (2.37)	**77.30 (0.12)**	76.78 (0.76)
	20%	79.97 (0.67)	80.23 (0.17)	**82.05 (0.09)**	81.72 (0.56)
Pastures	5%	72.70 (0.46)	59.41 (0.85)	**74.16 (0.29)**	73.61 (0.55)
	10%	75.95 (0.55)	74.78 (0.29)	78.04 (0.31)	77.85 (0.65)
	20%	80.31 (0.87)	80.42 (0.66)	**82.38 (0.17)**	82.10 (0.18)
Transitional Woodland/Shrub (TWS)	5%	57.12 (0.32)	51.30 (0.32)	58.21 (0.08)	**59.64 (0.27)**
	10%	60.24 (0.21)	52.45 (0.61)	61.82 (0.13)	**62.82 (0.61)**
	20%	64.98 (0.92)	55.47 (0.31)	67.15 (0.26)	67.24 (1.10)
Water Bodies (WB)	5%	76.52 (0.59)	54.29 (1.35)	77.47 (0.48)	**78.71 (0.33)**
	10%	80.81 (0.69)	57.23 (0.73)	81.76 (0.34)	**82.79 (0.35)**
	20%	85.27 (0.39)	86.11 (0.25)	85.46 (0.19)	**86.66 (0.31)**

References

[1] Harikrishna Narasimhan and Shivani Agarwal. Svmpauctight: a new support vector method for optimizing partial auc based on a tight convex upper bound. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 167–175, 2013.