Extension of Intersection over Union to Improve Small Object Detection in Few-Shot Regime

Pierre Le Jeune
COSE & L2TI
Université Sorbonne Paris Nord
pierre.le-jeune@cose.fr

Anissa Mokraoui
L2TI
Université Sorbonne Paris Nord
anissa.mokraoui@univ-paris13.fr

***Abstract*—Intersection over Union (IoU) is a widely used criterion that quantifies the overlap between two bounding boxes. It plays a crucial role in object detection, serving both as a cost function for training detection models and as a criterion for evaluating their performance. The IoU value between the ground truth and predicted boxes determines whether a detection is considered accurate, based on a predefined threshold. However, this approach poses challenges when detecting small objects, especially in situations where annotated data is scarce, such as in few-shot learning scenarios. The scarcity of supervision hinders the learning of robust localization, which is especially detrimental for small objects. A small discrepancy of just a few pixels between the predicted and annotated bounding boxes can result in a false detection for small objects. To address these issues, we propose Scale-adaptive Intersection over Union (SIoU), a new controllable and adaptive similarity criterion that adjusts based on object size. First, SIoU helps to find a better balance between small and large objects during the training of few-shot detection methods, for which small objects are extremely problematic. Experiments on four distinct datasets show superior detection performance when using SIoU as a cost function. Second, by being more lenient with small objects, SIoU aligns more closely with human perception than IoU, making it a more suitable evaluation criterion.**

***Index Terms*—Object Detection, Few-Shot Learning, IoU, GIoU, Small Objects.**

I. Introduction

Despite significant progress in object detection, locating small objects remains a persistent challenge (see e.g., [1], [2], [3], [4], [5]). This difficulty is exacerbated in scenarios where annotations are sparse, such as in few-shot learning, as demonstrated by [6]. In such cases, the limited supervision prevents the model from learning robust localization, and errors in localization become especially problematic for small objects [7]. Most detection pipelines, including few-shot object detection (FSOD), rely on Intersection over Union (IoU) as a regression loss [8], [9], for example selection [1], [2], [10], or as an evaluation metric. However, IoU is far from ideal when it comes to detecting small objects.

The main issue with IoU is that it is invariant to object size. Specifically, if two bounding boxes, b_1 and b_2, are offset by ρ pixels, scaling both the boxes and their coordinates by the same factor does not alter the IoU. While this invariance may seem desirable in some cases, it becomes problematic when dealing with small objects. IoU is typically used with a threshold, such as considering detections with an IoU greater than 0.5 as correct. However, this threshold is fixed for all object sizes, which means that for small objects, even a small localization error, ε_{loc} (i.e., the absolute pixel shift), can result in an incorrect detection. In contrast, larger objects can tolerate larger localization errors without failing the IoU threshold.

Furthermore, the ratio between localization error and object size increases significantly for small objects (see Figure 1 right). Despite being trained with IoU, existing networks are not size-independent and are much less effective at detecting small targets. To address this issue, we propose Scale-adaptive Intersection over Union (SIoU), an extension of IoU that adapts to object size, providing a more balanced and accurate detection performance across all object sizes.

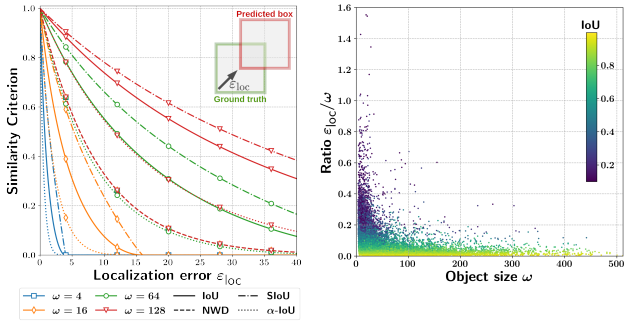

Fig. 1. **(Left)** Evolution of the IoU, NWD [15], SIoU with $\gamma = 0.5$ and $\gamma = -1$, and α-IoU [16] criteria as a function of the localization error ε_{loc} (in pixels) between a prediction and a label for different object sizes $\omega \in \{8, 32, 96\}$. **(Right)** Ratio between the localization error ε_{loc} and the object size ω for a detection model trained on the DOTA dataset. Each point represents the localization error of the model for an object in DOTA.

SIoU is configurable and allows for giving more or less importance to small objects while maintaining a behavior similar to IoU for large objects. The notion of object size is concretely defined in [11]: small objects are those whose area ω^2 does not exceed 32^2 pixels, medium objects satisfy $32 < \omega \leq 96$, and large objects have $\omega > 96$. SIoU can be used as a cost function to favor small objects during training. Our experiments show that this leads to significant improvements in detection under the few-shot regime (FSOD)

This work was funded by the company COSE and the National Research Agency (ANR) through the LabCom IRISER (ANR-21-LCV3-0004)

on natural images (on the Pascal VOC dataset [12] and MS COCO dataset [11]), as well as on aerial images (on the DOTA dataset [13] and DIOR dataset [14]). We focus our analysis on the few-shot learning regime, as it is more impacted by the presence of small objects and better reflects real-world application scenarios. However, consistent results for conventional detection are also presented. SIoU can also be used as a similarity criterion when evaluating models. A subjective study we conducted with 74 participants shows that, on average, humans are more lenient than IoU when it comes to small objects. Thus, SIoU is a more relevant criterion than IoU for evaluating detection models and building applications that are better suited to human users

II. SIoU: New proposed criterion

Before presenting the new criterion, let us first recall the definition of IoU and some of its variants.

A. IoU and its variants

The IoU between two bounding boxes $b_1 = [x_1, y_1, w_1, h_1]^T$ and $b_2 = [x_2, y_2, w_2, h_2]^T$ is calculated as the ratio of the area of their intersection to the area of their union. Here, x_i and y_i represent the coordinates of the center of the box b_i, while w_i and h_i represent its width and height. Many variants of IoU have been proposed in the literature. One of the most well-known is the *Generalized IoU* (GIoU) [17], which generalizes IoU when the union between b_1 and b_2 is empty. For this, GIoU subtracts from the IoU a term that measures the distance between the two boxes. Thus, $\text{GIoU}(b_1, b_2) \in [-1, 1]$ while $\text{IoU}(b_1, b_2) \in [0, 1]$. GIoU is particularly effective as a loss function: $\mathcal{L}\text{GIoU}(b_1, b_2) = 1 - \text{GIoU}(b_1, b_2)$ because it alleviates optimization issues with $\mathcal{L}$IoU when the boxes do not overlap. α-IoU [16] is another variant of IoU, which raises IoU to the power of α. α-IoU penalizes predictions with a large IoU with a ground truth box, aiming to improve detection accuracy in general. α allows adjusting the desired precision. More recently, [15] proposed an alternative to better detect small objects. Its principle relies on the calculation of a Normalized Wasserstein Distance (NWD) between two Gaussian distributions fitted to the boxes b_1 and b_2. This criterion is more lenient for small boxes. However, when b_1 and b_2 have the same dimensions, NWD becomes equivalent to an Euclidean distance between the centers of the boxes and thus loses its variable behavior depending on the object size.

B. Scaled-adaptative Intersection over Union

In order to address the difficulties of detecting small objects in a few-shot learning regime, we propose *Scale-Adaptive Intersection over Union* (SIoU), defined as follows:

$$\begin{aligned} &\text{SIoU}(b_1, b_2) = \text{IoU}(b_1, b_2)^p \\ &\text{where} \qquad p = 1 - \gamma \exp\left(-\frac{\sqrt{w_1 h_1 + w_2 h_2}}{\sqrt{2}\kappa}\right), \end{aligned} \tag{1}$$

where p is a function of the object size, so the IoU values are either amplified or reduced depending on the size of the objects. The parameters $\gamma \in]-\infty, 1]$ and $\kappa > 0$ allow for controlling the strength of the amplification and the speed with which the IoU behavior is restored for large objects:

$$\lim_{\tau \to +\infty} \text{SIoU}(b_1, b_2) = \text{IoU}(b_1, b_2) \quad (\tau = \frac{1}{2}(w_1 h_1 + w_2 h_2)).$$

Thus, SIoU enables a variable yet controlled behavior depending on the object size, while retaining properties similar to IoU. When $\gamma > 0$, SIoU assigns higher similarity values to smaller objects, making it better aligned with human perception (see Section III). Conversely, with $\gamma < 0$, SIoU generates lower values for small objects, which we will see in Section IV has benefits for training detection models. Figure 1 (left) shows the difference in behavior between the criteria discussed and proposed in the previous sections (IoU, SIoU, NWD, and α-IoU), for different object sizes.

SIoU can then be generalized as IoU in GIoU. It is simply necessary to replace IoU with GIoU (denoted $g(b_1, b_2)$ below) in equation 1:

$$\text{GSIoU}(b_1, b_2) = \begin{cases} \text{g}(b_1, b_2)^p & \text{if } \text{g}(b_1, b_2) \geq 0 \\ -|\text{g}(b_1, b_2)|^p & \text{if } \text{g}(b_1, b_2) < 0 \end{cases}. \tag{2}$$

Thus, we define the cost functions $\mathcal{L}_{\text{SIoU}}(b_1, b_2) = 1 - \text{SIoU}(b_1, b_2)$ and $\mathcal{L}_{\text{GSIoU}} = 1 - \text{GSIoU}(b_1, b_2)$. These cost functions can directly replace $\mathcal{L}_{\text{IoU}}$ and $\mathcal{L}_{\text{GIoU}}$. It is important to note the formal similarity between α-IoU and SIoU. However, there is a crucial difference between the two. The behavior of α-IoU does not change with the size of the objects, unlike SIoU (see Figure 1, left), and does not contribute to improving small object performance nor to a better alignment with human perception.

III. Alignment with human perception

In addition to its utility for training detection models, IoU is also necessary for their evaluation. These models are typically evaluated with the *mean Average Precision* (mAP) at an IoU threshold (for example, 0.5). This means that a bounding box prediction is considered correct only if its IoU with a ground-truth box exceeds the defined threshold. For small objects, this becomes problematic because a shift of just a few pixels is enough to turn a correct prediction into a false positive. Therefore, detection performance for small objects is often low, even though the associated predictions seem quite satisfactory to a human observer. Indeed, a human is likely to be more lenient than IoU when evaluating small objects.

We conducted a subjective study in which an observer must rate a prediction relative to a ground-truth box on a scale from 1 (very poor localization) to 5 (very good localization). Seventy-four participants (specialists and non-specialists in the field) took part in this study, resulting in a total of over 3000 ratings. The results show that, on average, a human assigns higher scores to small objects than to large ones when the ratio $\varepsilon_{\text{loc}}/\omega$ is fixed. In other words, the human eye is more lenient than IoU with small objects. For reference, IoU remains constant when $\varepsilon_{\text{loc}}/\omega$ is fixed. SIoU, on the other hand, relaxes

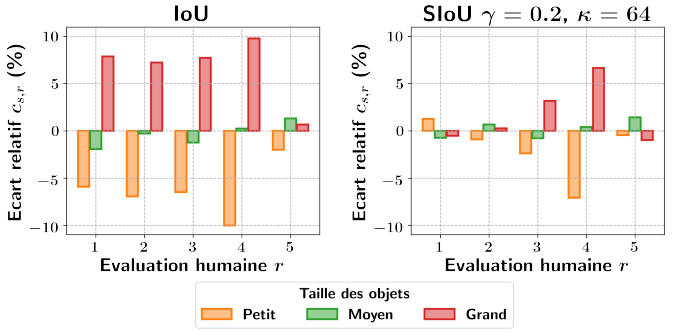

Fig. 2. Average value of IoU **(left)** and SIoU **(right)** for different object sizes s and rating $r \in \{1, 2, 3, 4, 5\}$. The scores are given as the relative deviation from the average score for each r.

this invariance and aligns better with human perception. This is visible in Figure 2, which shows the relative deviation of IoU (left) and SIoU (right) grouped by rating r and object size s, relative to the mean for each rating r. More specifically, we define:

$$c_{s,r} = \frac{\mathcal{C}s, r - \frac{1}{|\mathcal{S}|} \sum s \in \mathcal{S}\mathcal{C}s, r}{\frac{1}{|\mathcal{S}|} \sum s \in \mathcal{S}\mathcal{C}_{s,r}}, \tag{3}$$

where $\mathcal{C}_{s,r}$ represents the average IoU or SIoU for an object size s and a rating r. This figure first shows that to assign a rating r, a human has a lower IoU threshold for small objects than for large ones. It then shows that SIoU is much better suited to human perception, as the SIoU thresholds for a rating r are almost identical regardless of the object size. This demonstrates a better alignment of SIoU with human perception.

To reinforce this, Table I shows the correlation values between the different criteria studied here (IoU, SIoU, α-IoU, and NWD) and human perception. Again, SIoU is superior to the other alternatives. Here, we chose $\gamma = 0.2$ and $\kappa = 64$ to maximize alignment with human perception. Having a similarity criterion closer to human perception is crucial for the development of models better suited to human judgment, which is particularly useful for detection applications (photo-interpretation, radiology, etc.).

TABLE I. Correlation (Kendall's τ) between the values of different criteria and human ratings. For SIoU, $\gamma = 0.2$ and $\kappa = 64$, for α-IoU, $\alpha = 3$.

	IoU	**SIoU**	**α-IoU**	**NWD**
$\boldsymbol{r}$	0.674	**0.701**	0.674	0.550

IV. Experimental Results

To demonstrate the effectiveness of SIoU as a cost function, a series of experiments are presented in this section. We focus here on the few-shot regime, as it is more demanding for small objects and closer to real-world applications. The experiments are primarily conducted on aerial images (DOTA [13] and DIOR dataset [14]), but results on natural images (Pascal VOC [12] and MS COCO [11]) are also reported.

A. Implementation Details

In the few-shot learning regime, we mainly focus on performance for novel classes, i.e., classes for which very few annotated examples are available. Specifically, for all our experiments, we used 10 *shots* (i.e., 10 annotated images per class). Results on base classes are also included in the tables for reference. We based our experiments on XQSA [18] [7], a recent few-shot detection method that operates at multiple scales to improve small object detection. This method is based on the FCOS detector [3] and is trained episodically, first on base classes and then on novel classes.

B. Analysis of the experimental results

First, we compare the influence of the different criteria discussed and proposed on few-shot detection performance. These results are reported in Table II. We observe a clear dominance of SIoU and GSIoU (the criteria are separated into two groups based on whether they assign identical values or not when the boxes have an empty intersection). SIoU and GSIoU significantly improve performance on novel classes, especially for small objects.

TABLE II. Comparison of few-shot performance using different criteria: IoU, α-IoU, SIoU, NWD, GIoU, and GSIoU as the cost function. mAP is reported with an IoU threshold of 0.5 and according to object sizes: small (**S**), medium (**M**), large (**L**), and all sizes combined (**All**), with $\gamma = -3$ and $\kappa = 16$.

	Base classes				**Novel Classes**			
Loss	**All**	**S**	**M**	**L**	**All**	**S**	**M**	**L**
IoU	50.67	**25.83**	57.49	68.24	32.41	10.06	47.87	67.09
α-IoU	46.72	13.24	55.21	**69.94**	33.95	12.58	46.58	**74.50**
SIoU	**53.62**	24.07	**61.91**	67.34	**39.05**	**16.59**	**54.42**	**74.49**
NWD	50.79	19.19	58.90	67.90	41.65	28.26	50.16	65.06
GIoU	52.41	**26.94**	61.17	63.00	41.03	24.01	**52.13**	69.78
GSIoU	**52.91**	22.14	**61.19**	**66.02**	**45.88**	**34.83**	51.26	**70.78**

1) Impact of γ on the detection performance: It is important to note that the values of γ and κ must be chosen carefully when SIoU (or GSIoU) is used in the cost function calculation. Indeed, as seen in Section III, IoU is too strict for small objects, and by using $\gamma = 0.2$, we obtain a behavior closer to human perception (this is the case whenever $\gamma > 0$). However, when focusing on training, $\gamma > 0$ results in reducing the cost for small objects ($\mathcal{L}_{\text{GSIoU}} = 1 - \text{GSIoU}(b_1, b_2)$). This alters the optimization balance between small and large objects. The training then focuses on reducing the cost generated by large objects, resulting in poorer performance on small objects. Thus, when SIoU and GSIoU are used as cost functions, negative values of γ should be chosen. This increases the cost generated by small objects, and the training then focuses on detecting small targets. This can be clearly seen in Table III, which shows the performance on DOTA for different values of γ. κ has much less influence on performance and is set to 16 in our experiments.

2) Performance detection using four datasets: In order to demonstrate the robustness of GSIoU, we conducted experiments on four datasets: DOTA [13] and DIOR [14]

TABLE III. Evolution of few-shot detection performance on DOTA for different values of γ, with $\kappa = 16$ fixed, using GSIoU as a loss function, where the mAP is reported with a 0.5 IoU threshold.

	Base classes				Novel Classes			
γ	**All**	**S**	**M**	**L**	**All**	**S**	**M**	**L**
0.5	47.09	21.29	54.67	65.48	30.50	8.83	44.97	65.89
0.25	45.94	21.60	54.39	63.40	30.96	12.53	42.37	64.14
0	52.41	26.94	61.17	63.00	41.03	24.01	52.13	69.78
-0.5	52.80	**27.16**	61.19	64.61	41.06	25.20	50.18	**72.04**
-1	53.03	23.20	61.53	66.68	42.77	27.55	**52.01**	70.76
-2	**54.06**	23.68	**62.69**	66.62	43.67	30.04	51.69	69.66
-3	52.91	22.14	61.19	66.02	**45.88**	**34.83**	51.26	70.78
-4	53.59	22.50	62.48	66.18	42.43	27.56	51.79	68.70
-9	53.11	20.98	62.13	**67.00**	42.63	30.53	48.89	68.62

(containing aerial images), as well as Pascal VOC [12] and MS COCO [11] (natural images). Using GSIoU instead of GIoU as the cost function improves detection performance for novel classes, particularly for small objects (see Table IV). However, it is noteworthy that for natural images (Pascal VOC and MS COCO), the gains are slightly smaller compared to aerial images. This can be explained by the fact that objects are generally much smaller in aerial images than in natural images [6].

TABLE IV. Comparison of detection performance between GIoU and GSIoU on 4 datasets: DOTA, DIOR, Pascal VOC, and MS COCO, in a few-shot setting. $\gamma = -3$ and $\kappa = 16$ for DOTA and DIOR, but $\gamma = -1$ for Pascal VOC and MS COCO.

		Base classes				Novel Classes			
		All	**S**	**M**	**L**	**All**	**S**	M	L
DOTA	GIoU	52.41	**26.94**	61.17	63.00	41.03	24.01	**52.13**	69.78
	GSIoU	**52.91**	22.14	**61.19**	**66.02**	**45.88**	**34.83**	51.26	**70.78**
DIOR	GIoU	58.90	10.38	40.76	80.44	47.93	9.85	47.61	68.40
	GSIoU	**60.29**	**11.28**	**43.24**	**81.63**	**52.85**	**13.78**	**53.73**	**71.22**
Pascal	GIoU	51.09	**13.93**	**40.26**	62.01	48.42	18.44	36.06	59.99
	GSIoU	**54.47**	13.88	40.13	**66.82**	**55.16**	**22.94**	**36.24**	**67.40**
COCO	GIoU	19.15	**8.72**	22.50	30.59	26.25	11.96	23.95	38.60
	GSIoU	**19.57**	8.41	**23.02**	**31.07**	**27.11**	**12.81**	**26.02**	**39.20**

3) Additional results on non few-shot setting: Finally, we also conducted experiments for detection in the classic setting (i.e., with many annotations) on DOTA and DIOR (see Table V). Once again, using GSIoU as the cost function improves detection performance. However, the gains are less significant as the performance is already much better than in the few-shot setting.

TABLE V. Standard detection performance on DOTA and DIOR with GIoU and GSIoU ($\gamma = -3$ and $\kappa = 16$). Here, mAP is calculated as the average across multiple thresholds (from 0.5 to 0.95) as is the case in standard detection.

	DOTA				DIOR			
FCOS	**All**	**S**	**M**	**L**	**All**	**S**	**M**	L
GIoU	34.9	17.4	36.6	43.3	48.1	10.1	40.3	63.2
GSIoU	**36.8**	**17.5**	**40.4**	**45.2**	**49.2**	**11.0**	**41.2**	**66.1**

V. Conclusion

In this paper, we highlighted the suboptimality of IoU both for training and evaluating object detection models in images, particularly when it comes to detecting small objects. To address this issue, we proposed Scaled-adaptive Intersection over Union (SIoU), an extension of IoU that allows for better control of the behavior of this metric based on the size of the objects. SIoU can be easily parameterized to better align with human perception or to improve the training of detection methods. In the latter case, using SIoU as a cost function leads to significant performance gains on both natural and aerial images in the few-shot learning setting.

Acknowledgment

The authors would like to thank the COSE company and the Agence Nationale de la Recherche (ANR) for their financial support in the context of the LabCom IRISER project (ANR-21-LCV3-0004).

References

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[2] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[3] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9627–9636, 2019.

[4] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[6] Pierre Le Jeune and Anissa Mokraoui. Improving few-shot object detection through a performance analysis on aerial and natural images. In *Proceedings of the 30th European Signal Processing Conference (EUSIPCO)*, 2022.

[7] Pierre Le Jeune, Bissmella Bahaduri, and Anissa Mokraoui. A comparative attention framework for better few-shot object detection on aerial images. *Pattern Recognition*, 161:111243, 2025.

[8] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. UnitBox. In *Proceedings of the 24th ACM international conference on Multimedia*. ACM, oct 2016.

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[13] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018.

[14] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020.

[15] Chang Xu, Jinwang Wang, , Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS J P & RS)*, 2022.

[16] Jiabo He, Sarah Erfani, Xingjun Ma, James Bailey, Ying Chi, and Xian-Sheng Hua. Alpha-iou: A family of power intersection over union losses for bounding box regression. *Advances in Neural Information Processing Systems*, 34:20230–20242, 2021.

[17] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[18] Pierre Le Jeune and Anissa Mokraoui. A comparative attention framework for better few-shot object detection on aerial images. *arXiv*, 2022.