RefineLoc: Iterative Refinement for Weakly-Supervised Action Localization (Supplementary Material)

Alejandro Pardo[1*] Humam Alwassel[1*] Fabian Caba Heilbron[2] Ali Thabet[1] Bernard Ghanem[1]
[1]King Abdullah University of Science and Technology (KAUST) [2]Adobe Research
{alejandro.pardo,humam.alwassel,ali.thabet,bernard.ghanem}@kaust.edu.sa caba@adobe.com
http://humamalwassel.com/publication/refineloc

A. Additional Ablation Study

Here, we include the same ablation study presented in the main paper (Subsection 4.3) for three additional settings: ActivityNet v1.2 [2] using TSN [5] features and THUMOS14 [4] using TSN and I3D [3] features.

Effects of the Pseudo Ground Truth Generator and the Loss Trade-off Coefficient β. Tables 1a, 1c, and 1b summarize the best performance for the five generators and for five different β values on ActivityNet v1.2 using TSN, THUMOS14 using I3D, and THUMOS14 using TSN, respectively. The Segment Prediction-Based Generator consistently gives the best performance gain compared to the other generators in all settings.

(a) ActivityNet v1.2 with TSN features

Pseudo Ground Truth Generator	β = 0	1	2	4	8	16
Uniform Random	13.15	13.15	13.15	13.15	13.15	13.15
Distribution Aware	13.15	15.27	19.22	18.76	20.80	20.96
Class Activation	13.15	22.95	22.90	22.53	22.55	22.23
Attention	13.15	**23.15**	22.90	22.47	22.57	22.36
Segment Prediction	13.15	23.09	**23.16**	**22.98**	**23.02**	**22.88**

(b) THUMOS14 with I3D features

Pseudo Ground Truth Generator	β = 0	1	2	4	8	16
Uniform Random	19.45	21.12	20.20	19.78	19.45	19.45
Distribution Aware	19.45	20.69	20.32	19.45	19.45	19.45
Class Activation	19.45	20.18	20.11	20.10	20.21	20.34
Attention	19.45	19.45	19.45	19.45	19.45	19.45
Segment Prediction	19.45	**21.48**	**22.60**	**21.55**	**20.85**	**21.09**

(c) THUMOS14 with TSN features

Pseudo Ground Truth Generator	β = 0	1	2	4	8	16
Uniform Random	2.90	17.97	17.64	18.60	18.96	16.64
Distribution Aware	2.90	14.89	14.73	14.90	16.42	14.50
Class Activation	2.90	12.12	12.32	13.66	12.98	13.28
Attention	2.90	20.70	21.37	21.10	20.64	19.66
Segment Prediction	2.90	**20.92**	**21.87**	**22.63**	**21.13**	**20.64**

Table 1: **Effects of pseudo ground truth generator and loss trade-off coefficient β.** The metric is average mAP@tIoU=0.5:0.05:0.95 for ActivityNet v1.2 and mAP@tIoU$=0.5$ tIoU for THUMOS14. Bold represent the best generator for each β.

*indicates equal contribution.

(a) ActivityNet v1.2 using TSN features

Refinement Iteration	0	1	2	3	4	5
RefineLoc	13.27	21.62	22.76	23.09	22.68	**23.23**

(b) THUMOS14 using I3D features

Refinement Iteration	0	3	6	9	12	14
RefineLoc	19.45	20.96	21.36	22.46	21.87	**23.12**

(c) THUMOS14 using TSN features

Refinement Iteration	0	1	2	3	4	5
RefineLoc	2.90	11.13	18.73	20.60	**22.63**	20.12

Table 2: **Performance over refinement iterations.** The reported metric is average mAP@tIoU=0.5:0.05:0.95 for ActivityNet v1.2 and mAP@tIoU$=0.5$ tIoU for THUMOS14. We observe that even with a weak base model, our method has the capability to improve the performance over iterations.

Performance over Refinement Iterations. Tables 2a, 2b, and 2c show the evolution of RefineLoc's performance across refinement iterations on ActivityNet v1.2 using TSN, THUMOS14 using I3D, and THUMOS14 using TSN. In each setting, we consistently observe a significant performance increase over our baseline model $\mathcal{M}_0$ (iteration 0 in each table).

Diagnosing Detection Results. To further analyze the merits of the proposed refinement strategy, we conduct a DETAD [1] false positive analysis of RefineLoc on ActivityNet v1.2 and THUMOS14 using I3D and TSN (Figures 1a, 1b, 1d, and 1c). The false-positive profile analysis provides a fine-grained categorization of false-positive errors and summarizes the distribution of these errors over the top $5G$ model predictions, where G is the number of ground truth segments in the dataset. After refinement (right plot in each figure), we observe that RefineLoc generates more high-scoring true positive predictions (towards $1G$) and reduces background and localization errors. The DETAD results indicate that our iterative refinement encourages tighter temporal predictions, which we argue does occur primarily because of the snippet-level supervision injected in the form of pseudo ground truth.

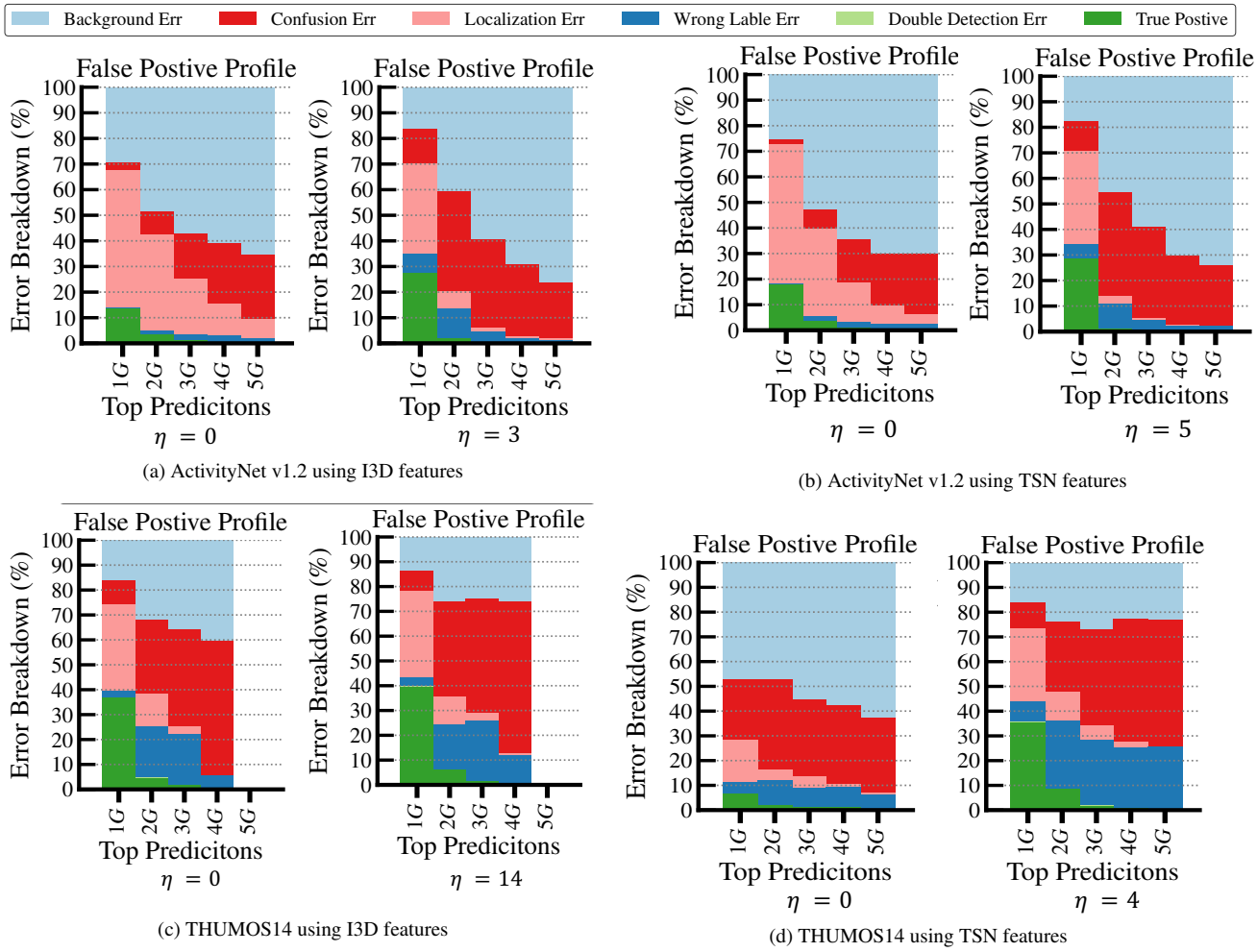
B. Logistic Regression vs Cross-Entropy

RefineLoc learns two values for the attention, instead of learning one single scalar. The motivation behind this design choice is to learn explicitly one value for background attention and one value for foreground attention. Besides, learning these two values trough a classification loss (*i.e*. cross-entropy) is an easier problem than learning one value through a regression loss (*i.e*. logistic regression). For ActivityNet, we found that our initial hypothesis is true. Indeed, when we learn only one scalar for attention, RefineLoc obtains only 22.2% average mAP using I3D features, a 1% drop in average mAP compared to the results obtained with cross-entropy. In contrast, the best result on THUMOS14 is obtained by learning only one scalar value. When learning two values for attention with cross-entropy, our model obtains only 19.95% mAP at tIoU 0.5.

B.1. Qualitative Results

ActivityNet v1.2. Figure 2 shows some RefineLoc qualitative detection results on ActivityNet. We present results across different refinement iterations. The top video shows our method not only enhances its coverage over iterations, but it is also able to detect a new instance at iteration 1 that was missed in the previous iteration. In the middle video, we see how RefineLoc manages to successfully merge different predictions over iterations. We also see erroneous predictions being cut off from iteration to iteration. The final example shows a failure case. Despite the starting point at iteration 0, our predictions diverge in later steps. We believe this confusion comes from the heavy context around the actions.

THUMOS14. Figure 3 showcases RefineLoc qualitative results from the THUMOS14 dataset. We present results for three different videos over mulitple refinement iterations. The top video shows our method not only enhances its coverage over iterations, but it is also able to detect a new instance at iteration 1 that was missed in the previous iteration. In the middle video, we see how RefineLoc manages to successfully cut off erroneous predictions from iteration to iteration. The final example shows a failure case. Despite starting with decent predictions at iteration 0, our predictions do not improve in subsequent steps. We believe this confusion comes from the heavy context around the actions.



(a) ActivityNet v1.2 using I3D features

(b) ActivityNet v1.2 using TSN features

(c) THUMOS14 using I3D features

(d) THUMOS14 using TSN features

Figure 1: **Diagnosing Detection Results.** We present DETAD [1] false positive profiles of RefineLoc at refinement iterations $\eta = 0$ (left) and $\eta =$ convergence (right). G represents the number of ground truth segments available in the dataset. Please refer to the DETAD paper [1] for the complete definition of each error type in the false positive profile.

References

[1] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *ECCV*, 2018.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[4] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 2017.

[5] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

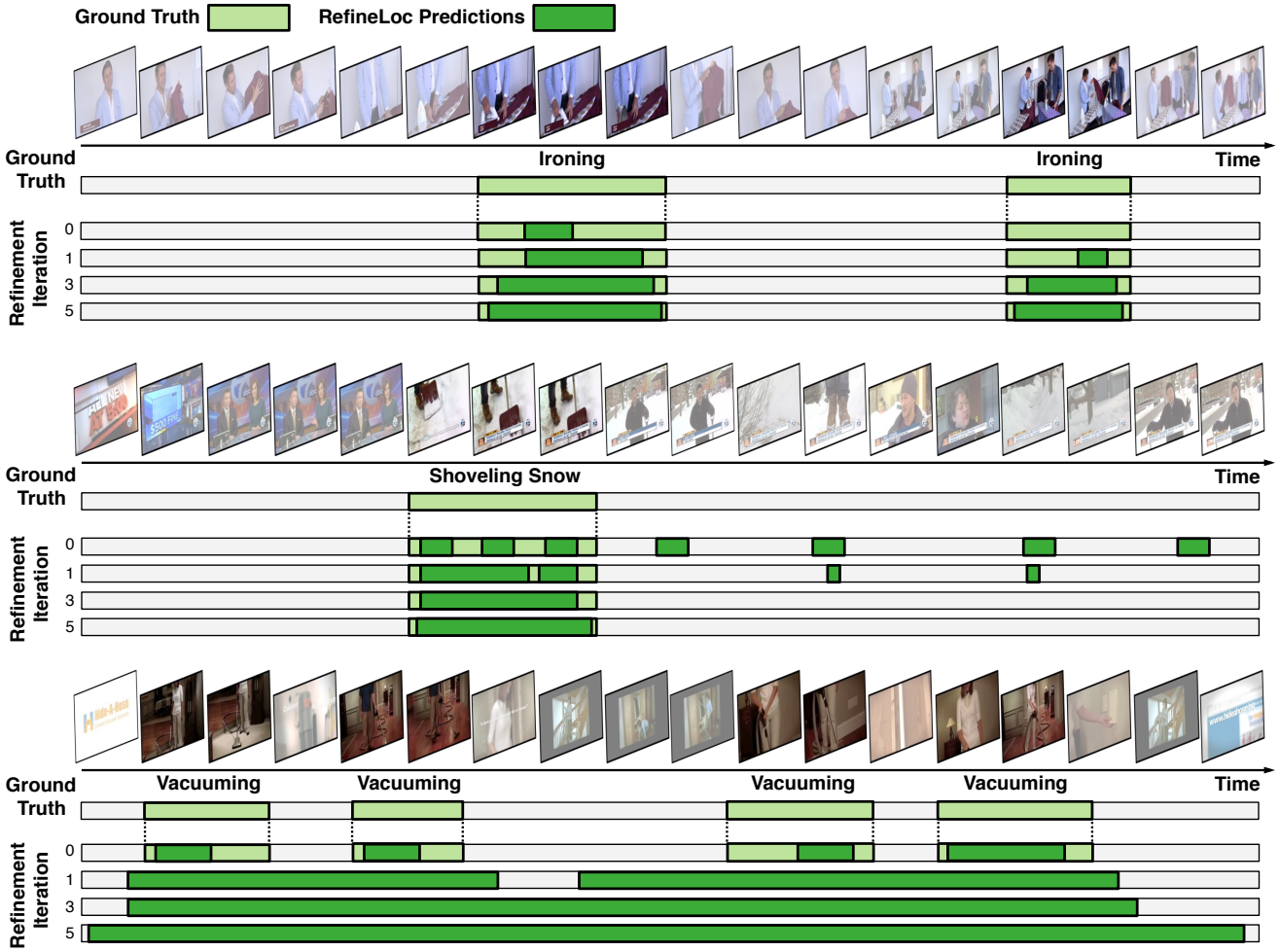


Figure 2: **Qualitative Results (ActivityNet v1.2).** *Top*: RefineLoc successfully enhances prediction coverage and detects missed instances as iterations evolve. *Middle*: RefineLoc manages to merge disjoint predictions and remove wrong background predictions from one iteration to the next. *Bottom*: In the presence of large context, iterative refinement can hurt RefineLoc predictions, as visual similarity between foreground and background confuses our attention model.

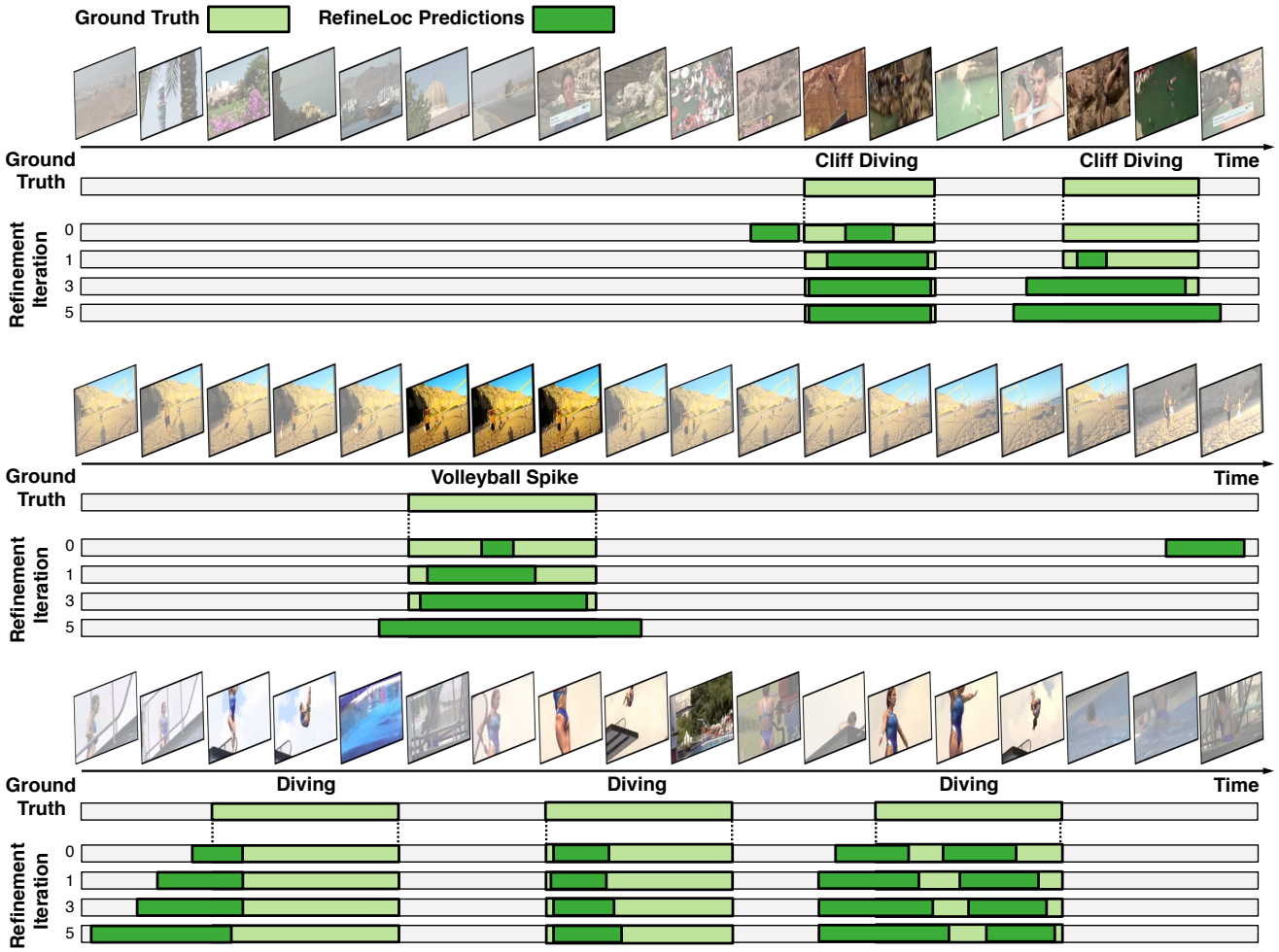


Figure 3: **Qualitative Results (THUMOS14).** *Top and Middle*: RefineLoc successfully enhances prediction coverage over iterations and is able to detect missed instances as iterations evolve. *Bottom*: In the presence of large context, iterative refinement can hurt RefineLoc predictions, as visual similarity between foreground and background confuses our attention model.