We would like to thank the reviewers for their comments. In the following, we address the points that were raised.

Analysis of Failure Cases (R2/R3). *(R2) Failure cases are missing. As the event camera only records intensity changes, the proposed method is mainly targeting moving objects. It will be better to highlight 'moving objects' in the title or somewhere else. (R3) It is better to discuss the limitations of the proposed method in extreme scenarios, such as the scenario with a large number of overlapping objects.* We agree with the reviewers that a more detailed discussion of failure cases will add insights into the robustness of the proposed approach. We will be happy to add such a section to the revised version of our submission. However, we would like to stress that our method is not limited to moving objects. As shown in Fig. 4, the memory mechanism introduced by the recurrent layers of our model, is able to consistently detect objects even when their motion does not generate events anymore.

Technical Novelty (R1/R4). *(R1) The authors borrow the convolutional architecture from 2D object detection, temporal loss/module from event analysis community.* Although the core blocks of our model (ConvLSTM, SSD head) are well-known layers from the frame-based community, the architecture we propose is original and designed to work specifically for event data. Our design choices make our method efficient enough to process high resolution event cameras and accurate enough to reach comparable accuracy as frame-based detectors. Moreover, even if we find inspiration in works such as [43, 44], to the best of our knowledge, our work is the first to introduce a double regression head and a temporal consistency loss, well adapted to the characteristics of event data. Finally, as also pointed by R3, we believe that the automated labeling protocol is an additional technical contribution that allows fast creation of large event-based datasets. We hope this will further extend the impact of our work for the community.

Details on Labeling Protocol (R1/R4). Due to space limitations, many details about the labeling protocol have been described only in the supplementary material. We will be happy to add them to the main body of the revised article. In particular, we will discuss in more depth quality assessment and failure cases.

Dataset and Code Release (R3/R4). We confirm that the dataset will be released upon acceptance of the paper. Moreover, we will be glad to also release the evaluation code together with the dataset. However, to date, we can not commit on the release of the training code, since this requires further internal discussions.

Additional comments. *(R2) more details about Events-RetinaNet. Based on the current version, it seems like use RetinaNet on the integrated event frame.* Events-RetinaNet uses Event Volumes as input, which are the input representation leading to the best results in our experiments and are also the same input used by our RED architecture.

(R2) Fig. 2, $h_{0,k+1}$ or $h_{0,k-1}$ at the bottom? Yes, there is indeed a typo in the figure caption. We thank the reviewer for pointing it out.

(R3) comparison of the amount of data [...] in terms of data storage efficiency. We report in Tab. 2 the amount of networks parameters. This shows that our network is smaller than Gray-RetinaNet. In terms of input data storage, our method does not take advantage of the input data sparsity yet. We briefly discuss this in Sec. 6. We will extend this analysis in the revised version of our submission.

(R3) process data from DAVIS and other event cameras to prove the compatibility. Indeed, our approach is not limited to a particular type of event camera and can be applied to any event sensor. However, since there is no available DAVIS dataset with object detection labels, we could not train our model on a DAVIS camera. Nevertheless, following the reviewer remark, we run some qualitative tests by applying a model trained on the ATIS dataset (which has similar resolution as the DAVIS) on the MVSEC dataset (DAVIS-346) [32]. Even if the model was trained on ATIS data, we observe that it generalizes well also on the DAVIS dataset. We will be happy to add these results in a revised version of the paper.

(R4) Have you considered using an event queue per pixel (Tulyakov et al.) [...] rather than training some RNN-like architecture. Our method is agnostic to the input representation and could be indeed combined with Tulyakov et al. We will add this remark and this reference to our submission. However, in our work, we decided to introduce the memory mechanism in the deeper layers of the network, rather than at the pixel level. The reason for this is that it would be too expensive to have per pixel memory states, especially for a 1Mpx event camera. Moreover, in this way, we can learn memory states corresponding to high-level object features, which vary slowly and allow smoother detections over time.

(R4) "Gray-RetinaNet" without the resolution and color alterations. In our experiments, we used gray-level images at the same resolution as the event camera. This was done to remove as much as possible a performance bias due to information which is not available to the event camera. However, we agree that a comparison with a typical frame-based camera is also interesting on its own. From some preliminary tests, it seems that color increase mAP by a few percentage points. We will add and clarify this aspect in a revised version of the paper.