Point-to-point Responses to Reviewers' Comments

Reviewer #1

Dear Lyu et al.,

This manuscript highlights a novel technique/device for the measurement of hail size and frequency, which are known to be difficult problems with limited robust solutions. The results show that the created device has impressive skill for both tasks. In particular, relative errors of only a few percentage points when measuring diameter (Figure 4) has the potential to be highly beneficial when populating hail size distributions, especially when considering many competing methods (such as human reporting) bin hail diameter into coarse categories of for example 5 mm. Overall, I am excited for what opportunities this device could bring to the atmospheric science community. Well done.

Response: We sincerely appreciate the reviewer's critical and insightful questions about model description and hail size distributions. We have revised manuscript according to these comments with point-to-point responses listed below.

I do have a few comments/suggestions that may help improve the overall quality of the manuscript. I have 2 main comments given below, along with a brief section for standard editing recommendations such as spelling or grammar etc.

Comment #1: This is perhaps my most pressing recommendation. It appears that section 2.2 titled "Deep Learning–Based Hailstone Segmentation and Characterization" could use some expansion or refinement. In the interest of EGU AMT review aspect #6 (Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists?), I would prefer to see more details/explanations on the deep learning architecture and feature engineering process. In particular, section 2.2.1 could use some thought.

Firstly, the description given in the paragraph between lines 110 – 124 and the contents of Figure 2 leave me confused about exactly the sequence of ML architectures I would have to deploy if I were to try and reproduce this work. Lines 112 through 116 imply the ConvNeXt-Tiny backbone's output is what is fed into the Feature Pyramid Network (FPN). However, the hatched rectangle in Figure 2 seems to imply the ConvNeXt-Tiny makes up the at least the downsampling portion of the FPN. It appears there is extensive overlap between these two architectures (rather than a sequence of two independent architectures), but this section can make this unclear. I would recommend more written descriptions clarifying this and perhaps a rework of the figure itself. For the figure, a legend may help, along with text explaining exactly where the ConvNeXt-Tiny comes into play, similar to the text given for

other features such as the FPN.

Response: We agree that the original description and schematic may have led to confusion regarding the model architecture. In the revised manuscript, we have clarified that ConvNeXt-Tiny is used strictly as the backbone feature extractor, while the FPN is a separate module that receives multi-scale feature maps extracted from ConvNeXt-Tiny. Specifically, ConvNeXt-Tiny generates hierarchical feature maps from multiple stages. These multi-scale feature maps are then passed as inputs into the FPN. The FPN performs lateral connections and top-down fusion to construct the final multi-scale feature representation used for downstream tasks.

We have revised Section 2.2.1 to explicitly describe this sequential relationship and avoid any overlap interpretation. In addition, we have updated Figure 2 by clearly separating the backbone (ConvNeXt-Tiny) and FPN blocks, adding arrows indicating feature map flow and including a legend defining each module and color coding. The revised texts are shown below, "*Second, the Feature Pyramid Network (FPN) is applied to the outputs of the ConvNeXt backbone. The FPN constructs a pyramid of feature maps by top-down lateral connections. This mechanism allows the fusion of strong semantic information from deeper layers with high-resolution spatial details from shallower layers. Specifically, each feature map* $\boldsymbol{C_i}$ *is first transformed using a* 1×1 *convolution to obtain a uniform channel embedding* $L_i = Conv_{1\times1}(C_i)$. *top-down pathway then progressively upsamples higher-level features and fuses them with corresponding lateral features via element-wise addition* $T_i = L_i + Upsample(T_{i+1})$. *A* 3×3 *convolution is subsequently applied to each fused feature map to suppress aliasing and refine spatial consistency, yielding the final pyramid features* $P_i = Conv_{3\times3}(T_i)$ *.The resulting feature pyramid* $\{P_1, P_2, P_3, P_4\}$ *provides strong multi-scale representations for both small and large hailstones.*"

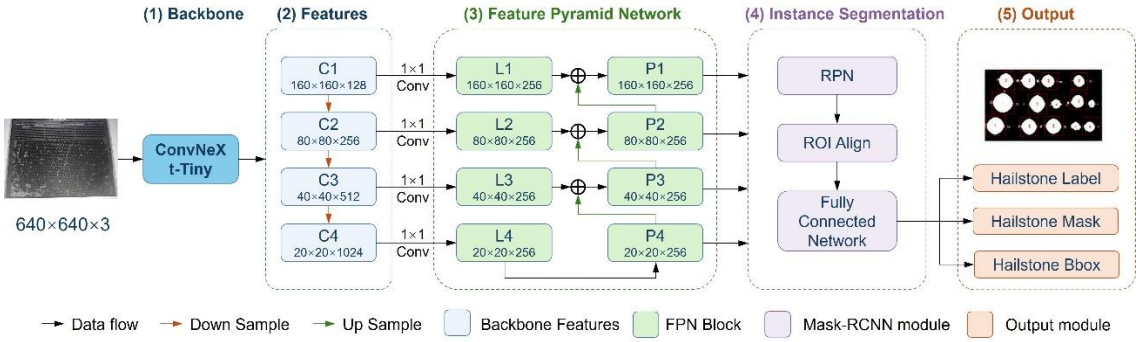

Figure 1: work flow of model detection algorithm.

Secondly, the instance segmentation performed by a R-CNN could be more clearly defined in Figure 2. I suspect this architecture runs on the fused feature map (yellow rectangle) and before the start of the 3 final models in the rightmost section of the figure, but this is not clear. It seems the text for this topic (ex: lines 119/120) is clearer, but the figure could use a legend/text for

this as well.

Response: We appreciate this observation and agree that the original figure lacked sufficient clarity.In the revised version, we now explicitly clarify related descriptions and Figures. The instance segmentation branch (Mask R-CNN -based module) operates on the fused feature map produced by the FPN. This module is applied after feature fusion and before downstream task-specific heads (hail size estimation, frequency counting, etc.). It produces instance-level hailstone masks, which are then used for subsequent quantitative characterization.

We have updated Figure 2 by adding a dedicated labeled block for the R-CNN segmentation branch, showing its input as the FPN-fused feature map and its output feeds into downstream processing modules. We also revised the text in Section 2.2.1 to explicitly state this processing order as follows,

"*Finally, the Mask R-CNN–style head operates on this fused feature pyramid to perform instance segmentation. The FPN outputs serve as inputs to a Mask R-CNN-based instance segmentation module. A region proposal network (RPN) generates candidate hailstone regions from the fused feature maps. For each proposal, RoI Align is applied to extract fixed-size region features while preserving spatial alignment. For each candidate region, the network branches into three parallel outputs: (1) a classifier to confirm the presence of hail, (2) a regressor to refine the bounding box coordinates, and (3) a Fully Convolutional Network (FCN) mask branch that predicts a binary segmentation mask for the hailstone at pixel resolution. This allows the system to distinguish individual hailstones even in scenarios of slight overlap, providing precise geometric boundaries for diameter calculation.*"

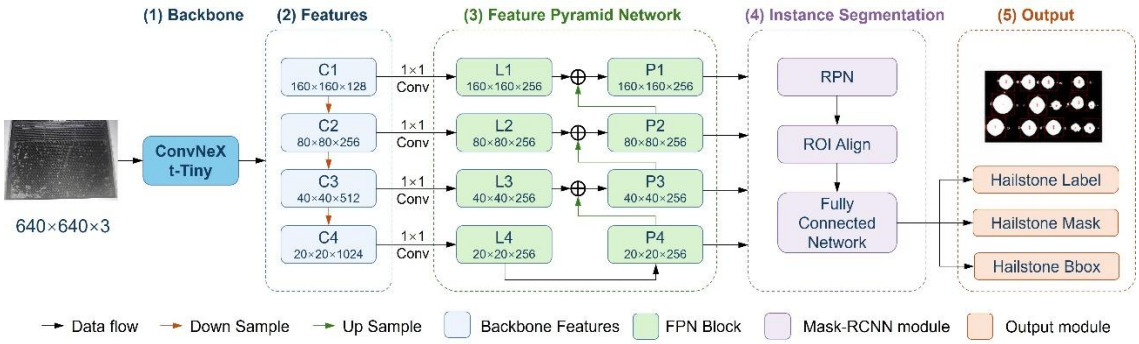

Figure 2: work flow of model detection algorithm.

Finally, (and most critically) I would like to see in section 2.2.1 more discussion on dataset organization (training/validation/testing splits) and the chosen hyperparameters used in the various ML architectures. Starting in line 129, the authors have noted 8742 images in a training dataset. However, there is no discussion on subsequent validation or testing datasets needed for

hyperparameter searches and result verification, respectively. It is not clear if the 8742 images were sliced into each of these splits, or if other images were used. Additionally, clarification on the validation split can give more confidence to readers that the optimal hyperparameters were discovered while clarification on the testing split would offer additional assurance to the integrity of any ML results.

Response: We thank the reviewer for identifying this important omission. To improve reproducibility, we have now added a detailed description of dataset partitioning. These details have been added to Section 2.2.2 to ensure compliance with AMT reproducibility standards as following, "*Specifically, the full dataset of 8,742 images is split into training set: 70% (6,119 images), validation set: 15% (1,311 images) and testing set: 15% (1,312 images). The split is performed at the image level with random stratification to ensure balanced hail size and density distributions across subsets. The validation set is used exclusively for early stopping, while the test set is strictly held out for final performance evaluation.*".

With regard to hyperparameters, some seem to be discussed in lines 133 to 136, but it is not clear if this is an exhaustive list for the various architectures. If it is exhaustive, something to note this would be helpful. If a hyperparameter search was performed, discussion on what the search spaces were would be helpful for reproducibility. Perhaps a table on all of the chosen hyperparameters and their spaces would solve this all at once, even if in supplemental material or appendices etc.

Response: We agree that the original manuscript did not sufficiently clarify the hyperparameter optimization procedure. We have now expanded this section to improve transparency. In the revised manuscript, we explicitly state that commonly used hyperparameters are used in this study. We have added a new table (Table 1 in main text) summarizing hyperparameters applied in this study.

"*The model hyperparameters are shown in the following Table 1. To enhance robustness, we applied photometric augmentations (e.g., random brightness/contrast shifts) and geometric transformations (rotation, scaling, elastic deformation) during training. The model was optimized using the AdamW optimizer with a cosine-annealed learning rate (initial = 3×10^{-4}) and a composite loss function combining focal loss for classification, smooth L1 loss for box regression, and Dice loss for mask refinement. Crucially, the deep learning approach eliminates the need for manual tuning of segmentation thresholds, adapts dynamically to changing environmental conditions, and achieves a high particle detection accuracy.*"

Table 1 The hyperparameters in our model.

Hyperparameter	*Ours*
Data augmentations	*photometric augmentations, geometric transformations*
Batch size	*4*
FPN channel size	*256*
Optimizer	*AdamW*
Scheduler	*Cosine Annealing,* $\eta_{max} = 3\times10^{-4}$, $\eta_{min} = 0$, $T_{max} = 50$
Training epochs	*up to 100 with early stopping patience = 10*
Weight decay	1×10^{-3}

Comment #2: The results highlighted in figure 8 (also lines 299 to 308) showing the differences between the disdrometer and HailCam hail size distributions (HSDs) seem to have a lot of potential important implications. A widely accepted mathematical estimate for HSDs is the gamma distribution (for example see doi: https://doi.org/10.1175/1520-0469(1987)044<1062:ATPROT>2.0.CO;2), so both systems showing a decreasing exponential is curious. Some further discussion noting the lack of this distribution may be warranted, even if it is just to acknowledge that the sample size of 2 events may not be large enough to see it.

It is also curious that although both systems show a negative exponential (and no statistically significant differences), the disdrometer has an increase in probabilities around the 8 mm mark, which is a characteristic of a gamma distribution. Is there perhaps something more happening? Such as the HailCam's under sizing bias coming into play or melting effects in the 60 second interval the HailCam processes in (perhaps hinted at in line 381 of conclusions)? I note that the authors included lots of discussion highlighting that the differences between HailCam and the disdrometer may be associated with mechanical differences between the devices, however, a few extra more detailed sentences acknowledging the potential for further biases could be helpful, especially when, for example, the authors state somewhat competing assertions in lines 308 to 310 and lines 313 to 316:

"This discrepancy suggests that the disdrometer's velocity-based classification algorithm may be misidentifying larger raindrops or graupel particles as hail, whereas HailCam's imaging-based segmentation strictly enforces morphological criteria for solid-phase identification."

"Despite the visual divergence in distribution shapes, particularly the disdrometer's enhanced probabilities in the 7–12 mm range, the cumulative distribution functions are sufficiently similar that the observed differences may plausibly arise from sampling variability rather than systematic measurement bias."

Response: We fully agree with the reviewer's suggestion to more discussions on gamma distribution theory and the observed deviation from theoretical expectations. We also added

more discussions on the bias from melting effects of HailCam. The expanded explanatory text are shown as below,

"*A widely adopted theoretical framework for describing hydrometeor size spectra, including hail, is the gamma distribution, which typically features a probability density function that rises to a distinct intermediate-size peak before decaying exponentially. Neither instrument's HSD fully conforms to this canonical gamma shape across the two observed hail episodes. HailCam produces a strictly unimodal, monotonically decreasing exponential distribution with a dominant peak confined to the smallest 5–6 mm bin and negligible probabilities for diameters >12 mm, while the Parsivel² disdrometer exhibits a weak secondary probability hump near 7–9 mm but still lacks a pronounced gamma-mode peak at larger sizes. A key limiting factor for recovery of a classical gamma HSD is the limited observational sample size: our analysis draws exclusively from two short-duration nocturnal hail pulses on a single storm night, with relatively limited total hail particle counts compared to multi-event, multi-storm climatological datasets used to derive gamma distribution parameters in prior literature. Beyond limited sampling, the fixed 60-second imaging duty cycle of HailCam introduces a melting bias: hailstones collected on the sampling grid sit exposed to ambient near-freezing air for up to 60 s before automated imaging, leading to partial surface ablation that reduces measured equivalent diameter and further shifts mass and particle counts toward the smallest 5–6 mm bin.*"

This edit eliminates the perceived logical conflict the reviewer identified and strengthens the internal consistency of our interpretation of Figure 8's statistical results.

Technical Corrections:

- The sentence between lines 31 to 34 seems to be close to a run-on sentence. Perhaps reword?

Response: We have revised the manuscript accordingly as follows, "*Accurate, high-resolution observational data on hail properties, including particle size distribution, number concentration, mass flux, and temporal evolution, are essential for understanding hail microphysics and storm dynamics. They also play a critical role in validating numerical weather prediction models (Wobrock et al., 2003), improving radar-based hail detection algorithms, and supporting risk assessment in agriculture, infrastructure, and insurance sectors (Allen et al., 2020; Martius et al., 2018).*"

- Line 35 could use rewording

Response: This sentence was reworded as follows, "*Hail observations are generally obtained through two complementary approaches: remote sensing techniques and in situ ground-based*

measurements."

- Lines 98 and 99 have some grammar issues. Fix: "detected by in the image" and "HailCam evacuates tray to"

Response: This sentence was fixed as follows, "*Once hailstones are detected by the image processing module, HailCam initiates evacuation of the tray to clear the sampling area*."

- The end of line 101 is missing a period

Response: Added.

- Many words seem to have added spaces between them. Ex: "defi ned" in line 349, "o r" in line 378, "syst em" in line 380, etc. But this may be caused by printing errors in the pdf. If this is not the source of the errors, please investigate and fix.

Response: This should be printing errors. We double-checked throughout the manuscript.

Once again, great work! I look forward to the final version of the manuscript.

Reviewer #2

We sincerely appreciate the reviewer's critical, physically grounded questions regarding hail impact deformation, shattering, inter-particle collision biases, and limitations of our current manual-placement laboratory calibration protocol. We fully acknowledge that static manual placement eliminates real-world dynamic impact artifacts that occur during natural hailfall, and we will expand error discussion in Section 3.3 Uncertainty Analysis, add new text addressing impact deformation/collision effects, and clearly outline targeted follow-up dynamic drop-test experiments as recommended. Below is structured response, reconciliation of existing results, and concrete manuscript revision plans.

How much would the differing velocities of hailstones hitting the grid surface impact what is imaged? I could see a case where a hailstone hits the grid surface and shatters and/or morphs shape. This would result in slightly inaccurate diameter measurements.

Response: We agree that impact dynamics can influence the post-impact morphology of hydrometeors and therefore potentially affect image-derived size estimation. This effect is physically well recognized in impact-based precipitation measurement systems.

In the case of HailCam, several design aspects mitigate this issue. First, the grid surface is designed as a low-impact interception interface, where hailstones are decelerated prior to imaging rather than undergoing high-energy collision. The segmentation algorithm estimates diameter based on the projected morphological boundary, which remains robust for mildly

deformed particles.

Nevertheless, we acknowledge that high-impact events involving large or fast-falling hailstones may introduce deformation or partial fragmentation, which could lead to small biases in reconstructed equivalent diameter. This limitation is now explicitly stated in the revised manuscript as a potential source of measurement uncertainty as follows,

"*Another unaccounted uncertainty source in static manual calibration is morphological distortion caused by terminal-velocity hail impacts on the collection grid. Natural hailstones fall with terminal velocities ranging from ~8 m s $^{-1}$ (5 mm hail) to >25 m s $^{-1}$ (45 mm hail); high-energy vertical impacts can temporarily flatten oblate hailstones or split brittle ice particles into multiple fragments. The HailCam funnel buffer partially mitigates this by decelerating hydrometeors prior to contact with the imaging grid, and sub-5 mm shattered ice fragments drain through the grid mesh and are excluded from imaging. Even so, partial flattening of intact hailstones widens their projected horizontal silhouette, creating a small positive bias in equivalent diameter retrievals—an error mode entirely absent in our static manual placement calibration. This impact deformation bias is not quantified in the present laboratory tests and represents an uncharacterized field uncertainty term.*"

I think it would be advantageous to drop the hailstones during calibration testing as opposed to simply placing them manually. When you manually place them, you remove variability associated with the shattering, breaking, and morphing of hailstones. These are all processes that occur in real life and are likely major sources of error.

Response: We fully agree with the reviewer's criticism and acknowledge that static manual placement constitutes a controlled idealized benchmark rather than a fully realistic reproduction of natural hail impacts. We provide two layers of response: interpretation of existing lab results, plus plans for follow-up dynamic drop testing, plus revised manuscript text to contextualize calibration limitations.

The static manual tests in Section 3.1 were designed to isolate intrinsic optical and deep learning algorithm errors (camera distortion, segmentation contour loss, pixel-to-mm conversion, overlapping particle merging) independent of mechanical impact artifacts. These tests quantify the instrument's inherent optical/AI measurement floor error, which acts as a lower bound on total observational uncertainty. However, we agree that this configuration does not fully represent real hailfall dynamics. We have revised the manuscript to explicitly clarify this limitation in Section 2.3.1,

"*Static manual placement of artificial ice and foam spheres was selected to isolate optical imaging and instance segmentation errors without confounding impact fragmentation or*

deformation effects. These laboratory results define the minimum achievable measurement uncertainty under idealized particle conditions; total field uncertainty will incorporate additional error contributions from high-speed hail impacts, shape distortion, and fragmentation unquantified in this static framework."

We also now explicitly state that dynamic drop-based calibration constitutes an important next step for future system validation, particularly to quantify impact-induced deformation, as follows,

"*Besides, to address the limitations of static manual calibration, a follow-up dynamic drop-test campaign would be helpful using a vertical fall tower capable of reproducing terminal fall velocities for 5–45 mm synthetic ice spheres, in order to develop a quantitative correction model for impact-induced sizing and counting biases to be integrated into the HailCam post-processing pipeline in future firmware updates.*"

Additionally, if images are taken every minute and the hailstones are then funneled out, there is potential for hailstones to collide with and impact one another. The calibration was performed entirely with manual placement, so I am struggling to see how collisions are accounted for in potential sizing errors.

Response: We agree that particle interactions can occur in dense hailfall conditions and may introduce secondary uncertainties in measurement systems relying on collection and imaging.

In the HailCam system, it reduces the likelihood and impact of such interactions that the system does not rely on long-term storage of accumulated hailstones within the imaging region. Even though 60-second imaging duty cycle was used in the filed test, it could be down to 20-second after revise configuration, which would further reduce the chances of particle collision. However, we acknowledge that under high-intensity hailfall conditions, transient overlaps or physical contact between particles may occur. Such events may lead to partial occlusion in images and overestimation or underestimation of individual particle size. We revised manuscript now explicitly includes this as a rare but non-negligible source of uncertainty, particularly during peak hail intensity periods.

"*It also should be noted that over the 60-second accumulation window prior to imaging, newly arriving hailstones collide with particles already resting on the collection grid. These collisions can displace settled hailstones, chip small fragments off particle edges, or press adjacent hailstones into tighter overlapping clusters. Dense overlapping from collision aggregation worsens Mask R-CNN instance merging, exacerbating the undercounting bias observed at high particle densities (Section 3.1.2). The instrument supports flexible configuration of sampling intervals ranging from a maximum of 60 s down to a minimum of 20 s. Shortening the imaging*

cycle to 20 s drastically cuts the total volume of hailstones accumulated on the tray within each capture window, which markedly reduces the frequency of sequential inter-particle collisions and particle stacking during accumulation."

Overall, I think more validation and testing are needed. It seems like the approach was to select the best-case scenario and proceed with the results that appeared most favorable. I would definitely recommend testing with synthetic hailstones being dropped rather than manually placed.

Response: We acknowledge the reviewer's overarching assessment that our current laboratory calibration framework isolates idealized static conditions and excludes key dynamic error modes (impact deformation, fragmentation, sequential inter-particle collision) present in real hailfall. We do not intend to present static lab results as fully representative of all field error sources—this misinterpretation arises from insufficient explicit discussion of calibration limitations in the original manuscript, which we will fully rectify via the following coordinated revisions as shown above, including:

(1) In Section 2.3.1 Laboratory Calibration, add clear text distinguishing static idealized testing (baseline optical/AI error) from untested dynamic impact/collision error sources;

(2) Expand Section 3.3 Uncertainty Quantification to separately discuss impact deformation bias, fragmentation bias, and sequential inter-particle collision bias, explicitly noting none are constrained by existing static calibration;

(3) Add a dedicated future work description outlining terminal-velocity dynamic drop tower testing as the required complementary validation campaign to address the reviewer's concerns;

We emphasize that while static calibration provides rigorous quantification of the instrument's core optical and deep learning segmentation performance, we fully recognize dynamic impact and collision artifacts constitute uncharacterized additional uncertainty terms that require the drop-test validation the reviewer recommends, and we commit to carrying out these experiments to advance the robustness of HailCam measurements in future work.