

We thank the reviewers for careful reading of the manuscript and much appreciate the detailed feedback. Please find out point-to-point answers below.

Reviewer 2:

The method description of Johnson et al. (2024) has been rewritten and is hopefully correct now (introduction lines 53-60):

“To avoid the massive manual labelling effort required for supervised learning, methods such as semi-supervised or self-supervised learning are being developed (Rani et al., 2023). In the recent study, Johnson et al. (2024) used a contrastive learning objective in analysing THEMIS all-sky images. First, latent representations of image contents were learned in an unsupervised fashion from roughly 910,000 images. In the second stage, the authors constructed a classifier to provide class labels to the latent representations. This was carried out by using supervised learning and (only) roughly 6,000 labelled images. The authors demonstrated the power of the methodology by analysing about 700 million THEMIS images taken at the auroral oval latitudes. This gave statistically meaningful results: auroral occurrence centred around magnetic midnight and arcs relating to milder geomagnetic activity than discrete and diffuse aurora.”

—

Reviewer 1:

Currently the readability of particularly section 3 is not sufficient.

Very true. This section is what we have struggled most with as the process itself has taken many years with many iterations making it challenging to document in a sensible way. We have done our best to improve the text of this section in the revised version. Comments related to the flow have been taken into account. The track-changed manuscript will only show the major additions and changes. In addition, there are small phrasing edits throughout the manuscript, which would have been too much to display.

Location of KHO: please add latitudes (and longitude)

At the beginning of Section 2, where KHO is introduced, we give the geographic coordinates and the geomagnetic latitude for the station.

Wording: "carefully construct" => Although the author who has long been worked on aurora image is one of the most reliable scientists (in terms of carefulness") to construct the "ground true" dataset, the word "careful" is unfortunately a subjective word. So, it is better to use "objective" numbers of iteration and numbers of year that the author has classified aurora. Similarly, best => best possible

Good point. These wordings will be changed to “construct through a number of iterations”, and “the best-scoring classifier”

Section 1

Solar cycle dependence: At Svalbard which is located far north of nightside aurora, no SSN dependence is reasonable. For nightside aurora, average size of the aurora oval (this depends on solar cycle) changes the probability of diffuse/discrete aurora at one fixed

station. This note should be already mentioned in the introduction and discussion.

Svalbard is located at the poleward boundary of the auroral oval, so as the size of the oval correlates with the solar activity, it would be reasonable to expect an anti-correlation of auroral occurrence with the solar activity, provided that other time sectors would not be strongly affected by the solar activity. However, this is not seen in our results. True enough, we are not separating diffuse or discrete aurora in this study, but particularly at the poleward boundary these two types separate less than at the equatorward boundary. There are little evidence for any of these things in the literature, which motivates our study. We have added a comment on the lack of solar cycle evolution of high-latitude aurora in the revised introduction (lines 31-33):

“the auroral occurrence evolution over a solar cycle time scale has not been systematically investigated beyond the studies by Nevanlinna and Pulkkinen (2001) and Pulkkinen et al. (2011).”

Section 3.2

"aurora expert": Did only Partamies checked manually, or any other too. Since Partamies is known as expert (readers can guess from reference list), I recommend to explicitly write as "The author, as the aurora expert"

Yes, that is a fair point and will be changed accordingly.

In example in Fig 2, only last column shows the moon case, and in Fig A1-A4, one two moon cases are found. When I looked at many aurora images for judging, the moon appears as often cloud because moons at second and third quarters are high during winter (dark month) at high latitude. How often is moon included in the training set and ML labelled result? I do not require exact number (not asking authors to classify all samples in Table 2, but just rough filling)

The moon does appear in the training set of all classes. To estimate how often that happens we define twilight as the time when the Sun is less than 12 degrees below the horizon. Then roughly 15% of the ground truth images represent twilight conditions. The Moon is above the horizon in roughly 31% of images and a combination of twilight and Moon in 42% of images. Here we consider only the times when the Moon is more than 50% illuminated. In the revised version of the manuscript, we state the presence of the Moon and twilight in the training set in connection to Figure 5 (line 220):

“all classes of the ground truth include about 15% of twilight images and images containing the Moon (about 30%).”

Line 151: "chose random images": from which years? Or is it from 2019-2020 unclear cases that was classified in the first manual labelling? Since this is not clear, I could not understand how to read Table 2. Is explanation in Section 4.1 (224-234) for Table 2?

For the two first rounds of manual labelling, the images were from 2019 and 2020. These correspond to the first two rows of Table 2. For all other iterations, random images were chosen from the entire dataset. We have add the years for all the rows in the revised Table 2 for clarity. The text on lines 224-234 (submitted version) describes the classification process and sets the scene for what we think is achievable. The labelled dataset mentioned there refers to the training data for the final KHOnet2026, which is an outcome of the iterations listed at the bottom row in Table 2.

Table 2 and 3 "title" (technical issue): Please change paragraph after the title (Unlike Figures, Tables do not have caption sentences). If you need caption, this should be moved to the footnote of table (with *1, *2, *3,, marking)

According to the submission guidelines at <https://www.annales-geophysicae.net/submission.html#manuscriptcomposition> , tables need concise but descriptive captions. We have streamlined the table captions for the revised version to be more concise.

Table 2

Also, where is "parentheses"? Is images for shaded part are all originally from 5000 images in the third row? How did you define ground truth? (classifier's result of which samples, and manually checked one last time?)

Apologies for the “parentheses”, that is a remnant from an earlier version and has been removed.

The images in the shaded cells of Table 2 are not from the previous set of 4x5000 random images. There is a new classification run between every iteration, meaning that the 5000 images per class were visually re-labelled, only the unambiguous images were added into the training set of the next classification round. Once the classifier was trained on the unambiguous dataset, the whole image dataset was classified with it, and another 2000 random images were chosen from each class for visual check and re-labelling. Therefore also the last visual check included 2000 random images from each of the four classes.

This part of the text has been reworked in the revised version (lines 160-161): “Each re-labelling round (each row) was followed by re-training of the classifier and a new selection of random images from each class for the next round of manual evaluation.”

Line 179 and Table 3: "2000 random images per class": If each category has 2000 samples, why Table 3 summation is not 2000? Could you explain more.

This is where we got ourselves tangled multiple times as well. The full training set for the final version of the classifier includes all the images counted in the shaded cells of Table 2. We have gathered all unambiguous examples for the four classes throughout the iterations to get a large enough of training dataset that represents the different sky conditions as well as possible. These are the ground truth numbers on the bottom row per class. From these images we used 70% for training and 30% for validation, and the numbers in Table 3 correspond to the validation part of the dataset.

This is mentioned in the connection to Table 3 (line 175), but we will try to rephrase it in a clearer way for the revised version (lines 158-159): “The bold numbers are the numbers of images, which together make the ground truth (bottom row). From the ground truth images, 70% was used for training and 30% for validating of the final KHOnet2026 classifier.”

Line 183-187: Unclear and incomplete sentences. Please fix. Errors are also found from section 4.

Incomplete sentences have been fixed and much of these text passages have been rephrased for clarity in the revised version. Note that the minor edits are not marked in the track-changed version because it would be unreadable.

It would also be better to switch the order of explanation (cloudy<->clear confusion is not surprising and better to explain first).

If this comment relates to the text on lines 182-187 in the submitted version of the manuscript, it roughly follows the order of the classes presented in the confusion matrix and mostly talks about cloudy versus clear conditions. We would therefore keep the order of the items.

Figure 4: Although this is a good example of "reason for wrong classification", I was bit surprised that you run your classification even for twilight (noon) case in the figure. The classification scheme drastically improves if you remove the twilight (SZA = 90-102 degree or 90-96 degree) cases. Could you explain why you include twilight cases?

The aim of the classification is to be robustly used on all our data, not just certain times of the day or certain seasons but on everything that we have images for. We recognise that there are conditions where the classifier fails, but these are the cases where human experts cannot tell the ground truth either, and that is what we need to accept as a classification error. This is admittedly a long-term study focussed view, but it seems artificial to us to start excluding data, which do not behave well.

Figure 4 actually shows that the classifier works surprisingly well in twilight conditions too, as the pre-noon part of the day is more cloudy and the post-noon part clear. Probabilities for the cloudy and clear classes ramp down and up respectively. If one wanted to do a temporal smoothing of the class probabilities as a post-processing step, it would give a very reasonable timeline. Here we merely just show the "raw data" as an example.

We have added an item into the discussion section related to the twilight/moon bias as well as other biases.

Line 194 "exposure time": The aurora judging scheme depends on exposure time. Isn't the exposure time just before the abrupt appearance of aurora a good value to stop operation?

Actually, often after the change of the exposure time, the image content becomes much easier to see as the contrast between aurora and the background increases. This is why the exposure times were changed in the first place. In reality, our ability (as humans or as ML classifiers) to detect aurora depends very much on the cloudiness level and the type of clouds, colour of the background sky and the level of auroral activity. There is therefore no one simple and obvious threshold or indicator on when to stop imaging with certain set of imaging parameters. The current imaging software does not contain exposure time changes at all. As explained in the previous answer, we want the entire dataset classified no matter what the imaging mode was, rather than implementing (and testing and validating) additional ways to pre-prune the data.

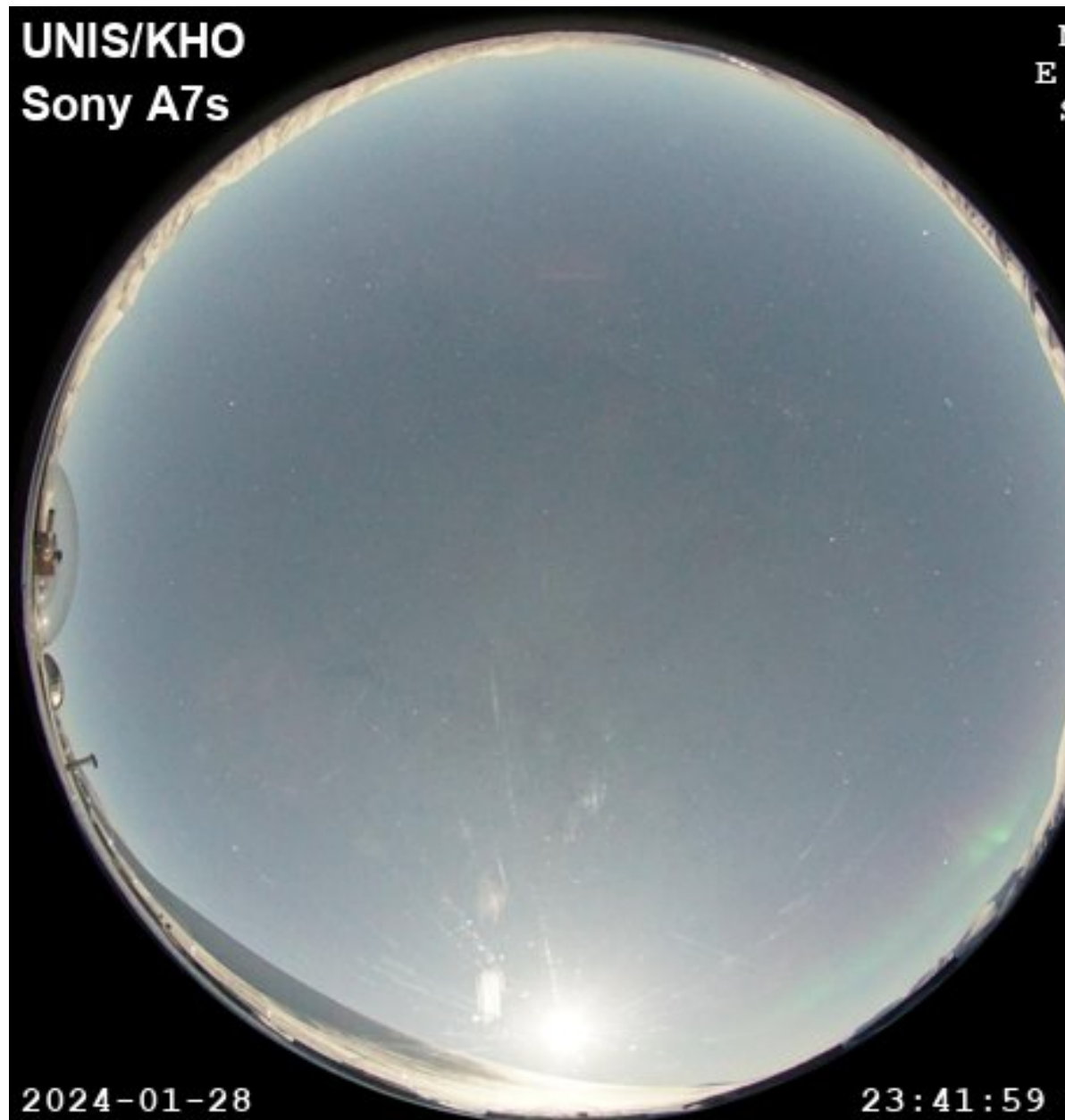
In the aurora quantification method (more explanation below), exposure time is used as one of the parameter to judge twilight time and moon (to adjust detected intensity). This case clearly show that "aurora could have existed but not detectable". From users viewpoint, such "twilight" should not marked as "no aurora".

We think it is fair to gather all detectable aurora into the class of aurora, while the rest is no aurora. The images that we as human experts or ML classifiers cannot classify correctly are the classification error we cannot improve. While the pixel values change drastically during twilight conditions, a CNN classifier will detect aurora when the contrast to the background sky is strong enough for the structures to be seen. This is comparable to the "visual aurora" concept introduced in Yamauchi et al. 2023.

Figure 4: The image is strongly affected by the moon after around 20 UT, and I suspect the some percentage of "ClearAurora" (particularly after 2330) could be no aurora. Could you show original images from around 2240 UT and 2340 UT?

There is actually some aurora at low elevation from about 22:40 to about 23:50 UT, as shown by the attached sample images. The class probability for ClearAurora only reaches 0.4 in that time frame, so this would not be counted as aurora in our statistics, where we assign the images to the class with the highest probability, as explained at the beginning of section 4.2.





It would be useful to mention the other types of automated aurora classification in addition to machine learning type. Oldest example is photometer count (filtered 5577) to classify intensity, and recently, pixel-level classification to quantify the aurora activity was introduced (Yamauchi+ <https://doi.org/10.5194/gi-12-71-2023>). While the ML type provides good classification of morphology, manual type provides quantified values, and combining both methods will improve in evaluating the aurora activity in future. This work is a good reference for such work.

This work is included in the revised version of the introduction as an example of other types of automatic detection tools:

“Another type of operational auroral detection method was developed by Yamauchi and Brändström (2023), based on thresholding and classifying RGB pixel values of the auroral all-sky images. This method is detecting and alerting on local auroral brightenings over Kiruna in northern Sweden in realtime with an estimated accuracy of 90%”

Line 210: 0.1 second per image is a good performance fitting to real-time operation of

burst-mode (high time resolution) observation. Please mention such advantage.

We have mention the realtime capability in the revised version at the end of section 3:

“With this processing speed, we can prune images in realtime even with a higher imaging rate than the current 12s.”

In fact, the classifier was internally test-run in realtime during the winter season 2025-2026, as mentioned in the last paragraph of section 3.

Section 4.1

Line 235-247 "Sony and Nikon": The RGB sensitivity and its balance are quite different between Sony and Nikon, and is also a big problem for automated aurora quantification method too (above reference). Could you make similar table (appendix is ok) as Table 2 and Figure 3 just for this purpose (1000 samples)? After reading your result, I am quite confident that statistic (like Figure 6) should be separated obtained for Nikon camera and Sony camera. In this respect, I could not catch what authors want to say this paragraph: Excuse why the statistics is limited to 2016-2025 (I agree), or it is feasible to correct the judging of Nikon camera result in terms of Sony camera (I personally do not agree)?

The sensitivity and the colour balance are indeed very different between Sony and Nikon, or any two different camera models. This was a test run for classifying Nikon data with the method we used for Sony. The results are not very bad, but the method clearly requires to be trained with some Nikon data in order to perform at a level that would be more comparable to the level of Sony classification. As the aim is just to prune all our colour image data for more detailed analysis and at the same time provide some statistical insights. Extending the AO/CO timeline back to 2008 will have its value for facilitating further auroral studies on the older data. The training of the classifier will allow it to learn the colour balance of the old data, while the sensitivity issue will remain and must be kept in mind, as we can only teach a classifier to detect aurora where we can see and detect the aurora ourselves.

We have include the image numbers for the Nikon camera test in the same format as we have for Sony images in the revised version of the manuscript, at the end of section 4.1. The human accepted image numbers from the 1000 random samples per class are: 785 for ClearAurora, 450 for ClearNoAurora, 632 for CloudyAurora, and 884 for CloudyNoAurora

Section 4.2

In the statistics (Fig 6), was twilight case (cf Fig 4 center part) included or excluded?

Twilight is included. All the images the camera took in 2016-2025 are included.

Figure 6: The "low" probability of AO at 12 MLT (which is close to local noon) and remaining double peak might be due to twilight that prevents recognising the aurora.

This was tested by excluding the twilight (solar zenith angles larger than -12) and moonlit (Moon illumination larger than 50%) conditions. The biggest change is that the data numbers drop to half and one daytime bin goes to zero. AO increases by 6-7%, while cloudiness and clear skies decrease by 7% and 3%. The double peak in AO still remains, as shown by the attached figure. These results are included in the discussion of biases in the revised discussion section:

“If the classifier judges these low-contrast conditions as cloudy more often than not, this

could lead to a bias towards cloudiness in our dataset. This was tested by excluding the twilight conditions (the Sun above -12 degrees below the horizon) and moon above the horizon with moon illumination larger than 50% from our dataset. As a result, CO and clear skies were reduced by 7% and 3%, respectively, while AO was increased by 6–7%. However, the diurnal behaviour of AO, CO and clear skies (seen in Figure 6) still persists. The drawback of this test is that the total number of images decreased by nearly 50%. Because the reduced dataset shows the same diurnal evolution, we have included the twilight and moon up conditions in our analysis, but added the solar elevation angle, the moon phase and the moon illumination in the classification results to facilitate the data reduction that best fits the purpose of any future study.”

Figure 6: It is useful figure but please add a comment that this is not for "solid AO probability", because part of the cloudy cases (thick cloud through which the aurora is not possible to see) should be removed from total number for such probability. Even $\text{ClearAurora}/(\text{ClearAurora}+\text{ClearNoAurora})$ will not give the probability because ClearAurora allows 3/8 of the sky to be covered by cloud)

Yes, the revised version emphasises that these occurrences include the cloudiness, which we cannot remove. This is rather the detectable ground-based AO (lines 277-279):

“AO is defined as a combination of classes of ClearAurora and CloudyAurora (~2 million images) to show an auroral evolution as it is seen from the ground. We cannot determine AO when the clouds block the visibility to the ionosphere, which reduces the number of samples in the statistics.”

Figure 7 and 8: It is easier to place 11-12-1-2 from top to bottom.

This would work well for a plot with one winter season, as in Figures 7 and 8, but we wanted to keep the format extendable to multiple winter seasons in Figure 9.

Figure 8: I worry that high cloud rate under daylight/twilight means that cloud detection depends on the sky brightness, so that you should remove the twilight case from the GroundTrue. Could you comment? Please also compare $\text{ClearAurora}/\text{ClearNoAurora}$ ratio and $\text{CloudyAurora}/\text{CloudyNoAurora}$

We will not remove twilight from the ground truth, because the images where it is possible to distinguish between clear and cloudy with and without aurora will help the classifier to learn the reality. We acknowledge that some bright images are too bright for this distinction, but that is by no means all twilight and moonlit images. These are the images where also the manual labelling becomes impossible, and the images were classified as Unclear.

It is not clear to us what the ratios of $\text{ClearAurora}/\text{ClearNoAurora}$ and $\text{CloudyAurora}/\text{CloudyNoAurora}$ would tell more than their occurrence variability.

Line 311 "sound agreement": The slope < 1 (ML method overestimate the cloud) is better explicitly written rather than just saying a "discrepancy".

True. The revised version says that the agreement is good. Because the cloudiness estimates are so very different, we should not compare the absolute numbers between the axes. Cloud sensor value 60 does not mean the same as 60% CO. This is why we think that a correlation of 0.86 is good. We emphasise the estimate difference in the new version of the manuscript (line 328):

“ It is worth noting, however, that the cloudiness value of e.g. 60 is not the same as 60%

CO, because the two measures are inherently different”

Line 330: For the next step, how about combining intensity of aurora (quantification of aurora using pixel size and integrated intensity of auroral pixels)? A general scheme is already presented in Yamauchi+ 2023.

This is also an option we can mention in the revised text. A possible combination of these methods could be to first prune out the clouds with our ML routine and then use the pixel-level method to detect the local brightenings. This is mentioned in the new version of the discussion.

Line 348 "Computer science point of view": What do you exactly mean? CPU time?

This means “from the methodology point of view”, and has been changed to say so, meaning that currently there are more options than just CNN. Vision Transformers seem particularly promising and should be tested.

Lin 358 "allow the Moon and twilight in": As described above, they causes change of exposure, and drastically decrease the chance of recognising aurora, particularly the diffuse one. If they are included, they should also be classified as "MoonNoAurora" "MoonAurora". I am sure that ratio of MoonAurora/MoonNoAurora is smaller than ClearAurora/ClearNoAurora.

The Moon in the images is not causing a change in the exposure time in our data. The twilight may do that depending on the imaging mode and the threshold value used for the solar zenith angle. The sky illumination conditions determine how dim aurora can be detected by human eye and by a classifier, and these two approaches can do about equally well. An ML approach is not sensitive to changes in the pixel-level colour balance, but rather the contrast between the aurora and the background. When a human cannot detect aurora, we call the conditions NoAurora acknowledging that there may still be faint diffuse aurora. In a long-term statistical approach, we still get good results, and the detection of the aurora is still helpful for further studies.

We have included the zenith angles for the Sun and the Moon, as well as the Moon illumination in our classification results, so that anyone using the classification results can decide on thresholding of those parameters as it best suits the purpose of their study.

Line 375 "This suggest...": Not necessary true according to Table 2. This illusion comes from different mother group "clearNoAurora" vs "cloudyNoAurora".

What we want to convey is that the largest class of images is the CloudyNoAurora class. Even with the twilight and moonlit conditions excluded, there is still more cloudy data than anything else. Therefore the biggest limitation for observing aurora is the cloudiness. While AO includes both ClearAurora and CloudyAurora, CO is CloudyNoAurora only, and clear skies is ClearNoAurora only. All these super-classes or mother-groups are then normalised the same way so that they are directly comparable. This is why we think that the statement is correct.

Line 389 "more than 20%": The difference also comes from twilight data (see comment on Figure 6). Also the purpose of the aurora classification is different from Nanjo+ 2022. This paper aims to search aurora with blind information for the other observation teams, whereas the previous works are rather for aurora scientists (including citizen scientists) to search more classified aurora.

The purpose of the classifier will define much of the structure of the implementation, which

is something that we have mentioned in the revised text, along with the inclusion/exclusion of the twilight data. In our data, the largest class is cloudy skies, with or without the twilight and moonlit conditions. We therefore want to prune out the cloudiness and detect the aurora and clear skies as the first analysis step to make more detailed classifications and studies more efficient. These differences do not allow direct comparison with our results and those from Nanjo et al. although their method is an impressive piece of work.

Line 404 "only one other": Actually the quantification of aurora activity from ASC (Yamauchi+ 2023) is already in operation, including ESA spaceweather site. Also, Nanjo's TromsoAI also includes "cloudy" probability

This statement was referring to ML methods, but can be extended to include other approaches and thus include the local brightening detection based on pixel values. A little downside with it is that ESA space weather site requires a log in. We were not aware of the IRF site, but it seems like the realtime classification can be followed through that interface as well, so the link has been included. We emphasise the different motivations guiding the different operational approaches in the revised version of the manuscript. Our motivation is to detect all aurora, so that this cleaner set of images containing aurora can be more efficiently analysed in more detail.

