Response to referee 1

First of all, we like to thank Darrel Baumgardner for his thorough and valuable feedback on our paper. In the following, we will address all his comments and show the according changes we made on the paper.

Comment 1:

How are holographic images currently being processed to separate ice from droplets? If other than those machine learning references given in the paper, then they need to be discussed and compared with the CNN.

Answer to comment 1:

We are not aware of any other machine learning technique apart from SVMs and decision trees being used by the holography community to classify holographic images.

Comment 2:

What are the techniques that are currently being used to separate ice from droplets in other imaging systems like optical array probes (OAPs)?

Answer to comment 2:

There are numerous approaches to separate ice particles from liquid droplets. To acknowledge them we added line 26 to line 31 on page 2 in the introduction:

"Imaging probes, which differentiate only ice from liquid usually extract features from the images that measure the circularity of the particles (e.g. Korolev and Sussman (2000); Crosier et al. (2011); Lawson et al. (2001)). Korolev and Sussman (2000) state an uncertainty for differentiating spheres from irregular particles of 20% to 25% for a pixel number between 20 and 60 and a few percents for higher pixel numbers. These values are comparable to our results for holographic images, which we will introduce later. However, the existing approaches are not suitable for holographic images since they do not account for artifacts. Finding good features for artifacts is difficult because they do not have a specific shape."

Comment 3:

If the techniques being used for processing images in OAPs can be used in holographic images (I can see no reason why they can't given the rendering of holographic images into 2D for the CNN), how do the error rates compare with those from CNN?

Answer to comment 3:

As already described in the answer to comment 2, we cannot use these algorithms due to the existence of artifacts in our datasets. Most of the mentioned studies do not give an uncertainty estimation and if they do, it is not easily possible to compare it to our results since they consider different sizes of particles than we do and have different pixel size.

Comment 4:

Nowhere in the introduction, or elsewhere, do the authors discuss how errors in discriminating liquid from ice will impact how measurements are interpreted with respect to scientific questions associated with mixed phase conditions. This is a critical omission when assessing the efficiency of one technique versus the other. Particularly when it comes to ease of implementing one technique versus the other.

Creating training sets for every data set is time consuming and one that is unnecessary if using one of the common techniques used in OPA analysis.

Answer to comment 4:

We added the section "Needed accuracy of cloud particle classification regarding scientific question" into the discussion section (line 5 to line 12 on page 20):

"Needed accuracy of cloud particle classification regarding scientific questions. How accurate the phase discrimination, the particle number or mass concentration has to be for a meaningful interpretation of the data highly depends on the scientific question. For example, in a model study, Young et al. (2017) showed that an overestimation of ICNC by only 17% (2.43 l^{-1} instead of 2.07 l^{-1}) led to cloud glaciation while the MPC was persistent for about 24h with the lower ICNC, while very few ice crystals (0.21 l^{-1} = -90%) may lead to cloud break-up. In theoretical calculation, Korolev and Isaac (2003) showed that the glaciation time of a MPC with an ICNC of only 1 l^{-1} is about four times as long as for 10 l^{-1} (+100%) at a temperature of -15°C. Comparing measurements with studies can therefore already lead to wrong conclusions with classification uncertainties of ±20%."

Comment 5:

Why hasn't the uncertainty in human-typing images been assessed, i.e. having at least two or more observers classify the same data set?

Answer to comment 5:

The human bias is now assessed with the results of three people classifiying the same dataset. To show our results we added line 11 to 16 on page 5 in the Experimental data section together with Figure 3. We also adapted line 33 on page 19 to line 2 on page 20 of the Discussion section "Applying the CNN to new datasets" accordingly.

"For the estimation of the human bias, three different people hand-labeled the same dataset consisting of 1000 particles. The number of particles hand-labeled as the considered class by at least one person are compared to the number of particles hand-labeled as the considered class by all three persons. Taking the average of these two numbers, the spread can be given as the percentage deviation to the two values (see Fig. 3). For liquid droplets, we have a deviation of ±4% and for ice crystal ±5%. However, this estimation does not take into account that in some cases humans might just not be able to recognize the correct class as outlined before."

"Other sources of uncertainty like the manual classification contribute with about ±5% (see Fig. 3) to the here considered size ranges. Therefore, the uncertainties using a fined-tuned CNN are of similar magnitude as uncertainties from other sources"

Comment 6:

In addition to the number of hours needed to create a training set, what are the computational times to analyze sample data sets of 10000 images by each technique?

Answer to comment 6:

We added line 19 to 22 on page 19 in the discussion section:

"Another important factor for the prediction performance is the time it takes to do the predictions. This highly depends on the dataset and the computational power of the computer. Classifying 10,000 particles takes about 15 s for the decision tree, about 30 s for the SVM and about 60 s for the CNN on a local server. None of these time scales is comparable to the time it takes to classify 10,000 particles by hand, which can vary between a few hours and a few weeks depending on the dataset."

Comment 7:

The list of references is missing a large number of studies on pattern recognition of cloud probe images. These have to be included.

Answer to comment 7:

To include studies on cloud pattern recognition of cloud probe images we added line 26 to line 31 on page 2 (the section is already written-out in the answer to comment 2) as well as line 8 to 14 on page 3 in the introduction:

"Deep learning (usually referred to as neural networks) has the potential to overcome transfer learning issues, which we will show in this work. For the classification of cloud particles, a feedforward neural network from Hagan and Menhaj (1994) was used by O'Shea et al. (2016) to classify CPI data into different ice particle shapes and liquid droplets. The network is fed by different features, which are calculated beforehand. Their results are promising with a total accuracy of 88% to classify the images into six habits including liquid droplets for particles larger than 50 µm. This type of a neural network also requires feature extraction and does not work for holographic images because it does not account for a class without a specific shape like artifacts."

Additional comment by authors about a change of a used metric

We want to point out that we changed the used metric "equivalent area particle diameter" for the evaluation of the CNN on different particle sizes to "major axis size". The reason for this is that we noticed that the equivalent area particle diameter was not calculated correctly. We, therefore, decided to use the major axis size instead, which is also a measure of size. The changes in the results are small and the interpretation of the results does not change.