Advanced Deep Learning–Based Supervised Classification of Multi-Angle Snowflake Camera Images

C. KEY,[a] A. HICKS,[a] AND B. M. NOTAROŠ[a]

[a] *Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado*

(Manuscript received 18 November 2020, in final form 1 June 2021)

ABSTRACT: We present improvements over our previous approach to automatic winter hydrometeor classification by means of convolutional neural networks (CNNs), using more data and improved training techniques to achieve higher accuracy on a more complicated dataset than we had previously demonstrated. As an advancement of our previous proof of concept study, this work demonstrates broader usefulness of deep CNNs by using a substantially larger and more diverse dataset, which we make publicly available, from many more snow events. We describe the collection, processing, and sorting of this dataset of over 25 000 high-quality Multi-Angle Snowflake Camera (MASC) image chips split nearly evenly between five geometric classes: aggregate, columnar crystal, planar crystal, graupel, and small particle. Raw images were collected over 32 snowfall events between November 2014 and May 2016 near Greeley, Colorado, and were processed with an automated cropping and normalization algorithm to yield 224 3 224 pixel images containing possible hydrometeors. From the bulk set of over 8 400 000 extracted images, a smaller dataset of 14 793 images was sorted by image quality and recognizability (Q&R) using manual inspection. A presorting network trained on the Q&R dataset was applied to all 8 400 0001 images to automatically collect a subset of 283 351 good snowflake images. Roughly 5000 representative examples were then collected from this subset manually for each of the five geometric classes. With a higher emphasis on in-class variety than our previous work, the final dataset yields trained networks that better capture the imperfect cases and diverse forms that occur within the broad categories studied to achieve an accuracy of 96.2% on a vastly more challenging dataset.

SIGNIFICANCE STATEMENT: Classification of precipitation, namely, deciding to which of the several typical classes of winter hydrometeors the observed particles belong, can enrich our understanding of polarimetric radar signatures of snow, as well as ice cloud processes and the resulting precipitation production. The high-resolution photographs of snowflakes collected by the Multi-Angle Snowflake Camera (MASC) are especially suitable for snowflake classification. However, classifying particle types from MASC photographs by visual inspection is not practical given the typical amount of MASC data. We present advanced automatic deep machine learning–based classification of MASC images using convolutional neural networks. This study demonstrates broad usefulness of our approach yielding trained networks that achieve extremely high classification accuracy on a large and diverse dataset from many snow events.

KEYWORDS: Snow; Ice particles; Optical properties; Data processing; In situ atmospheric observations; Surface observations; Classification; Machine learning; Neural networks

1. Introduction

Snowflake classification is important for improved weather radar, assessment of storm structure, and characterization of winter precipitation events from ground sensors (Zhang et al. 2011; Straka et al. 2000; Libbrecht 2017). Several types of in situ image capturing devices used for ground-based collection of data relevant to snowflake classification include the two-dimensional video disdrometer (Schönhuber et al. 2008), the Precipitation Instrument Package [an improved version of the system in Newman et al. (2009)], and the Multi-Angle Snowflake Camera (MASC). We focus on snowflake images collected by MASC systems in the present study. To allow researchers to study the microphysical characteristics of snowfall, relevant to a storm's composition, the MASC captures high-resolution images of falling hydrometeors from several angles. These images can be processed to extract images of individual snowflakes from a variety of perspectives, or even used to generate 3D models of hydrometeors automatically (Kleinkort et al. 2017). A MASC system is capable of capturing tens to hundreds of thousands of images during a single winter storm event, leading to datasets too large for manual classification. This has been a major motivation for accurate, automated snowfall classification.

Existing approaches to automated snowfall classification from MASC images vary and include the excellent work of Praz et al. (2017), our previous work (Hicks and Notaroš 2019), and an unsupervised technique (requiring no human input) from Leinonen and Berne (2020). The multinomial logistic regression (MLR)-based method described in Praz et al. (2017) has been demonstrated effective but requires careful definition and algorithmic extraction of several image features from which classifications are made. This approach has achieved an outstanding 95% classification accuracy, but may be somewhat rigid, relying on human-described features such as morphological skeleton statistics, rotational symmetry, and gray-level

Corresponding author: Branislav M. Notaroš, notaros@colostate.edu

DOI: 10.1175/JTECH-D-20-0189.1

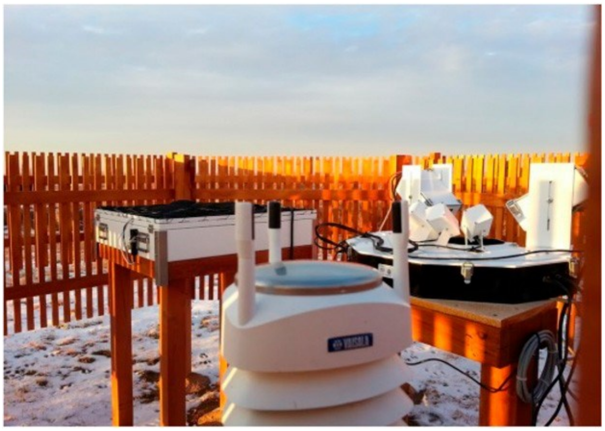


FIG. 1. MASCRAD snow field site at Easton Valley Airport, near Greeley, Colorado, under the umbrella of CSU–CHILL radar. MASC (top right of the photo), along with other surface instrumentation, is contained in the 2/3-scaled DFIR.

cooccurrence. Older supervised classification work in Lindqvist et al. (2012), similarly, applies principal component analysis coupled with Bayesian and weighted nearest-neighbor techniques to classify ice-cloud particles, typically achieving accuracies between 80% and 90%. We have previously presented convolutional neural networks (CNNs) as a robust alternative that can easily be applied and generalized in a black-box manner without expert definition of features. Both methods, of course, require manual input to generate training and test data labels. The work of Leinonen and Berne (2020), on the other hand, automatically classifies snowflake images by exploring the latent space of generative, as opposed to predictive, models. Such unsupervised approaches are extremely promising for discriminating and classifying different hydrometeor images in general, but an unsupervised method inherently produces its own categories, rather than directly assigning images to existing, known categories with which researchers are likely already familiar.

Accordingly, we offer improvements to our existing CNN-based, supervised approach (Hicks and Notaroš 2019), using more data and improved training techniques to achieve higher accuracy on a more complicated dataset than we had previously demonstrated. As an advancement of our previous proof of concept study, which used a geometric dataset focused on easily identifiable examples of each of the snowflake classes considered, a principal goal of this work is to demonstrate broader usefulness of deep CNNs for automated snowfall classification by using a larger dataset containing wider in-class variety. We present improved training methods and new, automated techniques for detection, cropping, and normalization of snowflake images as well as quality and recognizability preprocessing of image data. From these improvements, we demonstrate higher overall test accuracy on a vastly more challenging dataset than that used in our previous work. Together, these improvements constitute an accurate, efficient, and robust supervised machine learning approach to snowflake classification, using deep neural networks and images collected by the MASC or another image-based particle recording instrument or system.

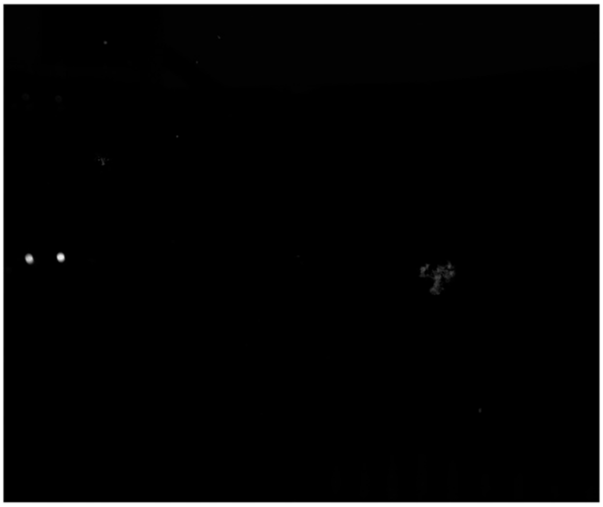


FIG. 2. Example of a normalized raw MASC image. Several snowflakes can be seen in addition to background glare (center left) and subtle ground and sky glow (top and bottom). Note that the ground and sky glow may not be visible in all prints or computer monitor settings.

2. Data collection and image processing

This section describes the collection of raw MASC images as well as the automated cropping and normalization performed on raw images to isolate potential snowflakes present in each image.

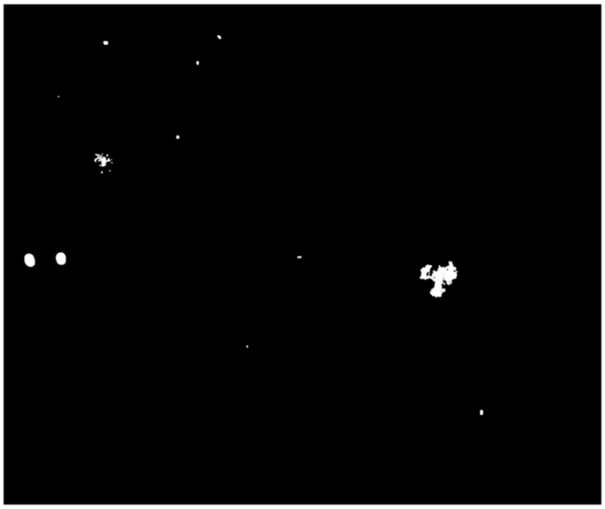


FIG. 3. Example of a binary image produced by application of a brightness threshold and five-pixel radius to the normalized raw image in Fig. 2. Possible snowflake silhouettes are now apparent. Background glare (center left) was rejected due to exceeding the mean brightness threshold. Dimmer glare cases are reliably assigned to the not-flakes Q&R category.

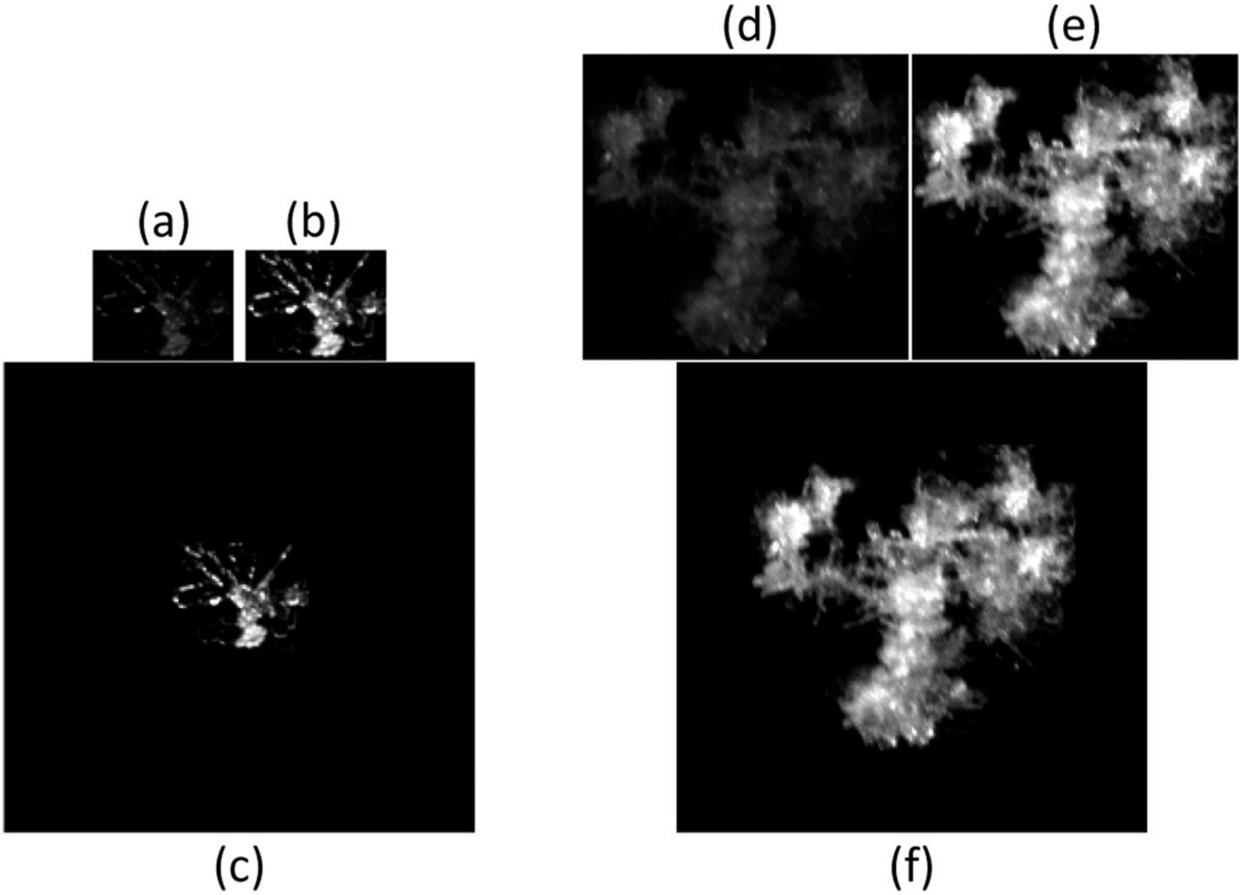


Fig. 4. Example crops and image chips extracted from the MASC image shown in Figs. 2 and 3. (a) Cropped image of a planar crystal. (b) Example crop from (a) after contrast scaling. (c) Final image chip produced from contrast scaled crop in (b). (d) Cropped image of an aggregate. (e) Example crop from (d) after contrast scaling. (f) Final image chip produced from contrast scaled crop in (e).

a. *Raw image collection*

The 3 458 848 raw images used to generate the training set were collected from several winter weather events between November 2014 and May 2016 using a modified MASC system. The system was located at a surface instrumentation field site established under MASC and Radar (MASCRAD) (Notaroš et al. 2016; Bringi et al. 2017; Kennedy et al. 2018). This is the same site and system used for data collection in Hicks and Notaroš (2019). The MASCRAD field site is located at the Easton Valley View Airport in La Salle, near Greeley, Colorado, shown in Fig. 1. The MASC system, along with other ground-level instrumentation at the site, is situated within a double fence intercomparison reference (DFIR). Raw images from both winter storm events used in Hicks and Notaroš (2019) constitute a subset of the total raw image set used in the present work. Details of the MASC system used are presented in Hicks and Notaroš (2019). Although the MASC allows for collection of snowflake imagery from multiple angles to help determine three-dimensional shape (Kleinkort et al. 2017), we did not make use of this feature directly for the present work. As described in Leinonen and Berne (2020), it is common that a given snowflake will only be captured at usable quality by a single camera of a multicamera system, the snowflake often out of focus or occluded in other fields of view, so limiting study to only snowflakes that appear at high quality in all fields of view substantially reduces the number of useable examples. By limiting study to single-view cases, we were able to manually classify thousands, rather than hundreds of snowflakes at a cost of increased ambiguity due to lack of multiangle data. Note that we did not explicitly remove cases where a single snowflake was imaged from multiple angles when forming the dataset for the present work.

b. *Detection, cropping, and normalization*

As the MASC produces raw, wide field of view images, typically containing many snowflakes, it is necessary to isolate individual examples for classification. All images were processed in grayscale (single channel). To detect possible flakes in each raw MASC image, we first normalized the entire grayscale image, dividing all pixel values by the maximum brightness value. An example of a normalized raw image is shown in Fig. 2. We then converted the grayscale image into a binary image by application of a threshold. Pixels in the grayscale image with brightness greater than or equal to the threshold

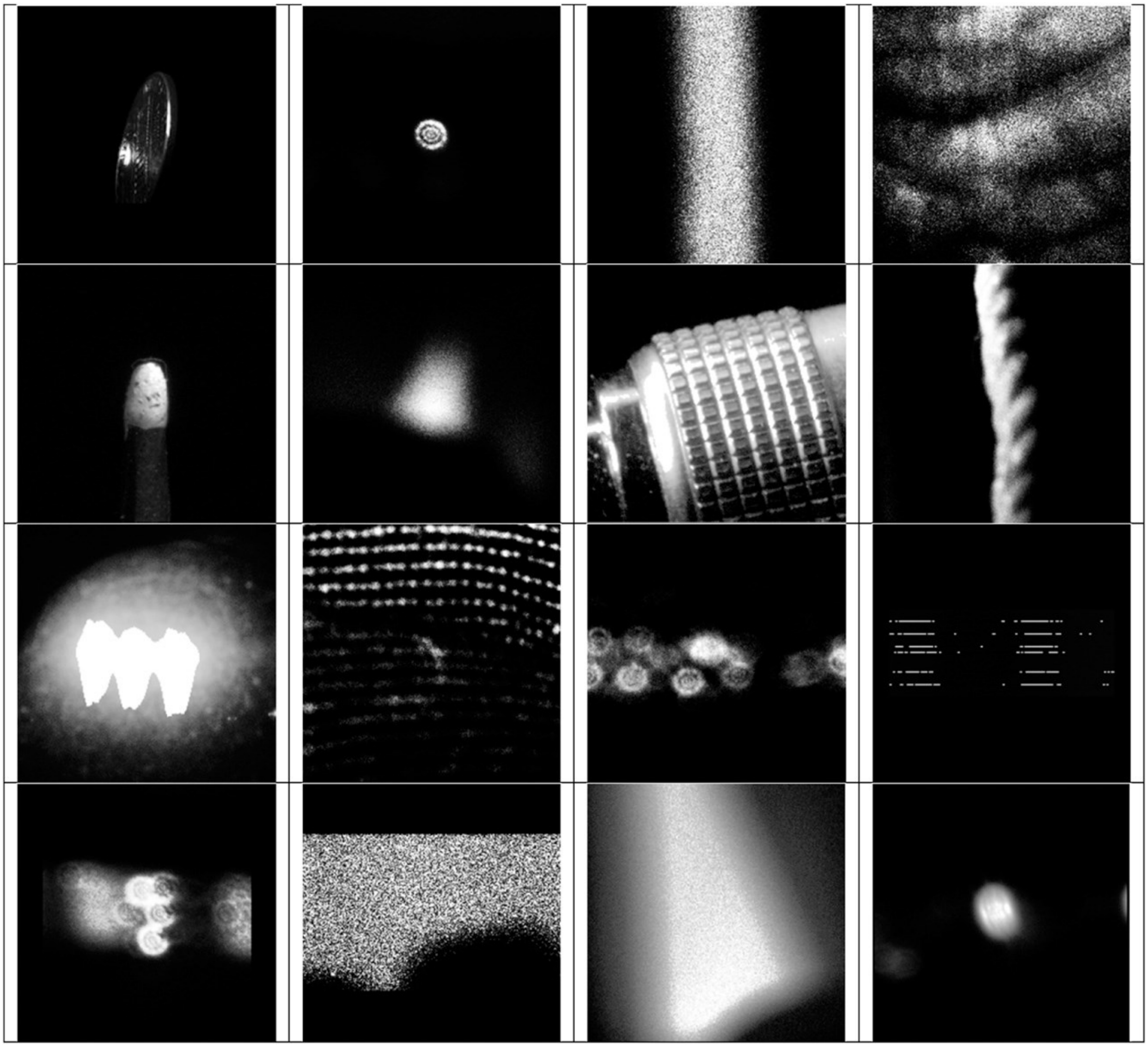


FIG. 5. Examples of image chips in the not-flakes quality and recognizability category. A description of this category is given in Table 1. (first row) (left to right) a coin, background glare, sky glow seen between fence posts, and a finger. (second row) (left to right) A sensor probe, an out of focus sensor probe, part of a pair of calipers, and a string. (third row) (left to right) A metal ball, part of a mitten, background glare, and amplified sensor noise. (fourth row) (left to right) Background glare, sky glow seen above fence posts, background glare, and background glare.

were assigned value 1, and pixels less than the threshold were assigned value 0. For the present work, this threshold was set to 0.1. We then set any pixels in the binary image with value 0 to 1 if they were within a two-pixel radius (using Chebyshev distance) of any pixel that had already been assigned a value of 1 in the previous thresholding step. The example image from Fig. 2 is shown after thresholding and application of the two-pixel radius in Fig. 3. This radius was chosen by hand as a reasonable value. Next, we computed sets of connected components in the binary image. A connected component is any group of active (value: 1) pixels that form an unbroken group. If a connected component contained fewer than 26 active pixels, it was discarded. For each connected component not discarded, we cropped a rectangular region from the original grayscale image corresponding to its bounding box. Two such examples produced from Fig. 3 are shown in Figs. 4a and 4d. Cropped images were then contrast scaled linearly such that the top 1% of brightest pixels were saturated. Figures 4b and 4e show cropped image examples from the previous step after contrast scaling. Note that contrast scaling destroys some information theoretically available in the images (by loss of absolute brightness and saturation of some pixels). However, we found that brightness variations between flakes were dominated by differing lighting conditions, rather than useful

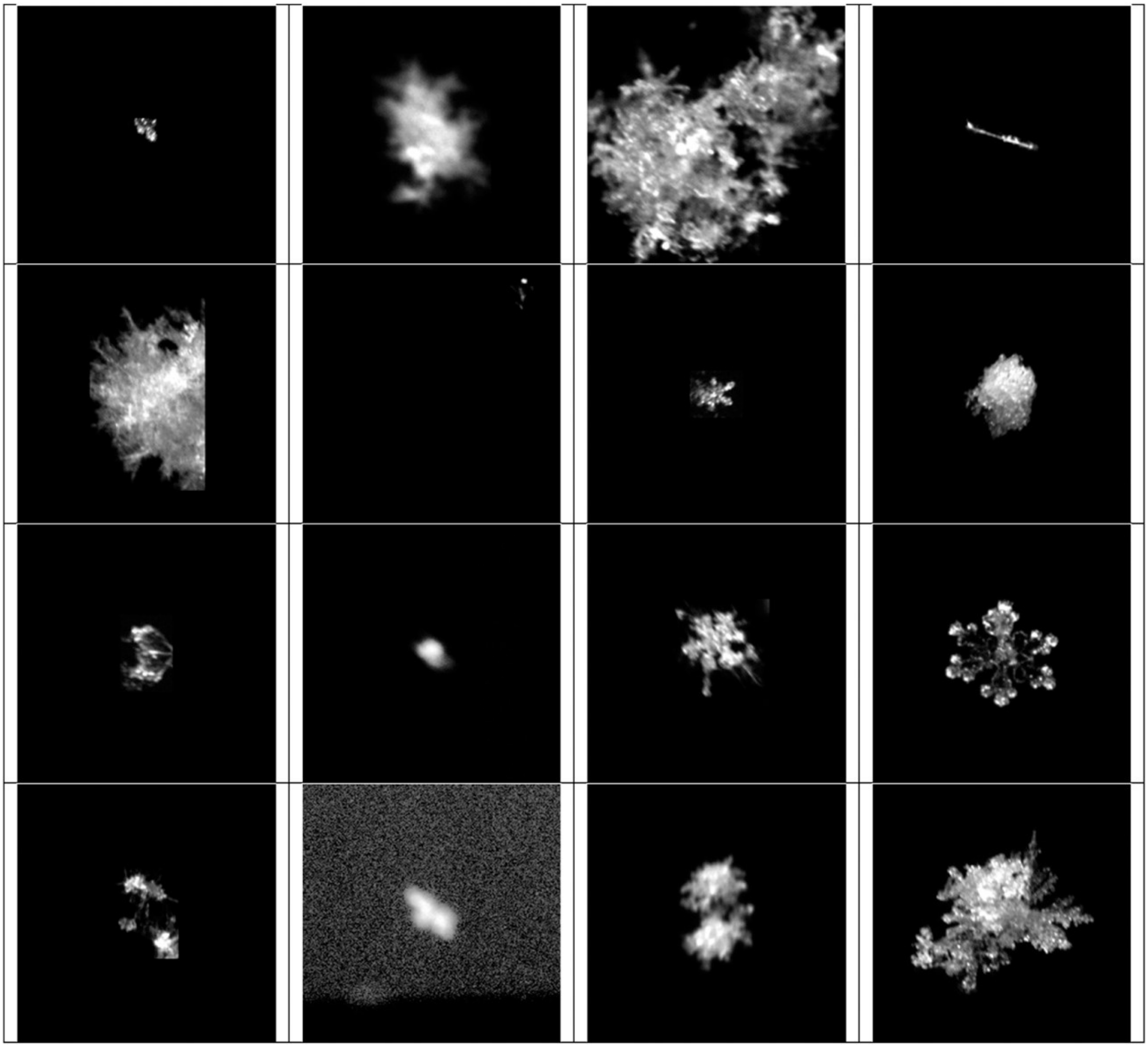


FIG. 6. Examples of (left to right) bad-crop, bad, okay, and good image chips, respectively. Category descriptions are given in Table 1.

information like snowflake class, so contrast scaling was performed to give the network input for which pixel brightness variations are dominated by microphysical characteristics rather than lighting conditions. After scaling, any cropped image was rejected if the mean value of its pixels was greater than 0.5. We then centered each remaining cropped, scaled image on a 224 3 224 black background to produce final image chips. Examples are shown Figs. 4c and 4f. Cropped images

TABLE 1. Category names, counts, and descriptions for the quality and recognizability dataset, a balanced subset of which was used to train a presorting network using the methods of Hicks and Notaroš (2019).

Category name	Count	Description
Not flakes	7020	Object other than a snowflake present in the image; examples: sensor noise, glare, sky/ground glow, and calibration probes
Bad crop	1500	Likely snowflake present, but poor cropping leaves a substantial portion of the snowflake out of the image chip, interfering with geometric classification
Bad	1977	Likely snowflake present, but poor lighting or focus prevent identification; image chips containing more than one disjoint (nonaggregated) snowflake are also assigned to this class, regardless of image quality
Okay	2796	Focus and lighting are good enough to identify coarse flake features, and likely geometric class, but are insufficient to capture microphysical characteristics
Good	1500	Lighting and focus are good enough to resolve microphysical characteristics and determine snowflake geometric class

TABLE 2. Number of examples in each class for the geometric dataset.

Class name	Count
AG	5038
CC	5021
GR	5000
PC	5014
SP	5126

that exceeded the 224 3 224 image chip sized were cropped to 224 3 224 pixels after centering.

This approach to cropping and normalization was arrived at for several reasons. In contrast to simply cropping a 224 3 224 pixel region centered on each connected component in an image (or similar), we found that the above method significantly reduced the number of image chips that contained multiple, physically disconnected snowflakes. In other words, during heavy snowfall events, we found it was common for two or more snowflakes to appear within 224 pixels of each other. By cropping a tight bounding box as above, we were able to recover far more closely spaced snowflakes into usable, unambiguous image chips. Rejection of cropped, scaled images with mean pixel value greater than 0.5 rejected most crops of the sky and background that did not actually contain a snow particle. By also rejecting connected components with pixel counts below 26, we avoided cases where a single bright pixel caused a false detection. In general, the described cropping

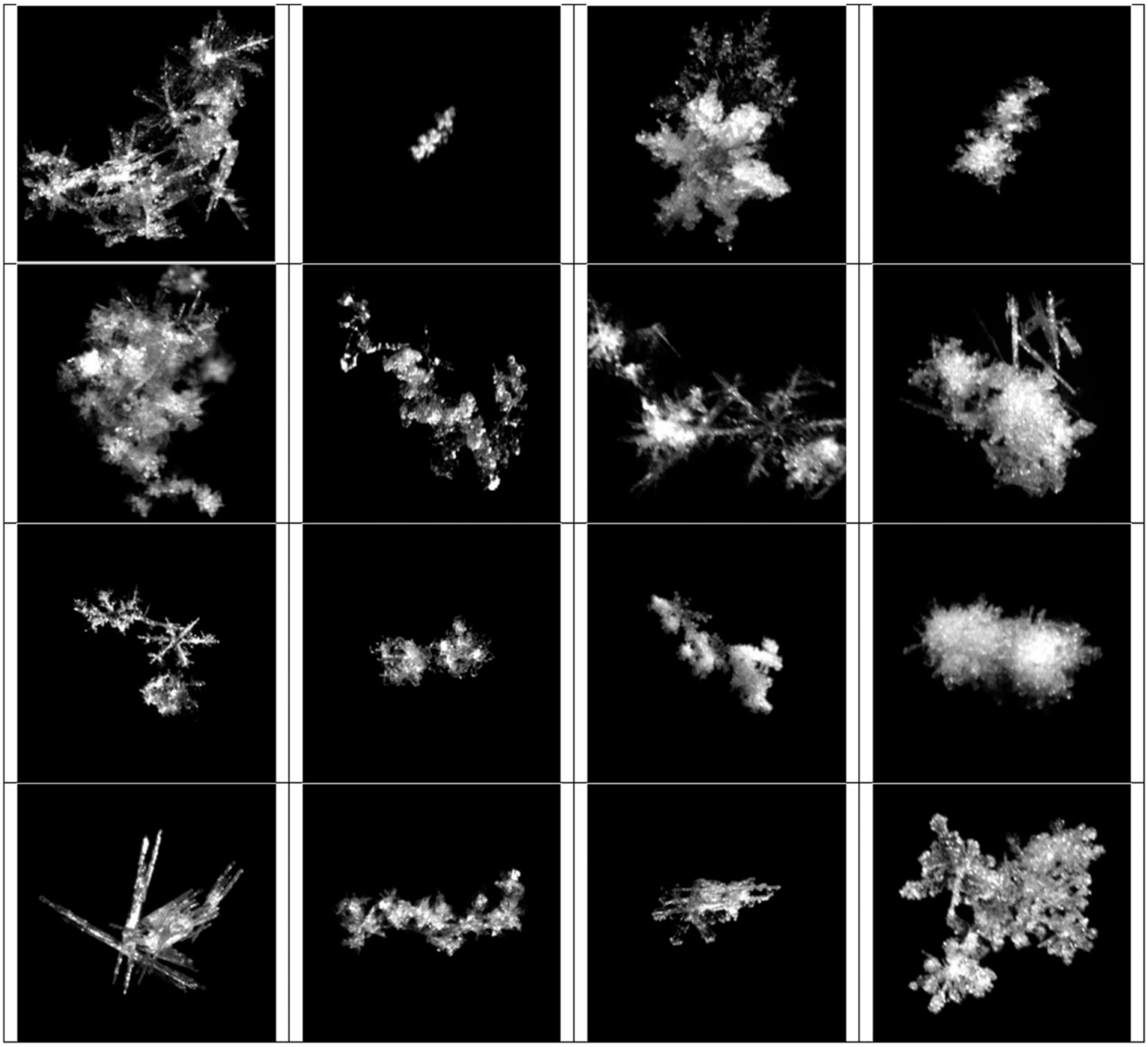


FIG. 7. Examples of image chips in the aggregate (AG) class of the final geometric dataset. All image chips in the final geometric dataset had been automatically categorized into the good Q&R category. We placed emphasis on collecting a wide variety of sizes and forms of aggregate with varying types of constituent particles.

and normalization approach was able to detect far more small particles, and dim, unrimed planar crystals than the approach used for Hicks and Notaroš (2019). Application of this cropping algorithm to all raw MASC images from November 2014 to May 2016 produced 8 441 563 image chips.

3. Hydrometeor classification scheme and training sets

This section describes how the 8 441 563 224 3 224 pixel image chips extracted from raw MASC images were automatically sorted to quality classes and how images from the best class were manually sorted into the five geometric categories studied. A total of 25 199 examples were manually sorted for the final geometric dataset covering 32 snowfall events, an event defined here as a period during which no more than 24 h passed between collection of any two image chips identifies as snowflakes during manual classification. All classification was performed by a single analyst who reviewed each image at least three times. Overall, we are confident the manual classifications used for training accurately represent the opinions of our analyst and have made this dataset available at Key et al. (2021). Note, however, that our use of only one human analyst has potential to introduce more bias relative to other work for which multiple humans performed analysis, such as Praz et al. (2017). We had originally planned to also produce an expanded riming dataset in addition to the presented geometric dataset, but we found that some riming degrees were insufficiently represented for production of a larger, balanced riming dataset from our current pool of raw images. We hope to contribute such a dataset in future work.

a. *Quality and recognizability preprocessing*

The snowflake detection, cropping, and normalization method described in section 2b remains imperfect. Therefore, many of the image chips produced contained bright points from a raw image that are not snowflakes. These included sources like glare, sensor noise, and sky/ground glow. In addition, operators of the MASC system occasionally forgot to turn off data collection while calibrating and testing the system after maintenance and redeployment. This led to captures of test probes, hands, coins, and other objects to occasionally appear in the raw image dataset. Several examples of image chips due to nonflake objects are shown in Fig. 5.

For image chips that contain snowflakes, there is an inherent range of quality. Some flakes appear out of focus in raw images. Others are poorly cropped, either due to overcropping by the image processing method in section 2b, or because they originally appeared partially out of field of view in a raw MASC image. We considered image chips containing snowflakes to fall into four recognizability categories: bad crop, bad, okay, and good. Image chips in the bad-crop category are those where unambiguous recognizability of the imaged snowflake is made difficult due to overcropping by the processing method described in section 2b or part of the flake appearing out of field of view in the raw image, leaving a substantial portion of the flake absent from the image chip. Note that cases where a flake was simply too large to fit in a single image chip were not included in the bad-crop category. In our manual exploration of the dataset, such flakes were almost exclusively in the AG class and easily identifiable despite cropping to 224 3 224 pixels. Rather, overcropping by the processing described in section 2b is typically due to poor or uneven illumination of the flake causing the rectangular bounding box of the resulting connected component to not contain most of the pixels covered by the snowflake. Four examples of bad-crop image chips are shown in the first column of Fig. 6. Bad image chips are those for which poor focus or poor illumination rendered the target snowflake unrecognizable. Image chips containing more than one disjoint (nonaggregated) snow particle are also included in the bad category, regardless of lighting and focus. We consider two snow particles disjoint if they were clearly identifiable as discrete, physically unconnected particles by our human analyst. Four such examples are shown in the second column of Fig. 6. Okay image chips were those that contained a recognizable snowflake but suffered from mild blur or high background noise that made examination of microphysical characteristics difficult. Four examples of okay image chips are shown in the third column of Fig. 6. Good image chips were those that were free of substantial overcropping and clear enough to identify relevant microphysical features. Column four of Fig. 6 shows four examples of good image chips.

TABLE 3. Test accuracy results of 10 independent training runs. Note that training runs 5 and 6 producing test accuracies identical to two decimal places occurred by chance and was verified not to be a mistake.

Run	Test accuracy
1	96.56%
2	96.04%
3	96.24%
4	95.88%
5	96.00%
6	96.00%
7	96.20%
8	96.08%
9	96.68%
10	96.64%

To avoid wasting human time visually inspecting images that did not contain flakes or were of quality too poor to use, we trained a preliminary quality and recognizability (Q&R) classifier on a small, manually sorted subset of the 8 441 563 image chips. This classifier was implemented by necessity to reduce the data volume needing manual inspection, and its results were not further analyzed or verified in the present work. To train the Q&R classifier, we collected at least 1500 examples for each of five categories: not flake, bad crop, bad, okay, and good, with an emphasis on variety within each class. Counts per category for the Q&R dataset are presented in Table 1 along with descriptions. When collecting example images, we included roughly equal numbers of examples from each geometric class in okay and good categories to avoid biasing the classifier against a given geometric type. The Q&R classifier was trained using the same

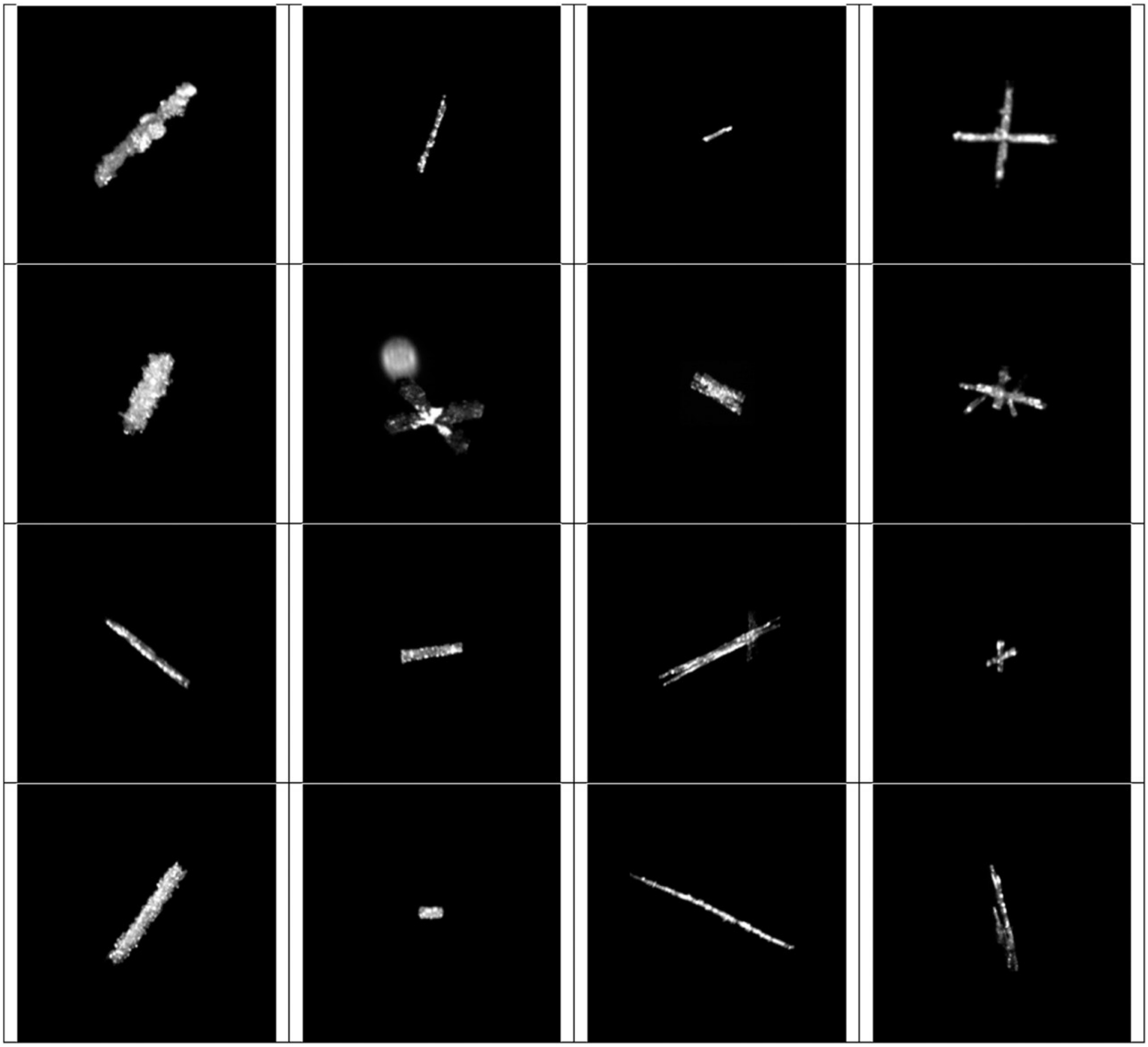


FIG. 8. Examples of image chips in the columnar crystal (CC) class of the final geometric dataset. All image chips in the final geometric dataset had been automatically categorized into the good Q&R category. We included a variety of sizes, forms, and degrees of riming. An example of a backlit snowflake is shown in row 2, column 2. Such cases were rare but were included whenever backlighting did not interfere with recognizability.

methodology used for the geometric classifier in Hicks and Notaroš (2019). For training, 1500 examples from each Q& R category were drawn randomly. The trained Q&R classifier was then applied to all 8 441 563 image chips to sort each into not-flake (3 791 326), bad-crop (723 550), bad (3 062 288), okay (582 333), and good (282 001) categories. Only image chips assigned by the Q&R network to the good category were examined to produce the geometric dataset for the present study.

b. Geometric classes

A variety of attempts have been made to classify snowflakes (Nakaya and Sekido 1936; Magono and Lee 1966; Korolev and Sussman 2000; Grazioli et al. 2014; Vazquez-Martin et al. 2020). As in our previous work (Hicks and Notaroš 2019), we chose to use the scheme adopted by Praz et al. (2017) for training and testing of their multinomial logistic regression snowflake classifier. We summarize this scheme here.

The scheme uses the nine categories of snowflakes defined in Magono and Lee (1966), with a few simplifications for data availability. Praz et al. (2017) additionally defined the aggregate and small particle classes. Aggregates are defined as single snowflakes that are the result of in-air collision of two or more particles. Small Particles are snowflakes whose features are too small to categorize. Note that

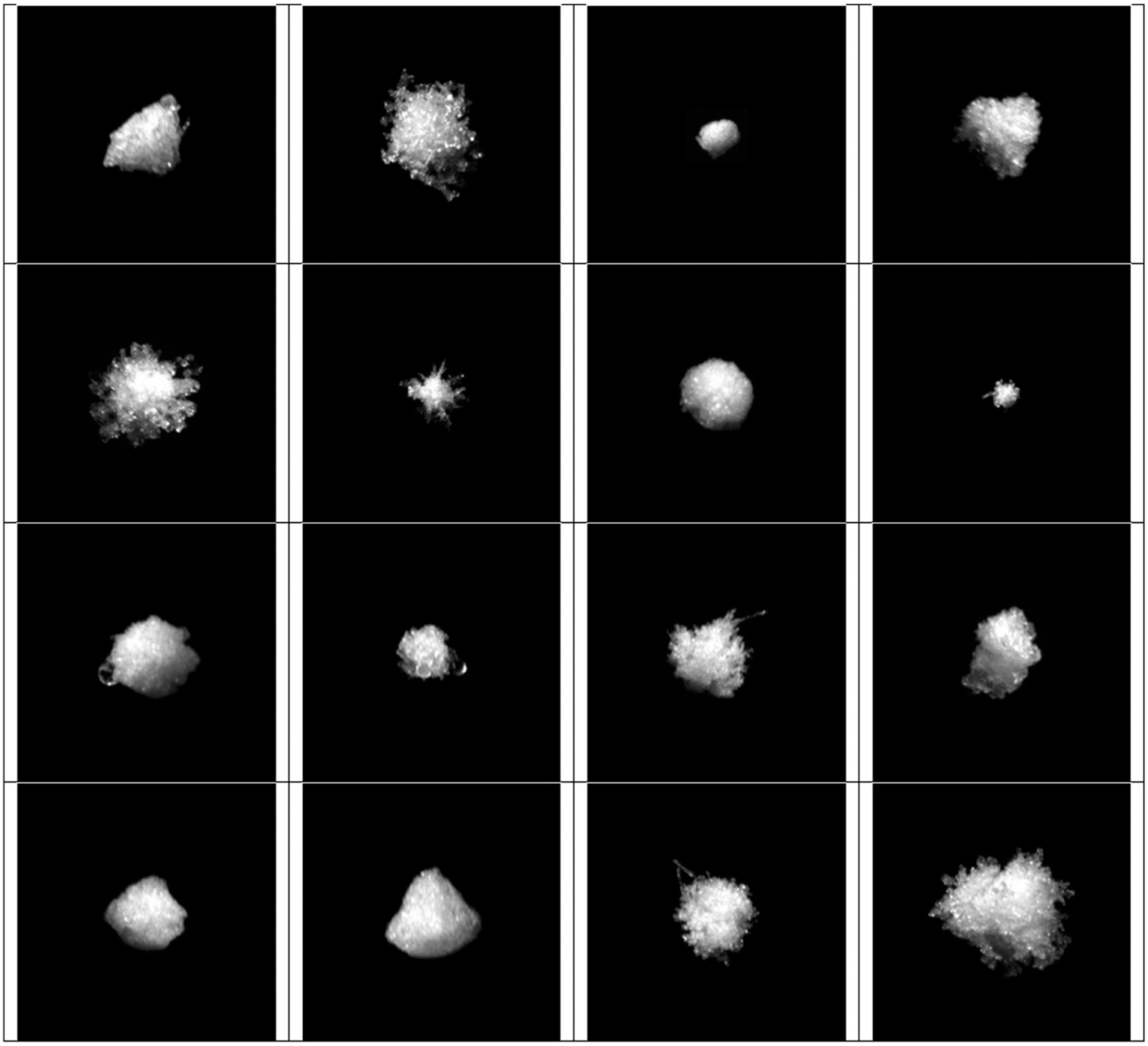


FIG. 9. Examples of image chips in the graupel (GR) class of the final geometric dataset. All image chips in the final geometric dataset had been automatically categorized into the good Q&R category. We included a variety of textures and sizes and also included melting examples when available.

this is based on the subjective opinion of the analyst, rather than a strictly defined size threshold. Simplifications from Magono and Lee (1966) and addition of AG and SP classes resulted in 10 individual categories, of which only 6 were used in Praz et al. (2017) due to data availability: aggregates (AG), small particles (SP), columnar crystals (CC), planar crystals (PC), combination of columnar and planar crystals (CPC), and graupel (GR). As in Hicks and Notaroš (2019), we chose to exclude the CPC class from the present study due to data availability. We found only a few hundred clear examples of CPC in the good Q&R class. CPC appeared far less commonly than the next rarest class, GR, which had several thousand good Q&R examples. Image chips that fell into unconsidered categories, like CPC, we simply omitted from consideration for the present work.

c. *Building the geometric dataset*

Our goal in collecting the geometric dataset for the present work was to establish a large, highly varied collection of image chips in each of the five categories considered. Deep neural networks, like that used in Hicks and Notaroš (2019) and the present work can achieve high accuracies but require substantial training data to avoid overfitting (Simonyan and Zisserman 2015; Szegedy et al. 2015). With tens of millions of parameters, deep CNNs like the ResNet-50 architecture (He et al. 2016) can store substantial quantities of information to learn highly complicated associations and trends (Zeiler and Fergus 2014).

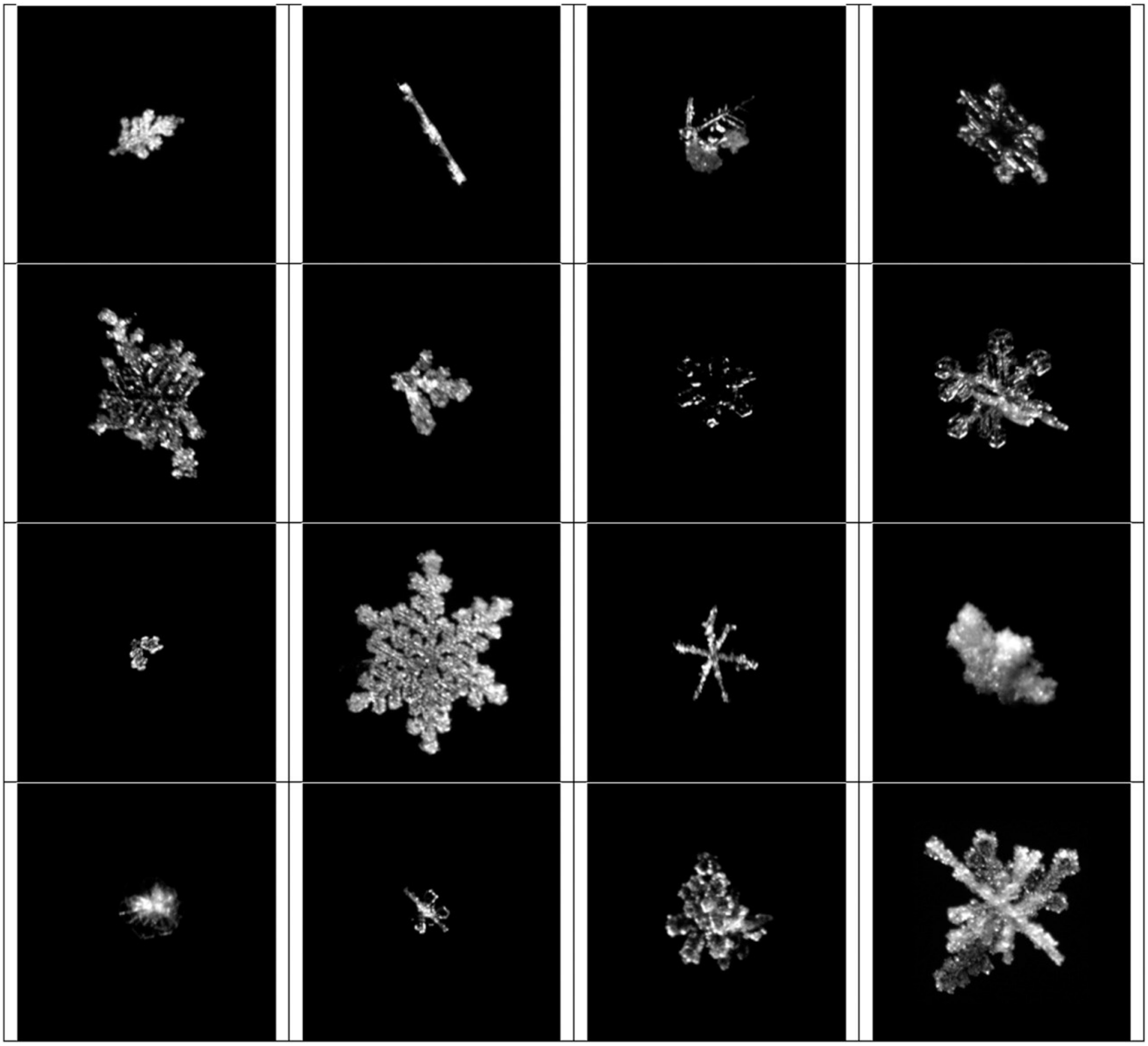


FIG. 10. Examples of image chips in the planar crystal (PC) class of the final geometric dataset. All image chips in the final geometric dataset had been automatically categorized into the good Q&R category. We included difficult examples like row 1, column 2 where possible to help differentiate such PC cases from CC examples. Emphasis was also placed on including examples that lacked easily identifiable sixfold symmetry.

Care must therefore be taken to train such networks on large enough datasets that they cannot simply memorize associations between specific images and their labels or extract spurious trends.

Another important consideration is balance between classes during training. Unless special precautions such as class-specific learning rates are used (not used in the present study), training a neural network on a dataset biased toward a particular class will often bias the network toward that class. As an extreme example, consider a network trained on a dataset of 900 GR images and 100 PC images; the network can attain 90% accuracy on the training set simply by learning to label every image as GR. It is therefore important to present the network with roughly equal numbers of examples in each class during training.

To account for these factors, we limited the number of examples in our geometric dataset for each class to the maximum number of good Q&R examples we could find for the rarest class considered. After CPC (not considered), GR was the rarest class, for which we could only find roughly 5000 examples. Accordingly, we collected roughly 5000 examples of each of the other classes considered, for a total of 25 199 examples. Exact image chip counts per class are presented in Table 2. Figures 7–11 show representative examples from the final AG, CC, GR, PC, and SP sets, respectively. When collecting examples for each class, we put emphasis not only on

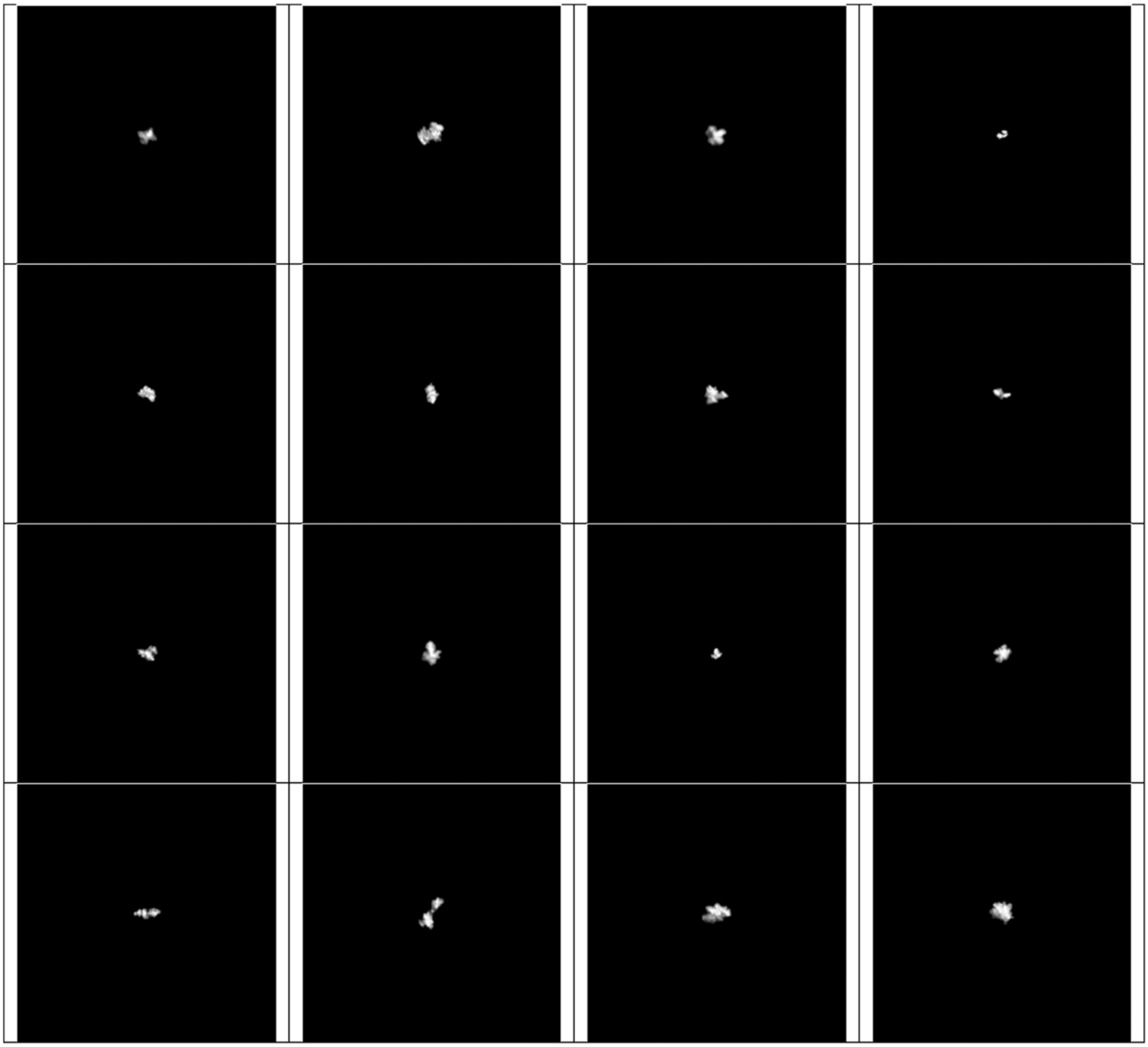


Fig. 11. Examples of image chips in the small particle (SP) class of the final geometric dataset. All image chips in the final geometric dataset had been automatically categorized into the good Q&R category. As small particles are, by definition, particles with features too small to classify, there is little interesting variety among the collected examples other than various shapes and degrees of riming.

archetypical examples, but also examples we considered good counterexamples to possible oversimplifications of each class: e.g., AGs are always large, PCs always have sixfold symmetry, or GR always has a smooth outline. Image chips were not included in the geometric dataset if we could not determine an appropriate label based on information present in the image chip alone, i.e., no multiangle information was used during manual sorting. We note overall that there is an inherent subjectivity in identification of snowflakes in single-view images, especially for classes like GR (Fig. 9), for which distinguishing from other heavily rimed particles is subjective, and SP (Fig. 11), for which deciding unrecognizability of features due to small size is highly subjective. We did not avoid using backlit examples where available, although these were rare, only occurring where a snow particle was imaged while falling in front of a sufficiently bright glare point in the background. Due to their rarity, inclusion of backlit cases likely did not have a substantial impact on accuracy of the trained network. Our analyst recollects seeing at most a dozen backlit cases during manual classification, but such cases were assigned no special designation or identifying information that would make quantification of their impact possible without another manual review of the dataset.

4. Convolutional neural networks methodology

A brief discussion of the network architecture is presented in this section. We also present a summary of the training method

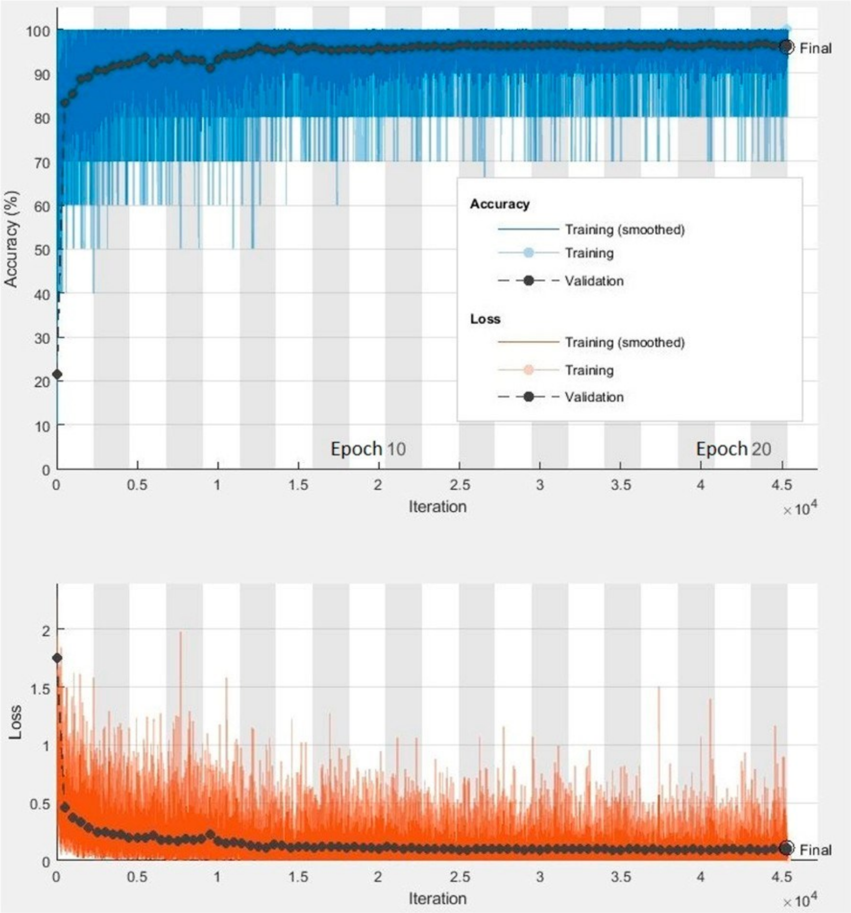


FIG. 12. Training progress for an example training run using the methods and hyperparameters described in section 4b.

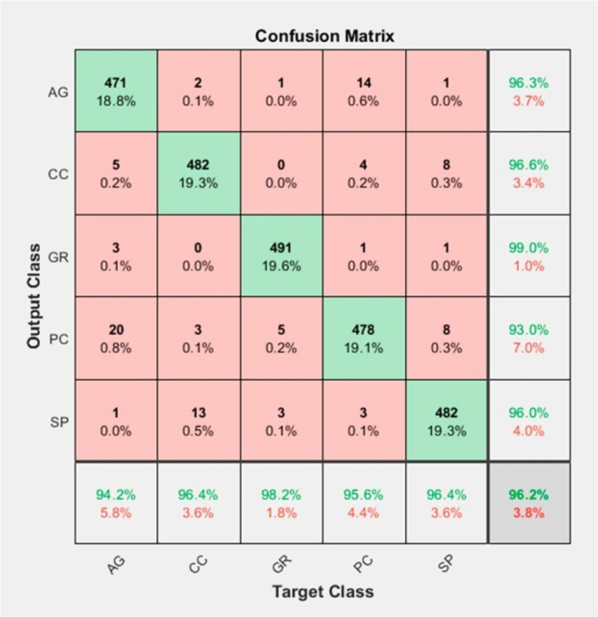
and hyperparameters used. Note that, although the network architecture remains the same as that in our previous work (Hicks and Notaroš 2019), hyperparameters for training differ.

a. Neural network architecture

We used an identical ResNet-50 architecture to that in Hicks and Notaroš (2019). The ResNet-50 architecture has been demonstrated as an excellent balance between speed and accuracy for image classification tasks and is described in detail in He et al. (2016). The residual approach, in general, was groundbreaking at the time of its publication, as it presented an elegant solution to the vanishing gradient problem that had previously limited scaling of CNN accuracy with increased depth. The use of residual connections (or similar), as described in He et al. (2016) has since been widely adopted by deep learning researchers and practitioners. As in Hicks and Notaroš (2019), we used a ResNet-50 model that had been pretrained for general image classification on the ImageNet database (Russakovsky et al. 2015). We also experimented with randomly initialized (no pretraining) versions of the same architecture but found no substantial benefit. We therefore chose to only focus on the pretrained model for the present work for easy comparison with Hicks and Notaroš (2019). A necessary change made to the architecture was reduction in the number of outputs of the final, fully connected layer for our substantially lower number of classes (the original ResNet-50 architecture trained on ImageNet had 1000 classes, not 5). Weights in the modified fully connected layer were initialized randomly.

b. Training method and hyperparameters

As in Hicks and Notaroš (2019), network performance was determined by cross-entropy error, and network weights and biases were optimized by stochastic gradient descent to minimize this loss function. For training, validation, and testing, we again limited the number of examples used in each class to the number of examples available in the smallest class (in the present work, GR with a total of 5000 hand-classified image chips available). The examples used from classes with raw counts larger than the minimum were drawn randomly. We again used a minibatch size of 10. Beyond this, we made several changes to the hyperparameters and training method used in Hicks and Notaroš (2019). Our dataset was also substantially larger; the testing



	AG	CC	GR	PC	SP	
AG	471 18.8%	2 0.1%	1 0.0%	14 0.6%	1 0.0%	96.3% 3.7%
CC	5 0.2%	482 19.3%	0 0.0%	4 0.2%	8 0.3%	96.6% 3.4%
GR	3 0.1%	0 0.0%	491 19.6%	1 0.0%	1 0.0%	99.0% 1.0%
PC	20 0.8%	3 0.1%	5 0.2%	478 19.1%	8 0.3%	93.0% 7.0%
SP	1 0.0%	13 0.5%	3 0.1%	3 0.1%	482 19.3%	96.0% 4.0%
	94.2% 5.8%	96.4% 3.6%	98.2% 1.8%	95.6% 4.4%	96.4% 3.6%	96.2% 3.8%

FIG. 13. Confusion matrix for the network trained in Fig. 12 applied to the test set. Each red or green cell corresponds to a target class (horizontal) and output class (vertical). Row 2, column 1, for instance, shows that five image chips in the test set with target class AG were assigned to the CC class by the trained network, and this corresponded to 0.2% of the entire dataset. The first five cells of the bottom row show accuracy (green) and error (red) for each target class. Row 6, column 1, for instance, shows that, of image chips in the test set with target class AG, 94.2% were classified correctly by the network while 5.8% were classified incorrectly. The first five cells of the rightmost column similarly show accuracy and error for each output class. Row 1, column 6, for instance, shows that, of image chips assigned by the network to the AG class, 96.3% were classified correctly while 3.7% were classified incorrectly. An overall network accuracy (all classes) of 96.2% is shown in the bottom right cell. AG and PC were the most confused classes.

set alone, in this case, was comparable in size to the entire geometric dataset used for Hicks and Notaroš (2019), roughly 1450 examples. In the present study, we randomly selected 500 examples from each class for a total of 2500 testing examples. The remaining 22 699 examples were randomly partitioned into a training set (;90%) and a validation set (;10%), both evenly distributed among the classes studied. The random partitioning between training and validation was unique to each training run. Only the training and validation sets were used for hyperparameter tuning, which was performed by a mix of expert hand tuning and small parametric sweeps and included tuning of the minibatch size, learning rate, and number of training epochs. We also trained for substantially longer than our previous work, training for a total of 20 epochs, as opposed to 10. The training set was shuffled (reordered) randomly every epoch. An epoch is defined as one complete pass through the training set, so, the present training dataset containing many more examples than that available in Hicks and Notaroš (2019), this corresponds to roughly a thirtyfold increase in training time. We were able to extend the training time substantially due to prevention of overfitting by the larger training dataset used in the present work. As opposed to the constant learning rate of 0.0003 used in Hicks and Notaroš (2019), we began with a learning rate of 0.001, which was then scaled by a factor of 1/10 every five epochs. We found this led to a small but noticeable improvement in final network accuracy. We expect improvements in network accuracy could be further improved with additional hyperparameter tuning using more compute resources for large parametric sweeps.

5. Results and discussion

This section presents and discusses the performance of the trained classification networks on the test dataset. The final mean test accuracy achieved was 96.23% with a standard deviation of 0.29% across 10 training runs, the individual test accuracies of which are presented in Table 3. Only the order in which images were presented to the network and random partitioning of nontest images between training and validation differed between training runs. We expect we could have achieved even higher accuracy if we had limited our dataset to only archetypal examples, but this would have diminished the usefulness of the dataset and resulting trained model for general snowfall classification tasks.

Figure 12 shows accuracy and loss of a typical trained network (test accuracy close to the mean) on the training and validation set with respect to training iteration (and epoch, indicated by alternating vertical bands) for a typical training run. There is no evidence of overfitting, and validation accuracy increased nearly monotonically with iteration count. Overfitting, if present, would be apparent in Fig. 12 as divergence of the black validation accuracy and blue training accuracy curves. For the training run shown, the network achieved a validation accuracy of 96.1% and a test accuracy of 96.2%. We suspect the much larger size of the geometric dataset is the dominant factor in improving performance over our previous work but did not have sufficient compute time to perform a full parametric sweep to confirm this. We found that network performance on the validation and training sets were comparable, indicating that training, testing, and validation datasets all sampled the underlying distribution of snowflake geometries well. The validation accuracy standard deviation for the 10 example runs shown in Table 3 was 0.42%, and their mean validation accuracy was 96.26%. We attribute the larger validation accuracy standard deviation, as compared to the test accuracy standard deviation, to random selection of the validation set for each training run (the test set did not change between runs). There was little variation between training runs, with the only nominal differences due to this random partitioning of the validation and training sets as well as random reordering of the training set during each epoch. Figure 13 shows a confusion matrix for the same network, the training progress of which is shown in Fig. 12.

In general, trained networks would confuse PC and AG classes most often. We included many difficult examples in the AG class that featured a prominent planar crystal with several

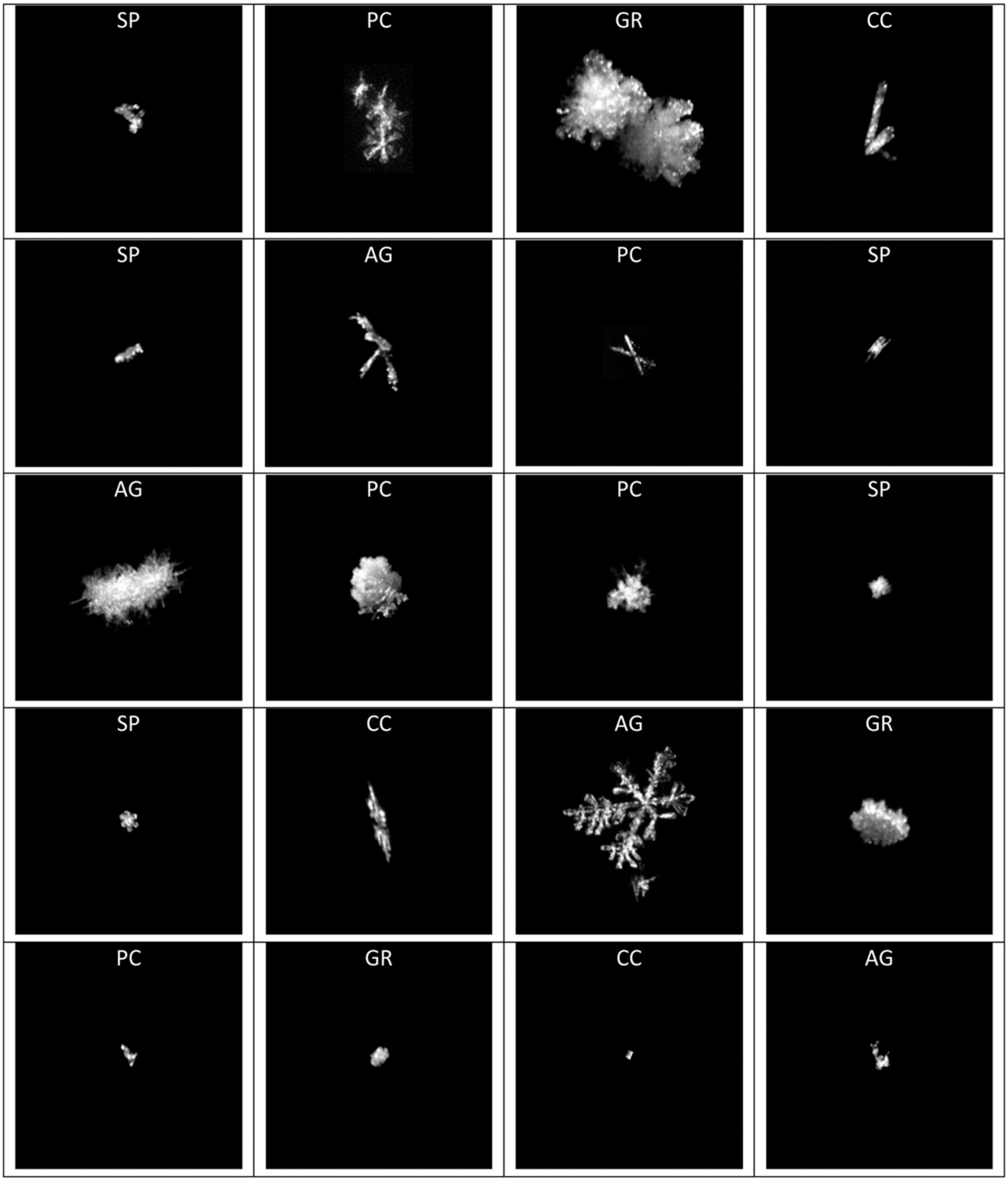


FIG. 14. Examples of image chips misclassified by a trained network: (top to bottom) misclassified aggregates, misclassified columnar crystals, misclassified graupel, misclassified planar crystals, and misclassified small particles are shown with the label assigned by the network overlaid for each image chip.

less-prominent particles that had adhered due to midair collisions, so confusion between the two classes seems understandable to us. Figure 14 presents examples of image chips misclassified by the typical network from Figs. 12 and 13. Overall, most misclassifications appear to be blatant errors due to imperfection of the trained model, but several stand out as ambiguous cases or possibly even human error. Figure 14, row 2, column 2, for instance, was assigned by the network to the AG class, having been human labeled as a columnar crystal. Further inspection indicates this snowflake may indeed be a simple aggregate or even a malformed planar crystal, suggesting this misclassification is due human error rather than network error. Figure 14, row 4, column 3, shows a clear planar crystal adhered to a small aggregate of columnar crystals. Although the planar crystal dominates the image chip, the aggregation present indicates the network is correct to assign this image chip to the AG class. Figure 14, row 3, column 4, and row 5, column 2, respectively, show a GR image chip misclassified as SP and a SP image chip misclassified as GR, respectively. These two cases show the ambiguity of the SP class and the difficulty of drawing a distinction between small GR flakes and relatively large, round SP flakes. Figure 14, row 5, column 3, shows another ambiguous case. Human classified as SP but network classified as CC, this particle shows possible CC-like features (dominant uniaxial crystal growth) but is barely too small for our analyst to assign confidently to the CC category.

6. Conclusions

This paper has presented improvements over our previous approach (Hicks and Notaroš 2019) to automated winter hydrometeor classification using deep convolutional neural networks. Using improved training methods and a substantially larger and more complicated dataset from many more snow events than in our previous study, we were able to achieve over 96.2% accuracy on a test set of 2500 images. We consider this result substantial for several reasons. The MASC is a high-throughput sensor, collecting tens to hundreds of thousands of detectable snowflake images during a winter storm event, so even small accuracy improvements lead to a substantial reduction in the total number of misclassified snowflake images. Namely, this is a ∼40% reduction in the fraction of incorrectly classified snowflakes relative to the already very high geometric classification accuracy result reported in our previous work and corresponds to a 2.8% increase in overall accuracy. Even more importantly, the dataset of 25 199 image chips sorted by geometric class used in the present study differs substantially from that developed for Hicks and Notaroš (2019). As a proof of concept study, Hicks and Notaroš (2019) used a geometric dataset focused on easily identifiable examples of each of the snowflake classes considered. To demonstrate the broader usefulness of deep CNNs for automated snowfall classification, the dataset used in the present study is not only larger but also contains wider in-class variety. In using such a dataset, we have shown that, with a few modifications to the network training process, the geometric classification method described in Hicks and Notaroš (2019) can achieve higher accuracy on a vastly more challenging dataset. Finally, the paper has presented several important components of the CNN-based, supervised approach to snowflake classification, including an improved training method and hyperparameters for training; new automated techniques for snowflake detection, cropping, and normalization of snowflake images; and new quality and recognizability preprocessing of image data. The described methodologies and techniques may be of great use to researchers and practitioners applying the same or similar approaches to hydrometeor classification based on the images collected by the MASC or another image-based particle recording instrument or system.

Acknowledgments. This work was supported in part by the National Science Foundation under Grants AGS-1344862 and AGS-2029806.

Data availability statement. The dataset of MASC images generated for and used in this study has been made publicly available at Key et al. (2021).

REFERENCES

Bringi, V. N., P. C. Kennedy, G.-J. Huang, C. Kleinkort, M. Thurai, and B. M. Notaroš, 2017: Dual-polarized radar and surface observations of a winter graupel shower with negative Z_{dr} column. *J. Appl. Meteor. Climatol.*, 56, 455–470, https://doi.org/10.1175/JAMC-D-16-0197.1.

Grazioli, J., D. Tuia, S. Monhart, M. Schneebeli, T. Raupach, and A. Berne, 2014: Hydrometeor classification from two-dimensional video disdrometer data. *Atmos. Meas. Tech.*, 7, 2869–2882, https://doi.org/10.5194/amt-7-2869-2014.

He, K., X. Zhang, S. Ren, and J. Sun, 2016: Deep residual learning for image recognition. *2016 IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, IEEE, 770–778, https://doi.org/10.1109/CVPR.2016.90.

Hicks, A., and B. M. Notaroš, 2019: Method for classification of snowflakes based on images by a Multi-Angle Snowflake Camera using convolutional neural networks. *J. Atmos. Oceanic Technol.*, 36, 2267–2282, https://doi.org/10.1175/JTECH-D-19-0055.1.

Kennedy, P., M. Thurai, C. Praz, V. N. Bringi, A. Berne, and B. M. Notaroš, 2018: Variations in snow crystal riming and Z_{DR}: A case analysis. *J. Appl. Meteor. Climatol.*, 57, 695–707, https://doi.org/10.1175/JAMC-D-17-0068.1.

Key, C., A. Hicks, and B. Notaroš, 2021: Colorado State University geometric snowflake classification dataset, version 1.0. Zenodo, accessed 5 March 2021, https://doi.org/10.5281/zenodo.4584200.

Kleinkort, C., G.-J. Huang, V. N. Bringi, and B. M. Notaroš, 2017: Visual hull method for realistic 3D particle shape reconstruction based on high-resolution photographs of snowflakes in free fall from multiple views. *J. Atmos. Oceanic Technol.*, 34, 679–702, https://doi.org/10.1175/JTECH-D-16-0099.1.

Korolev, A., and B. Sussman, 2000: A technique for habit classification of cloud particles. *J. Atmos. Oceanic Technol.*, 17, 1048–1057, https://doi.org/10.1175/1520-0426(2000)017<1048:ATFHCO>2.0.CO;2.

Leinonen, J., and A. Berne, 2020: Unsupervised classification of snowflake images using a general adversarial network and *K*-medoids classification. *Atmos. Meas. Tech.*, 13, 2949–2964, https://doi.org/10.5194/amt-13-2949-2020.

Libbrecht, K. G., 2017: Physical dynamics of ice crystal growth. *Annu. Rev. Mat. Res.*, 47, 271–295, https://doi.org/10.1146/annurev-matsci-070616-124135.

Lindqvist, H., K. Muinonen, T. Nousiainen, J. Um, G. McFarquhar, P. Haapanala, R. Makkonen, and H. Hakkarainen, 2012: Ice-cloud particle habit classification using principal components. *J. Geophys. Res.*, 117, D16206, https://doi.org/10.1029/2012JD017573.

Magono, C., and C. W. Lee, 1966: Meteorological classification of natural snow crystals. *J. Fac. Sci. Hokkaido Univ. Ser. 7*, 2, 321–335.

Nakaya, U., and Y. Sekido, 1936: General classification of snow crystals and their frequency of occurrence. *J. Fac. Sci. Hokkaido Univ. Ser. 2*, 1, 243–264.

Newman, A. J., P. A. Kucera, and L. F. Bliven, 2009: Presenting the Snowflake Video Imager (SVI). *J. Atmos. Oceanic Technol.*, 26, 167–179, https://doi.org/10.1175/2008JTECHA1148.1.

Notaroš, B. M., and Coauthors, 2016: Accurate characterization of winter precipitation using Multi-Angle Snowflake Camera, visual hull, advanced scattering methods and polarimetric radar. *Atmosphere*, 7, 81–111, https://doi.org/10.3390/atmos7060081.

Praz, C., R. Yves-Alain, and A. Berne, 2017: Solid hydrometeor classification and riming degree estimation from pictures collected with a Multi-Angle Snowflake Camera. *Atmos. Meas. Tech.*, 10, 1335–1357, https://doi.org/10.5194/amt-10-1335-2017.

Russakovsky, O., and Coauthors, 2015: ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vision*, 115, 211–252, https://doi.org/10.1007/s11263-015-0816-y.

Schönhuber, M., G. Lammer, and W. Randeu, 2008: The 2D video disdrometer. *Precipitation: Advances in Measurement, Estimation and Prediction*, S. Michaelides, Ed., Springer, 3–31.

Simonyan, K., and A. Zisserman, 2015: Very deep convolutional networks for large-scale image recognition. *Int. Conf. on Learning Representations*, San Diego, CA, ICLR.

Straka, J., D. S. Zrnić, and A. V. Ryzhkov, 2000: Bulk hydrometeor classification and quantification using polarimetric radar data: Synthesis of relations. *J. Appl. Meteor.*, 39, 1341–1372, https://doi.org/10.1175/1520-0450(2000)039<1341:BHCAQU>2.0.CO;2.

Szegedy, C., and Coauthors, 2015: Going deeper with convolutions. *2015 IEEE Conf. on Computer Vision and Pattern Recognition*, Boston, MA, IEEE, https://doi.org/10.1109/CVPR.2015.7298594.

Vazquez-Martin, S., T. Kuhn, and S. Eliasson, 2020: Shape dependence of falling snow crystals' microphysical properties using an updated shape classification. *Appl. Sci.*, 10, 1163, https://doi.org/10.3390/app10031163.

Zeiler, M. D., and R. Fergus, 2014: Visualizing and understanding convolutional neural networks. *European Conf. on Computer Vision*, Zurich, Switzerland, ECCV, 818–833, https://doi.org/10.1007/978-3-319-10590-1_53.

Zhang, G., S. Luchs, A. Ryzhkov, M. Xue, L. Ryzhkova, and Q. Cao, 2011: Winter precipitation microphysics characterized by polarimetric radar and video disdrometer observations in Central Oklahoma. *J. Appl. Meteor. Climatol.*, 50, 1558–1570, https://doi.org/10.1175/2011JAMC2343.1.