OPEN ACCESS
IOP Publishing
Measurement Science and Technology
Meas. Sci. Technol. **34** (2023) 105203 (16pp)
https://doi.org/10.1088/1361-6501/ace070

A deep neural network architecture for reliable 3D position and size determination for Lagrangian particle tracking using a single camera

M Ratz[1], S Sachs[1], J König[1] and C Cierpka[1,2,*]

[1] Institute of Thermodynamics and Fluid Mechanics, Technische Universität Ilmenau, 98693 Ilmenau, Germany
[2] Visiting research fellow, Department of Biomedical Engineering, Lund University, 22100 Lund, Sweden

E-mail: christian.cierpka@tu-ilmenau.de

Received 11 December 2022, revised 12 April 2023
Accepted for publication 21 June 2023
Published 30 June 2023

Abstract

Microfluidic flows feature typically fully three-dimensional velocity fields. However, often the optical access for measurements is limited. Astigmatism or defocus particle tracking velocimetry is a technique that enables the 3D position determination of individual particles by the analysis of astigmatic/defocused particle images. The classification and position determination of particles is a task well suited to deep neural networks (DNNs). In this work, two DNNs are used to extract the class and in-plane position (object detection) as well as the depth position (regression). The performance of both DNNs is assessed by the position uncertainties as well as the precision of the size classes and the amount of recalled particles. The DNNs are evaluated on a synthetic dataset and establish a new benchmark of DNNs in defocus tracking applications. The recall is higher than compared to classic methods and the in-plane errors are always subpixel accurate. The relative uncertainty in the depth position is below 1% for all examined particle seeding concentrations. Additionally, the performance on experimental images, using four different particle sizes, ranging from 1.14 µm to 5.03 µm is analyzed. The particle images are systematically rearranged to produce comprehensive datasets of varying particle seeding concentrations. The distinction between particles of similar size is more challenging but the DNNs still show very good results. A precision above 96% is reached with a high recall above 95%. The error in the depth position remains below 1% and the in-plane errors are subpixel accurate with respect to the labels. The work shows that first, DNNs can be trained with artificially rearranged data sets based on individual experimental images and are therefore easily adaptable to various experimental setups and applicable by non-experts. Second, the DNNs can be successfully adapted to determine additional variables as in this case the size of the suspended particles.

* Author to whom any correspondence should be addressed.

Keywords: astigmatism particle tracking velocimetry, Lagrangian particle tracking, size recognition, deep neural network, microfluidics, size classification

(Some figures may appear in colour only in the online journal)

1. Introduction

Flows with suspended particles of multiple sizes are important for many industrial applications. Examples for this are mixing or separation processes. In the case of microfluidics, this includes the manufacturing of pharmaceutical products or the analysis of chemical and biological samples [1]. To analyze the underlying processes, multiple techniques were developed that allow a classification of the individual particles based on their size. However, this is always only achieved with additional efforts, either by an increased manufacturing cost of the microchannels or additional equipment for the measurement [2–6]. In recent years, the measurement methods were extended to allow a classification of the particles based on their particle image from flow measurements [7, 8].

In microfluidics, the separation or mixing of particles is of common interest which is why all three components of the three dimensional flow field are often measured. Due to the limited optical access, methods using multiple perspectives for the position reconstruction [9] do not work or have high uncertainties [10]. Because of this, the astigmatism particle tracking velocimetry (APTV) was developed. It is a defocus based particle tracking measurement technique which allows the determination of all three velocity components in a volume. For this, a cylindrical lens is utilized to introduce astigmatism into the optical setup [11]. This lens distorts the particle image from a circle to an ellipse. The shape of this ellipse depends on the depth position of the particle, relative to the focal planes of the optical system. After a segmentation algorithm is applied to find potential candidates of particles, their center positions are found by means of a 1D Gaussian fit along both axis which results in a subpixel accurate detection [10]. The depth position is found by matching the particle image shape to a calibration function [12]. A common challenge of APTV measurements are spherical aberrations caused in particular by the cylindrical lens, which result in a further distortion of the intensity profile based on the position of the particle image within the image [13]. Hence, higher position uncertainties are expected that can only be compensated with more complex and time consuming algorithms, e.g. a correlation based evaluation approach [14].

This is a field in which deep learning can be applied because it requires no specific knowledge about the setup and the aberrations can be used if they are systematic. In recent years, machine learning and especially deep learning has received a surge of interest, starting with the ImageNet challenge [15] which is a large dataset containing examples of 1000 different classes of objects. Through the availability of stronger graphics processing units (GPUs) and more sophisticated algorithms, convolutional neural networks (CNNs) were used to great effect. CNNs apply multiple convolutions in so-called layers to extract features from the images and enhance the semantic information. Deep neural networks (DNNs) have since surpassed any classic image processing in classification tasks [16]. Large annotated datasets were then used to develop different architectures and training algorithms to achieve rapid improvements [17]. Since then, state-of-the-art results were achieved with deep learning in many other fields, such as the complex game of Go [18] or lip reading [19].

Different machine learning algorithms have also been applied to the field of defocus based particle tracking. For this, characteristics of the segmented particle images, such as the height and width of the semiaxes are extracted with classic image processing and then fed into a shallow neural network with few layers to extract the depth position [20, 21]. Another approach uses a DNN into which the cropped, individual images are fed to also obtain the depth position [22]. However, the field of deep learning offers another opportunity. The classic segmentation and determination of the center position is a task which is commonly known as object detection. Many different object detection pipelines have been developed, often using a DNN for the extraction of features which are then further processed. Object detection pipelines are commonly benchmarked on the COCO dataset [23], which consists of labeled images containing objects from 91 classes. Similarly, the detection and classification of particle images constitutes an object detection problem. Different object detection pipelines have been successfully applied to defocus particle tracking measurements as well [24, 25]. The depth position of the particles is then again found by matching them to the aforementioned calibration function. However, these investigations only used a monomodal size distribution, meaning all particles have the same size and no classification was done.

A different approach is to combine the deep learning method for both the in-plane localization of the particles and the determination of the depth position by cascading two DNNs. First, the in-plane position of the particles is found by means of an object detection network. Second, the individual particles are cropped and their depth position is determined by another DNN. This approach led to higher uncertainties in all three axes compared to classic methods [13, 26]. From a retrospective this is not surprising for the in-plane errors as the resolution of feature maps in later layers is strongly reduced compared to the original image. This downsampling leads to a lower spatial resolution which causes higher uncertainties. Regarding the uncertainty in the depth position, no definitive reason can be given. However, one surprising result was the fact that using particles of multiple different sizes did not increase the uncertainty in the depth position although a narrow monomodal size distribution is one assumption for classic

APTV [13]. A classification with DNNs was carried out in a recent study by Sachs *et al* [8], where the DNNs outperformed the classic methods in terms of size determination.

This work aims to improve the existing techniques for the particle position determination based on APTV measurements using two DNNs while also featuring a size classification of individual particles. Therefore, an object detection network is again used for the in-plane location and size classification of the particles and a regression network is employed to determine the depth position. The existing methods are extended to surpass the state-of-the-art results on synthetic images [26]. Furthermore, experimental images of particles of multiple different sizes are acquired to generate a feature-rich dataset that is used for the training of the two DNNs. The individual particle images are rearranged on synthetic images to generate comprehensive datasets with different particle seeding concentrations which present a complex task w.r.t. the localization and classification of the particles. This approach allows on the one hand a fast an flexible adaptation of the approach and the training data for different size distributions. On the other hand, limitations in terms of particle image density, signal-to-noise ratio and other experimental parameters can be assessed thoroughly.

The experimental setup used to acquire the images which are used to generate the training images is described in section 2. Section 3 is dedicated to the generation of the synthetic and the experimental datasets. The results of the particle localization and classification are given in section 4 and conclusions are drawn in section 5.

2. Experimental setup

As mentioned before, DNNs require lots of training examples to achieve state-of-the-art results. In previous studies, the training data was created with synthetically generated images [25–27]. However, the synthetic particle images rely on some model functions. Most often Gaussian intensity distribution [28] are assumed or ray-tracing is done using simplified optical systems as the real optical systems are not known in the case of proprietary lens design or complex lens shapes or manual distance adjustments [29]. For these reasons, the models often do not reflect the real images that show a much greater variety of distortions which stem from aberrations of the optical system. These distortions have to be taken into account for position correction using classical methods or even hinder classical approaches to be successfully used [12, 26, 30]. For neural networks they can be considered as features as they are not random noise but deterministic and show some dependence on the position within the image. In this respect the additional distortions even help to improve the results by the network. Therefore, a dataset of experimental particle images was acquired in addition to the synthetic datasets, which is described in section 3.1. Spherical, fluorescent particles (PS-FluoRed, MicroParticles GmbH) of four different diameters d_p were used; (1.14 ± 0.03) µm, (2.47 ± 0.08) µm, (3.16 ± 0.07) µm and (5.03 ± 0.07) µm.

The particles were suspended and sedimented in deionized water confined in a microfluidic chamber made of polydimethylsiloxane. The chamber was placed on a slide made of 128° YX $LiNbO_3$. This piezoelectric substrate is used in applications where the separation of particle mixes is of interest [4, 5, 31, 32]. The chamber was covered with a glass slide to avoid evaporation during the sedimentation process and the measurements.

For the image acquisition, the chamber was placed on top of an inverted microscope (Axio Observer 7, Zeiss GmbH) with a neofluar objective (M20x, NA = 0.4, Zeiss GmbH). A modulatable OPSL laser (tarm laser technologies tlt GmbH & Co. KG) was used to achieve a volumetric illumination of the microfluidic chamber. Reflected laser light was suppressed from the optical path to the camera with a dichroic mirror (DMLP567T, Thorlabs Inc) and a long pass filter (FELH0550, 550 nm, Thorlabs Inc). The $LiNbO_3$ substrate is birefringent, meaning that two particle images exist for one given particle. Thus, the second, undesired particle image was suppressed with a polarization filter. Astigmatism was introduced into the system with a cylindrical lens with a focal length of 250 mm that was placed approximately 40 mm in front of an sCMOS camera (LaVision GmbH, 16 bit, 2560 × 2160 px). For further details about the measurement setup, see [31].

The sedimentation process was carried out independently for all particle sizes except the smallest one, meaning that only one particle size is present in a given experimental image. In the case of the 1 µm particles, the sedimentation happened too slowly because the downward motion is superposed by Brownian motion. This lead to some particles never reaching the bottom of the chamber, which causes not well known particle positions (see section 3.2). Instead, the particles were dried off on the piezoelectric substrate. This was achieved by placing a small droplet with a volume of approximately 50 µl onto the substrate and then pulling it back into the syringe. The remaining film evaporated very quickly which mitigated an agglomeration of particles. The particle images of dried particles are brighter than the ones of sedimented particles. Thus, the power of the laser was reduced until the intensity of the particle images matched the intensity of particles suspended in water with the same diameter of 1.14 µm.

For each particle size, the microfluidic chamber was positioned by a motorized stage in the *xy*-plane and the focus was changed to generate a feature-rich dataset of particle images at multiple 3D positions. The stage was moved in the *xy*-plane in a region of 420 µm × 420 µm, to reduce the uncertainty in the *z*-position due to a possible inclination of the microscopic stage [13]. The focus was changed over different ranges z_{range} of *z*-positions as the intensity of smaller particles is weaker, in particular towards the margin of the measurement volume where the defocus gets too strong. This results in less images for the 1.14 µm, which is compensated by moving the stage in the *xy*-plane with a step size of $\Delta x = \Delta y = 60$ µm. The step size for the other three sizes was 70 µm. Important parameters as well as the total number of images for each size are summarized in the center column of table 1.

Table 1. Summary of the setup parameters for the image acquisition of each particle size.

d_p (μm)	Δx, Δy (μm)	# images	z_{range} (μm)	# particle images
1.14	60	5283	[−40, 40]	100 702
2.47	70	5438	[−55, 55]	157 533
3.16	70	5929	[−60, 60]	144 317
5.03	70	5929	[−60, 60]	89 692

3. Methods

The pipeline of the different steps from experimental images to trained DNNs is illustrated in figure 1. On the left hand side, the image acquisition, labeling and rearranging is described while the training of the DNNs is depicted on the right hand side.

3.1. Preparation of the synthetic datasets

Both DNNs are initially benchmarked on a synthetic dataset to show that the chosen hyperparameters achieve state-of-the-art results. For this, the challenge dataset proposed by Barnkob *et al* [26] is used. This specific dataset consists of synthetic particle images, thus no size classification is performed. For details about the image size, background noise and range of z-values, the reader is referred to the original publication [26].

As the images were generated with MicroSIG [29], their position in all three coordinates is well known. Thus, the left-hand part of the dataset preparation in figure 1 was not required. Instead, the rearranged images were directly created from the synthetic particle images. Datasets of different particle seeding concentrations were generated. The particle seeding concentration can be expressed with the source density N_s, defined as

$$N_s \approx N_p \frac{\overline{A}_p}{A_i}, \tag{1}$$

where N_p is the number of particles in each image, A_i the image area in px and $\overline{A}_p$ the average particle image area in px, taken to be the ellipse provided by MicroSIG. Six datasets of source densities with $N_s \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ were generated, each containing 500 images. These high source densities were used to test the new algorithms against the benchmark data presented in Barnkob *et al* [26] and available online[3]. However, for experimental data the source densities are lower for various reasons outlined in the next section.

For the object detection network, these datasets were used as they are, meaning that only 500 images are available. However, each of these images contains hundreds of labeled objects, resulting in a comprehensive and feature-rich dataset which is learned by the neural network. For the regression network, the individual particle images were cropped to a fixed size of 100×100 px which is large enough to ensure that even large defocused particles are fully contained within this crop. For each source density, a total of 100 000 particle images were cropped in this way. This number corresponds to the number of individual particle images for $N_s = 0.1$. For the other source densities, the number was reduced to obtain the same number of images for the object detection as well as the regression.

[3] https://defocustracking.com/datasets/.

3.2. Preparation of the experimental datasets

For the experimental images, the individual particle images had to be labeled before rearranging them on the synthetic images with labeled, overlapping particle images. The class label was known because experimental images were taken only with one particle size, either sedimented or dried off on the substrate. The depth position of the particles in a given image was taken to be determined by the focus of the microscope. The in-plane positions (x, y) and sizes of the semiaxes (a_x, a_y) of the particles were determined with subpixel accuracy using the classic image processing [11]. Separate calibration curves were obtained for each particle size.

False labels of agglomerated particles or overlapping particle images were removed by applying multiple filter criteria. First, particle images with a Euclidean distance to the calibration function in the $a_x a_y$-plane larger than a global threshold of 7.5 px were removed. In a second step, a local threshold for each depth position was adaptively determined until 90% of the particle images at a certain depth position were declared as valid. For details, see [8]. A final validation step considered the intensity of the particle image [24]. For this, a background subtraction was applied by subtracting 2D second order polynomial fit of the image intensity. An example is illustrated in figure 2. The filter method compares the sum of the intensities I_b within the proposed ellipse, defined by a_x and a_y (blue) to the sum of the intensities of the extended ellipse I_g, which is enlarged by an expansion factor $f = 1.25$ (green). Because of this extension, particle images whose extended ellipse exceeded the margins of the image were removed. The resulting histogram of the intensity ratio $I_r = I_b/(I_g - I_b)$ for all particle sizes is shown in figure 3. Proposed labels with an $I_r < 4$ were rejected, as indicated by the dashed line. This threshold changes slightly for different f. However, the local minimum in the histogram can be easily seen for different values and applied as a filter. The expansion factor has just to be large enough to cover a representative portion of the background and not to be larger than 1.5 so that neighboring particle images contribute to I_g. However, the shape of this histogram is almost independent from the actual particle sizes, so this threshold was applied to all experimental datasets. The resulting number of valid particle images in each experimental dataset is listed in the right column of table 1.

The validated particles were then rearranged on a synthetic new background to generate experimental images with labeled overlap. This has been shown to benefit the training process of neural networks [33]. For this, the particle images need to be cropped accurately to avoid artifacts as well as truncated

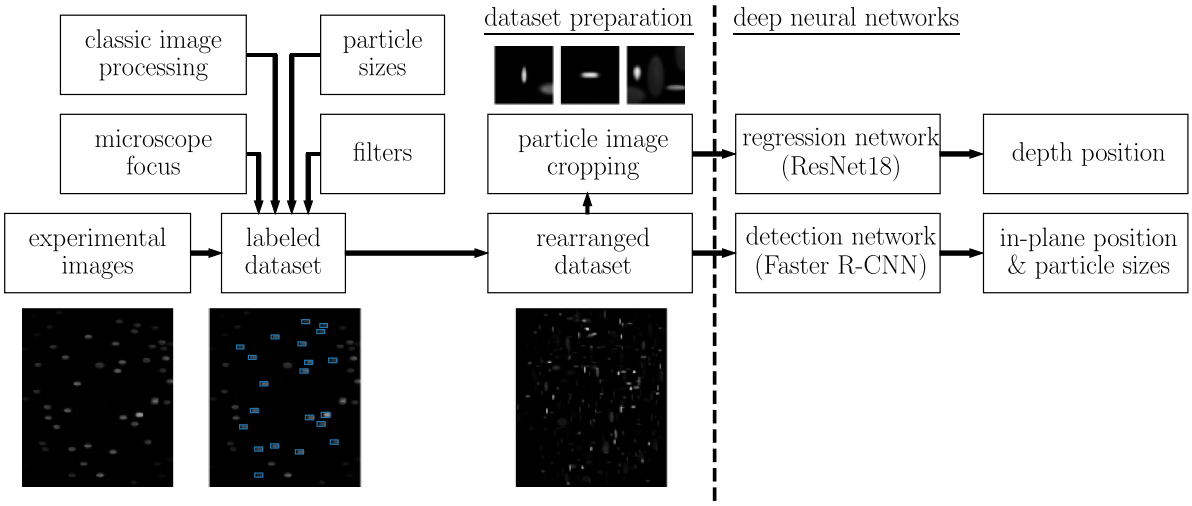


Figure 1. Flowchart explaining the processing pipeline for the 3D position determination and size classification. Particle images are acquired, labeled and then rearranged to generate datasets with different particle seeding concentrations. The in-plane position and class are obtained with an object detection pipeline and the depth position with a regression network.

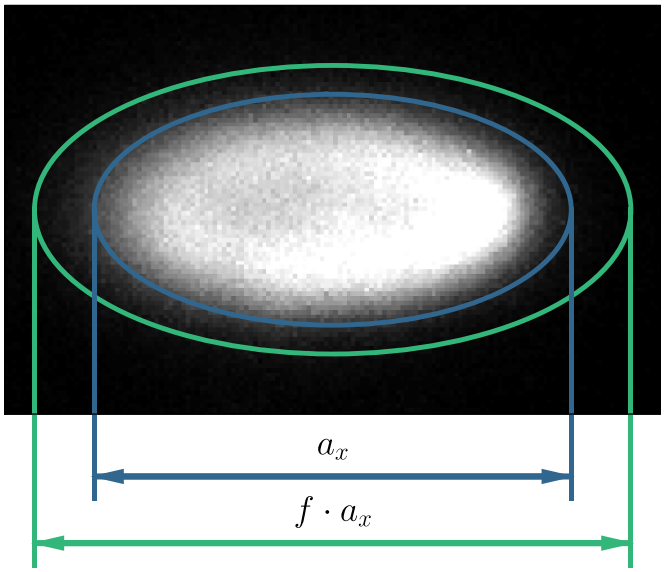


Figure 2. Example of an individual particle image with the associated ellipsis label from the classic image processing (blue). The particle image is cropped around the ellipsis, extended by the expansion factor f (green).

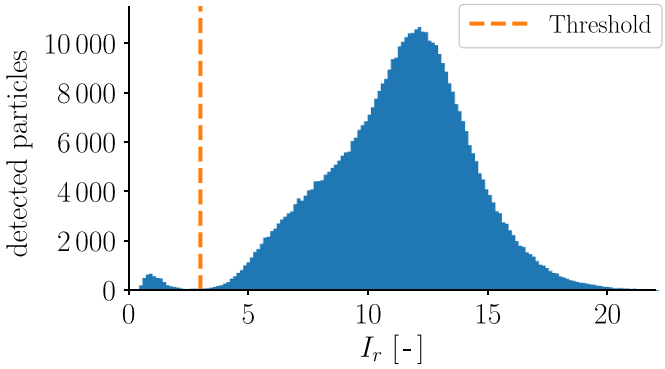


Figure 3. Histogram of the intensity ratio I_r for the multimodal dataset of all four particle sizes. The shown threshold of four is the same for all four particle sizes.

intensity distributions in the newly generated images [34]. This has also been applied to the detection of particle images to great effect [8, 24]. The particle images were cropped from the original image with an ellipse around the center point. This ellipse was again expanded with $f = 1.25$. The cropped particle image was then pasted onto a background of zero intensity at the same in-plane position to retain the local features stemming from optical aberrations [13]. The intensity of overlapping particle images was added. After all particle images were placed in the new image, Gaussian noise with a mean value of 0.5 and a standard deviation of 4.5 was added to simulate noise of the camera sensor.

The four different particle sizes allow a multitude of combinations for the rearranged particle images with multiple particle sizes. The number of possible cases was reduced to the seven cases which always contain the 2.47 µm particles (see table 2). These seven cases can further be divided into four cases which also contain the 3.16 µm particles and three cases which do not. The reason for this is that the distinction between these two narrow particle sizes is expected to be more difficult because of similar features in the particle images.

For each of the seven cases, datasets of four different source densities, i.e. $N_s \in \{0.05, 0.10, 0.15, 0.20\}$ were created. From the rearranged images, the individual particle images for the regression were again cropped based on the provided labels, at a fixed size of 250×250 px to fully capture the defocused particles of the largest size.

Because of the different combinations of test cases and number of particle sizes, the cases have a different amount of individual particle images. This results in a different number of training examples which can lead to different convergence

Table 2. Properties of the seven datasets of the experimental, rearranged images.

Name	Particle sizes	# individual particle images	# regression images
Case 1	1.14, 2.47, 3.16, 5.03 µm	492 244	984 488
Case 2	2.47, 3.16, 5.03 µm	391 542	783 084
Case 3	1.14, 2.47, 3.16 µm	402 552	805 104
Case 4	2.47, 3.16 µm	301 850	905 550
Case 5	1.14, 2.47, 5.03 µm	347 927	1043 781
Case 6	2.47, 5.03 µm	247 225	988 900
Case 7	1.14, 2.47 µm	258 235	774 705

speeds during the training process. To circumvent this, each particle image was placed multiple times in all of the datasets. For this, all particles were first placed once in the new images. Then, the list of particle images was shuffled and all of them were placed a second time to generate a different set of images. For a source density of 0.05, this was repeated until at least 750 000 particle images were placed. If the source density was doubled, each particle was placed twice in order to generate the same number of total images. For the regression network, the individual particle images were cropped from the images with the rearranged particle images. In theory, there are more particle images for higher particle seeding concentrations but the number was kept the same for all source densities because the data handling of millions of individual images is not feasible. The particle sizes, number of individual particle images and number of particle images for the regression are summarized in table 2.

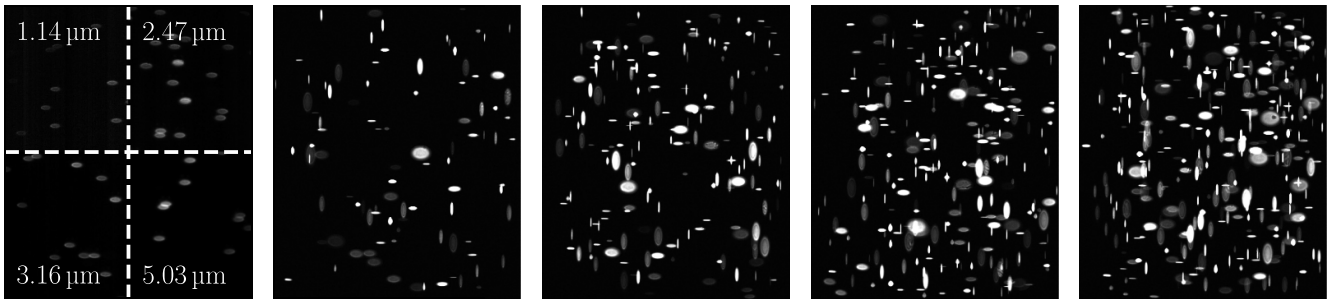
The rearranging of the particle images was carried out with a bit depth of 16 to accurately resolve the gradients when particles are overlapping. However, a large variety of data sets with millions of individual images is investigated in this work. Generating and storing all of these images with a bit depth of 16 is infeasible as it becomes computationally too expensive and the proposed algorithm shall work properly on conventional hardware. Therefore, all images were saved with a bit depth of 8, as this is also typically done for benchmark datasets, such as COCO [23] or ImageNet [15]. For this, the original intensities were rescaled by clipping values above 4000 counts to 4000 counts and applying a linear transformation in between. In this way, defocused particle images of the smallest particle size are still resolved by multiple intensity counts. A collage of raw experimental images of the four particle sizes is shown in figure 4(a). Corresponding, rearranged images for Case 1 are shown in figures 4(b)–(e) for the four different source densities. An increasing amount of overlap becomes visible, especially compared to the raw images. For visualization purposes, the intensity in the images is clipped to also show defocused particles. In comparison to the purely synthetic datasets, the source density was limited to $N_s = 0.2$ for the experimental case. There are three main reasons for this. Firstly, large particles images result due to the defocus. Therefore, they cover a large part of the image and can even obscure particles which are located behind them. If several particles are overlapping, this can become problematic and also lead to a saturation of individual pixels on the camera sensor. Secondly, if a particle tracking algorithm is later used to determine the velocity field, it is recommended to have a larger mean inter particle spacing in comparison to the displacement (for details see Cierpka *et al* [35]). Thirdly, the synthetic datasets present a theoretical benchmark which highlights the strengths and weaknesses of the different algorithms. They are not necessarily representative for images of actual measurements.

All of the synthetic and experimental datasets were split into 70%/10%/20% for the training/validation/testing of the DNNs. Care must be taken to avoid having one particle image in more than one of these sets as this could introduce a bias into the uncertainty estimation, since this particle image would be used during the training and the evaluation. Therefore, the particle images were initially split into these percentages and each particle image was then only used in one of these sets.

3.3. DNNs

In this work, the approach presented by König *et al* [13] is extended, meaning that two networks are again used to process the images. DNNs or more specifically convolutional neural networks (CNN) take images as input and apply convolutions in successive layers of the network to extract and enhance features of the input image on different scales. In the last layers, the information is typically compressed and then given to the output neurons to give the prediction of the network.

The first network is an object detection network that provides the position and class of objects in the image plane. The position is given in the form of a bounding box, which consists of the four edge coordinates in the *xy*-plane. Here, this box is described by the size of the semiaxes from the APTV toolbox. The box coordinates do not need to be integer values, they can also be float values which leads to subpixel accurate predictions of the DNN. The classes correspond to the sizes of the different particles. The network provides a score for each box, ranging from 0 to 1, which is a result of the softmax activation function in the last layer. A common postprocessing step is to remove predictions below a certain score threshold. A Faster R-CNN [36] is chosen as the object detection algorithm with a ResNet50 [37] as the feature extractor. The Faster R-CNN includes a feature pyramid network, which allows the combination of features with different scales and therefore enhances the semantic information in the last layers [38]. The loss function used considers the classification of the particles as well as the in-plane uncertainties of the proposed bounding boxes with respect to the labels.



(a) Original images (b) $N_s = 0.05$ (c) $N_s = 0.10$ (d) $N_s = 0.15$ (e) $N_s = 0.20$

Figure 4. Examples of calibration images (a). Each quadrant corresponds to a raw image of the given particle size. The intensity is normalized for each quadrant to visualize all particle images. Examples for newly generated images of case 1 for the four different source densities (b)–(e). The image intensity is clipped at 30 counts to visualize defocused particle images of smaller particle sizes.

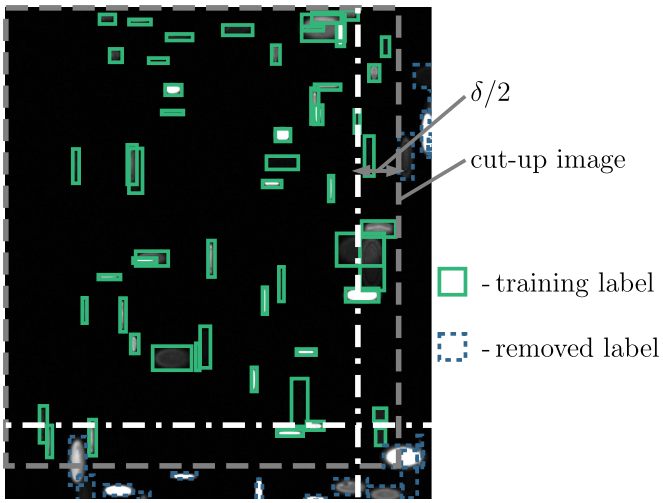


Figure 5. Example of cutting up the images for the object detection network. The symmetry line of the image is shown by the dashed white line. The dashed gray line shows the border of the image after cutting with an overlap δ to the neighboring images. The green boxes are valid labels that are used during the training while the dashed blue boxes are invalid labels that are not used.

The images are downsampled in the first layer to a resolution of 1333×800 px, to reduce the memory requirements for the GPU. Furthermore, if there are too many objects in the images, the performance significantly drops. To circumvent both problems, each image was cut up into four equally sized images, which have an overlap δ of 100 and 250 px for the synthetic and experimental dataset, respectively. An example of an experimental image extracted in that way is shown in figure 5. This overlap was chosen to be slightly larger than the maximum size of the extended ellipse of the largest particle size. In the labels for each image, boxes which are not fully contained within the cut-up image, are removed. If they were not removed, the network would also learn to detect cut-off particles. However, this can be disadvantageously if the network detects a single particle twice in different cut-off versions. The post-processing of the Faster R-CNN is then not always able to filter out these additional detections which leads to false positive (FP) samples. These removed labels are marked by the dashed blue boxes in figure 5, while the green boxes correspond to valid labels. The input resolution of the Faster R-CNN was set to match the resolution of the cut-up images.

Transfer learning was applied by using a model that was pretrained on the COCO dataset [23] which reduces the training time and reduces the risk of overfitting [39]. The network was trained with a batch size of four and the Adam optimizer [40]. For the synthetic datasets, the network was trained with two different configurations. For the first, denoted as V1, the network was trained for 20 epochs and the images were not manipulated. For the second one, denoted as V2, the network was trained for 60 epochs and the background was removed by subtracting a second order polynomial fit of the image intensity. For all rearranged, real datasets, the network was trained for seven epochs. To avoid overfitting for long training durations, an L2 regularization with 10^{-4} as a penalty was applied. Furthermore, the datasets were augmented by randomly multiplying the image intensity with a factor from [0.8, 1.2]. Also, the rearranging of the particle images can be seen as a form of data augmentation [41] which further reduces the risk of overfitting. The initial learning rate was set to 10^{-5} and linearly increased over the course of one epoch to a maximum learning rate of 10^{-4}. The learning rate was then decreased by a factor of 10 before the penultimate and last epoch to achieve a convergence of the network, as is common practice [37, 42].

During postprocessing, the Faster R-CNN relies on non-maximum suppression to filter out boxes which are overlapping. This potentially removes particles with a strong overlap and therefore, the threshold for the non-maximum suppression below which boxes are filtered out is set to 0.9. Furthermore, the predictions on the four cut-up images must be merged in a postprocessing step which is referred to as 'stitching'. First, predicted boxes located within 1 px of the image boundaries were removed, because these are boxes for particle images which were not fully contained within the image. The predictions on all images were then filtered to remove additional boxes in the overlapping regions. Every box was compared

with one another, removing the one with the smaller score if all following criteria apply: (i) the boxes have the same class prediction, (ii) the center points of both boxes are within 15 px, (iii) the relative difference of the aspect ratio is smaller than 15% and (iv) the relative difference of the area is smaller than 15%. These are user-defined thresholds that were found to work well for the present cases. A predicted box was marked as valid if it had a Euclidian distance of 5 px or less to a labeled ground truth box.

The performance of this DNN is evaluated by three metrics. The first metric compares the precision and the recall which are defined as:

$$\text{precision} = \frac{\text{TP}}{\text{TP}+\text{FP}},$$
$$\text{recall} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$

where the true positives are the valid predictions of the network with the correct class, the FPs are the valid predictions of matched particles with the incorrect class or unmatched predictions and the false negatives are the particles from the ground truth that were not matched to a network prediction. These values are shown in a diagram, called the precision-recall curve which plots both values for different score thresholds of the network. Here, score thresholds with a step size of 0.001 from 0.050 to 0.999 are used. The second and third metric are the in-plane uncertainties σ_x and σ_y defined as the root mean square between the center position prediction and label. Both uncertainties are always given at a score threshold of 0.8 and the uncertainties are always given w.r.t. to the position labels provided by the evaluation of the experimental images using the standard APTV evaluation approach [11].

The second DNN is solely used for the determination of the depth position. This DNN gets the individually cropped particle images as the input. The chosen network architecture for this is a ResNet18 [37] which has comparatively few layers. This is chosen because a cropped region around one particle is not expected to have many semantic features that are extracted at deeper layers, and a network with less layers has a smaller training time. The individual particle images are all cropped to the same size which depends on the dataset. Before the images are loaded into the network, they are cropped in a rectangle around the center, which is then zero padded to match the maximum size in the corresponding dataset. This rectangle is taken to be the size of the ellipse' semiaxes, again expanded by $f = 1.25$. This was found to result in a better performance than a constant crop around the particle center, which was used in previous works [13, 26]. The center points of the particle images for the cropping are taken from either MicroSIG for the synthetic images or the position labels of the evaluation of the experimental images using the standard APTV evaluation approach. This is done to avoid shifting potential uncertainties of the first DNN into the second one. For later measurement applications, the second DNN uses the predictions of the first DNN as input.

The chosen network is again available with a model pretrained on a different dataset, called ImageNet [15]. Therefore, transfer learning is again applied to reduce the training time. In the ImageNet dataset, the network makes a prediction between 1000 classes, thus the final layer has the same number of output neurons. Here, the network is only supposed to predict a single float value, which is why the final layer is replaced to have just one output neuron. The weights of the last layer are randomly initialized [16].

The pretrained weights are still able to mitigate the effects of this random initialization because the learning rate is linearly increased over the duration of one epoch from 10^{-5} to 10^{-4}. The network is trained for a total of 15 epochs with a batch size of 128, and the learning rate is reduced by a factor of 10 after eleven and 13 epochs to reach a converged state. The objective function, which is minimized during the training is the mean absolute error between the prediction and the ground truth of the z-position, and the Adam optimizer is used [40]. An L2 regularization with a penalty of 10^{-4} is applied again as well as data augmentation by multiplying the image intensity with a random factor in the range $[0.8, 1.2]$. Additionally, random crops of the particle images are used. For this, the center point (x_i, y_i) of the particle image is randomly varied in the range of $[-5, 5]$ px and the resulting image zero-padded to reach the size of the respective regression datasets. As is common in machine learning, the z-labels of the network are zero-centered and normalized with a factor of 10. For example, the 5.03 µm particles cover a depth range from −60 µm to 60 µm, the training labels for the network span the interval $[-6, 6]$. This range of values was found to yield the best results. The performance of the network is evaluated only by the uncertainty σ_z in the z-position, again taken to be the root mean square error between the prediction and the label.

The training for both networks was carried out on the high-performance cluster of the TU Ilmenau. The GPU used was an NVIDIA A100 Tensor-Core-GPU with 40 GB of RAM.

4. Results and discussion

4.1. Uncertainty for synthetic images

The results of the training on synthetic images are shown in figure 6. The arrangement of the diagrams is kept similar to the results of Barnkob *et al* [26]. The normalized out-of-plane uncertainty σ_z/h as a function of the source density is shown in figure 6(a). The depth range h amounts to 85 µm. Except for the smallest source density, the uncertainty appears to be almost constant, at a relative depth error of 0.8%. The reason for the higher uncertainty at $N_s = 0.1$ is not known. The almost constant uncertainty is in accordance with the DNN results from previous investigations but it was improved by an order of magnitude and also surpasses the classic methods [26].

The resulting in-plane error for the two different training configurations is shown in figure 6(b). Here, an increase of the in-plane uncertainty is noted with increasing source density, which is also in accordance with previous DNN results. Again, the uncertainties were improved by one order of magnitude and now match the results of classical evaluation methods. The background removal and longer training duration of V2 reduce the uncertainties by approximately 30%. A further reduction

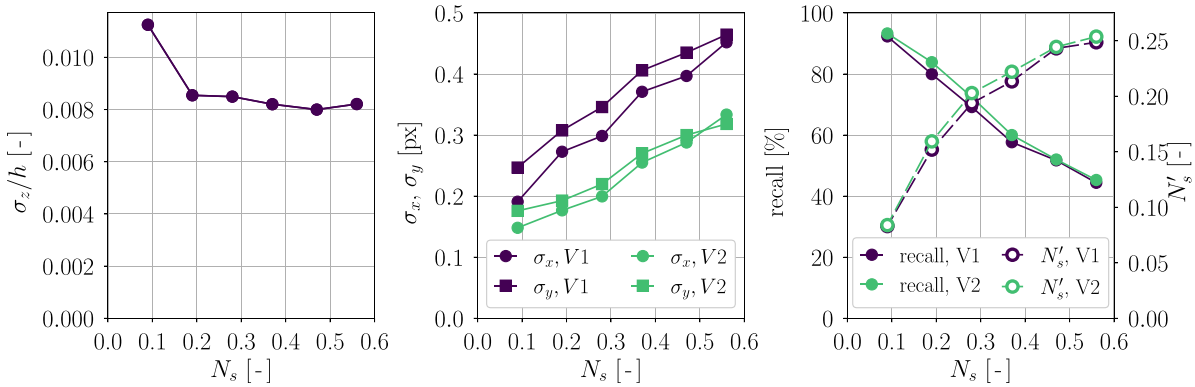


Figure 6. Uncertainties of the position determination for the synthetic particle images with (a) out-of-plane uncertainty σ_z/h, (b) in-plane uncertainty σ_x, σ_y, and (c) recall and apparent source density N_s' as a function of the source density. All quantities are given at a score threshold of 0.8. The in-plane uncertainties and recall are given for two different training configurations, denoted as V1 and V2.

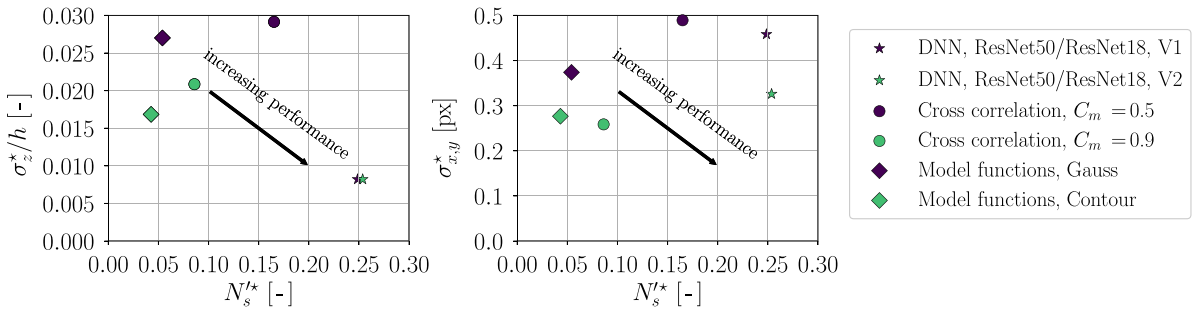


Figure 7. Comparison of different defocus particle tracking techniques. The out-of-plane and in-plane uncertainties are given at the maximum apparent source density $N_s'^\star$. Quantities for classic methods are drawn from [26].

of the uncertainties is expected for longer training durations at the expense of an increased computational cost.

The recall (solid line) and apparent source density $N_s' = \text{recall} \cdot N_s$ (dashed line) are shown in figure 6(c) for the two different training configurations. Compared to previous studies, the recall shows a strong improvement, especially for high source densities. The highest apparent source density of 0.27 is reached at $N_s = 0.6$ but the curve appears to flatten. A longer training only leads to a slight increase in the recall. The strong reduction in the recall down to 40% is suspected to stem from the large number of object instances in the images. However, cutting the images into nine individual subimages did not result in a further improvement of the recall, likely because the images are getting too small, so the deep network topology does not improve the extracted information.

A comparison with two methods based on classic image processing is shown in figure 7. The first one is based on the classical evaluation method. Assuming Gaussian-like images, a calibration function is used to relate the depth coordinate to the semiaxes of the particle images [8, 12]. The same method has been used to label particles in the experimental images of this work. The second method uses a cross correlation with calibration images to determine particle positions [14]. For all three methods, two variants are used, denoted by the different colors of the markers. All uncertainties are given at the highest apparent source density, denoted as $N_s'^\star$. On the left and right-hand sides, the relative depth error and in-plane errors at the corresponding source density are shown, respectively. In comparison to the classic methods, both versions of the DNNs have a much higher apparent source density which means they work reliably even for high seeding concentrations. This results in a higher spatial resolution and reduces the amount of images or the measurement time. Furthermore, the uncertainty in the depth position is less than half of the uncertainty of the classic methods. Important to highlight is that the network is easily applicable and no complex data processing is needed to obtain a calibration function. This makes the technique also interesting for non-expert users from different fields. The in-plane uncertainty of V2 is slightly larger than the best performing classic method. However, it is expected that a longer training would further decrease the error to also achieve the smallest in-plane uncertainty.

4.2. Uncertainty for real images

The resulting uncertainties of the real, rearranged datasets in x and y over the source density are shown in figures 8(a) and (b). The different cases are defined according to the legend. The first four cases are shown with different lines than the last three since the former contain both 2.47 and 3.16 µm particles and are expected to constitute a more difficult classification. The same trends are visible in both plots but the uncertainty in y is slightly larger than the uncertainty in x. A potential explanation for this is the aspect ratio of the camera sensor which results in non-square feature maps in the layers of the neural

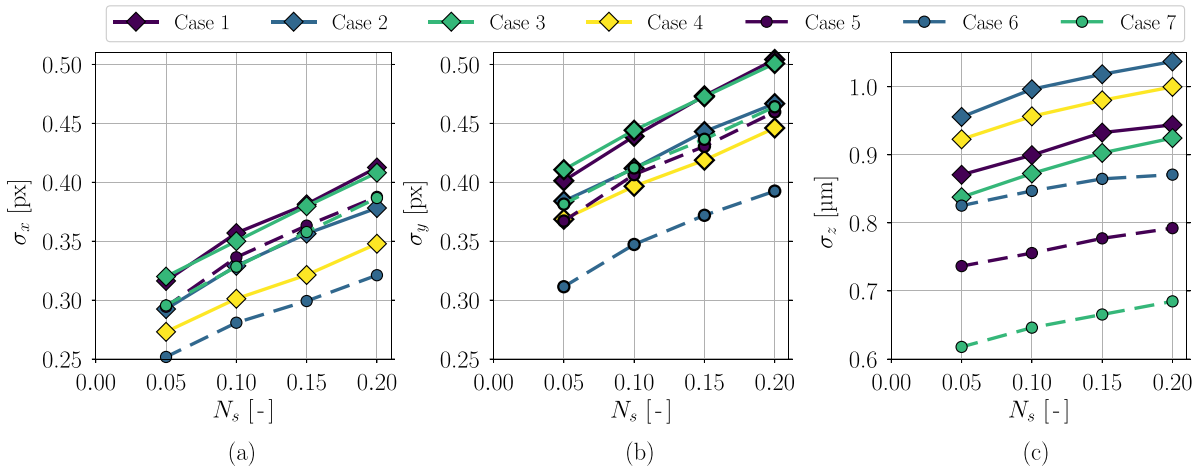


Figure 8. Uncertainties of the position determination for case 1–7. σ_x (a), σ_y (b) and σ_z (c) as a function of the source density. All quantities are given at a score threshold of 0.8. The legend defining each case applies to all three subfigures.

network. In this way, the information is extracted better in the x-direction as it has a smaller range of values.

For all cases, the in-plane uncertainty w.r.t. the labels is always below 0.5 px. Barnkob *et al* [26] estimate the in-plane uncertainty with classic image processing to be 0.7 px. However, this cannot be confirmed without knowing the ground truth of experimental images. This remains an open challenge in the field of defocus particle tracking which is why further discussion about the 'real' uncertainty is omitted. It is expected that the uncertainty of the ground truth adds to a certain extend to the uncertainty of the neural networks, as the uncertainty for the neural networks is a bit lower using purely synthetic data (see figure 6) for the same source density N_s. However, as the errors in the ground truth are assumed to be random and the data set is large, the effect is rather small. The in-plane uncertainty shows a maximum increase of about 25% at the largest source density which is only a small increase considering that four times as many particle are present.

Generally, very similar uncertainties are observed for all cases and only slight differences are visible. It appears that the cases with the 1.14 µm particles have a consistently higher uncertainty compared to other cases. The in-plane uncertainty for case 1 is the highest as this case features four different particles sizes. Furthermore, the cases with only two different particle sizes (4, 6, 7) show considerably lower uncertainties. It is interesting to note that also the uncertainty for case 4 is in the same range although the difference in size is with 2.47 and 3.16 µm particles the smallest. One possible explanation for this is that the DNN reaches the best possible classification and after that only a reduction of the in-plane error results in a smaller loss for the training and is thus further improved.

The resulting uncertainty in the depth position σ_z is shown in figure 8(c). The same legend applies again. The maximum uncertainty of 1.05 µm is observed for case 2 for the highest source density. This is remarkable, since it means that a maximum uncertainty of about 1 µm is expected, even for particle distributions where many different sizes are present. The requirement of APTV which only allows a narrow size distribution is made obsolete with the current approach, which widens the field of applications. Again, the uncertainty increases with larger source density although the relative gain is slightly smaller than for the in-plane uncertainty at approximately 15%.

In contrast to the in-plane uncertainty, case 5–7 have a smaller uncertainty than case 1–4. This is interesting, because the latter are expected to be hard to classify. The regression DNN has no knowledge about the size of each particle image and yet, the uncertainty is larger for these cases. One possible explanation for this is that the network performs an internal classification based on the features of the images. Akin to the different calibration functions of each particle size, the depth position is then given based on this internal classification.

An additional finding is that the cases with the 1.14 µm particles have a consistently smaller uncertainty compared to the ones that do not. This is different to what was observed for the in-plane uncertainty. This can potentially be explained by the different ranges of depth positions. The depth positions for this particle size only span 81 discrete values while the 5.03 µm ones have 121 values. Therefore, if the uncertainties were normalized by the depth position, they are expected to be much more similar between these two sizes.

To confirm these hypotheses, the uncertainties of the individual particle sizes are investigated. However, please note that the network was trained for all sizes at once. Case 1 is chosen for this analysis because it contains all particle sizes. The individual uncertainties over the depth position z are shown in figure 9, where the columns (a)–(d) correspond to the different particle sizes according to the caption. The different source densities are denoted by the colors and markers according to the legend. The individual uncertainties are shifted upwards in steps of 0.2 px or µm. This is purely for visualization purposes because otherwise, the curves would overlap, especially in the center.

The in-plane uncertainties of the first DNN are shown in the first and second row, for the x- and y-position, respectively. For both uncertainties, a clear minimum is visible at approximately +20 µm for σ_x and −20 µm for σ_y. This minimum

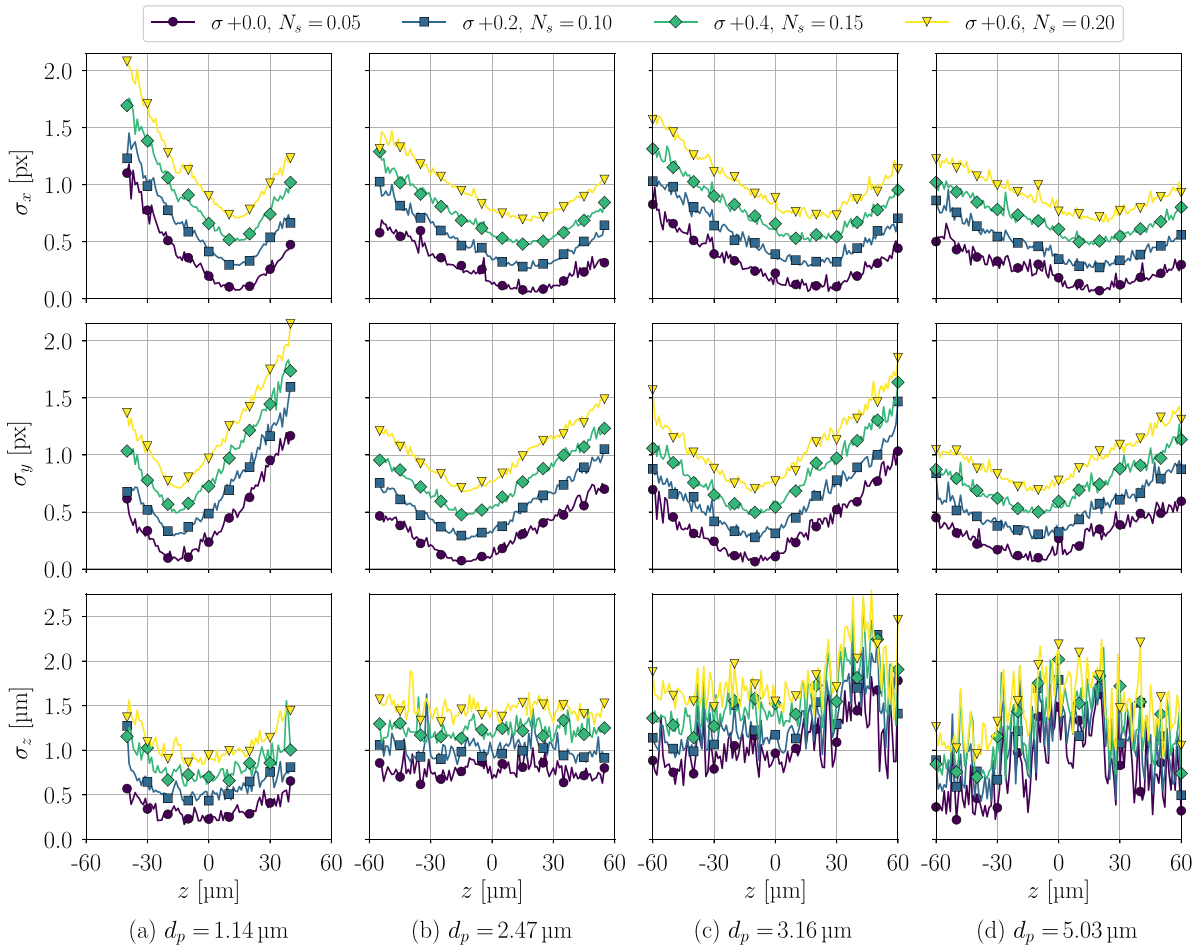


Figure 9. Uncertainties for each individual particle size of case 1. σ_x (top row), σ_y (middle row) and σ_z (bottom row) as a function of the depth position z. The scaling of each axis is the same but σ_x and σ_y are given in px while σ_z is given in µm. All quantities are given at a score threshold of 0.8. The legend defines the source density of each color and applies to all subfigures. For visualization purposes, the uncertainties are shifted upwards with increasing source density by 0.2.

coincides with the location of the focal planes, meaning that the uncertainty of the in-plane position is small when the size of the respective semiaxis is small and the particle image is sharp in this direction. There are two plausible explanations for this. The first is an error stemming from the labels, i.e. the fitting of the Gaussian is more accurate near the focal planes and thus, the labels are as well. The second possibility is that the particles near the focal planes have a small particle image. Therefore, this information is extracted at later layers which contain more information, therefore reducing the uncertainty.

The uncertainty increases towards the boundaries of the measurement domain and this increase becomes larger with increasing source density. As was already suspected based on the uncertainty of the individual cases, the smallest particles have the highest uncertainty. A maximum uncertainty in the y-direction of 1.5 px is reached for the highest source density at the margins of the investigated depth range. For the other three sizes, the maximum uncertainty is approximately 1 px, at the boundaries of the measurement domain. This increase in the uncertainty towards the boundaries can again be caused by one of the effects for the pronounced minimum given above. The increased uncertainty of the smaller particles can be explained by the rescaling of the images to 8 bit. For defocused 1.14 µm particles, the particle images are sometimes binned into just four different intensity values. This results in a flattening of gradients as well as a loss of information, thus leading to an increased uncertainty.

The resulting uncertainty in the depth position is shown in the third row of figure 9. Here, the shapes deviate for the individual particle sizes. For the 1.14 µm particles, the uncertainty is uniform in the center and then increases towards the boundaries. The 2.47 and 3.16 µm particles have a uniform uncertainty over the entire depth position although a spike can be observed for the 3.16 µm particles at approximately 40 µm. This is also were the largest uncertainty occurs for the 1.14 µm particles. For the largest particles, the uncertainty is smallest at the boundaries and increases in the center. These observations are present for all source densities and the same spikes and kinks in the course are observed.

For the smallest and largest particle size, the curves can once more be explained by the rescaling of the intensities. For the 1.14 µm particles, the strong binning results in an increased uncertainty for defocused particles. For the 5.03 µm particles, the intensities of particles near the focal planes is clipped

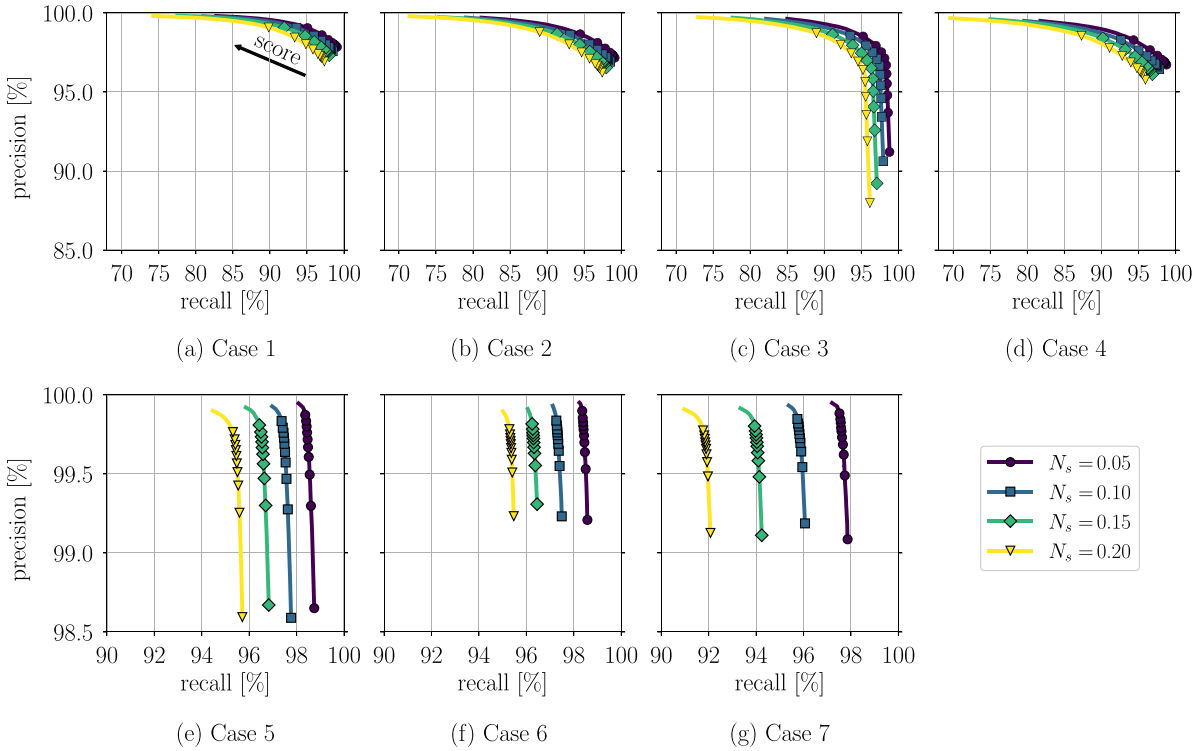


Figure 10. Precision-recall for different source densities. For case 1–4 (a)–(d), both the precision and the recall have the same scaling. For case 5–7 (e)–(g), the precision and the recall are the same again but they have a different scaling to (a)–(d). The legend defines the source density according to each color and applies to all seven subfigures. The direction of increasing score is defined by the arrow in figure (a) and also applies to all subfigures. The markers correspond to ten evenly spaced score thresholds from 0.05 to 0.95.

and therefore, gradients are clipped as well. This information appears to be important for the second DNN as this results in an increased uncertainty.

The reason for the spike in the 3.16 µm particles is not known even though it is consistently observed for all source densities. The uncertainty of the 2.47 and 3.16 µm particles is also consistently larger than for the other two particle sizes, as was already observed in the global uncertainties for each case. The network appears to perform an internal classification and as this is expected to be more difficult, the resulting uncertainties for particles with a similar particle image are larger. The resulting uncertainty of the 1.14 µm particles is smaller than that of the 5.03 µm ones. This is suspected to stem from the different range of depth positions; normalizing the uncertainty by the measurement height yields a very similar relative uncertainty.

4.3. Classification

The resulting precision-recall curves of all seven cases are shown in figure 10. Case 1–4 in the top row and case 5–7 in the bottom row, each use the same scaling of both the precision and the recall. The legend defines the source density and applies to all seven cases. The markers correspond to ten evenly spaced score thresholds ranging from 0.05 to 0.95.

Case 1, 2 and 4 show a very similar trend. The minimum precision and the maximum recall are approximately 96%, always for the smallest source density. For smaller source densities, the minimum precision and the maximum recall increase, which is expected. Reaching a higher precision is only possible by reducing the recall drastically, sometimes below 70%. Additionally, there is almost no difference for small score thresholds, the precision and the recall take very similar values for these small scores.

For case 3, the observations are slightly different. The recall spans a similar region but the shape of the curve is much different. For small score thresholds, the minimum precision is 86%, but it then shows a very strong initial increase, meaning that a higher precision can be reached at almost no additional cost. For higher score thresholds, the achieved precision and recall are again similar to the above discussed cases. Interesting to note is that the markers which indicate the different thresholds, are very similar above a score of 0.65. For these scores, the classification works equally well for Case 1–4. It was found that the drop in the precision for small score thresholds can be mitigated by training the network for a longer period of time but it can not be completely removed.

One possible explanation could be that this case percentagewise contains the most 1.14 µm particles of all four cases. These have the largest in-plane uncertainty as was shown in section 4.2. Therefore, the loss function used in the optimization process is dominated by the in-plane error, thus making the classification harder. This further explains, why case 4 in figure 8 has a very small uncertainty in the in-plane position. The classification between these two sizes is difficult and therefore, only the in-plane error is improved.

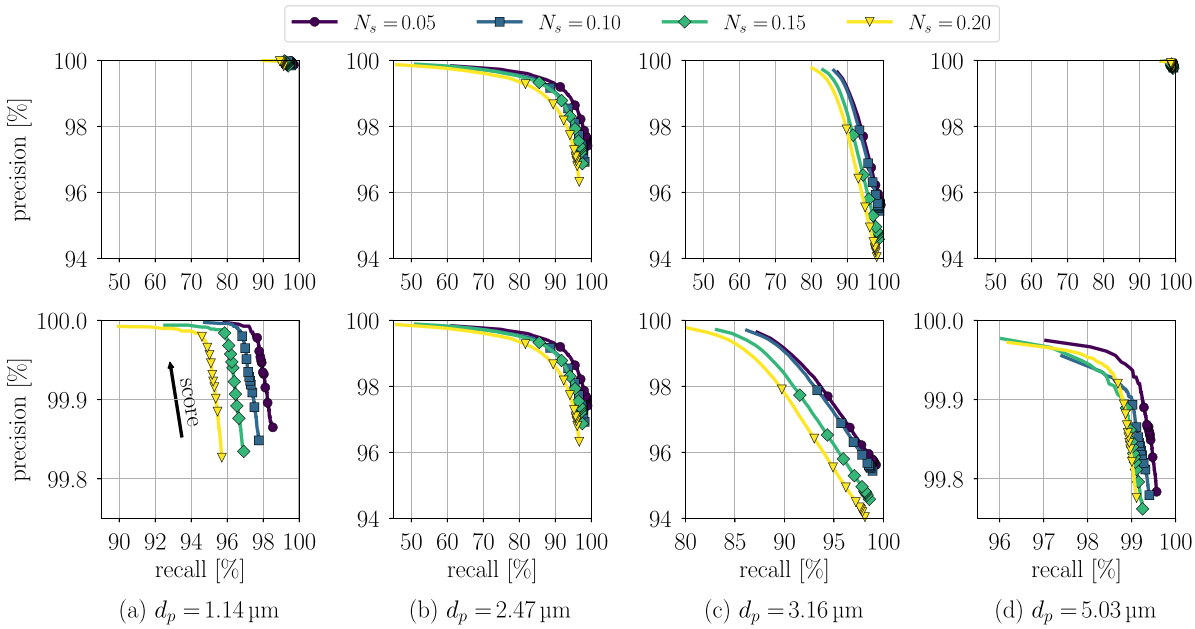


Figure 11. Precision-recall curve of each individual particle size for case 1. The legend defines the source density of each color and applies to all subfigures. The direction of increasing score is defined by the arrow in figure (a) and also applies to all subfigures. The markers correspond to ten evenly spaced score thresholds from 0.05 to 0.95. Top row: comparison of the four curves on the same scale of each axis. Bottom row: zoomed-in version on the relevant region of interest of each particle size.

A very different behavior is observed for case 5–7. The lowest precision of 98.5% is higher and also only present for case 5, the other two have a minimum of 99%. The precision increases for smaller source densities at the cost of an only slightly reduced recall. The strong difference to case 1–4 is that the curves show almost no flattening, even for higher score thresholds. Thus, almost no trade-off between the precision and the recall is observed. A precision of 100% can be reached while the recall is hardly reduced and only for score thresholds above 0.95. Furthermore, the precision shows almost no drop for larger source densities. The shape of the curve as well as the minimum and maximum value of the precision are very similar. A slight decrease in the recall is noted which is the strongest for case 7 where a worst value of 91% is reached. A similar decrease in the recall is also present for case 1–4 but the difference is not visible due to the scale of the axis.

A general observation is that the shape and values of the precision-recall curve do not appear to depend on the number of different sizes in the dataset. In fact, case 1 which has four particle sizes reaches better values than case 4 which has two. Instead, it seems more important that the individual particle sizes have different features in their particle image. The 2.47 and 3.16 µm particles included in case 4 have a very similar particle image shape and intensity distribution, resulting in a more difficult distinction and the worst precision among the examined cases.

To verify this proposed explanation, the precision-recall curves of the individual particle sizes are shown in figure 11. The markers again correspond to ten evenly spaced score thresholds from 0.05 to 0.95. The quantities are drawn from case 1 because it contains all particle sizes. In the top row of the figure, the curves are shown on the same scale for all four particle sizes. As expected, the precision of the smallest and largest particle size is very high and above 99.75% for all scores and source densities. In contrast, the 3.16 µm particles achieve a minimum recall of 80% with a precision as low as 94%. For the 2.47 µm particles, the minimum precision is slightly higher at 96% but the minimum recall drops significantly to below 50%.

The bottom row of figure 11 shows the relevant region of interest of the precision-recall curve for each particle size. For a better comparison, the ranges of the precision are the same in figures (a) and (d) as well as (b) and (c), respectively. A strong inclination is visible which flattens out for higher score thresholds. The maximum precision is reached at a smaller recall for the 1.14 µm particles compared to the 5.03 µm ones. However, for a score of 0.95, the precision is almost at the maximum for the 1.14 µm particles while improvements are still observed for the 5.03 µm ones. Additionally, the curves only show slight differences for the 5.03 µm particles with varying source density while a decrease in the recall can be noted for the smallest particles.

As mentioned before, the minimum precision of the 2.47 µm particles is much lower at 96%. Furthermore, the curve shows a much smaller initial increase and then flattens out much stronger, down to the aforementioned minimum recall below 50%. A different trend is visible for the 3.16 µm particles. The smallest precision is 94% and for smaller score thresholds, the curve has a linear inclination. The curve flattens out for precisions upwards of 99.9% but a smallest recall of 80% is still retained. Another interesting point is that for a score threshold of 0.95, a precision above 99.5% is reached for the 2.47 µm particles while the 3.16 µm particles only have a

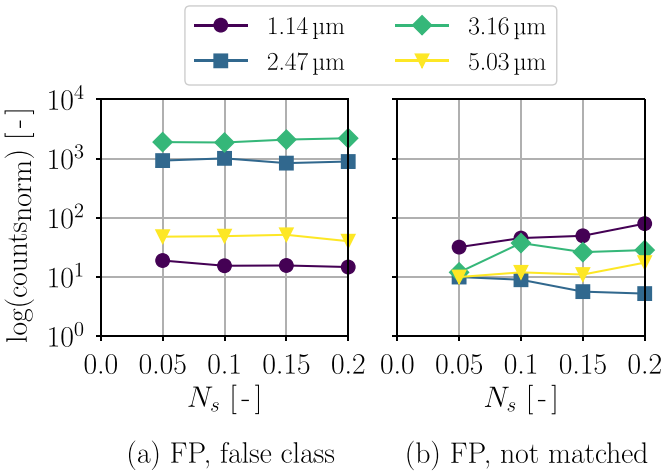


Figure 12. Absolute number of FPs due to false classification of a matched prediction (a) and due to an unmatched network prediction (b) with increasing source density. The values are normalized according to the source density to account for differently sized test sets.

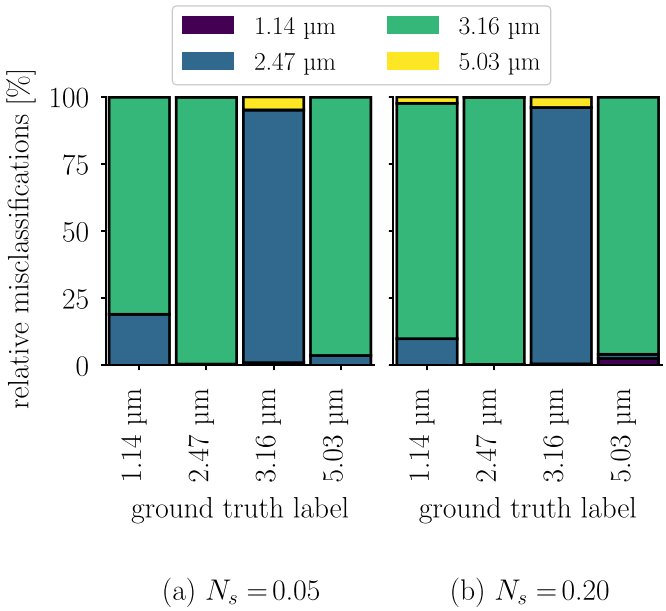


Figure 13. Misclassifications of each particle size for $N_s = 0.05$ (a) and 0.20 (b). The vertical axis shows the percentage with which each particle from the ground truth is misclassified. The legend relates the colors to the particle sizes and applies to all subfigures.

precision of 98%. For the latter, significant gains are observed for even higher thresholds, the reason for this is unknown.

In general, the classification of the 2.47 and 3.16 µm particles is more difficult, as is expected. They have a much smaller precision and recall. As was mentioned in section 3.3, the FPs consist of matched predictions which either have the false class or unmatched predictions for which no target box was found. The amount of these respective types for case 1 with a score threshold of 0.8 is shown in figures 12(a) and (b), respectively. The particle sizes are defined according to the legend. The number of FPs is 'normalized' because for a source density of 0.20, there are four times as many particles and hence, the number of FPs is divided by four for a better comparison.

As expected, the predictions with the false class are almost two orders of magnitude larger for the 2.47 and 3.16 µm particles compared to the 1.14 and 5.03 µm ones. On the logarithmic scale, the curves show almost no inclination which agrees with the observations in figure 11 where only the recall decreases with increasing source density.

The unmatched predictions show a similar trend, where only the 1.14 µm particles have a slight increase. For the other sizes, there are often less than ten unmatched predictions, stemming from random fluctuations. Furthermore, there are the most unmatched predictions for the smallest particle size, which constitutes a larger amount of FPs than due to misclassifications. This is expected to occur for defocused particle images which have a very small intensity, which can blend in with the background. The main takeaway is that the misclassification of 2.47 and 3.16 µm particles makes up the highest percentage of the FPs.

The expectation is that these particles are most commonly misclassified as one another. This is confirmed by looking at the misclassifications of each particle size which are shown in figures 13(a) and (b) for $N_s = 0.05$ and 0.20, respectively. The same trends are observed for all source densities, so the other two are omitted. The particle size of the bars refers to the label provided by the ground truth and the individual bars show the percentage of misclassifications as the respective size according to the legend. The 2.47 and 3.16 µm particles are misclassified as one another in more than 95% of the cases, meaning that the network has difficulties distinguishing between these two sizes which is expected given their similar particle image features.

Interesting to note is that the smallest particle size is misclassified as a 3.16 µm particle in 80% of the cases. This is unexpected but is likely a statistical problem due to the small number of 1.14 µm FPs. Also, there are some 5.03 µm particles which are misclassified as 1.14 µm ones for the highest source density. This is likely caused by a false matching of the prediction and target boxes.

5. Summary and conclusion

In this work, the application of DNNs to determine the particle position for any given defocus method using multiple particle sizes was analyzed and improved. Besides the 3D position determination, the percentage of correctly classified particles (precision) and the percentage of retained particles (recall) were of interest. Two CNNs were used to find and classify the particles in the image plane according to their size (object detection) and determine the depth position of individual particles (regression). The analysis of synthetic and experimental, rearranged particle images shows:

- A heavily improved performance with respect to the uncertainty in all three dimensions on a benchmark dataset of synthetic images with just one particle size. The uncertainty for

the position measurement was drastically reduced in comparison to previous approaches. The uncertainty was below 1% of the depth of the measurement volume and allows reliable measurements and a flexible adaptation form many different scenarios. The methodology itself is independent of the velocity field as it measures reliably the particle position in any flow. Hence, the results are relevant for many particle tracking methods, might this be classical approaches (e.g. [35, 43–45]) or based on neural networks (e.g. [46–48] or the overview and references herein [49]).

- The limitations of the current approach were analyzed and showed a decreased recall when the particle seeding concentration is very high because more than one thousand object instances are present in the image. Reducing the number of particle images yields an improved recall but this phenomenon can likely be traced back to the underlying network architecture and therefore requires further analysis.
- On rearranged, experimental images of up to four different particle sizes, the in-plane uncertainty is subpixel accurate for overlapping particle images w.r.t. the provided labels. The uncertainty in the depth position remains below 1 µm for all size combinations and source densities. The smallest particles have the highest in-plane uncertainty but the smallest out-of-plane uncertainty.
- A higher uncertainty in the depth position for particles of similar size was observed. It appears that the network still performs a sort of internal classification that is then used in the determination of the depth position.
- The distinction of particles of similar size is more difficult. Achieving a precision upwards of 99.5% results in a recall below 70% in the worst cases. Nevertheless, at the chosen operating point of the DNN, a recall above 95% is always retained with a precision above 96%.

Future work should extend the analysis to particles made of different materials and sizes. Additionally, other scalar values of interest can be learned by the network, e.g. pH-value, concentration, temperature [50, 51]. The current approach can be extended effortlessly to more particle sizes, shapes and materials and a good performance of the DNNs is expected based on the present results.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://defocustracking.com/datasets/.

Acknowledgments

The authors thank the German Research Foundation (DFG) for financial support within the priority program PP2045 'MehrDimPart' (CI 185/8-1), the Grant CI 185/14-1 and the Carl Zeiss Foundation's Grant: Deep Turb. Furthermore, support by the Center of Micro- and Nanotechnologies (ZMN) (DFG RIsources Reference: RI 00009) of TU Ilmenau as well as Henning Schwanbeck for the GPU support, is gratefully acknowledged.

Conflict of interest

There are no conflicts of interest to declare.

ORCID iDs

M Ratz https://orcid.org/0009-0008-8491-8367
S Sachs https://orcid.org/0000-0002-2839-7084
J König https://orcid.org/0000-0002-7832-9223
C Cierpka https://orcid.org/0000-0002-8464-5513

References

[1] Sajeesh P and Sen A 2014 Particle separation and sorting in microfluidic devices: a review *Microfluid. Nanofluidics* **17** 1–52
[2] Yamada M and Seki M 2005 Hydrodynamic filtration for on-chip particle concentration and classification utilizing microfluidics *Lab Chip* **5** 1233–9
[3] Dannhauser D, Romeo G, Causa F, De Santo I and Netti P 2014 Multiplex single particle analysis in microfluidics *Analyst* **139** 5239–46
[4] Sehgal P and Kirby B J 2017 Separation of 300 and 100 nm Particles in Fabry–Perot acoustofluidic resonators *Anal. Chem.* **89** 12192–200
[5] Ahmed H, Destgeer G, Park J, Afzal M and Sung H J 2018 Sheathless focusing and separation of microparticles using tilted-angle traveling surface acoustic waves *Anal. Chem.* **90** 8546–52
[6] Zhang W, Hu Y, Choi G, Liang S, Liu M and Guan W 2019 Microfluidic multiple cross-correlated coulter counter for improved particle size analysis *Sens. Actuators* B **296** 1–9
[7] Blahout S, Reinecke S, Kazerooni H, Kruggel-Emden H and Hussong J 2020 On the 3D distribution and size fractionation of microparticles in a serpentine microchannel *Microfluid. Nanofluidics* **24** 22
[8] Sachs S, Ratz M, Mäder P, König J and Cierpka C 2023 Particle detection and size recognition based on defocused particle images: a comparison of a deterministic algorithm and a deep neural network *Exp. Fluids* **64** 21
[9] Schanz D, Gesemann S and Schröder A 2016 Shake-The-Box: Lagrangian particle tracking at high particle image densities *Exp. Fluids* **57** 70
[10] Cierpka C, Rossi M, Segura R, Mastrangelo F and Kähler C J 2012 A comparative analysis of the uncertainty of astigmatism-µPTV, stereo-µPIV and µPIV *Exp. Fluids* **52** 605–15
[11] Cierpka C, Segura R, Hain R and Kähler C J 2010 A simple single camera 3C3D velocity measurement technique without errors due to depth of correlation and spatial averaging for microfluidics *Meas. Sci. Technol.* **21** 045401
[12] Cierpka C, Rossi M, Segura R and Kähler C J 2011 On the calibration of astigmatism particle tracking velocimetry for microflows *Meas. Sci. Technol.* **22** 015401
[13] König J, Chen M, Rösing W, Boho D, Mäder P and Cierpka C 2020 On the use of a cascaded convolutional neural network for three-dimensional flow measurements using astigmatic PTV *Meas. Sci. Technol.* **31** 074015
[14] Barnkob R, Kähler C J and Rossi M 2015 General defocusing particle tracking *Lab Chip* **15** 3556–60
[15] Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L 2009 Imagenet: a large-scale hierarchical image database *2009*

IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Miami, Florida, USA, 20–25 June)

[16] He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: surpassing human-level performance on imagenet classification *2015 IEEE Int. Conf. on Computer Vision (ICCV) (Santiago, Chile, 7–13 December)*

[17] Russakovsky O *et al* 2015 ImageNet large scale visual recognition challenge *Int. J. Comput. Vis.* **3** 211–52

[18] Silver D *et al* 2016 Mastering the game of Go with deep neural networks and tree search *Nature* **529** 484–9

[19] Assael Y M, Shillingford B, Whiteson S and de Freitas N 2017 LipNet: end-to-end sentence-level lipreading *5th Int. Conf. on Learning Representations (ICLR) (Toulon, France, 24–26 April)*

[20] Franchini S, Charogiannis A, Markides C N, Blunt M J and Krevor S 2019 Calibration of astigmatic particle tracking velocimetry based on generalized Gaussian feature extraction *Adv. Water Resour.* **124** 1–8

[21] Ichikawa Y, Kikuchi R, Yamamoto K and Motosuke M 2021 Determining particle depth positions and evaluating dispersion using astigmatism PTV with a neural network *Appl. Opt.* **60** 6538–46

[22] Zhang X, Wang H, Wang W, Yang S, Wang J, Lei J, Zhang Z and Dong Z 2022 Particle field positioning with a commercial microscope based on a developed CNN and the depth-from-defocus method *Opt. Lasers Eng.* **153** 106989

[23] Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L 2014 Microsoft COCO: common objects in context *13th European Conf. on Computer Vision (ECCV) (Zurich, Switzerland, 6–12 September)*

[24] Franchini S and Krevor S 2020 Cut, overlap and locate: a deep learning approach for the 3D localization of particles in astigmatic optical setups *Exp. Fluids* **61** 140

[25] Dreisbach M, Leister R, Probst M, Friederich P, Stroh A and Kriegseis J 2022 Particle detection by means of neural networks and synthetic training data refinement in defocusing particle tracking velocimetry *Meas. Sci. Technol.* **33** 124001

[26] Barnkob R, Cierpka C, Chen M, Sachs S, Mäder P and Rossi M 2021 Defocus particle tracking: a comparison of methods based on model functions, cross-correlation and neural networks *Meas. Sci. Technol.* **32** 094011

[27] Cierpka C, König J, Minqian C, Boho D and Mäder P 2019 On the use of machine learning algorithms for the calibration of astigmatism PTV *13th Int. Symp. on Particle Image Velocimetry (ISPIV) (Munich, Germany, 22–24 July)*

[28] Olsen M G and Adrian R J 2000 Out-of-focus effects on particle image visibility and correlation in microscopic particle image velocimetry *Exp. Fluids* **29** S166–74

[29] Rossi M 2019 Synthetic image generator for defocusing and astigmatic PIV/PTV *Meas. Sci. Technol.* **31** 017003

[30] Brockmann P, Symanczyk C, Ennayar H and Hussong J 2022 Utilizing APTV to investigate the dynamics of polydisperse suspension flows beyond the dilute regime *Exp. Fluids* **63** 129

[31] Sachs S, Baloochi M, Cierpka C and König J 2022 On the acoustically induced fluid flow in particle separation systems employing standing surface acoustic waves—Part I *Lab Chip* **22** 2011–27

[32] Sachs S, Cierpka C and König J 2022 On the acoustically induced fluid flow in particle separation systems employing standing surface acoustic waves—Part II *Lab Chip* **22** 2028–40

[33] Dwibedi D, Misra I and Hebert M 2017 Cut, paste and learn: surprisingly easy synthesis for instance detection *2017 IEEE Int. Conf. on Computer Vision (ICCV) (Venice, Italy, 22–29 October)*

[34] Dvornik N, Mairal J and Schmid C 2018 On the importance of visual context for data augmentation in scene understanding *IEEE Trans. Pattern Anal. Mach. Intell.* **43** 1

[35] Cierpka C, Lütke B and Kähler C J 2013 Higher order multi-frame particle tracking velocimetry *Exp. Fluids* **54** 1533

[36] Ren S, He K, Girshick R and Sun J 2015 Faster R-CNN: towards real-time object detection with region proposal networks *29th Int. Conf. on Neural Information Processing Systems (NIPS) (Montreal, Canada, 7–12 December*)

[37] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, California, USA, 26 June–1 July)*

[38] Lin T Y, Dollar P, Girshick R, He K, Hariharan B and Belongie S 2017 Feature pyramid networks for object detection *2017 Conf. on Computer Vision and Pattern Recognition (CVPR) (Honolulu, Hawaii, USA, 22–25 July)*

[39] Weiss K R, Khoshgoftaar T M and Wang D 2016 A survey of transfer learning *J. Big Data* **3** 1345–59

[40] Loshchilov I and Hutter F 2019 Decoupled weight decay regularization *7th Int. Conf. on Learning Representations (ICLR) (New Orleans, Louisiana, USA, 6–9 May)*

[41] Ghiasi G, Cui Y, Srinivas A, Qian R, Lin T Y, Cubuk E, Le Q and Zoph B 2021 Simple copy-paste is a strong data augmentation method for instance segmentation *2021 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Nashville, Tennessee, USA, 19–25 June)*

[42] Huang G, Liu Z and Weinberger K Q 2017 Densely connected convolutional networks *2017 Conf. on Computer Vision and Pattern Recognition (CVPR) (Honolulu, Hawaii, USA, 22–25 July)*

[43] Malik N A, Dracos T and Papantoniou D A 1993 Particle tracking velocimetry in three-dimensional flows *Exp. Fluids* **15-15** 279–94

[44] Cardwell N D, Vlachos P P and Thole K A 2011 A multi-parametric particle-pairing algorithm for particle tracking in single and multiphase flows *Meas. Sci. Technol.* **22** 105406

[45] Dou Z, Ireland P J, Bragg A D, Liang Z, Collins L R and Meng H 2018 Particle-pair relative velocity measurement in high-reynolds-number homogeneous and isotropic turbulence using 4-frame particle tracking velocimetry *Exp. Fluids* **59** 1–17

[46] Ouellette N T, Xu H and Bodenschatz E 2005 A quantitative study of three-dimensional Lagrangian particle tracking algorithms *Exp. Fluids* **40** 301–13

[47] Gan J, Liu P and Chakrabarty R K 2020 Deep learning enabled Lagrangian particle trajectory simulation *J. Aerosol Sci.* **139** 105468

[48] Han M, Sane S and Johnson C R 2022 Exploratory Lagrangian-based particle tracing using deep learning *J. Flow Vis. Image Process.* **29** 73–96

[49] Sciacchitano A and Discetti S 2021 Special issue on uncertainty quantification in particle image velocimetry and Lagrangian particle tracking *Meas. Sci. Technol.* **33** 010201

[50] Massing J, Kaden D, Kähler C J and Cierpka C 2016 Luminescent two-color tracer particles for simultaneous velocity and temperature measurements in microfluidics *Meas. Sci. Technol.* **27** 115301

[51] Deng Z, König J and Cierpka C 2022 A combined velocity and temperature measurement with an LED and a low-speed camera *Meas. Sci. Technol.* **33** 115301