

Enhancing defocused particle tracking with modified detection and localization

M. J. Yamakaitis^{1,*}, C. Fort¹, Y. Sawalhi¹, J.R. Rosen², P.M. Bardet¹

1: Dept. of Mechanical and Aerospace Engineering, The George Washington University, Washington, DC, United States

2: School of Public Policy and Urban Affairs, Northeastern University, Boston, MA, United States

*Corresponding author: myamakaitis@gwu.edu

Keywords: LPT, PTV Processing, Defocus PTV, Microfluidics

ABSTRACT

We adapt a recently published convolutional neural network architecture designed for conventional multi-camera image-based particle tracking to a three-dimensional single-camera imaging domain. Originally demonstrated for high particle densities where particle images are well approximated by Gaussian profiles only a few pixels wide, it is extended to ‘defocus’ particle tracking, where particle images are much larger and have arbitrarily complex shapes that encode depth. The single network performs simultaneous detection and localization in all three spatial dimensions for a single-image snapshot. We demonstrate simultaneous precision and recall of over 95% on previously published benchmark datasets, exceeding the performance of previous techniques. We extend the published dataset and test detection on images with 45 times more particles per pixel.

1. Introduction

Optical three-dimensional (3D) particle tracking is one of the most powerful diagnostic techniques in experimental fluid mechanics. Individual tracer particles provide high-resolution sampling of Lagrangian paths in fluid flow. Following the proliferation of 2D and 3D particle image velocimetry (PIV), it has seen increased interest because it is capable of overcoming many of the limitations of PIV while having similar hardware requirements. Most commonly applied for particle tracking velocimetry, it is capable of resolving 3D velocity fields. The ability to recover additional quantities from particle tracks has led to the adoption of the more general term Lagrangian particle tracking (LPT) in the literature (Schröder & Schanz, 2023).

Three-dimensional tracking is performed with either multiple cameras or with specialized setups and a single camera. Multi-camera approaches use multiple observations of the same particle from different cameras to triangulate its position. Two cameras are sufficient to solve for three-dimensional position, but additional cameras are necessary to resolve ambiguity when observing

multiple tracers. Single-camera approaches must uniquely encode the depth of a particle through alternate schemes. Approaches may work by modifying the imaging optics to create a point spread function (PSF) or optical transfer function (OTF), forming unique imaging patterns dependent on depth. This includes techniques such as defocus, astigmatic, coded aperture, plenoptic, or phase mask methods (Willert & Gharib, 1992; Kao & Verkman, 1994; Barnkob et al., 2015; Fahringer et al., 2015; Shechtman et al., 2015). Alternatively, depth may be encoded through an illumination scheme using spatially structured intensity or spectra (Aguirre-Pablo et al., 2019; Park et al., 2021).

A precursor step in image-based particle tracking algorithms is the effective detection of particles in recorded images. Intuitively, the resolution of the recovered velocity field depends on how densely the flow is sampled. High particle concentration is desirable for increased resolution, but complicates the task of tracking. Overlap of two or more tracer images diminishes detection capabilities and increases position estimate error. The probability of overlap scales with tracer concentration which is commonly quantified by two metrics: image density N_i usually reported as particles per pixel (ppp), and source density N_s (Adrian et al., 1991). Image density is straightforward to quantify, and it measures how efficiently the pixels in a sensor are being used. Source density is given by $N_i \bar{A}_p$ where $\bar{A}_p$ is the average image area of a single particle in pixels. This non-dimensional quantity more accurately reflects the particle detection difficulty, as larger particle images have an increased likelihood of image overlap. Single-camera approaches require a large particle area compared to multi-camera approaches to robustly encode depth. An appropriate definition of particle image area can make this quantity more ambiguous. At high enough source densities, individual particle images are no longer distinguishable, and tracking is impossible.

Because of the large variety in single-camera techniques, unique processing approaches are often devised for each unique optical configuration. To address the lack of software support, a modular software package has been proposed (Barnkob & Rossi, 2021). Alternatively, some single-camera approaches use splitting mirrors or plenoptic principles to mimic multi-camera systems by creating multiple non-overlapping views side-by-side on a single sensor. This allows triangulation to be used for 3D localization (Kreizer & Liberzon, 2011; Yamakaitis et al., 2025).

Recently, Godbersen et al. (2024) proposed PeakCNN, a novel convolutional neural network (CNN) architecture that detects particles in images with high concentration. In this case, the appearance of the particle is well described by a narrow Gaussian profile with known parameters. The particle images demonstrated were only a few pixels wide, as is typical for well-designed multi-camera setups. The approach showed significant performance improvement in detection and localization for both synthetic and experimental data. Here, we extend the approach and show that this CNN architecture naturally extends to defocus particle tracking (DPT) with similar performance improvements.

2. Synthetic defocus dataset

2.1. Data source

As discussed in Sect. 1, many methods have been developed and proposed for defocus tracking. Several recent works have explored deep learning approaches (König et al., 2020; Franchini & Krevor, 2020; Dreisbach et al., 2022). U-net architectures in particular have been applied for recovery of particle locations in plenoptic systems, but the U-net was used as an intermediate step to reconstruct 3D voxel intensities, not final particle positions (Cao et al., 2024; Zhu et al., 2024).

To the author's knowledge, only Barnkob et al. (2021) has pushed for the adoption of standard benchmark datasets. That work introduced nine synthetic datasets encompassing three different noise levels and three degrees of astigmatism. Several approaches, including a neural network, were tested in the initial work, and conventional approaches yielded the best results. Ratz et al. (2023) worked with the high-noise mild-astigmatism dataset from Barnkob et al. (2021) and utilized two separate detection and localization neural networks. The networks outperformed conventional techniques, but the densely seeded case remained challenging. Following this latter work, we use the high-noise mild astigmatism dataset as our benchmark.

2.2. Image generation

The original dataset images were generated using the open-source program MicroSIG (Rossi, 2020). The MicroSIG code is based on ray tracing and does not consider diffraction effects. Image generation can be slow as the number of particles increases, as each requires a large number of simulated rays. To decrease generation time, a different strategy was used. First, single-particle images with low noise and high resolution were generated for each depth z with the original MicroSIG code. This creates a calibration stack where the PSF $H(x, y; z)$ is sampled at discrete values. For each depth, the sampled PSF is normalized to have pixel intensities that sum to one. Images of the PSF are shown for five z -levels in Fig. 1. Synthetic images are generated with Eq. (1) consisting of two terms on the right-hand side: the summation of multiple PSF images and a noise term w .

$$F[x, y] = \sum_k^N A_k H(x - x_{c,k}, y - y_{c,k}; z_{c,k}) + w[x, y]. \quad (1)$$

The location parameters x_c, y_c, z_c and amplitude A are sampled for each particle k . The amplitude modulates signal intensity. Descriptions of how each parameter is randomly sampled are given in Table 1 where $\mathcal{N}$ and $\mathcal{U}$ denote normal and uniform distributions. The total number of particles N is sampled according to the desired image density. 5000 images with a resolution of 512×512 pixels were generated for three target image densities and two noise levels for a total of 30 000 images. This dataset was found to be sufficiently large to avoid overfitting while remaining compact

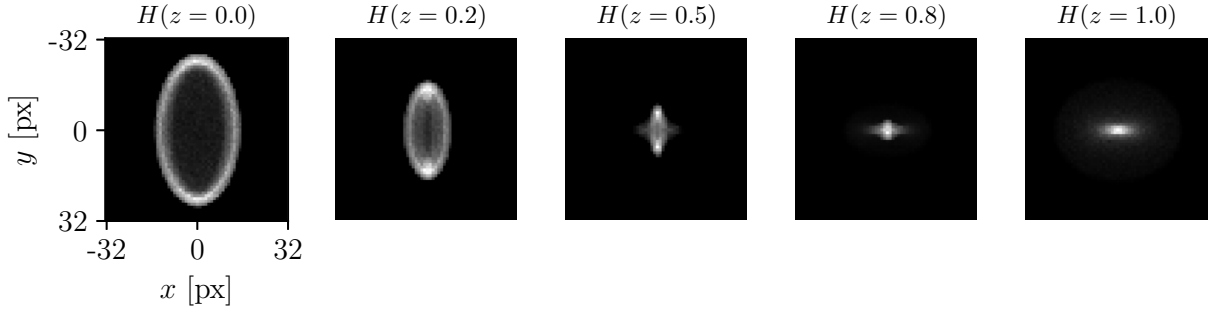


Figure 1. Defocus pattern with mild astigmatism at five discrete depths

Table 1. Parameters in synthetic image generation and their sampling rules

Parameter	Distribution	Parameters	Values
L	constant	—	512
N_i	grid value	—	$N_i = 0.001, 0.005, 0.025$
N	$ \mathcal{N}(\bar{N}, 0.2 \bar{N}) $	$\bar{N} = N_i L^2$	implied
x_c, y_c	$\mathcal{U}(0, L - 1)$	$L = 512$	continuous
z_c	$\mathcal{U}(0, 1)$	—	continuous
A	LogNormal	$\mu_A = 1, \sigma_A = e^{-1}$	continuous
w	$\mathcal{N}(0, \sigma_w)$	$\sigma_w = 2.5\text{e-}3, 6.25\text{e-}2$	continuous

enough to be stored entirely in CPU memory during training, avoiding slow disk operations. Representative generated images with labeled particle centers at four image densities and two noise levels are shown in Fig. 2. For each of fifteen image densities and two noise values, 32 images were generated and used for evaluation. Evaluation was also performed on images generated by the MicroSIG code (retrieved directly from ‘DefocusTracking.com’). Results for the MicroSIG images are denoted by the label ‘μSIG’ throughout the manuscript, and newly generated images are labeled by the noise level σ_ε used during their generation.

3. Network

3.1. Architecture

Here we repeat many of the details of Godbersen et al. (2024), but special attention is given to the changes introduced to improve network performance for strongly-defocused particle images. The primary difference in the network is the inclusion of a new per-pixel prediction channel in the output that allows for three-dimensional localization. In total, there are four output channels: p the

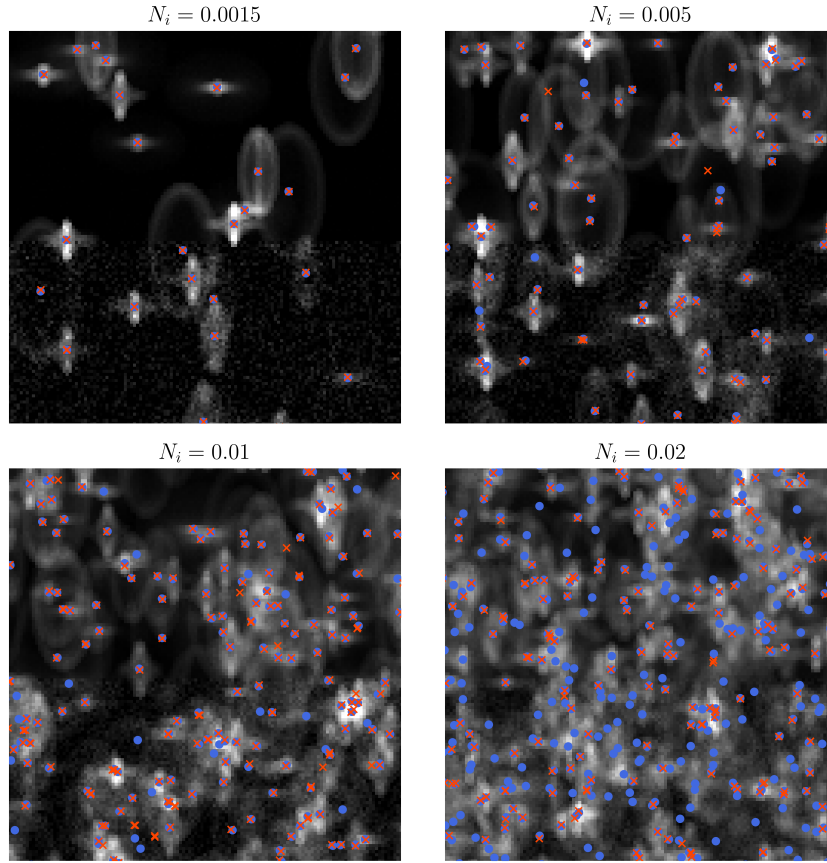


Figure 2. Synthetic images with ground truth (blue dots) and predicted (red crosses) particle centers

probability that a pixel contains a particle center, Δx and Δy the horizontal and vertical offsets of the particle center within that pixel, and z the absolute z position of the particle within that pixel. The latter three channels are only meaningful for pixels containing a particle center.

No parameters describing the PSF are explicitly given to the network. The only input to the trained network is the single-channel grayscale image. Structural information about the point spread function must be learned during the training phase from the labeled data. The parameter-free approach limits the network to the specific particle images it has been trained for, but allows the network to adapt to arbitrarily complex PSFs that are difficult to parameterize.

Each model follows a U-Net architecture with a constant width of 64 feature channels across all scales. The core building block is the ‘DoubleConv’ modules. At each resolution level, the DoubleConv module is applied, which contains two sets of 3×3 convolutions (without bias), batch normalization, and ReLU activation. Downsampling is performed using a ‘MaxPool’ operation to decimate, and upsampling is performed with a ‘ConvTranspose’ or fractional convolution to interpolate. To handle one-pixel mismatches due to odd-sized inputs, we include a bilinear interpolation fallback that ensures consistent tensor dimensions during decoding. The network in-

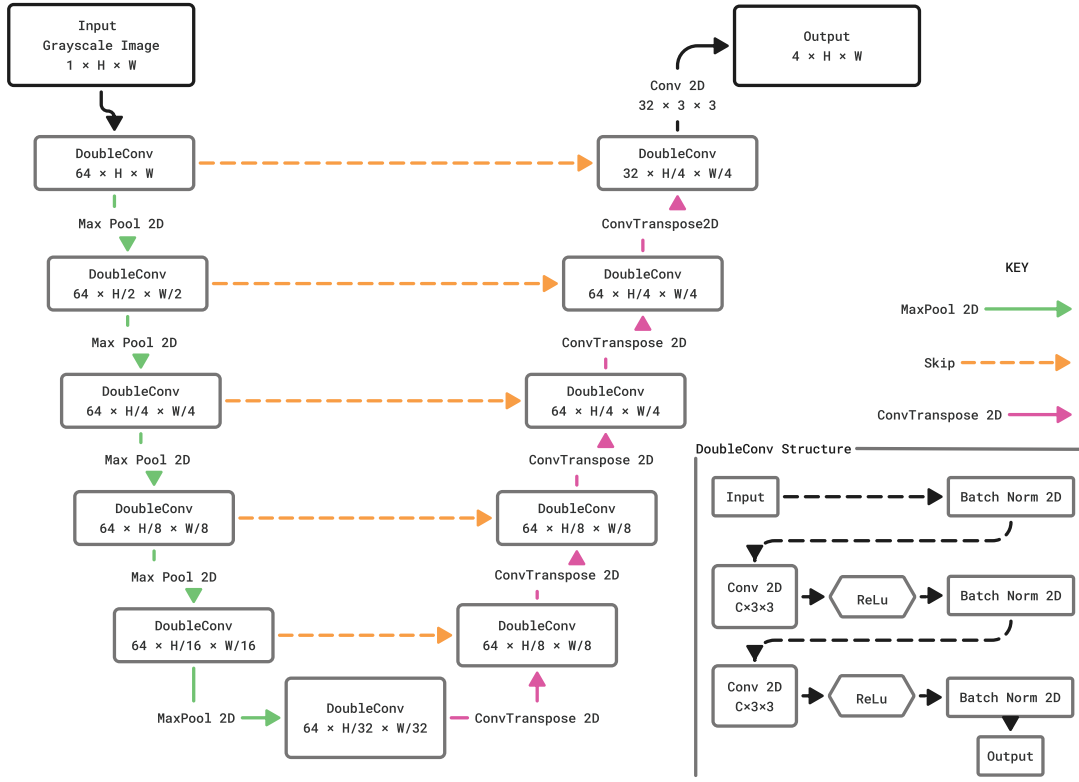


Figure 3. Architecture of the Convolutional U-net. Tensor dimensions are specified with the number of channels and as a fraction of the original images height H and width W in pixels.

cludes skip concatenations at each resolution level. A full network architecture diagram is given in Fig. 3. The total parameter count remains modest at approximately 8×10^5 . Here, a total of five down-sampling and up-sampling layers are used. The network bottleneck has a receptive field of 284×284 pixels, which is more than sufficient to capture long-range spatial correlations and depth cues encoded by the PSF, which has a maximum extent less than 64×64 pixels.

3.2. Training

The network is trained end-to-end using a compound loss function that combines a binary cross-entropy loss for dense pixel-wise probability channel and squared error for the localization channels. For a given pixel, let $(p, \mathbf{x})$ represent the ground-truth mask and regression targets, and $(\hat{p}, \hat{\mathbf{x}})$ their corresponding network prediction; the total loss is then given by Eq. (2).

$$\mathcal{L} = \lambda_{\text{cls}} \sum_i [p_i \ln \hat{p}_i + (1 - p_i) \ln(1 - \hat{p}_i)] + \sum_c \lambda_c \sum_i p_i (x_{c,i} - \hat{x}_{c,i})^2, \quad (2)$$

The index i indicates flattened pixel index and c indicates channel, corresponding to regression targets $\mathbf{x} = (\Delta x, \Delta y, z)$. Due to the large class imbalance even at the highest particle densities, the classification loss is set to $\lambda_{\text{cls}} = 1000$. Depth estimation is prioritized over lateral estimation, so

each is weighed by the following values,

$$\lambda_{\Delta x} = \lambda_{\Delta y} = 1 \quad \lambda_z = 100. \quad (3)$$

3.3. Performance metrics

Network performance needs to be evaluated in two categories: detection and localization. The first is characterized by precision P and recall R . These metrics are calculated using the number n of true positives TP , false positives FP , and false negatives FN as in Eqs. (4 & 5).

$$P = \frac{n_{TP}}{n_{TP} + n_{FP}} \quad (4)$$

$$R = \frac{n_{TP}}{n_{TP} + n_{FN}} \quad (5)$$

To count as a true positive, a predicted particle must be within 1 pixel (Euclidean distance) of a true particle that has not already been matched to any other predicted particles. A stricter requirement could also require a distance threshold in the z -direction; however, the choice of distance threshold would introduce a second arbitrary cutoff parameter. The prediction error in z is instead analyzed separately.

For a perfect detection scheme, both the precision and the recall values are 1. Since the network outputs a probability $\hat{p}$, the sensitivity of the network can be tuned arbitrarily by setting a threshold p_{thr} which must balance the precision and recall scores. To quantify detection efficacy as a single metric, we utilize F -score, which is given by Eq. (6) as the harmonic mean of precision and recall:

$$F = 2 \frac{PR}{P + R} \quad (6)$$

The maximum F -score for each test case is reported, and precision and recall are reported at the probability threshold that maximizes this score. A representative precision recall curve is given in Fig. 4.

The localization accuracy is estimated using the root mean square value between the predicted locations, and the actual location is taken for each component of the position.

4. Results

Qualitative network performance can be seen in Fig. 2, where predicted particle centers and true particle centers are overlaid on representative image samples. A representative precision-recall plot for the highest image density in the MicroSIG dataset is shown in Fig. 4. The network achieves over 95% precision and recall simultaneously.

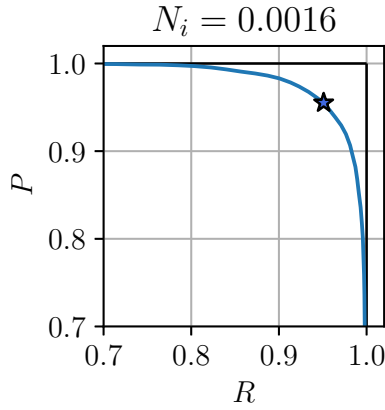


Figure 4. Precision-recall curve of the trained network for the highest image density in the MicroSIG dataset. Star denotes the threshold of the highest F -score.

Detection statistics as a function of image density are shown in Fig. 5 for both the MicroSIG dataset and the images generated through Eq. (1). At the seeding densities covered by both generation methods, the MicroSIG dataset falls between the two tested noise levels. This indicates that the network naturally extends to images generated with a ray-tracing model containing additional noise sources. The gray vertical bars represent the nominal image densities seen in training. As expected, network performance decreases as image densities increase. There is also a small decrease in network performance for low-density cases, not seen during training, even though these detection tasks are typically easier.

The localization error as a function of image density is shown in Fig. 6 for all true positive detections. All three directions show a consistent trend of increasing with image density.

Lateral errors in the x and y directions are apparently more sensitive to noise. At the lowest density, the higher noise case never decreases below 0.2 pixels. Both approach 0.3 pixels at the highest densities. This error is capped by the matching scheme that requires Euclidean distance between particles to be less than 1 pixels to be matched and included in the error statistics. The error in z begins to increase significantly at image densities beyond those seen in training. The depth-dependent nature of the PSF introduces non-homogeneous resolution. Figure 7 displays heat maps of predicted depth $\hat{z}$ against ground truth depth z for a single image density and two noise levels. The nearly linear trend demonstrates effective localization, but minor bias errors are noticeable. The precision error in z -localization dependence on z with minima near the center of the z -range. This indicates that boundary effects from the sharp cutoff in the z -range may have a prominent effect on localization error, especially for the low-noise case.

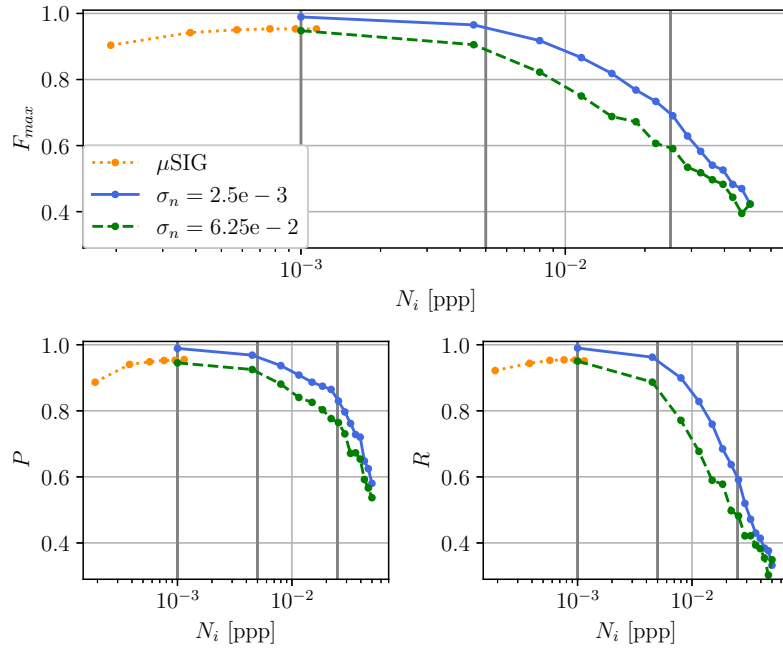


Figure 5. F -score (F_{max}), precision (P), and recall (R) as a function of image density

5. Conclusions

Convolutional U-nets have been demonstrated to be well-suited to the task of defocused particle detection and localization, even when particle concentration is high for a synthetic dataset. Once trained, the approach requires no time history or prior information beyond a grayscale image. The use of a synthetic dataset removes some complexity of real experimental data, but provides accurate benchmarks with an exact ground truth. The open availability of the chosen dataset allows for direct and reproducible comparison with other methods, and an extension of the dataset was required to accurately estimate performance limits, as results show significant improvement, achieving over 95% accuracy on challenging datasets.

Only a single defocus pattern is examined for detailed benchmarking. The method employed can easily be adapted to many other defocus or coded aperture style particle tracking setups since each shares the commonality of extracting repetitive, bright, and overlapped feature patterns, and the current approach makes no explicit assumptions about the parameterization of those patterns.

Future extensions could use the synthetic image generation introduced in Eq. (1) as a forward imaging model in the processing stage. Iterative strategies can remove detected particles from an image to enable iterative detection while also refining position estimates. Experimental results are the final test for validation. Early tests show promising results, but complete benchmarks are needed.

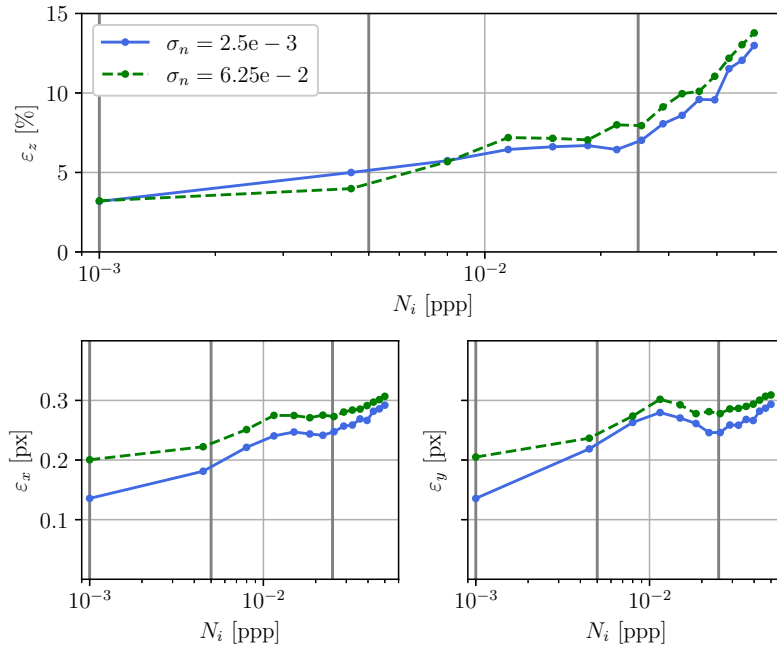


Figure 6. Localization error for true positive detections in all three dimensions. Gray vertical bars denote image densities used in training.

Acknowledgments

This work was supported by ONR grant N000142322001 managed by Dr. Yin-Lu (Julie) Young. M.J. Yamakaitis is a DoD SMART scholar. Joshua R. Rosen receives NSF Graduate Research Fellowship.

References

- Adrian, R. J., et al. (1991). Particle-imaging techniques for experimental fluid mechanics. *Annual review of fluid mechanics*, 23(1), 261–304.
- Aguirre-Pablo, A., Aljedaani, A. B., Xiong, J., Idoughi, R., Heidrich, W., & Thoroddsen, S. T. (2019). Single-camera 3D PTV using particle intensities and structured light. *Experiments in Fluids*, 60(2), 25.
- Barnkob, R., Cierpka, C., Chen, M., Sachs, S., Mäder, P., & Rossi, M. (2021). Defocus particle tracking: a comparison of methods based on model functions, cross-correlation, and neural networks. *Measurement Science and Technology*, 32(9), 094011.

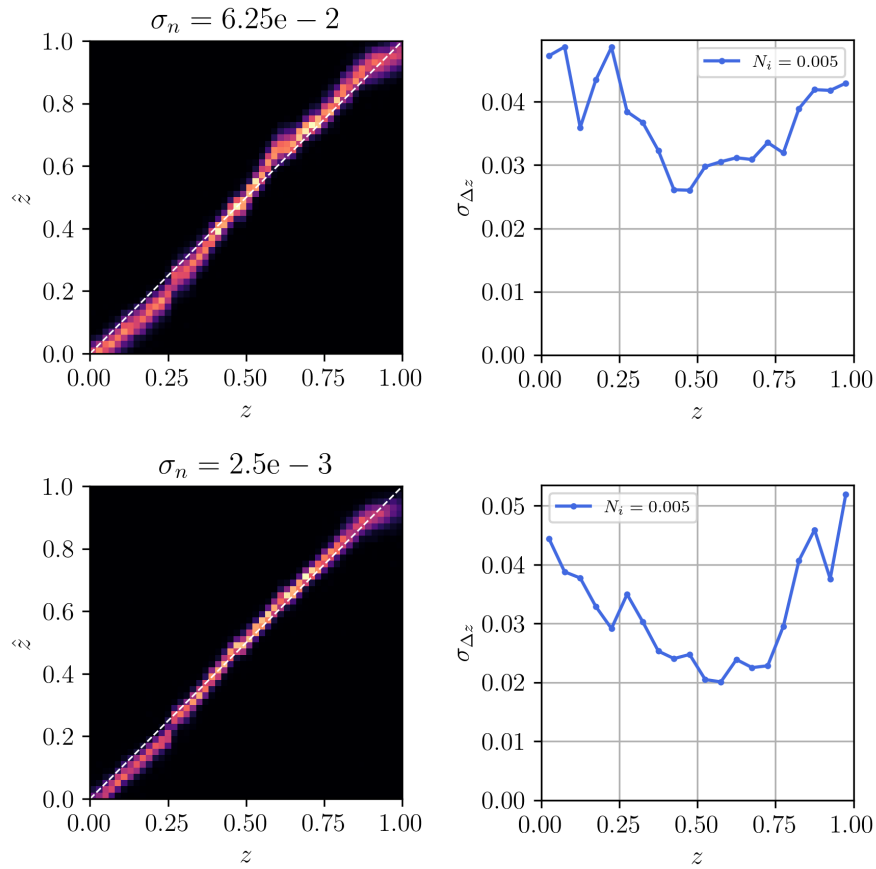


Figure 7. Heatmaps of predicted depth $\hat{z}$ against actual depth z , as well as average precision error $\sigma_{\Delta z}$ in prediction versus depth. Shown one image density $N_i = 0.005$ and two noise levels.

Barnkob, R., Kähler, C. J., & Rossi, M. (2015). General defocusing particle tracking. *Lab on a Chip*, 15(17), 3556–3560.

Barnkob, R., & Rossi, M. (2021). DefocusTracker: a modular toolbox for defocusing-based, single-camera, 3D particle tracking. *Journal of Open Research Software*, 9(1).

Cao, L., Hossain, M. M., Li, J., & Xu, C. (2024). Three-dimensional particle image velocimetry measurement through three-dimensional u-net neural network. *Physics of Fluids*, 36(4).

Dreisbach, M., Leister, R., Probst, M., Friederich, P., Stroh, A., & Kriegseis, J. (2022). Particle detection by means of neural networks and synthetic training data refinement in defocusing particle tracking velocimetry. *Measurement Science and Technology*, 33(12), 124001.

Fahringer, T. W., Lynch, K. P., & Thurow, B. S. (2015). Volumetric particle image velocimetry with a single plenoptic camera. *Measurement Science and Technology*, 26(11), 115201.

- Franchini, S., & Krevor, S. (2020). Cut, overlap and locate: a deep learning approach for the 3d localization of particles in astigmatic optical setups. *Experiments in Fluids*, 61, 1–15.
- Godbersen, P., Schanz, D., & Schröder, A. (2024). Peak-CNN: improved particle image localization using single-stage CNNs. *Experiments in Fluids*, 65(10), 153.
- Kao, H. P., & Verkman, A. (1994). Tracking of single fluorescent particles in three dimensions: use of cylindrical optics to encode particle position. *Biophysical journal*, 67(3), 1291–1300. doi: 10.1016/S0006-3495(94)80601-0
- König, J., Chen, M., Rösing, W., Boho, D., Mäder, P., & Cierpka, C. (2020). On the use of a cascaded convolutional neural network for three-dimensional flow measurements using astigmatic ptv. *Measurement Science and Technology*, 31(7), 074015.
- Kreizer, M., & Liberzon, A. (2011). Three-dimensional particle tracking method using FPGA-based real-time image processing and four-view image splitter. *Experiments in fluids*, 50, 613–620. doi: 10.1007/s00348-010-0964-3
- Park, H. J., Yamagishi, S., Osuka, S., Tasaka, Y., & Murai, Y. (2021). Development of multi-cycle rainbow particle tracking velocimetry improved by particle defocusing technique and an example of its application on twisted savonius turbine. *Experiments in Fluids*, 62(4), 71.
- Ratz, M., Sachs, S., König, J., & Cierpka, C. (2023). A deep neural network architecture for reliable 3D position and size determination for Lagrangian particle tracking using a single camera. *Meas. Sci. Technol.*, 34(10), 105203. doi: 10.1088/1361-6501/ace070
- Rossi, M. (2020). Synthetic image generator for defocusing and astigmatic piv/ptv. *Measurement Science and Technology*, 31(1), 017003.
- Schröder, A., & Schanz, D. (2023). 3D Lagrangian particle tracking in fluid mechanics. *Annual Review of Fluid Mechanics*, 55(1), 511–540.
- Shechtman, Y., Weiss, L. E., Backer, A. S., Sahl, S. J., & Moerner, W. (2015). Precise three-dimensional scan-free multiple-particle tracking over large axial ranges with tetrapod point spread functions. *Nano letters*, 15(6), 4194–4199. doi: 10.1021/acs.nanolett.5b01396
- Willert, C., & Gharib, M. (1992). Three-dimensional particle imaging with a single camera. *Experiments in Fluids*, 12(6), 353–358.
- Yamakaitis, M. J., Rezaigui, A., Fort, C., Capanna, R., & Bardet, P. M. (2025). Four-view single-camera plenoptic 3.0 imager for three-dimensional lagrangian particle tracking in an oscillatory laminar boundary layer. *Experiments in Fluids*, 66(9), 165. doi: 10.1007/s00348-025-04095-w

Zhu, X., Fu, M., Xu, C., Hossain, M. M., & Khoo, B. C. (2024). Volumetric reconstruction of flow particles through light field particle image velocimetry and deep neural network. *Physics of Fluids*, 36(7).