Contribution ID: **21** Type: **not specified**

PixelPatrol: Pre-validation for Scientific Image Datasets

Ensuring the quality and integrity of scientific image datasets upfront is crucial for reliable downstream analysis. This is particularly important with the increasing scale and complexity of data used in contemporary research, including large-scale training sets for applications such as foundation models. Discovering issues like inconsistencies or artifacts late in the workflow can lead to significant wasted effort and flawed results. To address this, we introduce PixelPatrol, an early version of a tool designed for the systematic pre-validation of scientific image datasets.

PixelPatrol provides researchers with capabilities to proactively assess their data before engaging in computationally intensive analysis. It offers dataset-wide visualization and interactive exploration, helping users gain a deep understanding of their data's structure and content across various dimensions. The tool generates detailed statistical summaries and visualizations covering aspects like image dimensions, sizes, pixel value distributions, and intensity statistics, often presented in plots and distributions per dimension. This enables the early identification of potential issues, discrepancies, or unexpected characteristics, including those related to metadata and acquisition parameters, and helps in finding outliers. Furthermore, the tool supports comparing these statistics and visualizations across different experimental conditions. Designed to be user-configurable, PixelPatrol empowers researchers to tailor checks to their specific needs and datasets. This pre-validation process not only flags problems but also empowers researchers to truly get to know their datasets, leading to more informed analytical choices and a better understanding of data variability and potential biases.

Built with future flexibility and scalability in mind, PixelPatrol is being developed to support diverse imaging modalities and large datasets. By integrating early validation and detailed data inspection into the research workflow, PixelPatrol helps improve data quality awareness, enhances the reliability of analyses, and supports the principles of reproducible science within the Helmholtz community.

Primary authors: BAHRY, Ella (Helmholtz Imaging / MDC); OTTO, Maximilian; ALBRECHT, Jan Philipp (MDC); SCHMIDT, Deborah

Session Classification: Data Curation & Data Handling