Multicore World 2026

Accelerating machine vision and computation tasks with Sparsity, Vulkan and SPIR-V

David Brebner

umajin

Speeding things up

Topics

- Sparsification
- Inverse rendering
- Vulkan and SPIR-V
- Wave Intrinsics
- Anomaly detection

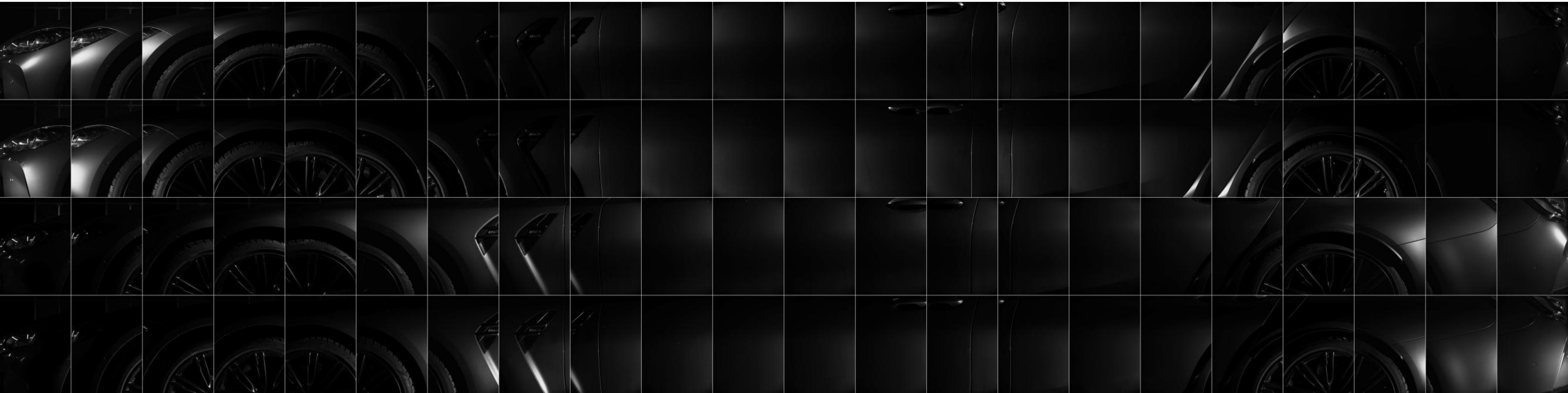
Data reduction & Sparsification

- 12.4 MP (4073 x 3055)
- 200 images per scan
- Over 2 gigabytes of raw data
- Deconvolution of photometric properties via inverse rendering reduces the data to <200 megabytes
- Feature detection means defects only transmitted < 5 megabytes

Why GPU?

- Designed to render pixels they contain thousands of small cores that handle math for graphics so ideal for inverse rendering

Automotive inspection context



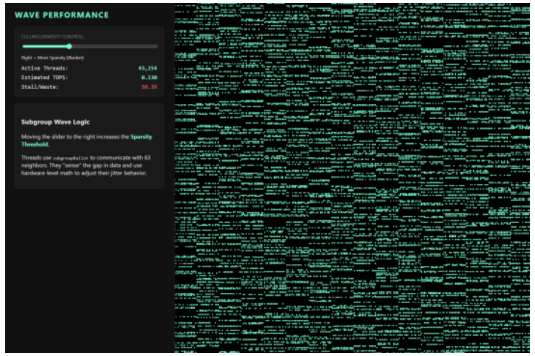
Data reduction & Sparsification

SPIR-V Wave Intrinsics (subgroup operations)

- A powerful feature for SIMD (Single Instruction, Multiple Data) processing on GPUs, allowing threads within a "wave" to exchange data without explicit barriers
- Waves/Subgroups act as a group of threads executed in lockstep (typically 32 or 64 lanes)
- The most critical part of wave organization is how the hardware handles **Sparsity** (if-statements)
- The registers are "sliced" across the wave. When you call a wave intrinsic like WaveActiveSum(), the hardware doesn't send data to/from RAM, it uses an internal Crossbar Switch or Shuffle Network

Grid	Total Task	The entire GPU
Workgroup	256–1024 Threads	Shared L1 Cache / Local Memory
Wave (Subgroup)	**32–64 Threads**	**Instruction Pointer / Shuffle Network**
Lane (Thread)	1 Thread	Individual Scalar Registers

Wave demo



WAVE PERFORMANCE

CULLING (SPARSITY CONTROL)

Right = More Sparsity (Blacker)

```
Active Threads:        43,254
Estimated TOPS:         0.130
Stall/Waste:            50.3%
```

Subgroup Wave Logic

Moving the slider to the right increases the Sparsity Threshold.

Threads use `subgroupBallot` to communicate with 63 neighbors. They "sense" the gap in data and use hardware-level math to adjust their jitter behavior.

SPIR-V normalising before integrating

bind a set of storage buffers

```
StorageBuffer_BIND(0, uint) uintMin;
StorageBuffer_BIND(1, uint) uintMax;
StorageBuffer_BIND(2, float) globalSum;
StorageBuffer_BIND(3, uint) globalCount;
Float32C1_BIND(0) map;
```

accessible group shared memory

```
groupshared float minData[MAX_THREADS];
groupshared float maxData[MAX_THREADS];
groupshared float sumData[MAX_THREADS];
groupshared uint countData[MAX_THREADS];
groupshared uint waveCount;
```

compute job is then dispatched with a thread invocation for every pixel in the input image

```
COMPUTE_WITH_SPEC(uint3 dispatchThreadID: SV_DispatchThreadID,
uint li: SV_GroupIndex)
{
    int2 coord = int2(dispatchThreadID.xy);
    float val = map.Load(coord);
```

find intrinsics over subgroup / wave

```
    bool valid = isfinite(val);
    float wMin = WaveActiveMin(valid ? val : POS_INF);
    float wMax = WaveActiveMax(valid ? val : NEG_INF);
    float sum = WaveActiveSum(valid ? val : 0);
    uint count = WaveActiveSum(valid ? 1u : 0u);
```

synchronises the memory of the groupshared variables

```
if (li == 0)
{
   waveCount = 0;
}
GroupMemoryBarrierWithGroupSync();

if (WaveIsFirstLane())
   {
       uint idx;
       InterlockedAdd(waveCount, 1, idx);
       minData[idx] = wMin;
       maxData[idx] = wMax;
       sumData[idx] = sum;
       countData[idx] = count;
   }
GroupMemoryBarrierWithGroupSync();
```

SPIR-V normalising before integrating

Once the group has made the data set of results from it's Subgroups/Waves, each invocation within the Group performs a parallel reduction of the set, to find the resulting values across that Group's pixels

```
 for (uint stride = (MAX_THREADS >> 1); stride > 0; stride >>= 1)
  {
    if (li < stride && (li + stride) < waveCount)
    {
      minData[li] = min(minData[li], minData[li + stride]);
      maxData[li] = max(maxData[li], maxData[li + stride]);
      sumData[li] = sumData[li] + sumData[li + stride];
      countData[li] = countData[li] + countData[li + stride];
    }
    GroupMemoryBarrierWithGroupSync();
  }
```

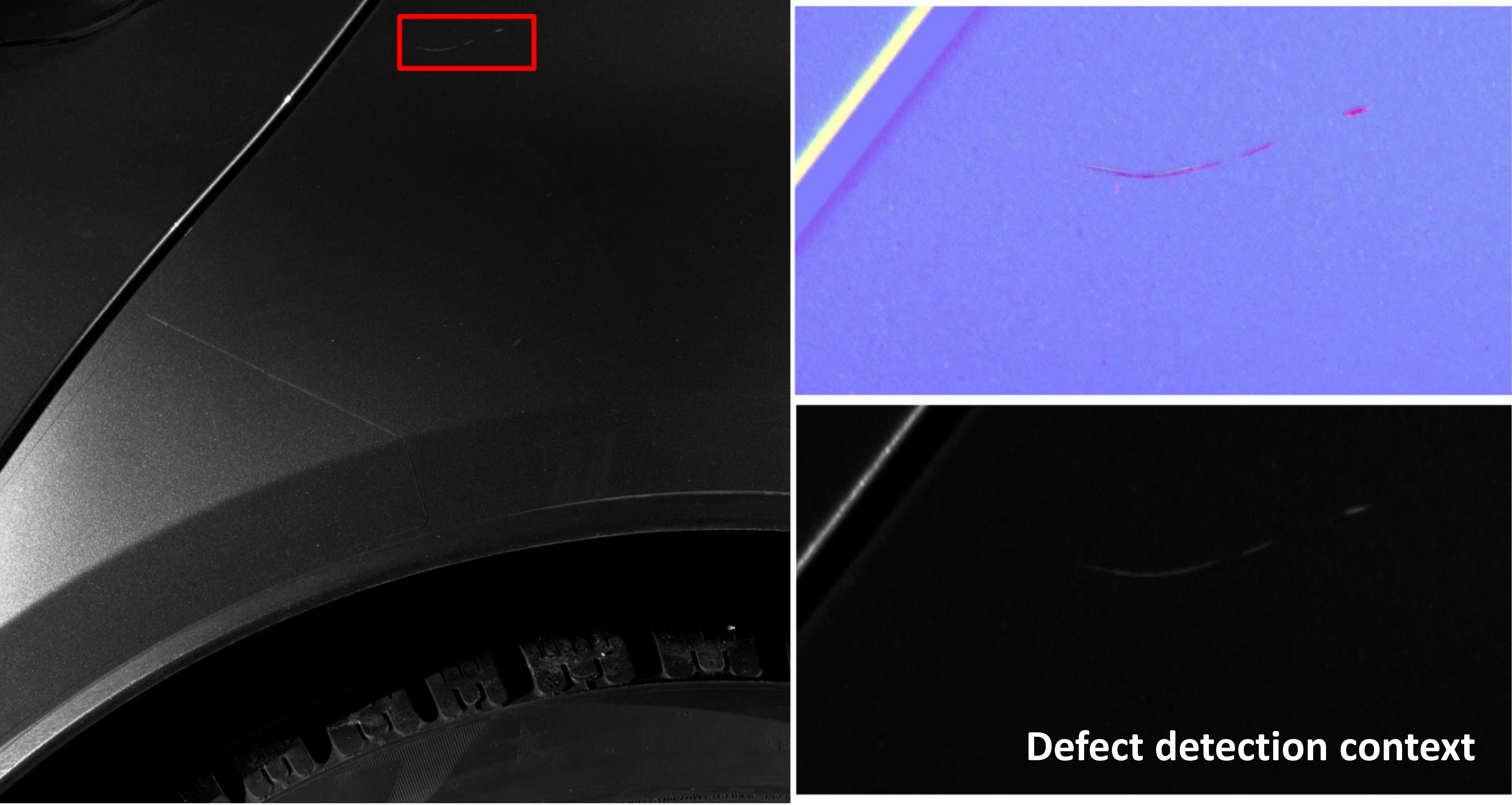
Finally, if an invocation is itself the first index of a Group, and that Group has a positive waveCount, it then contributes the Group's results to the global buffers

```
  if (li == 0 && waveCount > 0)
  {
    uint uMin = floatToOrderedUint(minData[0]);
    uint uMax = floatToOrderedUint(maxData[0]);
    InterlockedMin(uintMin[0], uMin);
    InterlockedMax(uintMax[0], uMax);
    InterlockedAdd(globalSum[0], sumData[0]);
    InterlockedAdd(globalCount[0], countData[0]);
  }
}
```

This approach means we can find relative values globally across a large image, while only performing blocking atomic operations once per invocation Group. While the exact numbers depend on hardware group sizes, this reduces the number of atomic calls across the whole input image by multiple orders of magnitude.

Summary: Finding global values

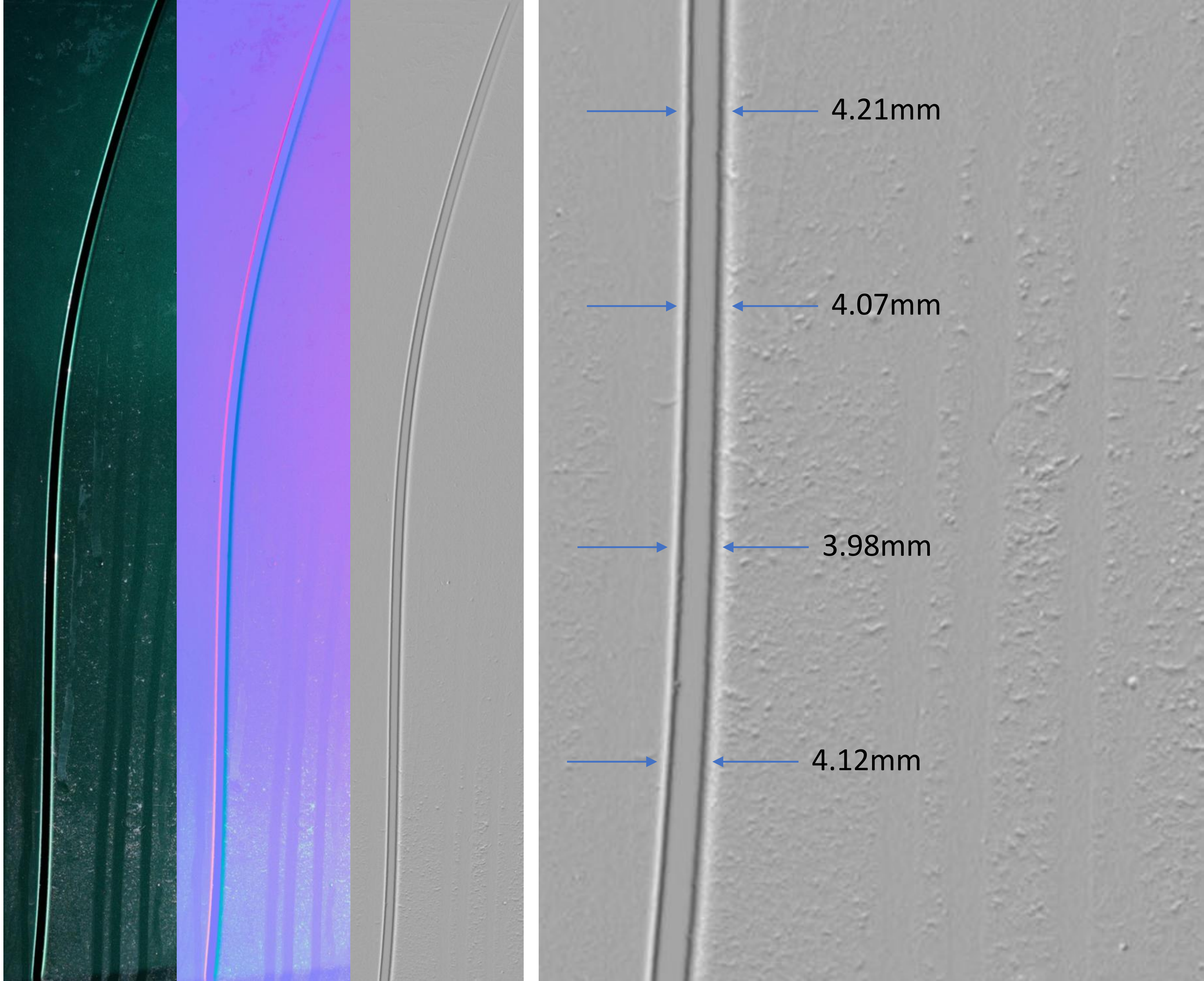
1. Subgroup/Wave Reduction
2. Group Reduction
3. Global Atomic



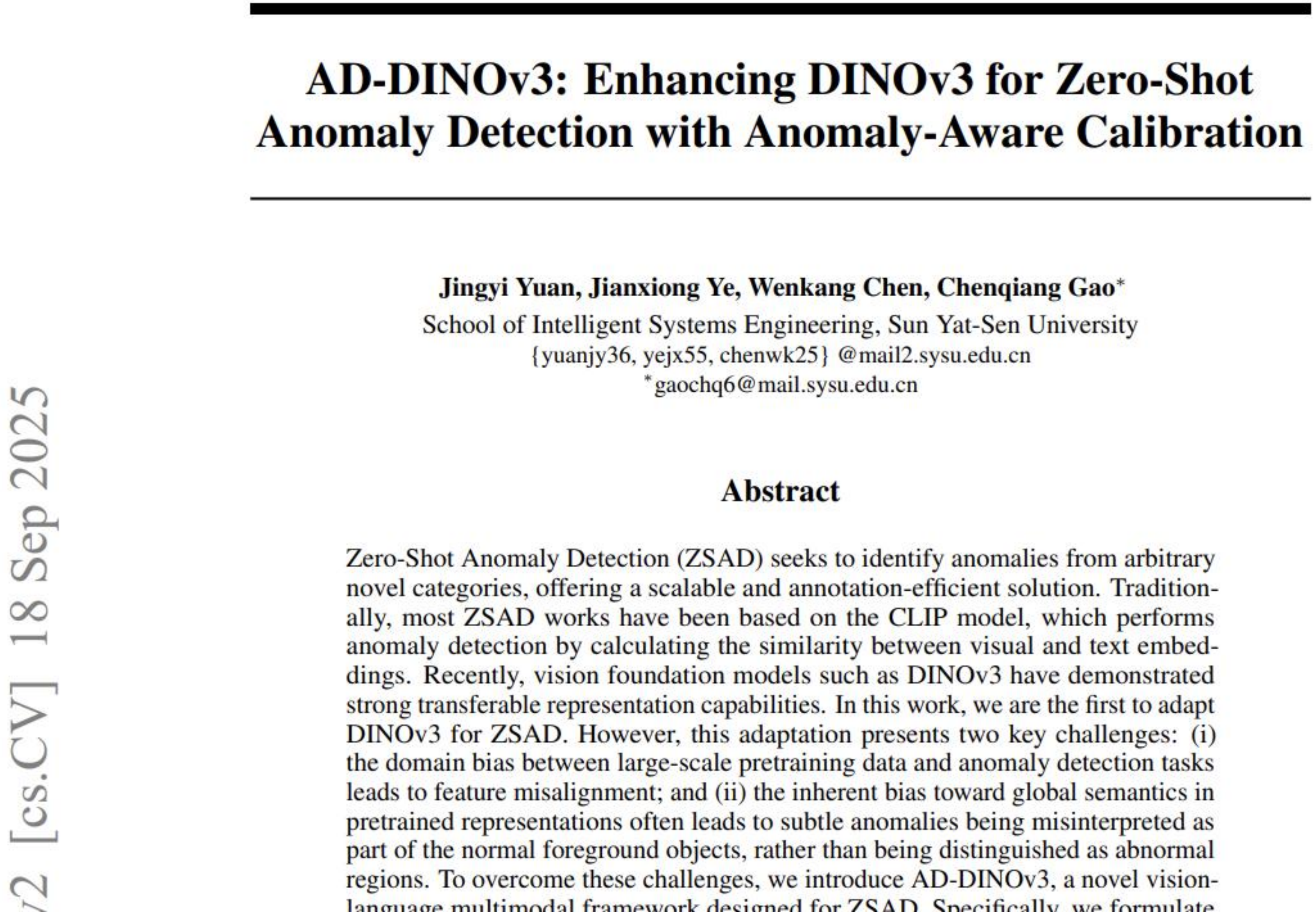
Defect detection context

Measurements

Precise measurement of panel gaps is performed using the normal map edge detection with the estimated depth map. This is able to be performed even with salt, dirt and moisture on the car surface.



Anomaly detection



AD-DINOv3: Enhancing DINOv3 for Zero-Shot Anomaly Detection with Anomaly-Aware Calibration

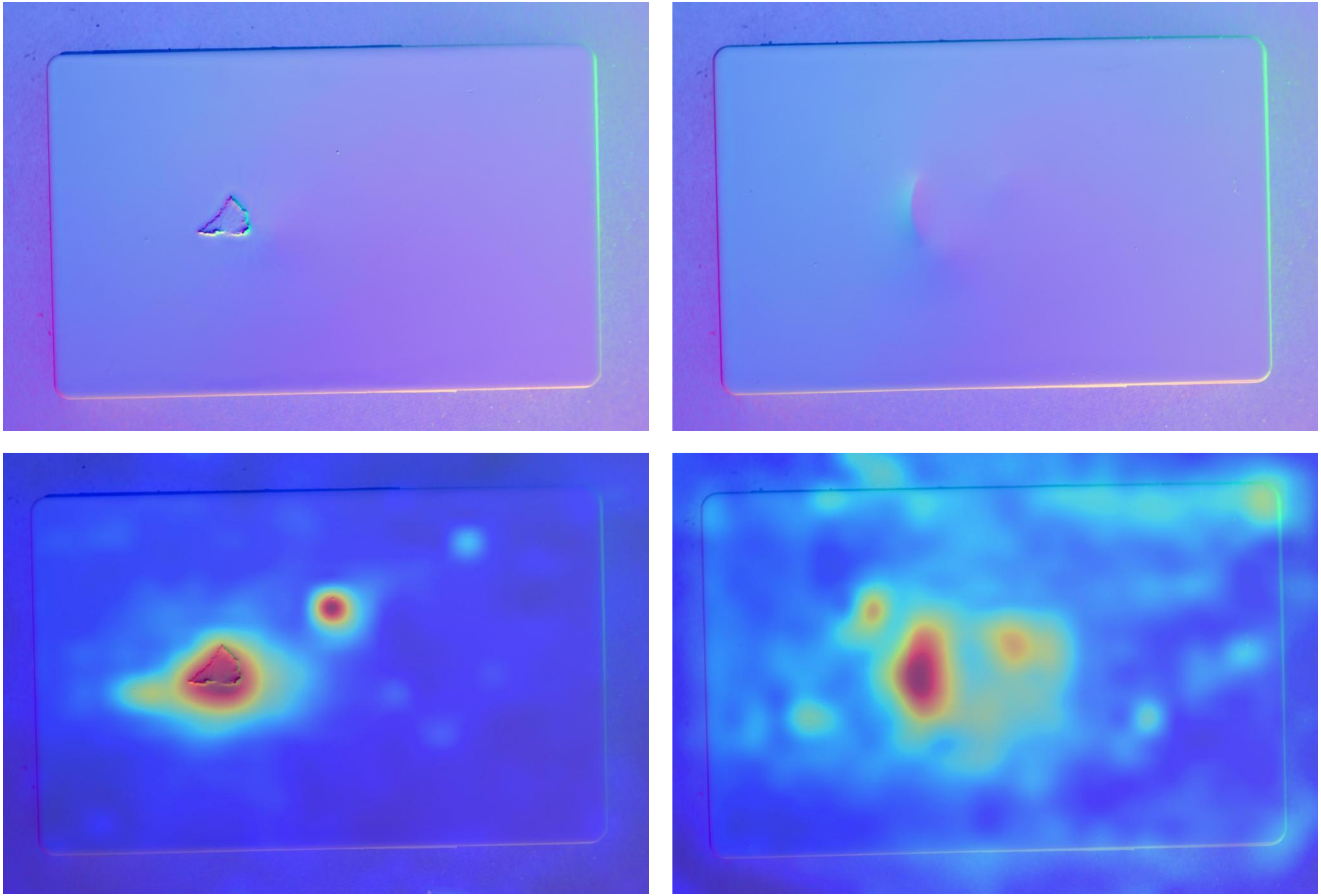
Jingyi Yuan, Jianxiong Ye, Wenkang Chen, Chenqiang Gao*
School of Intelligent Systems Engineering, Sun Yat-Sen University
{yuanjy36, yejx55, chenwk25} @mail2.sysu.edu.cn
*gaochq6@mail.sysu.edu.cn

Abstract

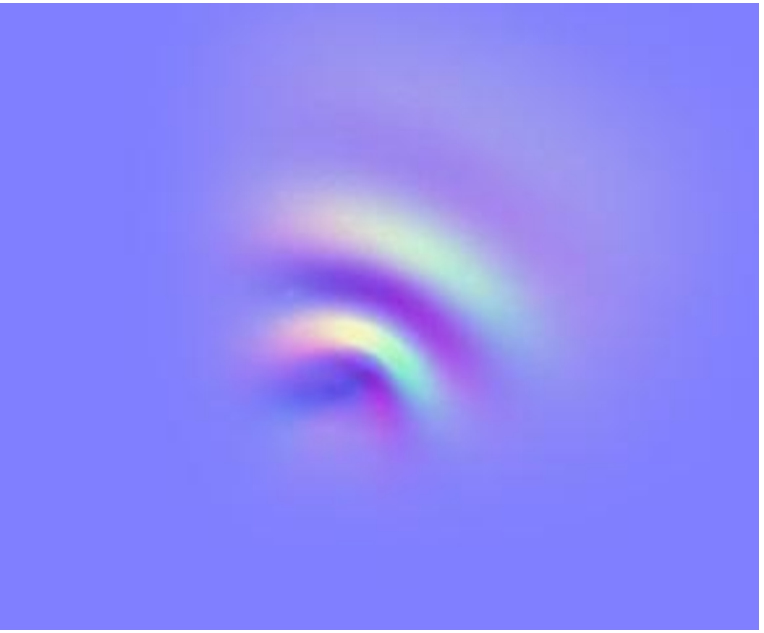
Zero-Shot Anomaly Detection (ZSAD) seeks to identify anomalies from arbitrary novel categories, offering a scalable and annotation-efficient solution. Traditionally, most ZSAD works have been based on the CLIP model, which performs anomaly detection by calculating the similarity between visual and text embeddings. Recently, vision foundation models such as DINOv3 have demonstrated strong transferable representation capabilities. In this work, we are the first to adapt DINOv3 for ZSAD. However, this adaptation presents two key challenges: (i) the domain bias between large-scale pretraining data and anomaly detection tasks leads to feature misalignment; and (ii) the inherent bias toward global semantics in pretrained representations often leads to subtle anomalies being misinterpreted as part of the normal foreground objects, rather than being distinguished as abnormal regions. To overcome these challenges, we introduce AD-DINOv3, a novel vision-language multimodal framework designed for ZSAD. Specifically, we formulate

1. **Train a zero shot model on high quality samples so it learns the natural variation in the sample**

2. **Run the model over a new sample image and represent the distance from the expected gold master set as a heatmap**



Defect classification

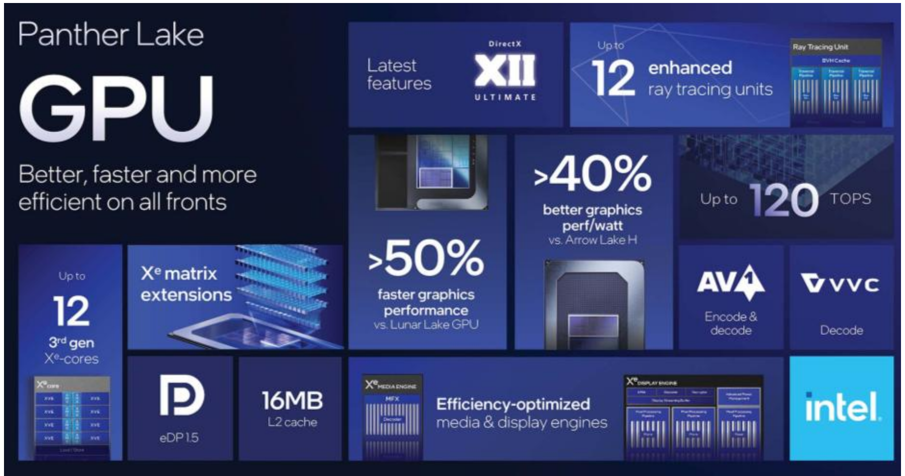


First pass is anomaly detection
(does not match the CAD based gold master)

Second pass is classification and quantification

Why NPU?

- Designed specifically for the math of deep learning—primarily Matrix-Vector Multiplication. Memory passes from cell to cell without going back to the shared memory bus



Panther Lake

GPU

Better, faster and more efficient on all fronts

Latest features DirectX XII ULTIMATE

Up to 12 enhanced ray tracing units

>40% better graphics perf/watt vs. Arrow Lake H

Up to 120 TOPS

Up to 12 3rd gen X^e-cores

X^e matrix extensions

>50% faster graphics performance vs. Lunar Lake GPU

AV1 Encode & decode

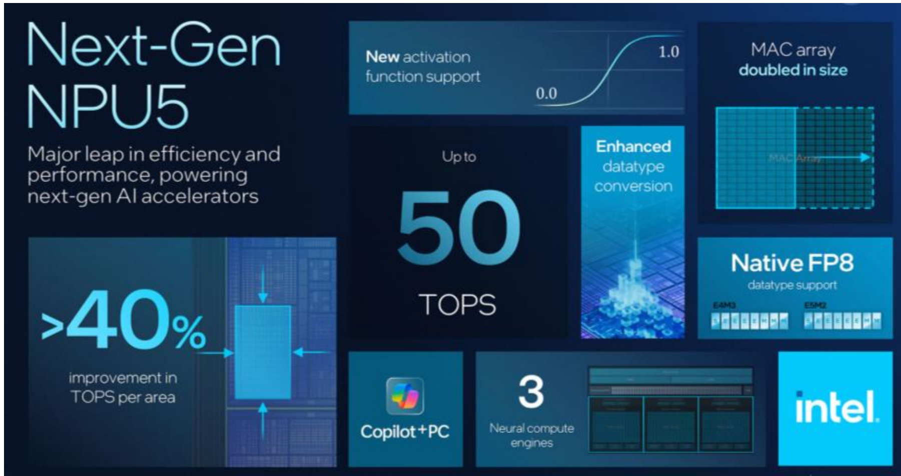
VVC Decode

eDP 1.5

16MB L2 cache

Efficiency-optimized media & display engines

intel.



Next-Gen NPU5

Major leap in efficiency and performance, powering next-gen AI accelerators

New activation function support

MAC array doubled in size

Up to 50 TOPS

Enhanced datatype conversion

Native FP8 datatype support

>40% improvement in TOPS per area

Copilot+PC

3 Neural compute engines

intel.

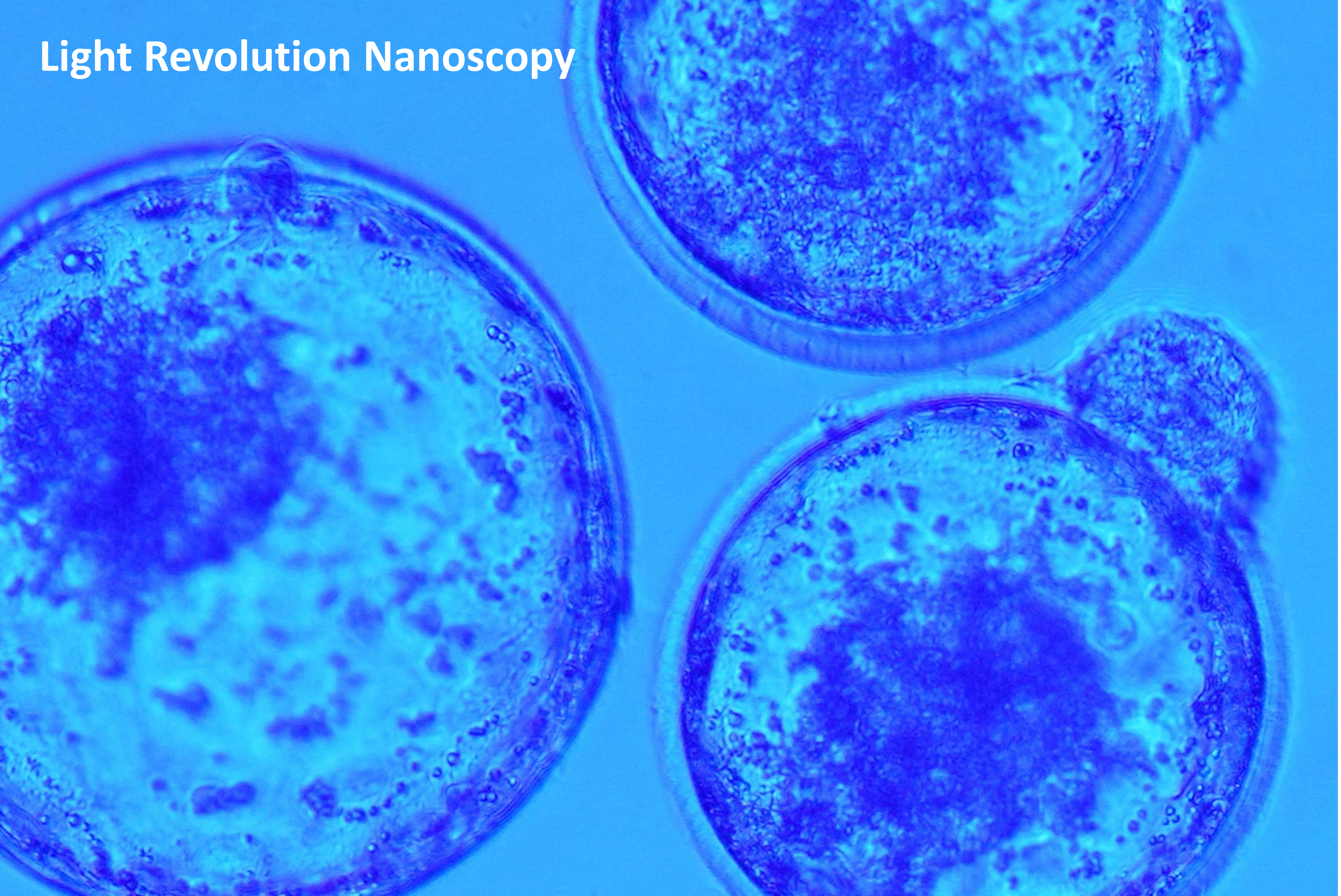
1.0
0.0
E4M3
E5M2

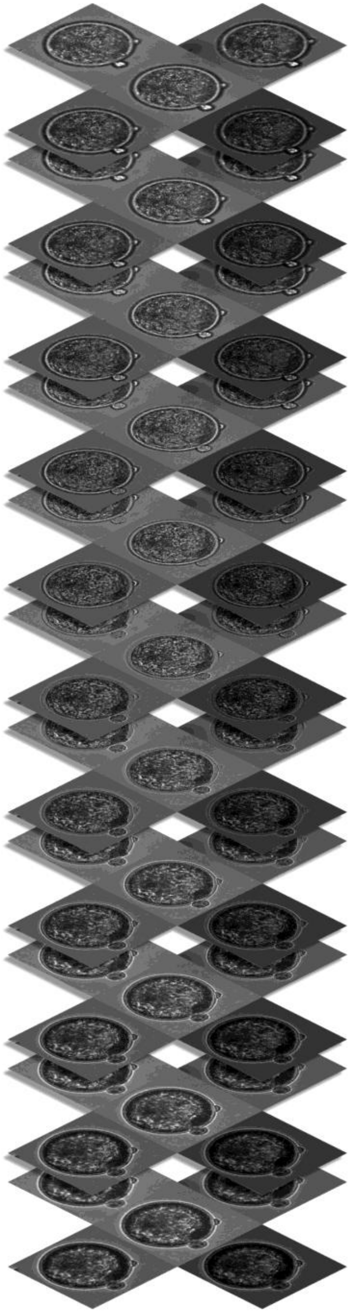
Umajin scales up inspection

Umajin provides huge benefits to scaling up inspection

- High speed coded image capture
- Sparse GPU accelerated inverse rendering to recover photometric data
- Advanced anomaly detection on manufactured parts using multi later photometric data
- High speed classification and measurement of anomaly's that are detected
- Integration with quality and manufacturing systems and production line PLC’s

Light Revolution Nanoscopy





Sparsity of illumination

Light Revolution: Sparsity of Illumination

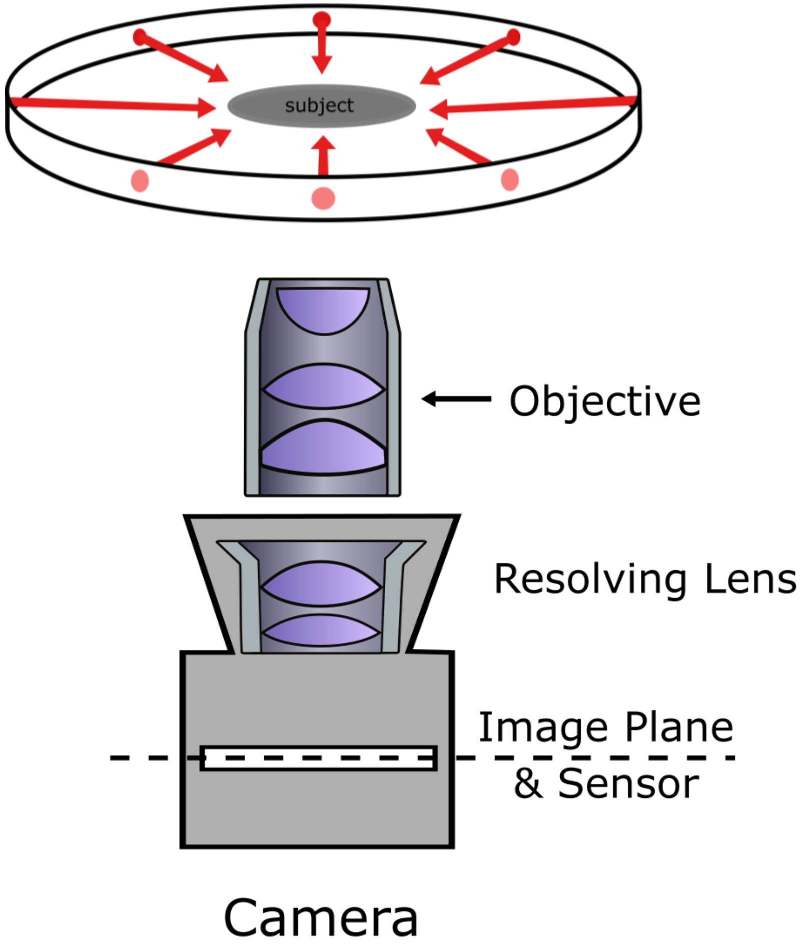
- Standard wide-field microscopy use Köhler bathing the subject in uniform light. This intentionally reduces the phase contrast of the system as it isolates the focal plane. The Light rays arriving from opposing angles mutually cancel the shadow casting and phase gradients that define high-frequency surface details
- By activating a single, incoherent point-light source (LED) at a time, we ensure that the photon ‘origin’ is largely spatially distinct. This sparsity results in maximum phase conversion at all depths of field. This allows the system to recover sharp edge responses even in regions that would be considered "out of focus" or low-contrast in a traditional bright-field arrangement
- This also results in an arrangement which is not light starved as the LED is able to illuminate the camera sensor and is only blocked where light is absorbed or refracted away allowing very high-speed capture with minimal shot noise

Sparsity of subject

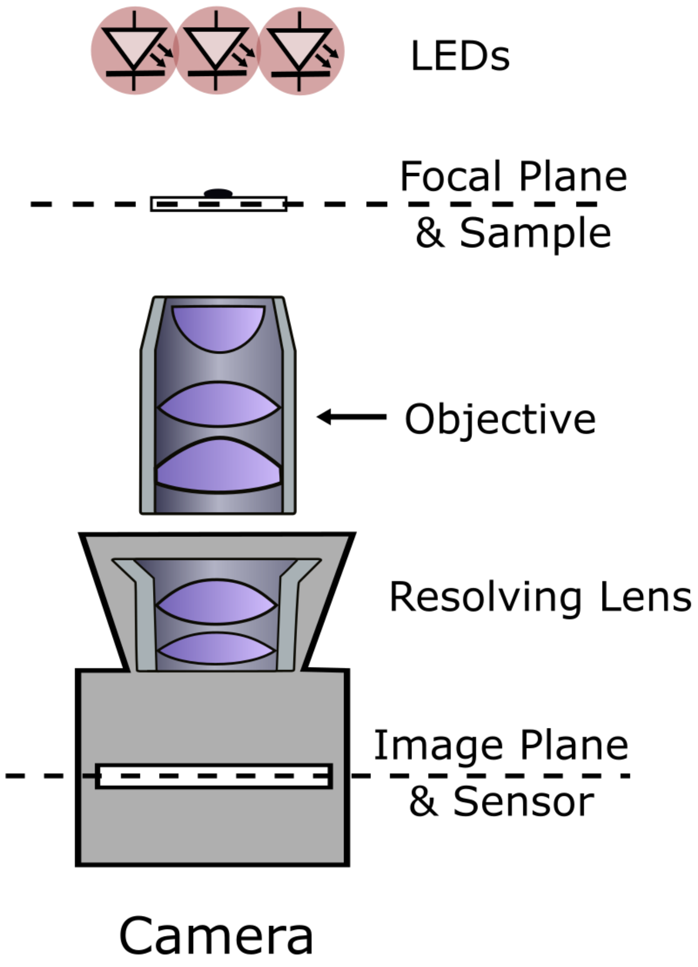
Light Revolution: Sparsity of the Subject

- The second governing principle is the sparsity of the subject relative to the detector. The system is designed to significantly oversample the optical resolution limit (100 nm pixel pitch vs. ≈ 980 nm Rayleigh limit)
- In many biological contexts and areas like photonics there are large homogeneous areas that are punctuated with "sparse" small features separated from neighbouring structures by distances exceeding the Rayleigh criterion (separated enough they do not blur together). When an isolated sub-resolution feature is imaged with this sparse illumination strategy, it behaves as a discrete signal source rather than part of a convoluted texture
- This capability is particularly evident in low-contrast biological samples. For example, visualizing the end-piece of a sperm flagella is typically too small and low contrast. However, because the feature is sparse even the weak gradient of the photometric response of the flagella tip can be isolated and reconstructed. The inverse rendering algorithm detects this sparse signal by its consistent orientation response across the different off axis lighting angles

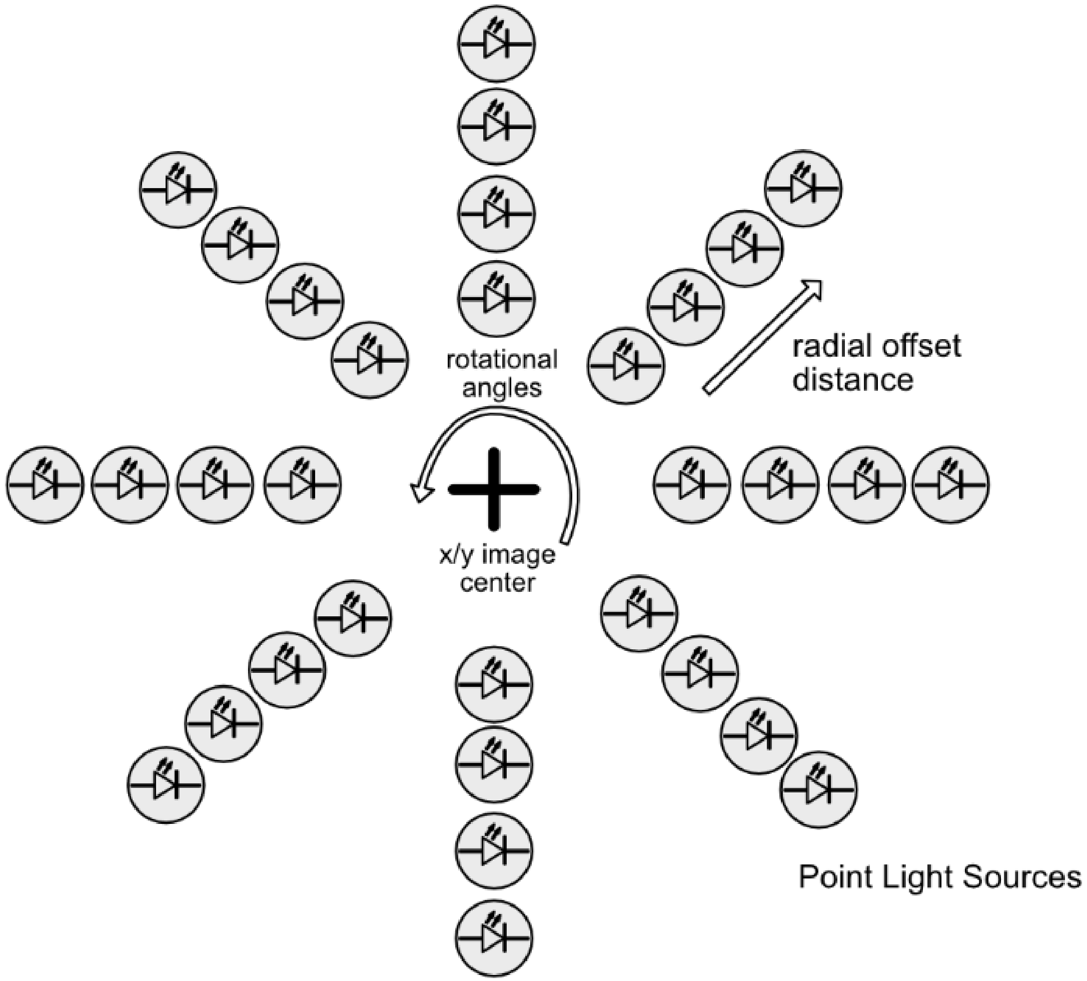
Specular Photometry

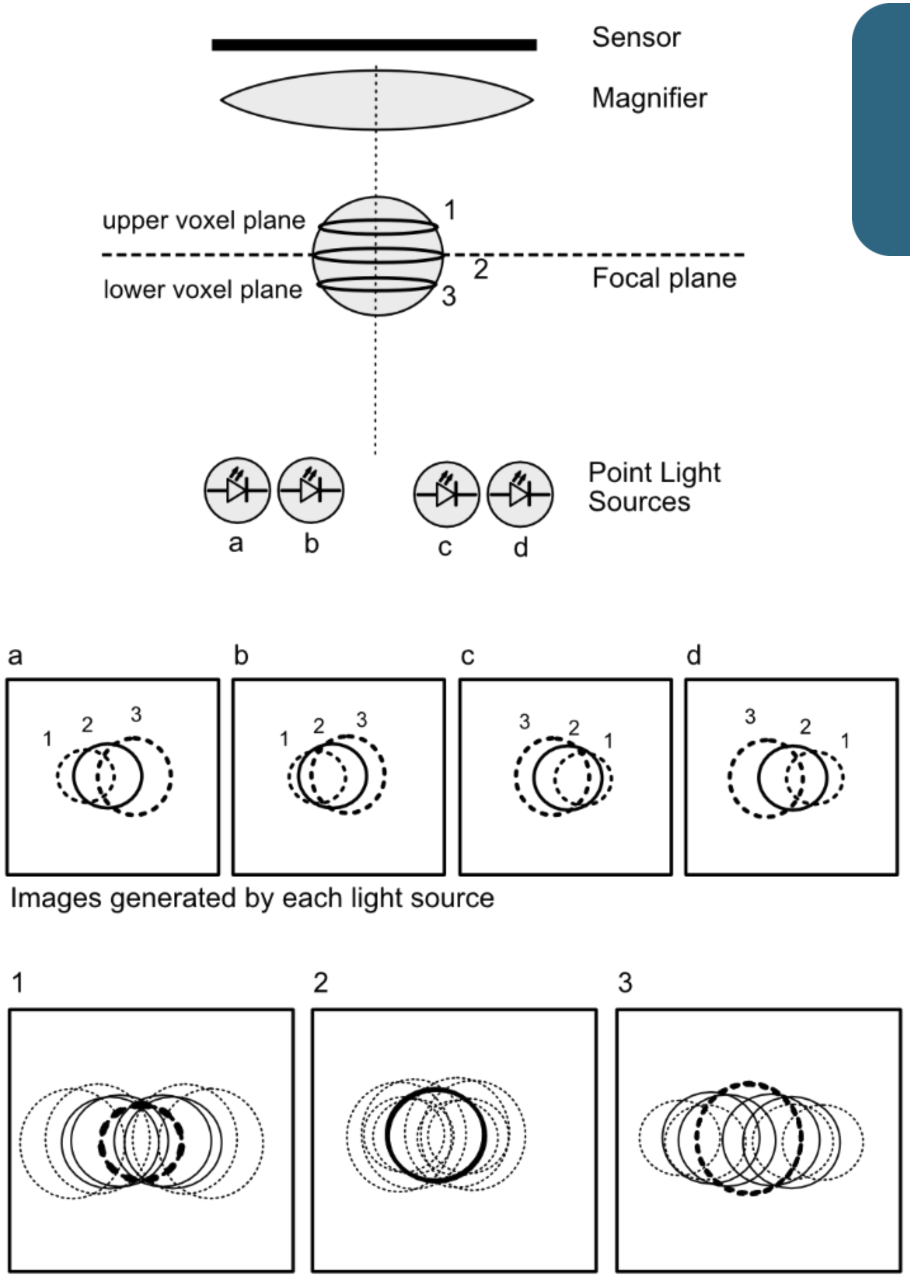


Light Revolution
Photon Tracing Microscope



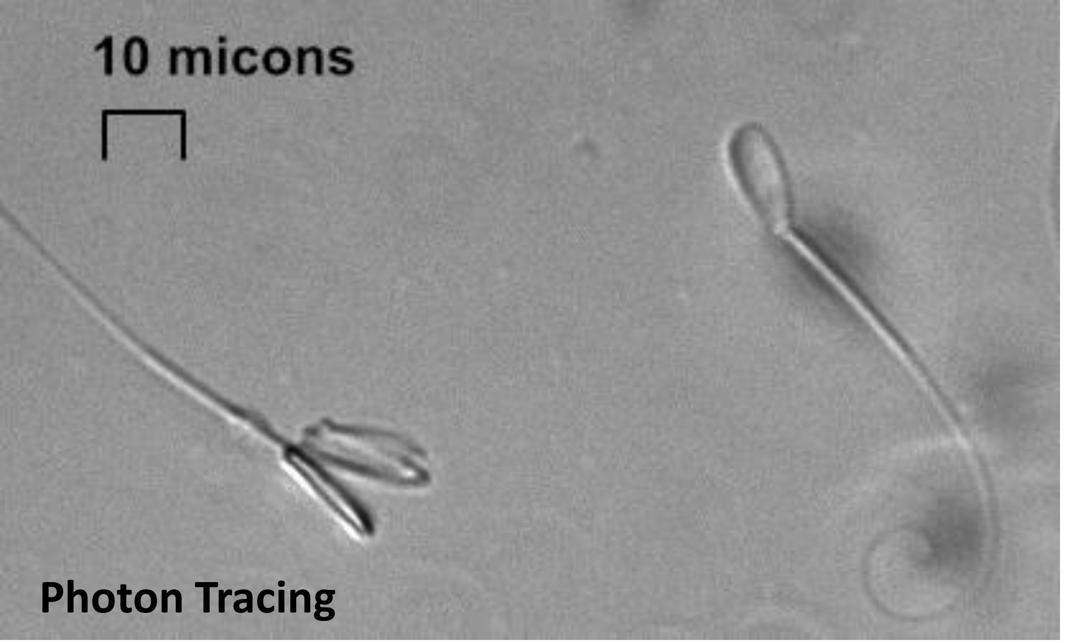
Light Revolution



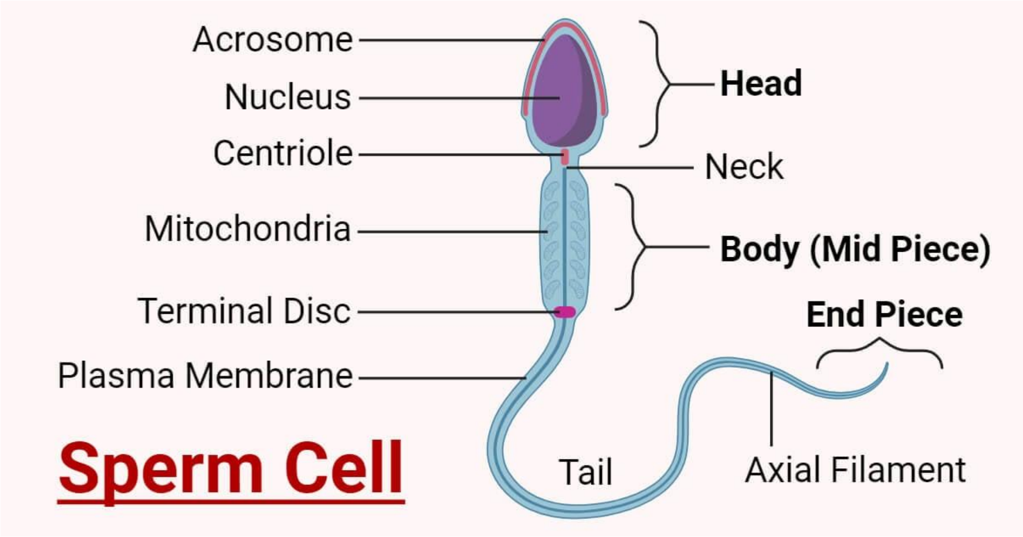


Images generated by each light source

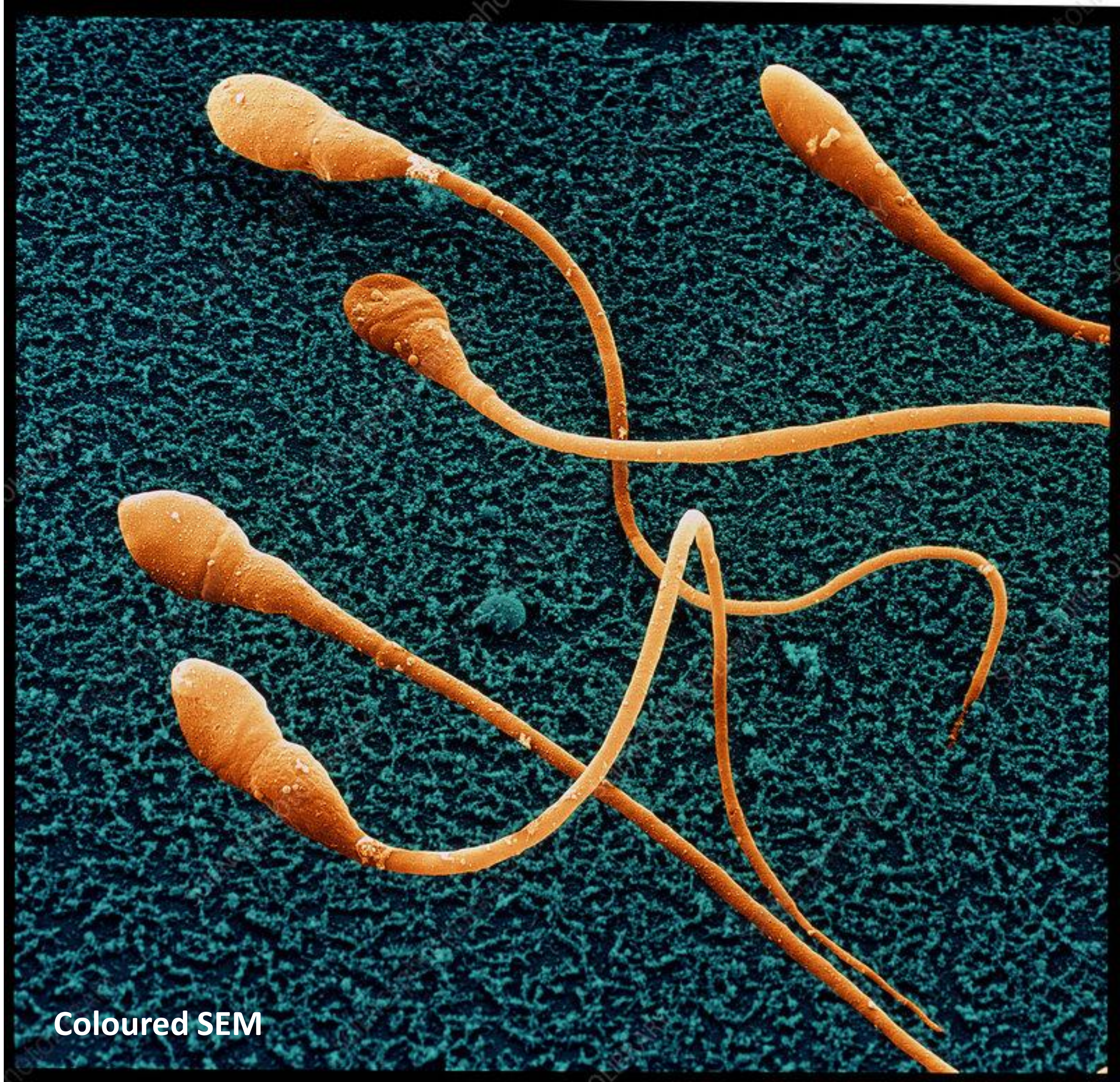
Images generated by stacking each light source image with suitable focal plane pivot x/y offsets applied to align on voxel slices 1, 2 and 3



Photon Tracing



Sperm Cell



Coloured SEM

Nuclides

Overall Concept

- Can we use alternative models of complex physics systems?
- By using a model which is geometry based – we can use a direct digital twin rather than having to simulating quantum states
- Speedup by 1,000,000 times or more

Nuclides

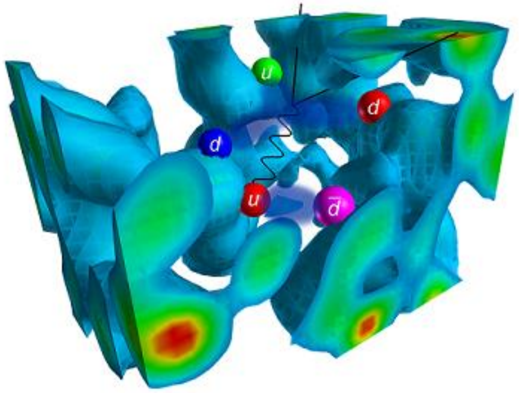
Quantum Field Dynamics (QFD) framework 2026

Tracy McSheery, Phasespace

- Generating nuclides is treated as a problem of geometric resonance within a stiff vacuum medium, rather than the traditional High-Performance Computing (HPC) approach of simulating n-body interactions

- QFD deconvolves the complexity of the 3,000+ known isotopes by separating them into three distinct geometric families based on their Harmonic Mode Index

- Standard HPC simulations for estimating the decay and stability of a single heavy isotope can be extremely time-intensive. Lattice QCD solves the nuclear many-body problem by summing 10^23 potential quark states



Nuclides

Quantum Field Dynamics (QFD) framework 2026

Tracy McSheery, Phasespace

- Rather than 19-23 parameters as in Lattice QCD, QFD treats the nucleus as a unitary Q-ball soliton—a single, continuous object. Because the "strong force" is reinterpreted as the hydrostatic pressure of the stiff vacuum ($\beta \approx$ 3.043) pushing inward, all nuclide properties can be calculated with just two further coefficients.

- All of these constants including β are derived from the fine-structure constant α (approximately 1/137.036) with no other needed constants (Arnold Sommerfeld described this as the strength of the electromagnetic interaction between elementary charged particles)

- QFD achieves an RMS accuracy of 0.11% across the periodic table without any adjustable constants. For the specific proton-to-electron mass ratio, it reaches a precision of 0.0023% (5 significant figures)

- QFD's Core Compression Law fits the stability line of over 5,800 isotopes with an R^2 = 0.979%, rising to R^2 = 0.998% for stable isotopes specifically
  (this shows a simple geometric resonance model can produce results in contrast to the challenges with complex, probabilistic N-body interactions)

Nuclides

Solving for heavy atoms such as lead 208 can consume days of HPC time

This geometric approach matches empirical measurements better than Lattice QCD but runs in real time

Nuclear decay is the process by which a soliton (stable wave) under "charge stress" relaxes toward its most efficient geometric configuration—the Zero-Stress Backbone

Every isotope has a measurable degree of stress relative to the Core Compression Law (Q = $c_1A^{2/3} + c_2A$). This stress defines its distance from equilibrium

Decay pathways are then not random(alpha, beta, fission). They are discrete geometric vectors that provide the shortest path back to the stable backbone

Resonance and Half-life: A nuclide's half-life is deterministically governed by its "Distance to the Nearest Decay Step." If an isotope’s stress is a near-perfect integer match for a decay vector, it is highly "resonant" and decays rapidly. Long-lived isotopes are those poorly aligned for a clean transition.

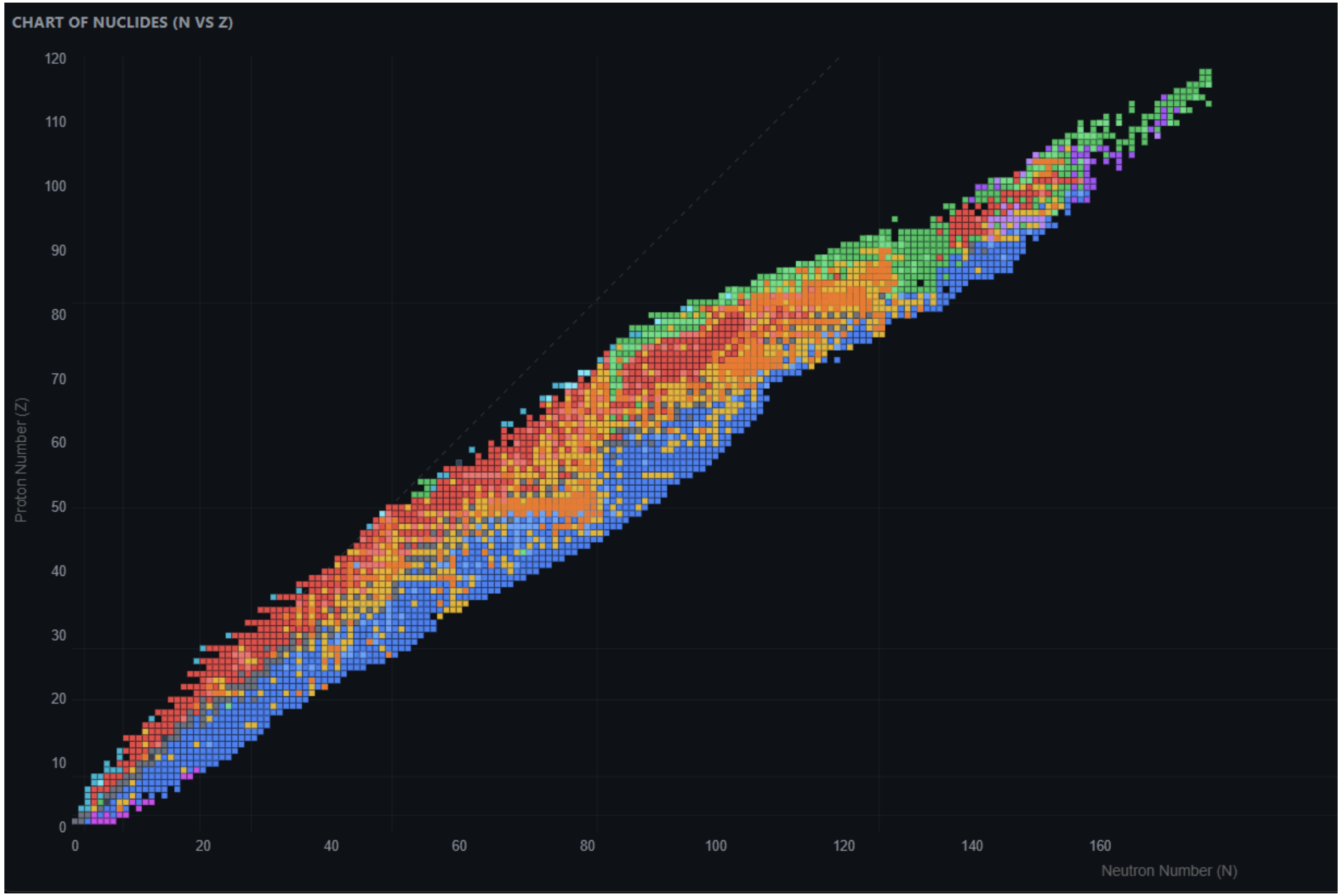


CHART OF NUCLIDES (N VS Z)
Proton Number (Z)
Neutron Number (N)