Diderot: a Domain-Specific Language for Portable Parallel Scientific Visualization and Image Analysis

Gordon Kindlmann  Charisee Chiw  Nicholas Seltzer  Lamont Samuels  John Reppy
Department of Computer Science, University of Chicago

Abstract—Many algorithms for scientific visualization and image analysis are rooted in the world of continuous scalar, vector, and tensor fields, but are programmed in low-level languages and libraries that obscure their mathematical foundations. Diderot is a parallel domain-specific language that is designed to bridge this semantic gap by providing the programmer with a high-level, mathematical programming notation that allows direct expression of mathematical concepts in code. Furthermore, Diderot provides parallel performance that takes advantage of modern multicores and GPUs. The high-level notation allows a concise and natural expression of the algorithms and the parallelism allows efficient execution on real-world datasets.

Index Terms—Domain specific language, portable parallel programming, scientific visualization, tensor fields.

1 INTRODUCTION

Scientific visualization and image analysis research seeks better ways to extract knowledge from scientific data. However, writing software is hard: it can be difficult to translate ideas and algorithms to working code, often due to a semantic gap between the mathematical concepts at the computational core of algorithms, and their actual expression in source code. This often happens with algorithms defined in terms of abstractions like fields or tensors, but implemented with library function calls or interfaces that do not reflect the basic mathematical structures. Another challenge is that visualization and analysis tools typically require parallel computing, either because new analysis methods may require more computation per iteration, or because meaningful evaluation of new methods requires their application to large real-world datasets. Datasets grow in size and complexity with continuing advances in scientific imaging modalities.

Domain-specific languages (DSLs) can address both of these challenges. By addressing a more narrowly-defined class of data types and operations than general-purpose languages, DSLs bridge the semantic gap by supporting mathematically idiomatic expression of algorithms. By targeting a specific class of algorithms, DSLs also facilitate compiling algorithms to efficient parallel execution, as well as achieving portable parallelism: mapping the same program to different parallel computing back-ends. Scientific visualization is an especially apt target for DSLs because many of its elementary abstractions are not directly supported by general-purpose languages, so the implementation of even basic methods obscures their underlying algorithmic simplicity. While many languages support vector- or matrix-valued variables, for example, we are not aware of languages that directly and idiomatically support the abstraction of a continuous scalar, vector, or tensor field, along with arithmetic and differential operations on fields. Fields, however, are fundamental to the basic definition of many core scientific visualization algorithms (e.g., ray-cast volume rendering, streamline integration, and fiber tractography).

Diderot is a portable parallel domain-specific language developed out of frustration with the lack of mathematical abstraction and the difficulty of parallel programming associated with research in scientific visualization and image analysis. Previous work [16] described an early version of Diderot, introducing its program structure, intermediate representations in compilation, computational abstraction of fields, and efficient parallel execution via pthreads. This paper presents a range of advances:

• Portable parallelism: With no more user intervention than setting a compiler command-line flag, Diderot can generate sequential code using SSE vector instructions, pthreads-based parallel code for shared-memory multicore systems, or OpenCL code for GPUs. Portable parallelism is increasingly useful in the context of rapidly evolving architectures for parallel computing.

• Strand creation and communication: The threads of execution (called strands) in Diderot can spawn new strands, strands can communicate with each other, and global reductions can be computed over the entire program. These features are important for algorithms such as particle systems that require creating dynamically new particles, and computing the interactions between particles.

• Higher-order mathematical functions: Fields in Diderot (scalar, vector, or tensor) can be operated on and combined as they would be in mathematics, creating new fields with expressions like “c = V × F” or “c = V(∇×F) • V.” While the previous compiler [16] had some higher-order operators, here we describe a general system for lifting tensor operations to fields and combinations of fields and tensors. Supporting idiomatic expressions of these operations simplifies implementing algorithms that depend on them.

• Compiling to a library: Rather than compiling to a self-contained executable, Diderot programs may be compiled to a C library with an API for setting global variables, defining the initial state of the computation, controlling program iterations, and accessing the computed output. This mechanism facilitates integrating Diderot into existing environments written in C or C++, as well as high-level languages that support C extensions.

• Other language conveniences: The Diderot programs shown here benefit from other new language features such as user-defined functions, dynamic sequences (e.g., for generating streamline geometry), and operations like clamp, mirror, and wrap on the convolution domains of fields.

2 BACKGROUND AND RELATED WORK

There are a number of approaches to supporting scientific visualization and biomedical-image processing applications. A common way is to build upon a domain-specific library or toolkit such as the Visualization Toolkit (VTK) [42], ParaView [2], and the Insight Toolkit (ITK) [53]. Related to libraries and toolkits are software tools that target a specific visualization method in a highly configurable way, such as the wealth of GPU-based volume rendering tools [11, 38, 34, 41].
An alternative to libraries and toolkits are domain-specific languages (DSLs), which trade generality and flexibility for expressiveness and ease of use [32]. There is a long and rich history of DSLs [50], but here we limit our comparison to DSLs that have some commonality with our rationale and application domain.

DSLs can either be implemented as libraries in a host language (embedded DSLs) or as standalone programming languages. For example, Delite is a framework for implementing embedded parallel DSLs on heterogeneous processors [15, 10]. Delite uses Scala as its host language and provides sophisticated compilation support for mapping DSL features to efficient code. The distinction between libraries and embedded DSLs is fuzzy; systems like Protovis [8] and DSL features to efficient code. The visualization and analysis methods in the target domain of Diderot are defined in terms of continuous scalar, vector, or tensor fields. Accordingly, Diderot programs typically start by defining a field as the convolution of discretely sampled data with a continuous reconstruction kernel. For example,

```plaintext
define field as field = ctmr @ image("hand.nrrd");
```

defines a C^1 continuous ("#1") scalar ("#1") field F over a two-dimensional ("2") world-space, by convolving ("@") the image data in hand.nrrd with the Catmull-Rom cubic spline ctmr. Fig. 1(c) shows a dense sampling of this field.

Diderot programs are decomposed in terms of strands, either autonomous or interacting, which can move throughout fields. Fig. 1(d) shows a dense sampling of an isocontour of F found via Newton-Raphson iteration. Isosurface sampling in Fig. 1(d) is computed by:

```plaintext
1  strand isofind (vec2 pos0) {  
2    output vec2 x = pos0;
3    int steps = 0;
4    update {  
5      // Stop after too many steps or leaving field  
6      if (steps > stepsMax || !inside(x, F)) {  
7        die;  
8        // one Newton-Raphson iteration
9        vec2 delta = -normalize(VF(x)) * F(x) / VF(x);  
10       x += delta;  
11       if ((|delta| < epsilon) stabilize;  
12       steps += 1;  
13     }  
14   }  
15  }
```

The heart of a Diderot program is the “update” function, which expresses one iteration of the algorithm. Strands can die to terminate without saving any output, or stabilize to save the most recently
computed value of the variables tagged with output. Here, output
position x (line 2) is saved when the Newton-Raphson iteration (line 9)
converges to within epsilon (line 12).

For a more complete example of a Diderot program, Fig. 2 lists a
particle-based isosurface sampler (Fig. 1(e)), a simple 2-D version of
the method of Meyer et al. [33], without curvature-dependent density.
At a high level, the program starts with (lines 1-6) the declaration of
local strands, to be set either on the command-line when compiling
to a stand-alone executable, or via API when compiling to a library.
After creating (line 8) the scalar field in which the isosurface \( F(x) = 0 \)
will be sampled by interacting particles (according to pairwise energy
\( \phi(x) \) and force \( \phi'(x) \)), the strand definition includes local strand state
(lines 15-19) and the update method, which proceeds in two stages.

```plaintext
1 input real radius; // particle interaction radius
2 input real epsilon; // convergence criterion
3 input int res; // initial grid sampling resolution
4 input real isoval; // isovalue
5 input real xmin; input real xmax; // sampling extent
6 input real ymin; input real ymax;
7 real limit = radius/2; // limit on motion
8 field(2)[1][0] F = ctmr @ image("hand.nrrd") - isoval;
9 // inter-particle energy, and its derivative
10 function real phi(real x) =
11 -1/[4*radius] if (x < radius) else 0;
12 function real phi'(real x) =
13 -4/radius*[1/[4*radius] if (x < radius) else 0; 0]
14 strand point(\tilde{vec} pos) {  // particle position
15 output \tilde{vec} pos = pos; // force particle position
16 vec2 delta = [0,0]; // change in position
17 bool foundIso = false; // initial isosurface search done
18 real hh = 1; // line search step size
19 int iter = 0; // iteration count
20 update {  // inter-strand energy computation
21 iter += 1;
22 if (!foundIso)  
23 if (iter > 15 || \text{inside}(x, F))
24 die;
25 // Newton-Raphson
26 delta = normalize(\n\frac{\nabla\phi(pos)}{\phi(pos)} \cdot \nabla F(pos) \nabla F(pos)) * F(pos)/|\n\nabla F(pos)|;
27 pos += delta;
28 if (!delta < epsilon*radius)  
29 foundIso = true;
30 } else { // we have foundIso
31 energyNew = 0; \quad vec2 force = [0,0]; \quad int ncount = 0;
32 foreach point P in sphere(radius) { // all particles
33 ncount += 1;
34 vec2 r_ij = pos - P.pos;
35 energy = \phi(|r_ij|);
36 force = normalize(r_ij)*\phi'(|r_ij|);
37 vec2 norm = normalize(F(pos));
38 if (ncount <= 1 \&\& iter % 10 = 0) {  // particle collision
39 vec2 opos = pos + 0.2*radius*[norm[1],-norm[0]];  // new point
40 ncount = ncount + 1;
41 } else {  // force particle convergence
42 \text{force} -= \phi(r_ij)*\phi'(r_ij)/r_ij;
43 vec2 posLast = pos;
44 pos += delta;
45 // take Newton-Raphson steps back to surface
46 pos = normalize(F(pos)); \quad F(pos) = \phi(pos);
47 energyNew = 0;
48 foreach point P in sphere(radius) { // record change in this iter
49 energyNew += \phi(pos) - \phi(\text{posLast});  // record change in this iter
50 // force particle convergence
51 if (\text{force} > 0) {  // take gradient descent step
52 \text{delta} -= \text{hh}*force;
53 } if (\text{delta} > limit)  // enforce speed limit
54 \text{hh} = limit/\text{delta};
55 \text{delta} = \text{hh} = force;
56 } \quad \text{vec2 posLast} = \text{pos};
57 \quad \text{pos} += \text{delta};
58 // take Newton-Raphson steps back to surface
59 pos = normalize(F(pos)); \quad F(pos) = \phi(pos);
60 energyNew += \phi(\text{posLast});  // record change in this iter
61 energyNew += \phi(\text{pos}) - \phi(\text{posLast});
62 \text{foundIso} = true;  // record change in this iter
63 if (energyNew > energy - 0.3*\text{force}) {  // backtrack
64 \text{pos} = \text{posLast};
65 } else {  // bigger step next time
66 \text{hh} = 1.1;  // bigger step next time
67 }
68 }
69 }
70 global{  // output strand
71 real motion = mean([\text{delta}]/radius) \quad \text{P in point.all});
72 if (motion < epsilon)  // stabilize
73 stabilize;
74 }
75 Initially (point[\text{min}(xmin, xmax, 0, \text{ui}, \text{res-1})],
76 \text{lerp}(\text{min}(ymin, ymax, 0, \text{vi}, \text{res-1})),
77 \text{vi in 0..(res-1), \text{ui in 0..(res-1)}];
78 Fig. 2. Complete Diderot program for particle-based isocontour sam-
79 pling of scalar field \( F \) at isovalue \( \text{isoval} \), producing Fig. 1(e).
80 First, while not \text{foundIso}, particles move towards the isocontour,
81 keeping only those that stay in the field domain and converge within
82 10 iterations. Second, particles query their radius neighborhood to
determine current energy and force (lines 33-38) due to nearby parti-
cles. If the particle has less than two neighbors, occasionally a new
particle will created nearby to fill in the gap. The force from neigh-
硼ors is projected to the tangent plane (line 45) and the particle position
is updated and re-projected onto the isosurface (lines 55, 56). The
new location is assessed by computing the new energy (lines 59, 60)
as part of a backtracking line search (line 62) [3]: the energy should
have gone down by an amount roughly predicted by the force, or else
we backtrack and try again on the next iteration (line 64).

In addition to the code executed in parallel by strands, Diderot
programs may also have global code that executes between iterations,
to monitor the state of computation. In this example, lines 73-75 find aver-
age motion of all the strands, and terminate when the particle system
has stabilized. The final part of the program is the initialization (line
77) of particles on a coarse grid covering the image domain.

3.2 Design rationale

The design of Diderot is motivated by a desire to speed the develop-
ment and refinement of new image analysis and visualization al-
gorithms, as evaluated on real-world datasets. Diderot addresses this
motivation with two design goals, which are reflected in the two major
facets of its design: ease of programming and parallel performance.

The ease of Diderot programming is fostered by notational and
structural aspects of the language. The mathematical notation we write
on whiteboards should work in programs, so Diderot supports tensor
fields as computational values, and uses Unicode to idiomatically
express mathematical operations. At a structural level, computation
in Diderot is organized into strands according to the structure of the al-
gorithm output. The output may be an image (in the case of volume
rendering), or a set of polylines (for streamlines), or simply a point set
(for particle systems). For these examples, the Diderot program states
the work required of each strand to produce one rendered pixel, one
streamline, or one particle, respectively. This increases the flexibilit
of Diderot relative to languages that structure parallelism around the
input data domain (pixels or voxels).

The second goal is to efficiently execute Diderot programs by using
modern parallel hardware, such as GPUs. Diderot’s execution model
attempts to achieve expressiveness while avoiding features that would
be outdated to map onto parallel hardware, or that would force pro-
grammers to worry about parallel programming challenges like data
distribution, synchronization, and locality. We have therefore bas
the semantics of Diderot on a simple deterministic parallel programming
model, described in Sec. 3.5. This model fits the iterative structure of
typical scientific visualization algorithms, while allowing some flexi
bility in how the model is mapped to different hardware targets.

3.3 Computing with tensors

Diderot strives to mimic the “direct” or coordinate-free notation for
vector and tensor expressions commonly used in continuum mechan-
ics [27]. This motivates using \( u \otimes v \) for the tensor product, instead
of something like “\( u \cdot v \cdot T \)”, which depends on recognizing \( u \) and \( v \)
as N-by-1 column vectors, and which does not generalize to higher-
order tensors. Expressing computations in a coordinate-free way tends
to clarify mathematical intent. The direct notation encouraged by
Diderot also fosters dimensionally general code. For example, to turn
Fig. 2 into a program to sample isosurfaces in three dimensions, the
only change required is replacing vec2 and field#1(2)[1] with vec3 and
field#1(3)[1], respectively, and adding a dimension to the strand ini-
3.4 Reconstructing and computing with fields

Fields are created by convolving (notated with \( \otimes \)) data on regularly
sampled grids with continuous reconstruction kernels. Diderot cur-
rently knows a variety of kernels that differ in their continuity and

1Unicode character entry is a possible concern, but the range of practical solutions is expanding: http://en.wikipedia.org/wiki/Unicode Input.
support (how many samples along each axis are needed in convolution). Besides the cubic Catmull-Rom $C^1$ cubic spline (4-sample support) used in Fig. 1, other possible kernels include $C^0$ linear interpolation $\text{tent}$ (2 samples), $C^2$ cubic B-spline $\text{bspline}$ (4 samples), and a $C^4$ piecewise hexic kernel $\text{ehexic}$ (6 samples) that accurately reconstructs cubic polynomials (a 4th-degree error filter or 4EF in the terminology of [35]).

Reconstruction of fields in Diderot is by separable convolution; the same kernel is convolved with each image data axis. For example, the convolution $F = V \otimes h$ of image data $V[i, j]$ with $s$-sample support kernel $h$ is evaluated at image-space location $(x_1, x_2)$ with

$$F(x_1, x_2) = \sum_{i,j} V[i, j]h(x_1 - i)h(x_2 - j);$$

(1)

$$n_i = [x_i] \Rightarrow q_i = x_i - n_i \in [0, 1] \Rightarrow$$

(2)

$$F(x_1, x_2) = \sum_{i,j=1-i/2}^{s/2} V[n_1 + i, n_2 + j]h(q_1 - i)h(q_2 - j).$$

(3)

The statement (1) of separable convolution typical in image processing and computer graphics [35] shows how the continuity order of field $F$ is the same as that of reconstruction kernel $h$, and that partial derivatives $\partial F / \partial h_i$ are reconstructed by convolving kernel derivatives $h'$ on one axis. Diderot uses (3) to limit the convolution to the kernel support. A field $F$ in Diderot is defined over world-space and evaluated with $F(p)$ at world-space position vector $p$. The field implementation includes the homogeneous coordinate matrix $M$, transforming index-space location $\vec{x}$ to world-space $\vec{p}$, to represent the image orientation learned from the image data file. Field evaluation $F(p)$ first computes $\vec{x} = M^{-1}\vec{p}$ (the inverse $M^{-1}$ is pre-computed once) before applying (3). Further details of the convolution implementation are in [16].

Diderot supports differential and arithmetic operators on fields in a mathematically idiomatic way. Evaluation of the field derivative $\nabla F(\vec{x})$ correctly returns the gradient in world space, according to the covariant transform of the index-space partial derivatives $\partial F / \partial x_i$. The derivative of a field $\phi$ of order-$n$ tensors ($n > 0$) is field $\mathcal{V} \otimes \mathcal{V}$ of order $(n+1)$ tensors. A compile-time error results from differentiating a field beyond its continuity order, like the (second derivative) Hessian $\mathcal{V} \otimes \mathcal{V} \otimes \mathcal{V}$ of a $C^1$ scalar field $\text{field} f(3) \otimes F$.

While thinking mathematically about visualization algorithms on fields, it is natural for programmers to consider expressions like $f = u + v \cdot w$ and the understanding that variables $u$, $v$, and $f$ could stand for constant values, or fields, or some mix of both. Diderot syntax supports this convention. As defined in Sec. A.10, arithmetic operators like addition, subtraction, scalar multiplication $\ast$, and inner product $\bullet$ can work on a mix of tensors and tensor fields. Two fields arising from image data with different orientations and resolutions may be added without restriction. Furthermore, as described in Sec. 5.2, differential operators like gradient $\nabla$, Jacobian $\mathcal{V} \otimes \mathcal{V}$, curl $\nabla \times$, and divergence $\nabla \bullet$ may be applied to such mathematically constructed fields, and the Diderot compiler determines how to compute their evaluation.

3.5 Bulk-Synchronous Parallel Strand Execution

The bulk-synchronous parallel (BSP) programming model of Valiant [49, 44] provides the basis of Diderot’s execution semantics, illustrated in Fig. 3. All active strands execute in parallel execution steps called super-steps, each with two phases: a strand update phase and an optional global computation phase. First, each active strand executes its update method once, to update its local state, including output variables. Strands do not directly communicate, but they may learn the state of nearby strands according to their position in world space, as used in the particle system code in Fig. 2. A strand makes a spatial query to get an iterable sequence of all strands within some world-space radius, providing (read-only) access to their state at the beginning of the super-step. This requires maintaining two copies of strand state, but we can optimize away the extra storage in many cases. Diderot is not intended for applications in which all pairs of strands need to directly communicate, and of the possible mechanisms for limiting strand communication we choose spatial proximity. The assumption is that each strand is responsible for analyzing or visualizing some local region of the data (as with particle systems), so any communication required between strands can likely be expressed in terms of region (or strand) neighborhoods.

Second, the global phase computes reductions over strands (such as finding average or extremal values of strand variables) to update global variables, or to terminate program execution by stabilizing all strands. At the start of the next super-step, strands created by new start executing, and all strands can read (but not write) values of globals set in the preceding global phase. Each strand is idle after it finishes executing its update method until the end of the strand update phase. Stable strands are idle for the entirety of their update. Dead strands are also idle during update, but do not produce any output. The program executes until all strands are either stable or dead. The BSP model is deterministic because the order of strand scheduling does not affect the computation. In practice, non-determinism may arise from the non-associativity of floating-point arithmetic in global reductions, but for debugging purposes this can be avoided at some performance cost.

3.6 Programs as libraries

One drawback of DSLs is their lack of infrastructure for I/O, graphics, networking, or other services required of complete applications. Such functionality requires significant implementation effort and may dilute the domain-specific focus of the language. One solution is to embed the DSL in a general-purpose host language, so the embedded DSL can use all the features of the host language. Delite takes this approach, for example [15, 10]. We did not embed Diderot because we did not want to commit to a single host language, and be constrained by its syntax and type system. Supporting idiomatic expression of mathematical operators required Unicode, and field properties like continuity order and tensor shape are not readily encoded in other languages.

To address the infrastructure problem, we compile Diderot programs to libraries with a C-language API. This allows programmers to develop applications in any language that can call C code. The API includes functions for setting inputs, reading outputs, and controlling program execution. A Diderot program can either be run to completion or for some number of super-steps. Snapshots of program state can be accessed to create animations of program execution. The compiler can additionally generate stand-alone executables (with command-line options to set inputs), which link against the C library created by the compiler. Appendix B demonstrates the API generated from a program for isosurface sampling (Fig. 1d)).

4. DEMONSTRATIONS AND RESULTS

4.1 Volume rendering

Volume rendering of scalar datasets is a well-studied research area, with many efficient and configurable tools [11, 38, 34, 41]. We show volume renderings here that go well beyond the capabilities of existing...
tools, by exploiting Diderot’s ability to operate on continuous fields as first-class mathematical objects. The quantities volume rendered in these examples (scalar field Canny edges, flow field vortex structure indicators, and tensor invariants) are all defined with a few lines of Diderot code, and the subsequent volume rendering code remains the same. Even though the code works from an underlying scalar-, vector-, or tensor-valued volume dataset, the rendering is always of a scalar field analytically derived from the data (not pre-computed on a grid) and numerically evaluated at each sample along each ray. The Diderot language permits a clean separation between the code that implements the core visualization algorithm, and the code that defines the quantity of interest to visualize.

```diderot
input real isoval; input real thick;
input vec3 camEye; input vec3 camAt; input vec3 camUp;
input real camFOV, input int irealV; input int irealU;
input real refStep; input real rayStep;
input vec3 LightVec; // light direction in view-space
input real phongKa, input real phongKd;

field[4][] V = bspln5 @ image("ctscan-prefilt.nrrd");
field[4][] F = V - isoval; // isovalues is (x|F(x)|>0)
function real mask(vec3 x) = 1.0;
field[4][] cmap = vec3 @ clamp(image("bow.nrrd");
function vec3 color(vec3 x) = cmap(V(x));
function real alpha(real v, real g) = 1 - pow(1-a, rayStep*|rayVec|/refStep);
if (transp < 0.01) { rgba = [rgb[0]/a, rgb[1]/a, rgb[2]/a, a];
    if (a > 0) rgba = [rgb[0]/a, rgb[1]/a, rgb[2]/a, a];
}
rgb += transp*a*depth*(phongKa + phongKd*shade)*color(x);

rayU = lerp(-camUmax, camUmax, -0.5, ui, iresU-0.5);
rayV = lerp(camVmax, -camVmax, -0.5, vi, iresV-0.5);
rayVec = camN + (rayU*camU + rayV*camV)/camDist;
rayN = camNear;
light = transpose([camU,camV,camN])
camUmax = camVmax*iresU/iresV;
camU = camV = camU;
camDist = (camAt-camEye);
vec3 camNorm = normalize(camAt-camEye); // away
vec3 camN = normalize(camN*camAt); // right
vec3 camN = camN + camNorm;
vec3 camFar = tan(camFOV*rad/360)*camDist;
vec3 camNear = camFar - norm(camN)*rayStep*

if (inside(ui, vi)) {
    real val = F(x);
    vec3 grad = -∇F|x| = X
    real = alpha(val, |grad|)*mask(x);
    if (a > 0) {
        a = 1 - pow(1-a, rayStep*|rayVec|/refStep);
        real shade = max(0, normalize(|grad|)*light);
        rgb += transp*|depth|*(phongKa + phongKd*shade)*color(x);
        transp = 1 - a;
    }
}
if (transp < 0.01) { // early ray termination
    transp = 0; 
    // undo pre-multiplied alpha
    stabilize();
    if (rgb > camFar) stabilize();
    rayN = rayU + rayStep;
}

initially(raycast(ui, vi) | vi in 0..iresV-1, ui in 0..iresU-1);
```

Fig. 4. Complete Diderot program for volume rendering an approximate isocontour at isovalue isoval with thickness thick in scalar field V.

Fig. 4 shows the complete listing of a Diderot program for ray-cast volume rendering. While the code is mostly self-explanatory, lines 8 to 14 merit further explanation, since they will be changed in subsequent examples. The CT scan volume dataset ctscan.nrrd used on line 8, and all subsequent volume datasets used in these volume rendering examples, are pre-filtered so that convolution with the 5-order B-spline bspln5 (piece-wise quintic, C^4 continuous, 6-sample support) interpolates in [48]. The scalar field in which to render the zero isocontour is (line 9) F = V - isoval. The per-sample color and opacity in the ray-casting are defined in terms of the cmap colormap (1-D field of 3-vectors, line 11) and the alpha bivariate opacity function (13) based on Levoy [30].

Fig. 5 shows a CT scan of a Cebus apella (capuchin) head, for which an image analysis goal involves recovering the bone surface. Diderot can visualize the bone surface by volume rendering. The left side of Fig. 5 shows isocontours rendered with the code in Fig. 4, at three isovalues that span the range of CT values for bone. At too low an isovalue (V=1300) some of the soft tissue is visible, but at a higher isovalue (V=2150) suitable for most of the skull, holes incorrectly appear at the frontal sinuses (above and between the eye sockets). The tooth surface is cleanest at yet another isovalue (V=3000). The colormap is approximately isoluminant.

A classical principle of edge detection proposed by Canny is that edge points are where the image gradient magnitude is maximized with respect to motion toward the (normalized) gradient direction [14]. Algorithmically, Canny edge detection also involves finding optimal smoothing, and hysteresis thresholding of edge components based on gradient magnitude. These are outside the scope of Diderot, but we can still capture Canny’s principle of edge localization. In scalar volume data field V(x), we seek locations where |∇V(x)| is maximized with respect to motion along V(x)/|∇V(x)|, which means that |∇V(x)| (the gradient of quantity being maximized) is orthogonal to V(x)/|∇V(x)|. In Diderot, we write:

```
field[2][] F = -V(|∇V|) • V/|∇V|
function real mask(vec3 x) = 1.0 if (|∇V(x)| > gmin) else 0.0;
```

in place of lines 9 and 10 in Fig. 4. Strong edges (as opposed to minima of gradient magnitude) are selected by user-specified threshold gmin. The right of Fig. 5 shows the results, with no other changes to the code of Fig. 4. Note that the Canny edges are shaded correctly, based Vr. The isocontour shading also used Vr (line 41), but now Vr is some expression involving third derivatives of the volume data V, which the Diderot compiler generates automatically, based on its understanding of vector calculus. The smoothness apparent in the bone surface is thanks to the C^4 continuous reconstruction of the data field[4][] V (line 8). The Canny function field[2][] F is C^2 continuous (based the second derivative of V), and its gradient Vf will be C^1 continuous. With no prior segmentation or parameter tuning, Fig. 5 correctly shows the skull surface over the fronto sial s, as well as the very thin bones of the orbital walls (within the eye socket), neither of which are visible with any isoalue reliably higher than all the soft tissue. Diderot has thus greatly simplified the successful visual exploration (by volume rendering) of a first-principles approach (Canny edges) to bone surface extraction in real-world three-dimensional image data.

Flow visualization is another domain of scientific visualization that involves mathematically sophisticated consideration of vector fields and their derivatives. For example, one statement of vortex structure
identifies them with locations where the direction of flow $\frac{\mathbf{v}}{||\mathbf{v}||}$ is aligned with that of its curl $\nabla \times \mathbf{v}$ [20, 37]. This is equivalent to saying that the normalized helicity $\frac{\mathbf{v} \cdot (\nabla \times \mathbf{v})}{||\mathbf{v}||^2}$ is at an extremum (near +1 or -1). We show here how volume rendering in Diderot can directly visualize the mathematical elements of this type of flow field analysis.

The renderings in Fig. 6 use a single time-step from a Navier-Stokes simulation of flow (from left to right) past a square rod, creating a train of vortices [13, 51]. The Fig. 4 volume rendering code is re-used, but with $\mathbf{v}$ defining a 3-D vector field:

$$\begin{align*}
\text{field}[4](3)[3] \mathbf{v} &= \text{bspln5 @ image("flow.nrrd")};
\end{align*}$$

and with a diverging colormap of the Jacobian discriminant (in lieu of lines 11 and 12):

$$\begin{align*}
\text{field}[4](3)[3] \mathbf{J} &= \mathbf{v} \times \mathbf{v};
\text{field}[4](3)[3] \mathbf{F} &= \nabla \mathbf{v};
\text{field}[4](3)[3] \mathbf{A} &= -\text{trace}(\mathbf{J});
\text{field}[4](3)[3] \mathbf{B} &= (\text{trace}(\mathbf{J}) \times \text{trace}(\mathbf{J}) - \text{trace}(\mathbf{J} \times \mathbf{J}))/2;
\text{field}[4](3)[3] \mathbf{C} &= -\text{det}(\mathbf{J});
\text{field}[4](3)[3] \mathbf{Q} &= (\mathbf{A} + \mathbf{B} + \mathbf{C})/3.0;
\text{field}[4](3)[3] \mathbf{R} &= (-2.0 \times \mathbf{A} + \mathbf{A} + \mathbf{A} + 27.0 + \mathbf{A} + \mathbf{B} + 3.0 - \mathbf{C})/2.0;
\text{field}[4](3)[3] \mathbf{D} &= \mathbf{R} \times \mathbf{Q} - \mathbf{Q} \times \mathbf{Q}; \quad // \text{the discriminant}
\text{field}[4](3)[3] \text{dmap} &= \text{clamp}(\text{tent@image("diverg.nrrd")});
\text{function vec3 color(vec3 x) = dmap(D(x));}
\end{align*}$$

After forming a field of the Jacobian $\nabla \times \mathbf{v}$, fields $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are the coefficients of the characteristic polynomial $p(\lambda) = \lambda^3 + \mathbf{A} \lambda^2 + \mathbf{B} \lambda + \mathbf{C}$, and the discriminant field $\mathbf{D}$ is the pre-cursor to analytically solving $p(\lambda) = 0$ for $\lambda$. There are two complex-conjugate eigenvalues (indicating rotational flow) when the discriminant $D(x)$ is negative [37].

The only substantial difference in how the various renderings in Fig. 6 were computed was in the statement of the derived scalar field $F$ being rendered. Fig. 6(a) shows simple isosurfaces of flow magnitude, created by using

$$\begin{align*}
\text{field}[4](3)[3] \mathbf{F} &= ||\mathbf{v}||;
\end{align*}$$

Fig. 6(b) visualizes isosurfaces (at $\pm 0.99$) of a quantity related to the extremal lines of the field: places where the vector magnitude is extremal with respect to motion perpendicular to the vector, or, where $\nabla ||\mathbf{v}||$ is aligned with $\mathbf{v}$ [46, 37].

$$\begin{align*}
\text{field}[4](3)[3] \mathbf{F} &= (\nabla ||\mathbf{v}||) \cdot (\nabla ||\mathbf{v}||/||\nabla ||\mathbf{v}||||);
\end{align*}$$

Fig. 6(c) visualizes isosurfaces (at $\pm 0.99$) of normalized helicity, using

$$\begin{align*}
\text{field}[4](3)[3] \mathbf{F} &= (\nabla ||\mathbf{v}||) \cdot (\nabla \times \mathbf{v}/||\nabla \times \mathbf{v}||);
\end{align*}$$

Note that in all cases, the Diderot compiler is determining how to analytically differentiate field $F$ so that it can be shaded correctly. These renderings use two-sided lighting:

$$\begin{align*}
\text{real shade} &= \text{normalize}(\text{grad}) \cdot \text{light};
\end{align*}$$

It should be emphasized that although a scalar field is being volume rendered to make these images, there is no pre-computed scalar dataset: the scalar field is defined symbolically in terms of a vector field (reconstructed by convolution), and the Diderot compiler generates the necessary instructions to probe the scalar field at each ray sample position during rendering.

Two examples of volume rendering tensor fields further demonstrate the simplicity of creating mathematically and computationally sophisticated visualization tools with few lines of Diderot code. Research in tensor field topology often investigates lines of degeneracy: locations in a tensor field where two of the tensor eigenvalues are equal, or equivalently, places where the tensor mode is $\pm 1$ [19, 47].

$$\begin{align*}
&\text{(a) Lines of degeneracy in a stress tensor field revealed by volume rendering isosurfaces of tensor mode +1 (yellow) and -1 (blue).}
\end{align*}$$

$$\begin{align*}
&\text{(b) Volume rendering a diffusion tensor field of the human brain at FA=0.3.}
\end{align*}$$
The RGB coloring is provided by the traditional map of the principal implementation of the original LIC method [12], in which samples the background image of Fig. 10. The program is a straightforward solution (LIC) visualization of a two-dimensional vector field, seen as

4.2 LIC and Streamlines

invariants, expanded in terms of the spatial derivatives of the individual and numerically computing the spatial derivatives of the tensor velocity at the start point

steps or until the streamline leaves the domain. Field evalu-
istions falling outside the domain are gracefully handled by the border field. Two strands are used for each output pixel; one upstream (line 32 stepNum)

and restart it with a larger preallocation. For example, if the host side preallocated memory for 30 strands, used to computation, but an execution step might exceed the allocation. For GPU memory be preallocated on the host side before running the GPU needed to support strand motion and creation. OpenCL requires that presents additional challenges for the dynamic memory management value for the world-space coordinate at each level of the tree. The GPU version of the median-of-medians algorithm [7] to select the splitting

gets. To account for strand motion, and the creation of new strands, the tree must be maintained at the end of each super-step, prior to ex-

proportional to

along the stream-

are assigned to nodes in the tree by comparing their coordinate on the (three dimensions) through the axes as one descends the tree. Strands

field stores the start indices for each streamline. The polylines were rendered (not shown) to a filled triangle. In the program output, all polyline vertices are concatenated into a single array, and a second array stores the start indices for each streamline. The polylines were post-processed to produce the vector graphics seen in Fig. 10.

5 IMPLEMENTATION

5.1 Strand communication and global reduction

Strands may communicate with neighboring strands according to their proximity in world space. We accelerate the spatial queries with a k-d tree [6, 24], a binary tree that at each internal node splits world space along one axis. The split axis alternates (for two dimensions) or cycles (three dimensions) through the axes as one descends the tree. Strands are assigned to nodes in the tree by comparing their coordinate on the splitting axis to the splitting value.

Spatial queries pose an implementation challenge for parallel targets. To account for strand motion, and the creation of new strands, the tree must be maintained at the end of each super-step, prior to executing the next strand updates. Maintaining the spatial data structure must be parallelized to avoid sequential bottlenecks. We use a parallel version of the median-of-medians algorithm [7] to select the splitting value for the world-space coordinate at each level of the tree. The GPU presents additional challenges for the dynamic memory management needed to support strand motion and creation. OpenCL requires that GPU memory be preallocated on the host side before running the GPU computation, but an execution step might exceed the allocation. For example, if the host side preallocated memory for 30 strands, used to store the states of 20 active or stable strands, then the current execution step cannot successfully create more than 10 new strands. Fortunately, Diderot’s execution model preserves the starting state of strands in a given execution step, so we can abort the step if we run out of memory and restart it with a larger preallocation.
This computed the streamlines drawn over the LIC result in Fig. 10.

Fig. 9. Diderot code for streamlines (with arrowheads) in a vector field. Initially stabilize \( p = p@(x-b); p = p@(x+a); p = p@(x+b); p = p@x; \)

\begin{verbatim}
vec2{} x0s = load("seeds.txt"); // list of seedpoints
real h = 0.02;
int stepNum = 200;
field3d(2)(2) V = bspln3 @ image("flow.nrrd");
real arrow = 0.1; // scale from |V(x)| to arrow size
strand sline(vec2 x0) {
  int step = 0;
  vec2 x = x0;
  output vec2[] p = (x0); // start streamline at seed
  update {
    if (inside(x, V)) { // finish streamline with triangular arrow head
      vec2 a = arrow*V(x); // length of arrow head
      vec2 b = 0.4*(-a[1],a[0]); // perpendicular to a
      p = p@x+a; p = p@x+b; p = p@x;b; p = p@x;
      stabilize;
    }
    // append new point to streamline
    x += h*V(x + 0.5*h*V(x));
    if (step == stepNum) { // scale from |V(x)| to arrow size
      x += h*V(x);
    }
    step += 1;
  }
  \}
\end{verbatim}

Fig. 10. A comparison of parallel speedup on the benchmark programs from [16] on a 16-core Intel Xeon E5-2687W.

where \( t \) is the variable being assigned, \( \lambda params(e)_{\alpha} \) is an EIN operator with formal parameters \( params \) and body \( (e)_{\alpha} \), and \( args \) are the arguments to the EIN operator. Once a surface language operation is mapped to an EIN operator, the compiler can handle the computations generically, by systematically applying EIN operators to one another, normalizing, and optimizing. The compiler can then break the EIN operator apart into simple and direct scalar and vector operations to generate code.

A family of operations (like tensor plus tensor, field plus tensor, field plus field) that previously [16] required several type-specific operations and case analysis can be captured with a single EIN operator.

All tensor operators are lifted to tensor fields as consequence of their expression in EIN. The previous compiler generated code for a fixed set of tensor types, and a fixed set of operations between them, which impeded adding new operators to Diderot. EIN represents tensors and tensor fields of arbitrary shape. The new compiler simplifies complicated and large EIN operators into operations on scalars and vectors. It needs only to generate code for scalar and vector operations, avoiding the explosion of shape-specific operators in the compiler.

The compiler uses a rewrite system to optimize and lower the EIN representation. The Diderot programmer can then define fields with expressions like \( \nabla(V \bullet G) \) without worrying about how to expand and simplify this in terms of derivatives of \( F \) and \( G \), and their individual scalar components. Rewrite rules include derivative identities such as \( \nabla(V + G) \Rightarrow (\nabla V + \nabla G) \), the product rule \( \nabla(f \bullet g) \Rightarrow g \nabla f + f \nabla g \), the quotient rule, and the chain rule. EIN also includes the permutation tensor \( \delta_{ijk} \) and the Kronecker delta \( \delta_{ij} \). The compiler understands that \( \delta_{ijk} \delta_{jk} \Rightarrow 0 \), where \( \delta_{jk} \) indicates partial derivatives along axes \( j \) and \( k \), which produces well-known identities like \( \nabla \times \nabla \varphi \Rightarrow 0 \) and \( \nabla \bullet (\nabla \times F) \Rightarrow 0 \). The compiler also understands that \( \delta_{ijk} \delta_{lm} \Rightarrow \delta_{jlm} - \delta_{jml} + \delta_{jml} \) and \( \delta_{ij}T_j \Rightarrow T_i \), which produces identities such as \( (a \times b) \times c \Rightarrow b(a \bullet c) - a(b \bullet c) \). Being able to automatically find identities as part of optimization provides the benefits of domain-specific optimizations without having to encode a lot of domain-specific knowledge in the compiler.

6 PERFORMANCE EVALUATION

While space does not permit an in-depth performance analysis, we revisit previously published benchmarks [16] to demonstrate significant performance improvements provided by the new compiler. These
results address our second design goal (Sec. 3.2) of combining good parallel performance with our high-level programming model.

Our test machine is a dual Intel Xeon E5-2687W system (16 cores) running Ubuntu 12.04 Linux. The OpenCL measurements were taken on the same system using a NVIDIA Tesla K20c with NVIDIA’s CUDA 6.0 driver. All code was compiled with optimization level -O3. For each benchmark, we report the average wall-clock time for the computational part from 10 runs on a lightly-loaded machine.

The benchmarks presented here are from our earlier work, and were originally chosen to represent typical workloads:

- **textbf:** a simple volume renderer with Phong shading, like Fig. 4.
- **illus-vr:** A more complex volume renderer with curvature-based illustrative rendering [29], including tensor calculations that are awkward to express in other languages.
- **lic2d:** Line integral convolution visualization of a synthetic 2D vector field [12], like in Fig. 8 but simpler.
- **ridge3d:** An initial uniform distribution of points within a portion of CT scan of a lung is moved iteratively towards the centers of blood vessels, using Newton optimization to compute ridge lines [22], requiring eigenvalues and eigenvectors of the Hessian.

The most striking improvement of our current system over our previously reported work is in the performance scaling on SMP hardware. Fig. 11 shows the speedup curves for our benchmarks using both the current version of Diderot (solid lines) and the previous version (dashed lines). We now get excellent speedup for all four benchmarks at 16 cores, whereas the previous version scales poorly and does significantly worse on three benchmarks. Furthermore, the baseline performance of the previous version is worse, often significantly, so we are seeing better scaling even when compared to a better baseline.

Table 1 presents selected numeric results from our experiments. For each benchmark, we report several sets of numbers: the execution time for the Teem version (hand-coded C), sequential and parallel (1, 6, 12, and 16-core) execution times for the PLDF’12 version of the system, sequential and parallel (1, 6, 12, and 16-core) execution times for the current version of the system, and execution times for the OpenCL version of the program. The sequential performance of our compiler has improved — 15% faster for three of the four benchmarks.

The performance of the OpenCL target is benchmark dependent. For a program like illust-vr, which has significant arithmetic intensity, the GPU is comparable to 9-core performance. On the other hand, it performs much worse on lic2d, which is more memory-bound. Appendix C compares Diderot with hand-written OpenCL code for two kinds of volume rendering. These limited results suggest that non-trivial programs are more easily expressed in Diderot than in OpenCL, with a performance penalty that is acceptable when the coding skill and time of researchers is the actual bottleneck. Nonetheless, there is room for significant improvement in our OpenCL implementation.

### 7 Discussion and Ongoing Work

This paper presents a significant step towards creating a powerful and portable parallel DSL for scientific visualization and image analysis. In particular, supporting fields as a fundamental abstraction in Diderot frees programmers from worrying about the computational details of convolving on the underlying sampling grid and the mathematical details of differentiating arithmetically constructed fields.

The utility of Diderot would be enhanced by further work on its mathematical basis. We hope to expand the kinds of fields it can handle. Sec. 4.1, for example, used the same rendering code to visualize a variety of fields (all based on convolution on regular grids). The same rendering code could in principle be re-used to visualize fields arising from higher-order basis functions on unstructured meshes, point clouds with radial basis function, or analytic closed-form expressions. More ambitious would be a way of defining the updated values method not with imperative program statements, but with a declarative statement of mathematical intent. The volume rendering code of Fig. 4 would be simplified by stating its goal as solving a volume rendering integral, if the sampling and numerical integration schemes could be automatically generated by the compiler.

Diderot will also benefit from work in less mathematical areas. Most important are improvements to our implementation on GPUs. These include developing a better scheduler for strands on the GPU; our current scheduler does well with irregular workloads, but its base-line performance is not very good. We are also looking to implement a virtual memory strategy akin to that already developed for GPU-based rendering [25]; such a scheme will allow us to scale to data sets that are larger than the available memory on GPUs. Further work on code generation may lead to other GPU performance improvements.

Our compiler currently generates OpenCL code, but generating CUDA code should allow better performance on NVIDIA hardware. Targeting OpenCL or CUDA, instead of assembly code or low-level GPU code (e.g., SPIR or PTX), greatly simplifies our Diderot compiler implementation, while benefiting from the optimizing compilers specialized for the target hardware. Finally, automatically generating GUIs for the input and output variables of a Diderot program would accelerate the exploration of the parameter space of different algorithms.

The source code for the Diderot compiler is available and a release is upcoming, which will merge features currently split across different branches. Scripts for regenerating representative images from freely available datasets will be made available from our web page.

### Acknowledgments

We gratefully acknowledge the anonymous reviewers for their constructive comments. We also thank the providers of data seen in the figures. Fig. 1: University of Utah SCI group, NIH NIGMS grant P41GM103545. Fig. 5: Callum Ross, University of Chicago. Fig. 6: Resampling by Tino Weinkauf of Navier-Stokes simulation by S. Camarri, M.-V. Salvetti, M. Buffoni, and A. Iollo [28]. Fig. 7(a): Xavier Tricoche, Purdue University. Fig. 7(b): Centre for Functional MRI of the Brain, John Radcliffe Hospital, Oxford University. Fig. 10: Wolfgang Kollmann, UC Davis. Portions of this research were supported by National Science Foundation award CCF-1446412. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of these organizations or the U.S. Government.

### References


http://diderot-language.cs.uchicago.edu
A  THE DIDEROT LANGUAGE

This appendix provides a concise, but complete tour of the Diderot language.

A.1 Types

Diderot is a strongly-typed language with a novel, dependent-type system that statically tracks important mathematical properties of the program. The following table summarizes the types of the language:

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bool</td>
<td>booleans</td>
</tr>
<tr>
<td>int</td>
<td>integers</td>
</tr>
<tr>
<td>string</td>
<td>strings</td>
</tr>
<tr>
<td>tensor</td>
<td>tensor with shape (\sigma)</td>
</tr>
<tr>
<td>S</td>
<td>reference to a strand (S)</td>
</tr>
<tr>
<td>(\tau{d})</td>
<td>known-size sequence ((1 \leq d))</td>
</tr>
<tr>
<td>(\tau())</td>
<td>unknown/dynamic length sequence</td>
</tr>
<tr>
<td>(\text{image}(d)[\sigma])</td>
<td>dimension-(d) image ((1 \leq d \leq 3))</td>
</tr>
<tr>
<td>(\text{kernel}k)</td>
<td>reconstruction kernel of continuity (k) ((0 \leq k))</td>
</tr>
<tr>
<td>(\text{field}k(d)[\sigma])</td>
<td>tensor field with continuity (k), dimension (d), and shape (\sigma)</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>tensor shape ((0 \leq d))</td>
</tr>
</tbody>
</table>

Diderot also defines synonyms for common tensor shapes, such as \(\text{real}\) for \(\text{tensor}[\] and \(\text{vec}3\) for \(\text{tensor}[3]\).

A.2 Top-level definitions

A Diderot program consists of a sequence of top-level definitions. At a minimum, the top-level definitions include a strand definition and a global \(\text{initially}\) block, but they may also include function definitions, global variable declarations, input declarations, and a global update block.

A.3 Global variables

Global variables, which include input variables, provide a mechanism to communicate information to all of the strands in the program. They are typically declared and initialized at the beginning of the program and may be updated in the global update block.Globals may be annotated as \(\text{input}\) variables, in which case they can be initialized from outside the Diderot program (either from the command-line in a standalone executable or using a library call).

A.4 Functions

Diderot allows the definition of user-defined functions at top level. There are two syntactic forms. For a function whose definition can be given as an expression, we write

\[
\text{function type Name (parameters) = expression;}
\]

whereas if the function implementation requires local variables and/or statements, we write

\[
\text{function type Name (parameters) { statements }
}
\]

A.5 Strand definitions

Every Diderot program has a single strand definition, which has the form:

\[
\text{strand Name (parameters) { state variables update (statements) stabilize (statements) }
}
\]

The (optional) strand parameters are used to pass instance-specific data to the strand initialization. The strand state variables hold the strand’s state; some variables may be annotated as \(\text{output}\) variables, which means that there final values are included in the result of the program. The strand state variables are initialized with the strand is created and are modified by the strand methods. Strands have one or two methods: the required \(\text{update}\) method that is invoked at each step and an optional \(\text{stabilize}\) method that is invoked when the strand stabilizes.

A.6 Global update block

Diderot programs may optionally include a global \(\text{update}\) block that is run at the end of each execution step. The code in this block may perform global reductions over the strands in the program and is also allowed to update global variables.

A.7 Global initially block

The global \(\text{initially}\) block is used to create the initial set of strands. Diderot uses a comprehension syntax, similar those of Haskell or Python, to define the initial set of strands. For example, the following code specifies a grid of initial ray positions:

\[
\text{initially [ RayCast(ui, vi) | vi in 0 .. imgResV-1, ui in 0 .. imgResU-1 ]; }
\]

When the strands are initialized as a grid, it implies that the strands will all stabilize \((\text{i.e., they do not die})\). The grid structure is then preserved in the output.

Diderot also allows one to specify an initial \(\text{collection}\) of strands by using \(\{ \}_{;}\) as the brackets around the comprehension (instead of \(\{\}_{;}\)). In this case, the program’s output will be a one-dimension array of values for each \(\text{output}\) variable in a stable strand.

A.8 Statements

Statements in Diderot are a subset of C statements and include local-variable definitions, assignments, conditionals, and blocks that define a nested scope. In addition, there are several statement forms that are special to Diderot. The \(\text{new}\) statement is used to dynamically create a new strand. The \(\text{stabilize}\) and \(\text{die}\) statements terminate the calling strand; in the case of \(\text{stabilize}\), the strand’s state is preserved and may contribute to the remaining computation and to the result. Lastly, Diderot has a looping statement form (\(\text{foreach}\)) that provides iteration over dynamic sequences.

A.9 Expressions

Diderot’s expression syntax is fairly standard, with infix binary operators, unary operators, function application (including field application), etc. Conditional expressions follow the Python syntax. We use Unicode characters for operators and literals to support traditional mathematical notation. Tensor values can be constructed from lower-order tensors using the notation \([e_1, \ldots, e_n]\). We also provide the special tensor literals \(\text{id}entity(d)\) for the \(d \times d\) identity matrix, \(\text{zeros}(\sigma)\) for the zero tensor of the given shape, and \(\text{nan}(\sigma)\) for the tensor of the given shape initialized to NaN.

A.10 Operators and built-in functions

Diderot provides a rich set of unary and binary operations, as well as various built-in functions. These can be be overloaded (\(e.g., + \) works on integers, tensors, fields, etc.) and polymorphic (\(e.g., \) operates on tensors of any shape). A major use of overloading is to support \(\text{lifted}\) versions of tensor operations. We first present the overloaded operators and functions, along with the types at which they are defined. In this description, we use the type synonyms \(\text{real, vec}2\), and \(\text{vec}3\).

A.10.1 Overloaded operators and functions

Comparison operators: \(<,<=,=>,=\>
\[
\text{int} * \text{int} \rightarrow \text{bool}
\]
\[
\text{real} * \text{real} \rightarrow \text{bool}
\]

Equality operators: \(=\), \(!=\)
int * int → bool
real * real → bool
bool * bool → bool
string * string → string

Minimum and maximum (including reductions): min, max
int * int → int
real * real → real
real() → real

Unary negation: -
tensor[s] → tensor[s]
field[k](d)[σ] → field[k](d)[σ]
in field[k](d)[σ]
in tensor[s][σ]
Scalar multiplication:
int * int → int
tensor[s] ← tensor[s]
tensor[s] ← tensor[s]
real * field[k](d)[σ] ← field[k](d)[σ]
real * tensor[s] ← tensor[s]
tensor[s] * tensor[s] ← tensor[s]
tensor[s] * tensor[s] ← tensor[s]
tensor[s] * field[k](d)[σ] ← field[k](d)[σ]
field[k](d)[σ] * tensor[s] ← field[k](d)[σ]
field[k](d)[σ] * tensor[s] ← field[k](d)[σ]
field[k](d)[σ] * tensor[s] ← field[k](d)[σ]
Scalar division:
int * int → int
tensor[s] ← tensor[s]
tensor[s] ← tensor[s]
real * field[k](d)[σ] ← field[k](d)[σ]
real * tensor[s] ← tensor[s]
tensor[s] * tensor[s] ← tensor[s]
tensor[s] * tensor[s] ← tensor[s]
tensor[s] * field[k](d)[σ] ← field[k](d)[σ]
field[k](d)[σ] * tensor[s] ← field[k](d)[σ]
field[k](d)[σ] * tensor[s] ← field[k](d)[σ]
field[k](d)[σ] * tensor[s] ← field[k](d)[σ]
Scalar exponentiation:
real * int → real
real * real → real

Inner product:
tensor[s1](d1) * tensor[d2](d2) → tensor[s1](d2)
tensor[s1](d1) * field[k](d1)[d2] → field[k](d1)[d2]
field[k](d1)[d2] * tensor[s1](d1) → field[k](d1)[d2]
field[k](d1)[d2] * tensor[s1](d1) → field[k](d1)[d2]
field[k](d1)[d2] * tensor[s1](d1) → field[k](d1)[d2]

Double-dot product:
tensor[s1](d1,d2) * tensor[d2](d1) → tensor[s1](d2)
tensor[s1](d1,d2) * field[k](d1)[d2] → field[k](d1)[d2]
field[k](d1)[d2] * tensor[s1](d1,d2) → field[k](d1)[d2]
field[k](d1)[d2] * tensor[s1](d1,d2) → field[k](d1)[d2]
field[k](d1)[d2] * tensor[s1](d1,d2) → field[k](d1)[d2]

Cross product: ×
vec2 * vec2 → real
vec3 * vec3 → vec3
field[k](d)[2] * field[k](d)[2] → real
field[k](d)[3] * field[k](d)[3] → field[k](d)[3]

Outer product: ⊗
tensor[d1] * tensor[d2] → tensor[d1,d2]
field[k](d)[d1] * field[k](d)[d2] → field[k](d)[d1,d2]

Convolution:
image(d)[σ] * kernel[k] → field[k](d)[σ]
kernel[k] * image(d)[σ] → field[k](d)[σ]

Gradient: V
field[k](d)[1] → field[k](d)[d]

Tensor derivative: V⊗
field[k]+1(d)[σ] → field[k](d)[sh,d]

Curl: V×
field[k]+1(2)[2] → field[k](2)[σ]
field[k]+1(3)[3] → field[k](3)[3]

Divergence: V·
field[k]+1(d−1)[sh,d2] → field[k](d)

Norm: 1:1
tensor[s] → real
field[k](d)[σ] → field[k](d)[σ]

Normalization: normalize
tensor[s] → tensor[s]
field[k](d−1)[σ] → field[k](d)[σ]

Trace of a matrix: trace
tensor[d][d] → real
field[k](d−1)[sh,d2] → field[k](d−1)[σ]

Transpose of a matrix: transpose
tensor[d1,d2] → tensor[d2,d1]
field[k](d)[d1,d2] → field[k](d)[d2,d1]

Eigenvalues: evals
tensor[2,2] → real[2]
tensor[3,3] → real[3]

Eigenvectors: evvec
tensor[2,2] → vec2[2]
tensor[3,3] → vec3[3]

Determinant: det
tensor[2,2] → real
tensor[3,3] → real
field[k](d)[2] → field[k](d)[d]
field[k](d)[3] → field[k](d)[d]

Clamping: clamp
real * real * real → real
tensor[σ] * tensor[σ] * tensor[σ] → tensor[σ]
Lerp: \texttt{lerp}

\[
\text{tensor}[\sigma] \times \text{tensor}[\sigma] \times \text{real} \to \text{tensor}[\sigma] \\
\text{tensor}[\sigma] \times \text{tensor}[\sigma] \times \text{real} \times \text{real} \to \text{tensor}[\sigma]
\]

Trigonometry: \texttt{arccos, arcsin, cos, sin}

\[
\text{real} \to \text{real} \\
\text{field}^k(d)[\sigma] \to \text{field}^k(d)[\sigma]
\]

Sequence concatenation: \texttt{@}

\[
\tau \times \tau() \to \tau() \\
\tau() \times \tau \to \tau() \\
\tau() \times \tau() \to \tau()
\]

A.10.2 Other operators and functions

Pointwise vector multiplication: \texttt{modulate}

\[
\text{tensor}[d] \times \text{tensor}[d] \to \text{tensor}[d]
\]

Testing the domain of a field: \texttt{inside}

\[
\text{tensor}[d] \times \text{field}^k(d)[\sigma] \to \text{bool}
\]

Sequence length: \texttt{length}

\[
\tau() \to \text{int}
\]

Boolean-sequence reductions: \texttt{all, exists}

\[
\text{bool}() \to \text{bool}
\]

Scalar-sequence reductions: \texttt{mean, product, sum, variance}

\[
\text{real}() \to \text{real}
\]

Loading images: \texttt{image}

\[
\text{string} \to \text{image}(d)[\sigma]
\]

Image dimensions: \texttt{load}

\[
\text{image}(d)[\sigma] \to \text{int}(d)
\]

Loading sequences: \texttt{load}

\[
\text{string} \to \tau[\sigma]()$
\]
**B Embedding Diderot into an Application**

In addition to compiling to a self-contained executable, a Diderot program may be compiled to a C library, which allows Diderot programs to be embedded in any language that supports calling C code. We demonstrate this here with the program that produced the dense iso-contour sampling seen in Fig. 1(d). The complete Diderot code is below.

```c
1 // Initialize inputs
2 ISO_Init(wrld);
3
4 // Global initialization
5 ISO_Initially(wrld);
6
7 // Main loop
8 while (ISO_NumActive(wrld) > 0) {
9     // get and render the state
10     if (ISO_Snapshot_pos(wrld, nout)) {
11         // handle error...
12     }
13     Draw(nout);
14     // step the computation
15     ISO_Run(wrld, 1);
16 }
17
18 // get and render final state
19 if (ISO_OutputGet_pos(wrld, nout)) {
20     // handle error...
21 }
22 Draw(nout);
23 // shutdown the world
24 ISO_Shutdown(wrld);
25 ...
26 }
27
28 Lines 4–25 start the Diderot program. Line 5 allocates a new execution context, called a *world*. This context contains all runtime state of the program. Line 8 requests 8 cores for parallel execution, but only when the program was compiled for the SMP target. Line 12 allocates the computational resources needed to support the program’s execution (such as memory and threads). Lines 15 through 19 initialize those program input variables lacking defaults (the same ones set via command-line options above). Line 23 initializes the program, including allocating the initial set of strands. Line 25 allocates an Nrrd struct to contain the program output.
29
30 The while loop (lines 28–36) runs the program until all of the strands have stabilized or died. For each iteration of the loop, we grab (line 30) a snapshot of the program state (the union of all active and stable strands) and display it (line 33) with some draw function, the details of which are not important here. For this program, which computes a collection of strands (due to the initially (...) in lines 30 and 31 of Diderot code above), the output allocated in nout by ISO_Snapshot_pos will be a two-dimensional 2-by-N array, where N is the number of output values, and the (faster) first axis is length 2 because the output variable pos is a vec2. Had this program initialized on grid of strands (with initially (...) ) the output would be a higher dimensional array, with one axis for each dimension of the grid, and an additional axis if the output variable was not real.
31
32 Each while loop iteration executes (line 35) one super-step of the Diderot program. It is also possible to run all strands to completion (without being able to access snapshots of intermediate state) with ISO_Run(wrld, 0). The while loop terminates when there are zero active strands (line 28). The final output of the program execution is grabbed by line 39 and displayed in line 42. The resources allocated by program initialization and execution are released by line 45.
```

The executable has command-line options to set all the input variables; values must be given for those lacking defaults. Inputs stepsMax and epsilon do have defaults (lines 6 and 7). The final “-n 8” option sets the number of cores for parallel execution, but only when the program was compiled for the SMP target. Line 25 allocates a `ISO_World_t * wrld = ISO_New();` context containing all runtime state of the program. Line 8 requests 8 cores for parallel execution, but only when the program was compiled for the SMP target. Line 12 allocates the computational resources needed to support the program’s execution (such as memory and threads). Lines 15 through 19 initialize those program input variables lacking defaults (the same ones set via command-line options above). Line 23 initializes the program, including allocating the initial set of strands. Line 25 allocates an Nrrd struct to contain the program output.

The while loop (lines 28–36) runs the program until all of the strands have stabilized or died. For each iteration of the loop, we grab (line 30) a snapshot of the program state (the union of all active and stable strands) and display it (line 33) with some draw function, the details of which are not important here. For this program, which computes a collection of strands (due to the initially (...) in lines 30 and 31 of Diderot code above), the output allocated in nout by ISO_Snapshot_pos will be a two-dimensional 2-by-N array, where N is the number of output values, and the (faster) first axis is length 2 because the output variable pos is a vec2. Had this program initialized on grid of strands (with initially (...) the output would be a higher dimensional array, with one axis for each dimension of the grid, and an additional axis if the output variable was not real.

Each while loop iteration executes (line 35) one super-step of the Diderot program. It is also possible to run all strands to completion (without being able to access snapshots of intermediate state) with ISO_Run(wrld, 0). The while loop terminates when there are zero active strands (line 28). The final output of the program execution is grabbed by line 39 and displayed in line 42. The resources allocated by program initialization and execution are released by line 45.

The executable has command-line options to set all the input variables; values must be given for those lacking defaults. Inputs stepsMax and epsilon do have defaults (lines 6 and 7). The final “-n 8” option sets the number of cores for parallel execution, but only when the program was compiled for the SMP target. Line 25 allocates a `ISO_World_t * wrld = ISO_New();` context containing all runtime state of the program. Line 8 requests 8 cores for parallel execution, but only when the program was compiled for the SMP target. Line 12 allocates the computational resources needed to support the program’s execution (such as memory and threads). Lines 15 through 19 initialize those program input variables lacking defaults (the same ones set via command-line options above). Line 23 initializes the program, including allocating the initial set of strands. Line 25 allocates an Nrrd struct to contain the program output.

The while loop (lines 28–36) runs the program until all of the strands have stabilized or died. For each iteration of the loop, we grab (line 30) a snapshot of the program state (the union of all active and stable strands) and display it (line 33) with some draw function, the details of which are not important here. For this program, which computes a collection of strands (due to the initially (...) in lines 30 and 31 of Diderot code above), the output allocated in nout by ISO_Snapshot_pos will be a two-dimensional 2-by-N array, where N is the number of output values, and the (faster) first axis is length 2 because the output variable pos is a vec2. Had this program initialized on grid of strands (with initially (...) the output would be a higher dimensional array, with one axis for each dimension of the grid, and an additional axis if the output variable was not real.

Each while loop iteration executes (line 35) one super-step of the Diderot program. It is also possible to run all strands to completion (without being able to access snapshots of intermediate state) with ISO_Run(wrld, 0). The while loop terminates when there are zero active strands (line 28). The final output of the program execution is grabbed by line 39 and displayed in line 42. The resources allocated by program initialization and execution are released by line 45.
C COMPARISON WITH HAND-CODED OPENCL

We show here comparisons between Diderot and hand-coded OpenCL code, for two simple volume renders: a maximum intensity projection (MIP), and isosurface rendered colormapped by the second derivative. These examples are somewhat favorable to OpenCL due to their simplicity compared to the previous examples in the paper, but the properties of the code, and changes required for a more sophisticated algorithm, nonetheless demonstrate benefits of Diderot.

1 // coefficients of piece-wise cubic bspln3
2 __constant float4 h[4][4] = {
3 { 1.333333f, 0.0f, 1.0f, 0.166667f }, // -2 .. -1
4 { 0.666667f, 0.0f, -1.0f, -0.5f }, // -1 .. 0
5 { 0.666667f, 0.0f, -1.0f, -0.5f }, // 0 .. 1
6 { 1.333333f, -2.0f, 1.0f, -0.166667f }, // 1 .. 2
7 };
8 __constant int ha = 2; // half of support=4
9 #define FLOAT4(p) (float4)(p[0], p[1], p[2], p[3])
10
11 static inline float probVal(float global ushort = volData,
12 int sx, int sy, float4 volPos) {
13 float4 nf, t;
14
15 float4 d = (float4)(h[3][0],h[2][0],h[1][0],h[0][0]);
16 float4 c = (float4)(h[3][1],h[2][1],h[1][1],h[0][1]);
17 float4 b = (float4)(h[3][2],h[2][2],h[1][2],h[0][2]);
18 float4 a = (float4)(h[3][3],h[2][3],h[1][3],h[0][3]);
19
20 float4 t = modf(volPos,sf);
21 int n = convert_int4(nf);
22
23 t = (float4)(f.x + 1, f.x, f.x - 1, f.x - 2);
24 float4 hx = d + t * (c + t * (b + t*a));
25 float4 hy = d + t * (c + t * (b + t*a));
26 float4 hz = d + t * (c + t * (b + t*a));
27 float4 hv4[4], vy4[4]; // 4 == kernel support
28
29 for (int k = 1; ha; k++) { // slowest
30 for (int j = 1; ha; j++) { // medium
31 int index = n.*x + 1 + sy.*y + j + sy.*y.*z+k);
32 float4 v = FLOAT4(volData + index);
33 vec3 h[3][3] = dot(v, h[3]);
34 }
35 vec3 vy[4] = dot(FLOAT4(vy), hvy);
36 }
37 return dot(FLOAT4(vy), hz);
38 }
39
40 __constant float4 eye = (FLOAT4)(-240.538f, 396.133f, 64.680f, 1.0f);
41 __constant float4 orig = (FLOAT4)(0.170942f, 0.146426f, -0.0851239f, 0.0f);
42 __constant float4 du = FLOAT4(0.170942f, 0.146426f, -0.0851239f, 0.0f);
43 __constant float4 dv = FLOAT4(0.049961f, 0.0720342f, -0.224239f, 0.0f);
44 __constant float raystep = 0.3f;
45 __constant float lemmax = 130;
46
47 __kernel void kern(global ushort = volData,
48 float4 w2lPos,
49 float4 w2lGrad, // (unused)
50 global int = volSize,
51 int imgSizeU, int imgSizeV,
52 global float = out) {
53 int ui = get_global_id(0), vi = get_global_id(1);
54 int sx = volSize[0];
55 int sy = volSize[1];
56 float4 pos0 = orig + ui*du + vi*dv;
57 float4 raydir = normalize(pos0 - eye).x012;
58 float mip = 8;
59 for (float rlen = 0.0; rlen < lemmax; rlen += raystep) {
60 float4 ipos = pos0 + rlen*raydir; // Transform position from world to index
61 float4 ipos = FLOAT4(ipos+ (w2lPos, w2lPos, 0.0123),
62 dot(w2lPos, w2lPos, 0.4567),
63 dot(w2lPos, w2lPos, 0.89ab),
64 dot(w2lPos, w2lPos, 0.123456789ab),
65 if ((ipos.x > 1) && (ipos.y < sx - 2)
66 && (ipos.z > 1) && (ipos.z < sz - 2)) {
67 float val = probVal(volData, sx, sy, ipos);
68 mip = fmax(val, mip);
69 }
70 }
71 out[ui + imgSizeU*vi] = mip;
72 }
73
Above is hand-written OpenCL code for a simple ray-casting maximum intensity projection volume renderer. The code begins (lines 2-7) with a hard-coded definition of cubic B-spline kernel, in terms of its four piecewise cubic segments over $[-2,2]$. The probVal function (lines 12-42) implements convolution-based reconstruction with this kernel. The evaluation of the B-spline kernel at the grid sample locations is four-way parallelized in the setting of $h_x$, $h_y$, and $h_z$ (lines 25, 27, 29). The Diderot compiler generates the same parallelization for kernels with support four. After defining simple camera and ray-casting parameters (lines 44-53), the remainder of the code defines the computational kernel kern, which computes the maximum intensity projection for a single ray, indexed by $ui$ and $vi$. Rays are traversed in world-space, and converted (lines 73-76) to index space, to determine (lines 77-79) if the sample location ipos falls within the volume data, before evaluating the field with probVal. There are about 400 additional lines of hand-written C driver code (not shown) for compiling and running the kernel, and for saving its output.

1 vec3 eye = [-240.538, 396.133, 64.680];
2 vec3 orig = [-47.5009, 48.5999, 71.5282];
3 vec3 du = [0.170942, 0.146426, -0.0851239];
4 vec3 dv = [0.049961, 0.0720342, -0.224239];
5 real raystep = 0.3;
6 real lemmax = 130;
7
8 field[2][2][2] f = bspln3 @ image ("vrhand.nrrd");
9 input int imgSizeU;
10 input int imgSizeV;
11
12 strand raycast (int vi, int ui) {
13 vec3 pos0 = orig + ui*du + vi*dv;
14 vec3 raydir = normalize(pos0 - eye);
15 real rlen = 0;
16 output real mip = 0;
17
18 update |
19 vec3 pos = pos0 + rlen*raydir;
20 if (inside(pos, F)) {
21 mip = max(F(pos), mip);
22 }
23 rlen += raystep;
24 if (rlen >= lemmax) {
25 stabilize;
26 }
27
28 }
29 initially { raycast(vi, ui) | vi in 0..imgSizeV-1,
30 ui in 0..imgSizeU-1; }

Above is the same MIP renderer written in Diderot. The main difference is that the Diderot code is expressed in terms of world space, and that the details of convolution are hidden behind "$\ast$" and "$f(pos)$". The code for update (lines 12-28) is similar to the OpenCL code. This is expected: Diderot is something like a shader language for tensors and fields. However, to change the kernel in Diderot, only the kernel name needs changing on line 8. In OpenCL, the kernel coefficients would be different, and the details of the OpenCL probeVal function would be significantly different for a kernel with smaller or larger support, or higher or lower polynomial order. The Diderot compiler automatically generates all the necessary code for convolution in OpenCL, as well as the driver code needed to compile and execute the computational kernels. Both programs produce the same rendering, seen in Fig. 12.

Fig. 12. 640x480 maximum intensity projection of right hand of Visible Human female CT scan.
The Diderot-generated OpenCL code is not has fast as the handwritten code, although it is more flexible. On the same platform as used for benchmarking (Sec. 6), the OpenCL code took on average 0.53 seconds to run (excluding compilation and disk I/O), while the Diderot code took on average 1.46 seconds. As noted in Sec. 6, our OpenCL implementation does not currently perform as well as might be hoped; work on it continues. On the other hand, as noted in Sec. 1, in creating Diderot we are not seeking to optimize execution time so much as human implementation time, with the long-term goal of assisting the rapid development and application of novel visualization algorithms.

The next rendering example illustrates this. Below is Diderot code for a kind of shaded pseudo-isosurface volume rendering. The rendering algorithm was designed to help assess whether an isosurface coincides with a material edge, considered as the zero-crossing of the second-directional derivative along the gradient vector. This is complementary to the Canny edge criterion visualized in Sec. 4.1, and requires only second (not third) derivatives.

```diderot
valTol = 150;
input int imgSizeU;
input int imgSizeV;
real raystep = 0.1;
real lenmax = 130;
real valIso = 1440;
float4 f = bspln3 @ image ("vsfhand.nrrd") - valIso;
input int imgSizeW;
in world space. For this more complex rendering task, however, the performance hit from Diderot (relative to hand-written OpenCL code) is smaller than in the MIP example. The hand-written OpenCL code took on average 1.84 seconds, while the Diderot code took on average 3.30 seconds.

The OpenCL code is below. The computational kernel (starting line 194) is about as legible as the Diderot update code, though the double contraction of the Hessian (lines 232-234) is not as clear. Also, whereas the Diderot compiler can easily optimize repetitions of "(VF(pos))" with common subexpression elimination, the OpenCL programmer is more likely to call functions like probeGrad once (line 224) and save the results in a suggestively named variable. Given the mathematical limitations of OpenCL, it is hard to correctly implement probeGrad (lines 44-86) for reconstructing the world-space gradient, and probeHess (lines 88-166) for the Hessian, especially the conversion of the index-space Hessian to world-space (lines 144-163). Even with these utility functions written for scalar data, however, the OpenCL code is no closer to having code for evaluating \(\nabla (VF)\) than in the MIP example. The hand-written OpenCL code used for rendering Canny edges, which Diderot generated automatically. Additional OpenCL code would be needed for doing convolution-based reconstruction of vector- and tensor-valued data. Handling derivatives in vector and tensor fields (including conversions from index to world space) would be challenging even for the most meticulous programmer. We suggest that this is exactly the kind of tedious error-prone work that should be automated by the compiler of a high-level language like Diderot.

```diderot
real alpha = atxf(F(pos), |F(pos)|) && |F(pos)| <= valTol) {
    sdd = norm x F(pos);
    real lit = lerp(0, 1, -1, normOfLight, 1/2);
    real addn = norm x (VF(pos) * normOfLight); // half of support=4
    real col = lerp(gray, [0.5, 0.5, 0.5], addn); // 0 .. 1
    rgb = transp*alpha*lit*col;
    transp = 1 - alpha;
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}
}```

Fig. 13. 640 × 480 rendering of isosurface at 1440 on right hand of Visible Human female CT scan, colored by second directional derivative (green: positive, gray: zero, magenta: negative).
float4 hx = dtx + (c + t * (b + t * a));
float4 hxx = dtx + (c + t * (b + t * a));
float4 hxh = dtx + (c + t * (b + t * a));
float4 hxxh = dtx + (c + t * (b + t * a));

float vx(x, y) = dot(FLOAT4(volData + index), ihes1.s0);

static inline float3 probGrad(float4 i2wGrad, volData,
int sx, int sy, float4 volPos) {
}

static inline float3 probGrad(float4 i2wGrad, volData,
int sx, int sy, float4 volPos) {
int n = convert_int4(nf);
int j = 1 - hs; j <= hs; j++) { // medium

// Transform position from world to index

// convert hessian to world space, 1st index

// convert hessian to world space, 2nd index

// hessian in index space

// transform hessian from index to world space

// convert position from world to index

// convert position to world

// transform hessian from index to world space

// convert position from world to index

// convert position from world to index

// transform hessian from index to world space
float3 col = (sdd < 0) ? mix(gray, (float3)(1.0, 0.0, 0.7), -sdd) : mix(gray, (float3)(0.3, 1.0, 0.3), sdd);
rgb += transp*alpha*lit*col;
transp *= 1 - alpha;
}
global float *opix = out + 4*(ui + imgSizeU*vi);
float aa = 1 - transp;
if (aa > 0) {
opix[0] = rgb.s0/aa;
opix[1] = rgb.s1/aa;
opix[2] = rgb.s2/aa;
opix[3] = aa;
} else {
opix[0] = 0;
opix[1] = 0;
opix[2] = 0;
opix[3] = 0;
}