Compiling Embedded Languages

Conal Elliott[1], Sigbjørn Finne[1] and Oege de Moor[2]

[1] Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
[2] Oxford University Computing Laboratory,
Wolfson Building, Parks Road,
Oxford, OX1 3QD, England

Abstract. Functional languages are particularly well-suited to the implementation of interpreters for domain-specific embedded languages (DSELs). We describe an implemented technique for producing *optimizing compilers* for DSELs, based on Kamin's idea of DSELs for program generation. The technique uses a data type of syntax for basic types, a set of smart constructors that perform rewriting over those types, some code motion transformations, and a back-end code generator. Domain-specific optimization results from chains of rewrites on basic types. New DSELs are defined directly in terms of the basic syntactic types, plus host language functions and tuples. This definition style makes compilers easy to write and, in fact, almost identical to the simplest embedded interpreters. We illustrate this technique with a language *Pan* for the computationally intensive domain of image synthesis and manipulation. [1]

1 Introduction

The "embedded" approach has proved an excellent technique for specifying and prototyping domain-specific languages (DSLs) [8]. The essential idea is to augment a "host" programming language with a domain-specific library. Modern functional host languages are flexible enough that the resulting combination has more the feel of a new language than a library. Most of the work required to design, implement and document a language is inherited from the host language. Often, performance is either relatively unimportant, or is adequate because the domain primitives encapsulate large blocks of work. When speed is of the essence, however, the embedded approach is problematic. It tends to yield inefficient *interpretive* implementations. Worse, these interpreters tend to perform redundant computation.

We have implemented a language *Pan* for image synthesis and manipulation, a computationally demanding problem domain. A straightforward embedded implementation would not perform well enough, but we did not want to incur

[1] This paper will appear at the *Semantics, Applications and Implementation of Program Generation (SAIG 2000)* workshop as part of PLI 2000, and is © Springer-Verlag. See `http://www.springer.de/comp/lncs/index.html`.

the expense of introducing an entirely new language. Our solution is to embed an *optimizing compiler* rather than an interpreter. Embedding a compiler requires some techniques not normally needed in embedded language implementations, and we report on these techniques here. Pleasantly, we have been able to retain a simple programming interface, almost unaffected by the compiled nature of the implementation. The generated code runs very fast, and there is still much room for improvement.

Our compiler consists of a relatively small set of domain definitions, on top of a larger domain-independent framework. The framework may be adapted for compiling other DSLs, and handles (a) optimization of expressions over numbers and Booleans, (b) code motion, and (c) code generation. A new DSL is specified and implemented by defining the key domain types and operations in terms of the primitive types provided by the framework and host language. Moreover, these definitions are almost identical to what one would write for a very simple interpretive DSL implementation.

Although a user of our embedded language writes in Haskell, we do not have to parse, type-check, or compile Haskell programs. Instead, the user *runs* his/her Haskell program to produce an optimized program in a simple target language that is first-order, call-by-value, and mostly functional. Generated target language programs are then given to a simple compiler (also implemented in Haskell) for code motion and back-end code generation. In this way, the host language (Haskell here) acts as a powerful macro (or *program generator*) language, but is completely out of the picture at run-time. Unlike most macro languages, however, Haskell is statically typed and higher order, and is more expressive and convenient than the underlying target language.

Because of this embedded compiler approach, integration of the DSEL with the host language (Haskell) is not quite as fluid and general as in conventionally implemented DSELs. Some host language features, like lists, recursion, and higher-order functions are not available to the final executing program. These features may be used in source programs, but disappear during the compilation process. For some application areas, this strict separation of features between a full-featured compilation language and a less rich runtime language may be undesirable, but in our domain, at least, it appears to be perfectly acceptable. In fact, we typically write programs without being conscious of the difference.

The contributions of this paper are as follows:

- We present a general technique for implementing embedded *optimizing* compilers, extending Kamin's approach [10] with algebraic manipulation.
- We identify a key problem with the approach, efficient handling of sharing, and present techniques to solve it (bottom-up optimization and common subexpression elimination).
- We illustrate the application of our technique to a demanding problem domain, namely image synthesis and manipulation.

While this paper mainly discusses embedded language compilation, a companion paper goes into more detail for the Pan language [4]. That paper contains many more visual examples, as does [2].

2 Language embedding

The embedding approach to DSL construction goes back at least to Landin's famous "next 700" paper [12]. The essential idea is to use a single existing "host" programming language that provides useful generic infrastructure (grammar, scoping, typing, function- and data-abstraction, etc), and augment it with a domain-specific vocabulary consisting of one or more data types and functions over those types. Thus the design, implementation, and documentation work required for a new "language" is kept to a minimum, while the result has plenty of room to grow. These merits and some drawbacks are discussed, e.g., in [3, 8].

One particularly elegant realization of the embedding idea is the use of a modern functional programming language such as ML or Haskell as the host. In this setting, the domain-specific portions can sometimes be implemented as a simple denotational semantics, as suggested in [11, Section 3]. For example, consider the problem domain of image synthesis and manipulation. A simple semantics for images is function from continuous 2D space to colors. The representation of colors includes blue, green, red, and opacity ("alpha") components:

$$\begin{aligned}
&\textbf{type}\ \mathit{Image} &&= \mathit{Point} \to \mathit{Color}\\
&\textbf{type}\ \mathit{Point} &&= (\mathit{Float}, \mathit{Float})\\
&\textbf{type}\ \mathit{Color} &&= (\mathit{Float}, \mathit{Float}, \mathit{Float}, \mathit{Float})
\end{aligned}$$

It is easy to implement operations like image overlay (with partial opacity), assuming a corresponding function, *cOver*, on color values:

$$a\ \text{`}\mathit{over}\text{`}\ b = \lambda p \to a\ p\ \text{`}\mathit{cOver}\text{`}\ b\ p$$

Another useful type is spatial transformation, which may be defined simply as a mapping from 2D space to itself:

$$\textbf{type}\ \mathit{Transform} = \mathit{Point} \to \mathit{Point}$$

This model makes it easy to define some familiar transformations:

$$\begin{aligned}
&\mathit{translate}\ (dx,\ dy) &&= \lambda (x, y) \to (x + dx,\ y + dy)\\
&\mathit{scale}\ (sx,\ sy) &&= \lambda (x, y) \to (sx * x,\ sy * y)\\
&\mathit{rotate}\ \mathit{ang} &&= \lambda (x, y) \to (x * c - y * s,\ y * c + x * s)\\
&\quad\textbf{where} &&\\
&\qquad c = \mathit{cos}\ \mathit{ang} &&\\
&\qquad s = \mathit{sin}\ \mathit{ang} &&
\end{aligned}$$

While these definitions can be directly executed as Haskell programs, performance is not good enough for practical use. Our first attempt to cope with this problem was to use the Glasgow Haskell compiler's facility for stating transformations as rewrite rules in source code [15]. Unfortunately, we found that the interaction of such rewrite rules with the general optimizer is hard to predict: in particular, we often wish to inline function definitions that would normally

not have been inlined. Furthermore, there are a number of transformations (if-floating, certain array optimizations) that are not easy to state as rewrite rules. We therefore abandoned use of the Haskell compiler, and decided to build a dedicated compiler instead. We will discuss this decision further in Section 10.

3 Embedding a compiler

In spite of our choice to implement a dedicated compiler, we would like to retain most of the benefits of the embedded approach. We resolve this dilemma by applying Kamin's idea of DSELs for program generation [10]. That is, replace the *values* in our representations by *program fragments* that represent these values. While Kamin used strings to represent program fragments, algebraic data types greatly facilitate our goal of compile-time optimization. For instance, an expression type for *Float* would contain literals, arithmetic operators, and other primitive functions that return *Float*.

```
data FloatE =
    LitFloat Float
  | AddF FloatE FloatE | MulF FloatE FloatE | ...
  | Sin FloatE | Sqrt FloatE | ...
```

We can define expression types *IntE* and *BoolE* similarly.

What about tuples and functions? Following Kamin, we simply adopt the host language's tuple and functions, rather than creating new syntactic representations for them. Since optimization requires inspection, representing functions as functions poses a problem. The solution we use is to extend the base types to support "variables". Then to inspect a function, apply it to a new variable (or tuple of variables as needed), and look at the result.

```
data FloatE = ... | VarFloat String    — named variable
```

These observations lead to a hybrid representation. Our *Image* type will still be represented as a function, but over *syntactic* points, rather than actual ones. Moreover, these syntactic points are represented not as expressions over number pairs, but rather as pairs of expressions over numbers. Similarly for colors. Thus:

```
type ImageE     = PointE → ColorE
type TransformE = PointE → PointE
type PointE     = (FloatE, FloatE)
type ColorE     = (FloatE, FloatE, FloatE, FloatE)
```

The definitions of operations over these types can often be made identical to the ones for the non-expression representation, thanks to overloading. For instance *translate*, *scale*, and *rotate* have precisely the definitions given in Section 2 above. The *meaning* of these definitions, however, is quite different. The

arithmetic operators and the functions *cos*, *sin* as well as several others have been overloaded. The *over* function is also defined exactly as before. Only the types *BoolE*, *IntE*, and *FloatE* of expressions over the usual "scalar" value types *Bool*, *Int*, and *Float*, are represented as expressions, using constructors for their primitive operations. Assuming that these base types are adequate, a DSL is just as easy to define and extend as with a simple, non-optimizing embedded interpreter. Otherwise new syntactic types and/or primitive operators may be added.

As an example of how the hybrid technique works in practice, consider rotating by an angle of $\pi/2$. Using the definition of *rotate* plus a bit of simplification on number expressions (*FloatE*), the compiler simplifies $rotate\ (\pi/2)\ (x, y)$ to $(-y, x)$.

Admittedly, the picture might not always be this rosy. For instance, some properties of high-level types require clever or inductive proofs. Formulating these properties as high-level rules would eliminate the need for a generic compiler to rediscover them. So far this has not been a problem for our image manipulation language, but we expect that for more substantial applications, it may be necessary to layer the compilation into a number of distinct abstract levels. In higher levels, domain types and operators like *Image* and *over* would be treated as opaque and rewritten according to domain-specific rules, while in lower levels, they would be seen as defined and expanded in terms of simpler types like *Point* and *Color*. Those simpler types would themselves be expanded at lower levels of abstraction.

4 Inlining and the sharing problem

The style of embedding described above has the effect of *inlining* all definitions, and β-reducing resulting function applications, before simplification. This inlining is beneficial in that it creates many opportunities for rewriting. A resulting problem, however, is that uncontrolled inlining often causes a great deal of code replication. To appreciate this problem, consider the following example spatial transform. It rotates each point about the origin, through an angle proportional to the point's distance from the origin. The parameter r is the distance at which an entire revolution (2π radians) is made.

$$
\begin{aligned}
&swirling\ ::\ FloatE\ \rightarrow\ TransformE\\
&swirling\ r\ =\ \lambda\, p\ \rightarrow\ rotate\ (distO\ p\ *\ (2\,\pi\ /\ r))\ p
\end{aligned}
$$

$$
\begin{aligned}
&distO\ ::\ PointE\ \rightarrow\ FloatE\\
&distO\ (x, y)\ =\ sqrt\ (x\ *\ x\ +\ y\ *\ y)
\end{aligned}
$$

Evaluating $swirling\ r\ (x, y)$ yields an expression with much redundancy.

$$
\begin{aligned}
&(\ x\ *\ cos\ (sqrt\ (x\ *\ x\ +\ y\ *\ y)\ *\ 2\,\pi\ /\ r)\\
&\quad -\ y\ *\ sin\ (sqrt\ (x\ *\ x\ +\ y\ *\ y)\ *\ 2\,\pi\ /\ r)\\
&,\ y\ *\ cos\ (sqrt\ (x\ *\ x\ +\ y\ *\ y)\ *\ 2\,\pi\ /\ r)
\end{aligned}
$$

$$+\ x\ *\ sin\ (sqrt\ (x\ *\ x\ +\ y\ *\ y)\ *\ 2\,\pi\ /\ r)\ )$$

The problem here is that *rotate* uses its argument four times (twice via each of *cos* and *sin*) in constructing its results. Thus expressions passed to *rotate* get replicated in the output. In our experience with Pan, the trees resulting from inlining and simplification tend to be enormous, compared to their underlying representation as graphs. If *swirling* r were composed with *scale* (u, v) before being applied to (x, y), the two multiplications due to *scale* would each be appear twice in the argument to *sqrt*, and hence eight times in the final result.

In an interpretive implementation, we would have to take care not to evaluate shared expressions redundantly. Memoization is a reasonable way to avoid such redundance. For a compiler, memoization is not adequate, because it must produce an external representation that captures the sharing. What we really want is to generate local definitions when helpful. To produce these local definitions, our compiler performs common subexpression elimination (CSE), as described briefly in Section 8 and in more detail in [5].

5 Static typing

Should there be one expression data type per value type (*Int*, *Float*, *Bool*, etc) as suggested above, or one for all value types? Separate expression types make the implementation more statically typed, and thus prevent many bugs in implementation and use. Unfortunately, they also lead to redundance for variables, binding, and polymorphically and overloaded expression operators (e.g., if-then-else and addition, respectively), as well as polymorphic compiler-internal operations on terms (e.g., substitution and CSE).

Instead, we use a single all-encompassing expression data type *DExp* of "dynamically typed expressions":

```
data DExp =
    LitInt Int | LitFloat Float | LitBool Bool
  | Var Id Type | Let Id Type DExp | If DExp DExp DExp
  | Add DExp DExp | Mul DExp DExp | ...
  | Sin DExp | Sqrt DExp | ...
  | Or DExp DExp | And DExp DExp | Not DExp | ...
```

It is unfortunate that the choice of a single *DExp* type means that one cannot simply add another module containing a new primitive type and its constructors and rewrite rules. For now we are willing to accept this limitation, but future work may suggest improvements.

The *DExp* representation removes redundance from representation and supporting code, but loses type safety. To combine advantages of both approaches, we augment the dynamically typed representation with the technique of "phantom types" [13]. The idea is to define a type constructor (*Exp* below) whose parameter is not used, and then to restrict types of some functions to applications of the type constructor. For convenience, define abbreviations for the three supported types as well:

```
data Exp α  = E DExp

type BoolE  = Exp Bool
type IntE   = Exp Int
type FloatE = Exp Float
```

For static typing, it is vital that *Exp* α be a new type, rather than just a type synonym of *DExp*.

Statically typed functions are conveniently defined via the following functionals, where typ_n turns an n-ary *DExp* function into an n-ary *Exp* function.

```
typ1 :: (DExp → DExp) → (Exp a → Exp b)
typ2 :: (DExp → DExp → DExp) → (Exp a → Exp b → Exp c)

typ1 f (E e1)         = E (f e1)
typ2 f (E e1) (E e2)  = E (f e1 e2)
```

and so on for typ_3, typ_4, *etc.* The type-safe friendly names $+$, $*$, *etc.*, come from applications of these static typing functionals in type class instances:

```
instance Num IntE
  where
    (+)         = typ2 Add
    (*)         = typ2 Mul
    negate      = typ1 Negate
    fromInteger = E . LitInt . fromInteger
```

Type constraints inherited from the *Num* class ensure that the newly defined functions be applied only to *Int* expressions and result in *Int* expressions. For instance, here

```
(+) :: IntE → IntE → IntE
```

The important point here is that we do not rely on type inference, which would deduce too general a type for functions like "+" on *Exp* values. Instead we state restricted type signatures.

Other definitions provide a convenient and type-safe primitive vocabulary for *FloatE*. Unfortunately, the *Bool* type is wired into the signatures of operations like $\geq$ and ||. Pan therefore provides alternative names ending in a distinguished character, which is "*E*" for alphanumeric names (e.g., "*notE*") and "$*$" for non-alphanumeric names (e.g., "$<*$").

6 Algebraic optimization and smart constructors

An early Pan implementation was based on the Mag program transformation system [1]. Generation in this implementation was much too slow, mainly because Mag redundantly rewrote shared subterms. To avoid this problem, we

```
— Type-safe smart constructor
(&&∗) :: BoolE → BoolE → BoolE
(&&∗) =typ2 andD

— Non-type-safe smart constructor
andD :: DExp → DExp → DExp
— Constant folding
andD (LitBool a) (LitBool b) = LitBool (a && b)
— If-floating
andD (If c a b) e2 = ifD c (andD a e2) (andD b e2)
andD e1 (If c a b) = ifD c (andD e1 a) (andD e1 b)
— Cancellation rules
andD e (LitBool False) = false
andD (LitBool False) e = false
andD e (LitBool True) = e
andD (LitBool True) e = e
— Others
andD (Not e) (Not e′)      = notE (e || ∗ e′)
andD e e′ | e == e′        = e
andD e e′ | e == notE e′ = false
— Finally, the data type constructor
andD e e′ = And e e′
```

Fig. 1. Simplification rules for conjunction

now do all optimization *bottom-up*, as part of the construction of expressions. Then the host language's evaluate-once operational semantics prevents redundant optimization. Non-optimized expressions are never constructed. The main drawback is that optimization is context-free. (An optimization can, however, delve arbitrarily far into an argument term.)

Optimization is packaged up in "smart constructors", each of which accomplishes the following:

- constant-folding;
- if-floating;
- constructor-specific rewrites such as identities and cancellation rules;
- data type constructor application when no optimizations apply; and
- providing a statically typed interface.

As an example, Figure 1 shows a smart constructor for conjunction over expressions. In fact, because all smart constructors perform constant folding and if-floating, the real definition is more factored, but it does the same work.

Because if-then-else is not overloadable, Pan uses *ifE* for syntactic conditionals, based on an underlying dynamically typed *ifD*.

```
ifD :: DExp → DExp → DExp → DExp
```

$$
\begin{array}{lcl}
\mathit{ifD}\ (\mathit{LitBool}\ \mathit{True}\ )\ a\ b & = & a \\
\mathit{ifD}\ (\mathit{LitBool}\ \mathit{False})\ a\ b & = & b \\
\mathit{ifD}\ (\mathit{Not}\ c)\ a\ b & = & \mathit{ifD}\ c\ b\ a \\
\mathit{ifD}\ (\mathit{If}\ c\ d\ e)\ a\ b & = & \mathit{ifD}\ c\ (\mathit{ifD}\ d\ a\ b)\ (\mathit{ifD}\ e\ a\ b) \\
\mathit{ifD}\ c\ a\ b & = & \mathit{ifZ}\ c\ a\ b
\end{array}
$$

The function *ifZ* simplifies redundant or impossible conditions.

The statically typed *ifE* function is overloaded.

$$\textbf{class}\ \mathit{Syntactic}\ a\ \textbf{where}\ \mathit{ifE}\ ::\ \mathit{BoolE}\ \rightarrow\ a\ \rightarrow\ a\ \rightarrow\ a$$

$$\textbf{instance}\ \mathit{Syntactic}\ (\mathit{Exp}\ a)\ \textbf{where}\ \mathit{ifE}\ =\ \mathit{typ}_3\ \mathit{ifD}$$

Other overloadings include functions and tuples. In the latter case, conditions are pushed downward. Later when the resulting tuple is consumed to form a single (scalar-valued) expression, if-floating typically causes the redundant conditions to float, to form a cascade of redundant conditionals, which are coalesced by *ifZ*.

As an example of if-floating, consider the following example (given in familiar concrete syntax, for clarity):

$$\mathit{sin}\ ((\textbf{if}\ x\ <\ 0\ \textbf{then}\ 0\ \textbf{else}\ x)\ /\ 2)$$

If-floating without simplification would yield

$$\textbf{if}\ x < 0\ \textbf{then}\ \mathit{sin}(0/2)\ \textbf{else}\ \mathit{sin}(a/2)$$

Replacement followed by two constant foldings ($0/2$ and $\mathit{sin}\ 0$) results in

$$\textbf{if}\ x < 0\ \textbf{then}\ 0\ \textbf{else}\ \mathit{sin}(a/2)$$

If-floating causes code replication, sometimes a great deal of it. CSE factors out the "first-order" replication, i.e., multiple occurrences of expressions, as with e_2 for the first if-floating clause in Figure 1. There is also a *second-order* replication going on, as seen above before simplification. The context $\mathit{sin}\ (\bullet/\ 2)$ appears twice. Fortunately for this example, one instance of this context simplifies to 0. In other cases, there may be little or no simplification. We will return to this issue in Section 10.

We should stress at this point that we intend the algebraic optimizations to be *refinements*: upon evaluation, the optimized version of an expression e should yield the same value as e *whenever evaluation of* e *terminates*. It is possible, however, for simplified version to yield a well-defined result when e does not. This could happen for example when a boolean expression $e\ \&\&\ *\ \mathit{false}$ would raise a division-by-zero exception, while the simplified version would instead evaluate to *false*.

7 Adding context

More optimization becomes possible when the usage context of a DSL computation becomes visible to the compiler. For instance, after composing an image,

```
type TimeE = FloatE
type Anim = TimeE → ImageE
type DisplayFun = TimeE → VTrans → VSize → IntE → ActionE
type VSize = (IntE, IntE)   — view size: width & height in pixels
type VTrans = (FloatE, FloatE, FloatE)   — view transform: pan XY, zoom

display :: Anim → DisplayFun
display anim = λ t (panX, panY, zoom) (w, h) output →
  loop h (λ j →
    loop w (λ i →
      setInt (output + 4 * (j * w + i)) (
        toBGR24 (
          anim t (
            zoom * i2f (i − w `div` 2) + panX,
            zoom * i2f (j − h `div` 2) + panY )))))
```

Fig. 2. Animation display function

a user generally wants to display it in a window. The representation of images as *PointE* → *ColorE* suggests iteratively sampling at a finite grid of pixel locations, converting each pixel color to an integer for the display device. (For a faithful presentation, images need to be antialiased, but that topic is beyond the scope of the present paper and not yet addressed in our implementation.) Our first Pan compiler implementation took this approach, that is it generated machine code for a function that maps a pixel location to a 32-bit color encoding. While this version was much faster than an interpretive implementation, its efficiency was not satisfactory. For one thing, it requires a function call per pixel. More seriously, it prevent any optimization across several pixels or rows of pixels.

To address the shortcomings of the first compiler, we made visible to the optimizer the two-dimensional iteration that samples and stores pixel values. In fact, to get more use out of compilation, we decided to compile the display of not simply static images, but animations, represented as functions from time to image. (We go even further, generating code for nearly arbitrarily parameterized images, with automatic generation of user interfaces for the run-time parameters.)

The main function *display*, defined in Figure 2, converts an animation into a "display function" that is to be invoked just once per frame. A display function consumes a time, window size, viewing transform (zoom factor and XY pan), and a pointer to an output pixel array. It is the job of the viewer to come up with all these parameters and pass them into the display function code.

The critical point here is that (a) the *display* function is expressed in the embedded language, and (b) *display* is applied to its *anim* parameter (of type *TimeE* → *Image*) *at compile time*. This compile-time application allows the

code for *display* and *anim* to be combined and optimized, and lets some computations be moved outside of the inner or outer loop. (In fact, our compiler goes further, allowing focused recomputations when only some display parameters change, thanks to a simple dependency analysis.)

The *ActionE* type represents an action that yields no value, much like Haskell's type *IO* (). It is supported by a small number of *DExp* constructors and corresponding statically typed, optimizing wrapper functions. The first takes an address (represented as an integer) and an integer value, and it performs the corresponding assignment. The second is like a for-loop. It takes an upper bound, and a loop body that is a function from the loop variable to an action. The loop body is executed for every value from zero up to (but not including) the upper bound.

$$\begin{aligned} &\mathit{setInt} :: \mathit{IntE} \rightarrow \mathit{IntE} \rightarrow \mathit{ActionE} \\ &\mathit{loop} \quad :: \mathit{IntE} \rightarrow (\mathit{IntE} \rightarrow \mathit{ActionE}) \rightarrow \mathit{ActionE} \end{aligned}$$

According to *display*, a generated display function will loop over Y and X, and set the appropriate member of its output array to a 32-bit (thus multiplication by four) color value. Aside from calculating the destination memory address, the inner loop body samples the animation at the given time and position. The spatial sampling point is computed from the loop indices by placing the image's origin in the center of the window (thus the subtraction of half the window width or height) and then applying the user-specified dynamic zoom and pan (using $i2f$ for int-to-float conversion). In fact, the optimized code is much more efficient, thanks to code motion techniques described briefly in Section 8 and illustrated in Appendix A.

8 Code motion and code generation

Once context is added and all of the above optimizations have been applied, the result is an expression tree (of type *DExp*). As explained in Section 4, this tree contains a great deal of sharing, mostly because of the inlining and rewriting process. The next step in compilation is to make the sharing structure explicit using let-bindings, i.e., performing a common subexpression elimination (CSE). Another very important form of code motion is hoisting evaluation out of loops when independent of the loop variable. Finally, we also sometimes synthesize arrays of values that depend on an inner loop variable but not an outer one. For details see [5], where some subtle strictness issues are also discussed.

Having performed code motion and loop hoisting, we are in good shape to start generating some code. The output of the code motion pass could either be interpreted or compiled, but we choose to compile. The resulting *DExp* is converted into a C function. This translation is reasonably straightforward, but requires a little bit of care in places, to account for the fact that C does not have expression level variable binding support or array initialization. The generated C code is then compiled and linked into a viewer that displays the specified image effect.

9 Related work

There are many other examples of embedded DSLs, for music, two- and three-dimensional geometry, animation, hardware design, document manipulation, and many other domains. See [8] for an overview and references. In almost all cases, the implementations were interpretive. Several characteristics of functional programming languages that lend themselves toward the role of host language are enumerated in [3].

Kamin's work on embedded languages for program generation is in the same spirit as our own [10]. As in our approach, Kamin uses host language functions and tuples to represent the embedded language's functions and tuples, and he uses overloading so that the generators look like the code they are generating. His applications use a functional host language (ML) and generate imperative programs. The main difference is that Kamin did not perform optimization or CSE. Both would be difficult, given his choice of strings to represent programs.

Leijen and Meijer's HaskellDB [13] provides an embedded language for database queries and an implementation that compiles search specifications into optimized SQL query strings for further processing. After trying several unsuccessful designs, we imitated their use of an untyped algebraic data type and a phantom type wrapper for type-safety.

Our approach to compiling embedded languages can be regarded as an instance of *partial evaluation*, which has a considerable literature (see, e.g., [7, 9]). In this light, our compiler is a handwritten *cogen* (as opposed to one generated automatically through self-application). The main contrasting characteristic of our work is the embedding in a strongly typed meta-language (Haskell). This embedding makes particular use of Haskell type-class-based overloading so that the concrete syntax of meta-programs is almost identical to that of object-programs, and it achieves inlining for free (perhaps too much of it). It also exploits meta-language type inference to perform object-language type inference (except on the optimization rules, which are expressed at the type-unsafe level). Another closely related methodology is multi-stage programming with explicit annotations, as supported by MetaML [14], a polymorphic statically typed meta-language for ML-style programs.

FFTW is a successful, portable C library for computing discrete Fourier transforms of varying dimensions and sizes [6]. Its numerical procedures are generated by a special purpose compiler, *fftgen*, written in Objective Caml and are better in almost all cases than previously existing libraries. The compiler has some of the same features as our own, performing some algebraic simplification and CSE. One small technical difference is that, while fftgen does memoized simplification, our compiler does bottom-up simplifying construction. It appears that the results are the same. Because the application domain is so specialized, fftgen is more focused than our compiler.

Veldhuizen and others have been using advanced C++ programming techniques to embed a simple functional language into C++ *types* [16, 17]. Functional evaluation is done by the C++ compiler during type-checking and template in-

stantiation. Code fragments specified in inlined static methods are chosen and combined at compile-time to produce specialized, optimized low-level code.

10 Future work

More efficient and powerful rewriting. Our optimizer uses a simple syntactic approach to rewriting. To obtain better results, rewriting and CSE should make use of associative-commutative (AC) matching and comparison, respectively, while still exploiting representation sharing, which is critical for compile-time efficiency.

CSE cleans up after inlining, recapturing what sharing still remains after rewriting. However, where inlining does *higher-order* substitution (in the case of functions), CSE is only first-order, so higher-order redundance remains. Ideally, inlining, if-floating, and CSE would all work cooperatively and efficiently with rewriting. Inlining and if-floating would happen only where rewarded with additional rewrites. Fundamentally, this cooperation seems precluded by the embedded nature of the language implementation, which forces full inlining as the first step, before the DSEL compiler gets to look at the representation.

Invisible compilation. The techniques described in this paper turn compositional specifications into efficient implementations. Image editing applications also allow non-programmers to manipulate images by composing operations. Imagine that such an application were to use abstract syntax trees as its internal editable representation and invisibly invoke an incremental optimizing compiler in response to the user's actions. Then a conventional point-and-click user interface would serve as a "gestural concrete syntax". The display representation would then be one or more bitmaps augmented by custom-generated machine code.

Embeddable compilation. By embedding our language in Haskell, we were able to save some of the work of compiler implementation, namely lexing, parsing, type checking, supporting generic scalar types, functions and tuples. However, it should be possible to eliminate still more of the work. Suppose that the host language's compiler were extended with optimization rules so that it could work much like the one described in this paper. We tried precisely this approach with GHC [15], with partial success. The main obstacle was that the compiler was too conservative about inlining and rewriting. It takes care never to slow down a program, whereas we have found that it is worth taking some backward steps in order to end up with a fast program in the end. Because we do not (yet) work with recursively defined images, laziness in a host language appears not to be vital in this case. It might be worthwhile to try the exercise with an ML compiler.

11 Conclusions

Embedding is an easy way to design and implement DSLs, inheriting many benefits from a suitable host language. Most such implementations tend to be

interpretive, and so are too slow for computationally intensive domains like interactive image processing. Building on ideas from Kamin and from Leijen and Meijer, we have shown how to replace embedded interpreters with optimizing compilers, by using a set of syntax-manipulating base types. The result is much better performance with a very small impact on the languages. Moreover, given a reusable DSL compiler framework such as we have implemented, an embedded DSL interpreter can be turned into a compiler with very small changes (thanks to overloading). In our Pan compiler, the rewriting-based optimizations helped speed considerably, as of course does eliminating the considerable overhead imposed by interpretative implementation.

We have produced many examples with our compiler, as may be seen in [2, 4], but more work is needed to make the compiler itself fast and producing even better code. We hope that the compiler's speed can be improved to the point of invisibility so that it can be used by non-programmers in image editors.

12 Acknowledgements

Brian Guenter originally suggested to us the idea of an optimizing compiler for image processing, and has collaborated on the project. Erik Meijer helped to sort out the many representation possibilities and suggested the approach that we now use.

References

1. Oege de Moor and Ganesh Sittampalam. Generic program transformation. In *Proceedings of the third International Summer School on Advanced Functional Programming*, Springer Lecture Notes in Computer Science, 1999. `http://-users.comlab.ox.ac.uk/oege.demoor/papers/braga.ps.gz`.
2. Conal Elliott. A Pan image gallery. `http://research.microsoft.com/~conal/-pan/Gallery`.
3. Conal Elliott. An embedded modeling language approach to interactive 3D and multimedia animation. *IEEE Transactions on Software Engineering*, 25(3):291–308, May/June 1999. Special Section: Domain-Specific Languages (DSL). `http://-research.microsoft.com/~conal/papers/tse-modeled-animation`.
4. Conal Elliott. Functional images. `http://research.microsoft.com/~conal/-papers/fip`, unpublished, March 2000.
5. Conal Elliott, Sigbjørn Finne, and Oege de Moor. Compiling embedded languages (extended version). Technical report, Microsoft Research, May 2000. `http://-research.microsoft.com/scripts/pubs/view.asp?TR_ID=MSR-TR-2000-52`.
6. Matteo Frigo. A fast Fourier transform compiler. In *Proceedings of the ACM SIGPLAN '99 Conference on Programming Language Design and Implementation*, pages 169–180, 1999. `http://www.acm.org/pubs/articles/proceedings/pldi/-301618/p169-frigo/p169-frigo.pdf`.
7. John Hatcliff, Torben Mogensen, and Peter Thiemann, editors. *Partial Evaluation: Practice and Theory*, volume 1706. Springer-Verlag, 1999.

8. Paul Hudak. Modular domain specific languages and tools. In P. Devanbu and J. Poulin, editors, *Proceedings: Fifth International Conference on Software Reuse*, pages 134–142. IEEE Computer Society Press, 1998.
9. Neil D. Jones, Carsten K. Gomard, and Peter Sestoft. *Partial Evaluation and Automatic Program Generation.* Prentice Hall International, International Series in Computer Science, June 1993. `http://www.dina.kvl.dk/ sestoft/pebook/pebook.html`.
10. Samuel Kamin. Standard ML as a meta-programming language. Technical report, University of Illinois at Urbana-Champaign, September 1996. `http://www-sal.cs.uiuc.edu/~kamin/pubs/ml-meta.ps`.
11. Samuel Kamin and David Hyatt. A special-purpose language for picture-drawing. In USENIX, editor, *Proceedings of the Conference on Domain-Specific Languages, October 15–17, 1997, Santa Barbara, California*, pages 297–310, 1997. `http://www-sal.cs.uiuc.edu/~kamin/fpic/doc/fpic-paper.ps`.
12. Peter J. Landin. The next 700 programming languages. *Communications of the ACM*, 9(3):157–164, March 1966. Originally presented at the Proceedings of the ACM Programming Language and Pragmatics Conference, August 8–12, 1965.
13. Daan Leijen and Erik Meijer. Domain specific embedded compilers. In *2nd Conference on Domain-Specific Languages (DSL)*, Austin TX, USA, October 1999. USENIX. `http://www.cs.uu.nl/people/daan/papers/dsec.ps`.
14. Walid Taha and Tim Sheard. MetaML and multi-stage programming with explicit annotations. *Journal of Theoretical Computer Science*, 2000. To appear. `http://www.cs.chalmers.se/~taha/publications/journal/tcs00.ps`.
15. GHC Team. The Glasgow Haskell compiler. `http://haskell.org/ghc`.
16. Todd Veldhuizen. Expression templates. *C++ Report*, 7(5):26–31, June 1995. `http://extreme.indiana.edu/~tveldhui/papers/pepm99.ps`. Reprinted in *C++ Gems*, ed. Stanley Lippman.
17. Todd Veldhuizen. C++ templates as partial evaluation. In *Workshop on Partial Evaluation and Semantics-Based Program Manipulation (PEPM'99)*. ACM Sigplan, 1999. `http://extreme.indiana.edu/~tveldhui/papers/pepm99.ps`.

A Optimization example

To illustrate the compilation techniques described in this paper, Figure 3 shows snapshots of a sample animation whose specification and supporting definitions are given in Figure 4. Note that *ImageE* is really a type *constructor*, parameterized over the "pixel" type. Visual images have type *ImageE ColorE*, while what one might call "regions" have type *ImageE BoolE*.

As a building block, *checker* is a Boolean image checker that alternates between true and false on a one-pixel checkerboard. The trick is to convert the pixel coordinates from floating point to integer (using the floor function) and test whether the sum is even or odd.

The *checkerBoard* image function takes a square size s and two colors c_1 and c_2. It chooses between the given colors, depending on whether the input point, scaled down by s falls into a true or false square of *checker*.

To finish the example, *swirlBoard* swirls a black and white checker board, using the *swirling* function defined in Section 4.

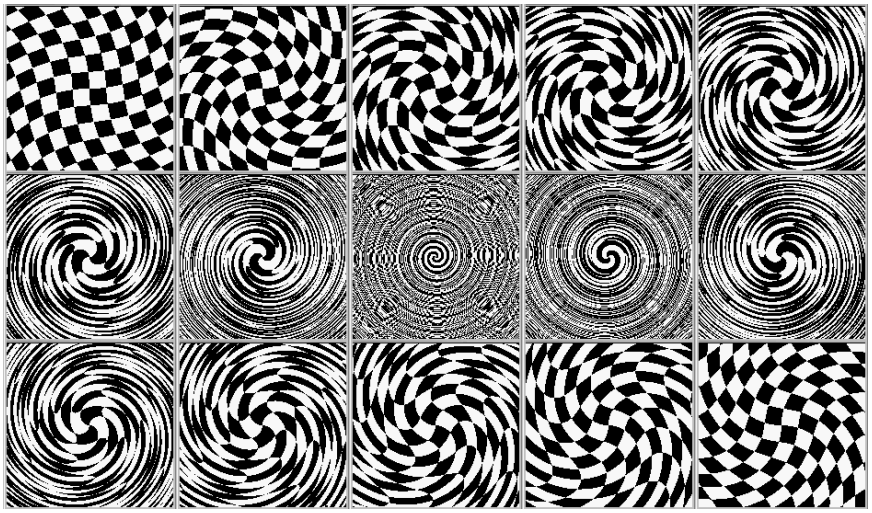

Fig. 3. snapshots of *swirlBoard*, defined in Figure 4

As a relatively simple example of compilation, Figure 5 shows the result of *display swirlBoard* after inlining definitions and performing CSE, but without optimization.

Simplification involves application of a few dozen rewrite rules, together with constant folding, if-floating, and code motion. The result for our example is shown in Figure 6.

Note how the CSE, scalar hoisting, and array promotion have produced three phases of computation. The first block is calculated once per frame of the displayed animation, the second once per line, and the third once per pixel. As an example of the potential benefit of AC-based code motion, note that in the definition of n in Figure 6, the compiler failed to hoist the expression $e * 6.28319$. The reason is simply that the products are left-associated, so this hoisting candidate is not recognized as a sub-expression.

```
swirlBoard :: TimeE → ImageE ColorE
swirlBoard t = swirl (100 * tan t) (checkerBoard 10 black white)

swirl :: Syntactic c ⇒ FloatE → ImageE c → ImageE c
swirl r im = im . swirling r    — Image swirling function

checker :: ImageE BoolE    — Unit square boolean checker board
checker = λ(x, y) → evenE (⌊x⌋ + ⌊y⌋)

checkerBoard :: FloatE → α → α → ImageE α
checkerBoard sqSize c1 c2 =
  ustretch sqSize (cond checker (const c1) (const c2))

— Some useful Pan functions:

cond :: Syntactic a ⇒ BoolE → Exp a → Exp a → Exp a
cond = lift3 ifE    — pointwise conditional
— uniform image stretch
ustretch :: Syntactic c ⇒ FloatE → ImageE c → ImageE c
ustretch s im = im . scale (1/s, 1/s)
```

Fig. 4. Definitions for Figure 3

```
λ t (panX, panY, zoom) (width, height) output →
loop height (λ j →
  loop width (λ i →
    let
      a = 2 π / (100 * sin t / cos t)
      b = −(height `div` 2)
      c = zoom * i2f (j + b) + panY
      d = c * c
      e = −(width `div` 2)
      f = zoom * i2f (i + e) + panX
      g = sqrt (f * f + d) * a
      h = sin g
      k = cos g
      m = 1 / 10
      n = m * (c * k + f * h)
      p = m * (f * k − c * h)
      q = if (⌊p⌋ + ⌊n⌋) .&. 1 == 0 then
            0
          else
            1
      r = ⌊q * 255⌋
      s = 0 <<< 8
      u = output + 4 * j * width
    in
    setInt (u + 4 * i)
           (((s .|. r) <<< 8 .|. r) <<< 8 .|. r)))
```

Fig. 5. Inlined, unoptimized code for Figure 4

```
λ t (panX, panY, zoom) (width, height) output →
let
  a = −(width `div` 2)
  b = mkArr width (λ c → zoom ∗ i2f (c + a) + panX)
  d = −(height `div` 2)
  e = recip (sin t / cos t ∗ 100.0)
in
loop height (λ j →
  let
    f = j ∗ width
    g = zoom ∗ i2f (j + d) + panY
    h = g ∗ g
  in
  loop width (λ i →
    let
      k = (f + i) ∗ 4 + output
      m = readArr b i
      n = sqrt (m ∗ m + h) ∗ e ∗ 6.28319
      p = sin n
      q = cos n
      r = g ∗ q + m ∗ p
      s = m ∗ q + g ∗ −p
    in
    if (⌊s ∗ 0.1⌋ + ⌊r ∗ 0.1⌋) .&. 1 == 0 then
      setInt k 0
    else
      setInt k 16777215))
```

Fig. 6. Optimized version of code from Figure 5