Composition-Aware Scene Optimization for Product Images

Tianqiang Liu
Princeton University
and
Jim McCann and Wilmot Li
Adobe System
and
Thomas Funkhouser
Princeton University

Increasingly, companies are creating product advertisements and catalog images using computer renderings of 3D scenes. A common goal for these companies is to create aesthetically appealing compositions that highlight objects of interest within the context of a scene. Unfortunately, this goal is challenging, not only due to the need to balance the trade-off among aesthetic principles and design constraints, but also because of the huge search space induced by possible camera parameters, object placement, material choices, etc. Previous methods have investigated only optimization of camera parameters. In this paper, we develop a tool that starts from an initial scene description and a set of high-level constraints provided by a stylist and then automatically generates an optimized scene whose 2D composition is improved. It does so by locally adjusting the 3D object transformations, surface materials, and camera parameters. The value of this tool is demonstrated in a variety of applications motivated by product catalogs, including rough layout refinement, detail image creation, home planning, cultural customization, and text inlay placement. Results of a user study indicate that our system produces images preferable for product advertisement compared to previous approaches.

Categories and Subject Descriptors:

■

1. INTRODUCTION

A growth application for computer graphics is creation of images for product advertisements and catalogs. As photorealistic rendering algorithms have improved, it has become practical to synthesize images that are indistinguishable from photographs for many types of scenes commonly found in product advertisements (e.g., kitchens, bathrooms, living rooms, etc.). As a result, several furniture and home goods companies are beginning to create product images for their catalogs by rendering 3D models rather than photographing physical objects [Southern 2012]. For example, IKEA has reported that 25% of scenes shown in its 2013 catalog were rendered from 3D models, and that they expect the percentage to increase dramatically in upcoming years [Souppouris 2012].

The advantages of creating catalog images from 3D models are numerous: e.g., less expense, less space, more flexibility, and higher customizability [Enthed 2012]. If libraries with accurate models of 3D surfaces, materials, and lights are available for most objects in scenes of interest, then it is possible to create advertising images with off-the-shelf 3D scene modeling tools and photorealistic rendering software, which are relatively cheap, require little physical space, and allow incremental edits. Virtual photography avoids the need to build physical sets in large photo studios, store physical objects in large warehouses, and/or schedule actors, stylists, and photographers to meet for photo shoots. Moreover, it provides increased opportunities for customization – for example, to adapt the selections, colors, and placements of objects based on demographics (or even individual preferences), to produce multiple images of the same scene with different objects of interest, to adapt scene composition to the resolution and aspect ratio of the display device, and/or to adapt the size and placement of text labels depending on language. Companies currently achieve these goals (partially) with manual effort at great expense.

Despite these advantages, producing good advertisement images from 3D models is difficult: it requires optimizing large numbers of often-competing design constraints. First and foremost, the images must highlight certain "objects of interest" within a scene (the ones being advertised), which implies constraints on the positions, sizes, color contrasts, and visibilities of those objects (e.g., Figure 7). Second, they must include scene context to provide cues about where the objects of interest might be found, how they might be used, who might be using them, etc., requiring contextual objects to be arranged with plausible 3D positions, sizes, and support relationships. Finally, they should follow well-established rules of composition and aesthetics, which further dictate the screen-space positions, balance, and colors of objects in the scene. Typically, 3D artists spend multiple days tweaking the positions, orientations, and materials of objects to optimize the composition of each new image *after* an initial approximate scene layout has been chosen. This effort must be duplicated for every variant of the image, which may be adapted for different cultures (e.g., IKEA makes 62 versions for 43 countries), different display devices (e.g., iPhone, PC, print, etc.), and/or different objects of interest (for zoomed views of certain objects).

The goal of our work is to develop a tool to assist in the creation of product images by optimizing 2D compositions of 3D scenes by adjusting object transformations, materials, and cameras. Our tool starts with an approximate scene description provided by a stylist[1] that includes which objects should appear in the scene, which objects rest upon which other objects, and which materials can be used for which objects, plus an initial plausible configuration for object transformations, surface materials, lighting parameters, and (optionally) camera views. Our tool then optimizes the scene description to improve its estimated composition and aesthetic quali-

[1] A person who designs scenes for product images

ties, optionally satisfying additional design constraints common in product image formation (e.g., highlight these objects of interest, use a landscape image format, leave space for a text box, etc.).

To achieve this goal, we define an energy function that encodes a large variety of image composition rules and user-provided constraints. We then take object positions and orientations, surface materials, and camera parameters as free variables and optimize the scene description by minimizing the energy function. Our system is able to refine scenes automatically to produce better compositions for many aspects of product catalog production.

We have made the following contributions in this work. First, we introduce a scene optimization tool to assist in the creation of product images. Second, we propose a method for optimizing 2D compositions by simultaneously manipulating object positions, materials, and cameras. To the best of our knowledge, we are the first to have combined all these degrees of freedom in a single optimization. Third, we demonstrate the utility of our tool for reducing the human effort to create scenes providing good image compositions for a variety of use-cases.

2. RELATED WORK

Our work draws upon previous work in image composition and aesthetics, image analysis and optimization, virtual camera control, and automatic scene synthesis.

Image composition and aesthetics: Our work is inspired by composition “rules” that have been established to guide photographers and graphics designers towards better scene compositions and aesthetics [Arnheim 1988][Bethers 1956][Clifton 1973] [Grill T. 1990][Krages 2005][Martinez and Block 1988] [Taylor 1938]. Well-known examples include visual balance, diagonal dominance, and color contrast. Additional guidelines recommend ways to choose the position, size, aspect ratio, coincidence, visibility, and background for important objects – for example, the “rule of thirds” suggests that important objects should appear at the intersections of one-third lines.

Image analysis and optimization: Several papers have used these guidelines to quantify and enhance the compositional and aesthetic quality of images. For example, [Datta et al. 2006] learned a regression model to predict the aesthetic quality of an photo from computed image features, including saturation, hue, aspect ratio, rule of thirds, etc. [Kao et al. 2008] developed a model of image composition quality based on horizon balance, intensity balance, locations of regions-of-interests, and merger avoidance and then used the model to automatically rotate and crop images. [Lok et al. 2004] provided an algorithm to move objects in a 2D composition to improve visual balance. [Bhattacharya et al. 2010] provided ways to adjust the locations of salient foreground regions and the image aspect ratio to improve images according to the rule of thirds and visual balance. [Wong and Low 2011] adjust pixel luminance, saturation, and sharpness to improve the salience of important regions. [Cohen-Or et al. 2006] adjust pixel colors to improve color harmony. [Liu et al. 2010; Jin et al. 2012] optimize visual balance, diagonal dominance, rule of thirds, and salient-region sizes by cropping, warping, and retargeting images. All these methods operate only on edits to 2D images – they do not optimize 3D scene parameters, such as cameras, materials, and/or object transformations, as our system does.

Camera optimization: Several methods have incorporated principles of image aesthetics and composition in optimization algorithms for camera control in 3D rendering systems [Christie et al. 2008; Gooch et al. 2001]. For example, [Olivier et al. 1999] optimized camera parameters to match user-prescribed screen-space positions, sizes, and spatial relationships of rendered objects. [Abdullah et al. 2011] extended this approach to also consider visual balance, diagonal dominance, rule of thirds, and depth of field. [Bares et al. 2000; Bares 2006] added consideration for object visibility depth order in automatic camera control for virtual environments. While these papers provide motivation for our work, they consider only camera control – we additionally optimize object transformations and surface materials, which can significantly improve image compositions, but require solving a more difficult optimization problem.

Scene optimization: Several recent papers have proposed methods for automatically placing objects in scenes. For example, [Yu et al. 2011; Fisher et al. 2012] and [Merrell et al. 2011] proposed systems to produce plausible furniture layouts based on examples and design guidelines, respectively. These methods focus on scene plausibility without concern for any particular camera viewpoint and/or image composition principles. As a result, they produce scenes that may not support generation of aesthetic images from any camera viewpoint. Our framework works synergistically with these systems: we employ plausibility constraints similar to [Merrell et al. 2011] to ensure plausible object arrangements as we optimize image compositions in our system.

Camera and scene optimization: In their pioneering work on “Through the Lens Camera Control,” [Gleicher and Witkin 1992] discussed the possibility of simultaneously adjusting camera and object parameters to satisfy screen-space constraints in rendered images. However, they discussed only low-level constraints (e.g., keep the object at a prescribed screen-space position) and provided no implementation or investigation of the idea – they left the topic as a suggestion for future work.

Joint optimization of camera parameters and scene content has been considered in some domain-specific applications. For example, [Bell et al. 2001] proposed an approach that maintains visual constraints of objects in screen space to support interactive update of labels in a virtual and augmented reality system as the user changes the camera viewpoint. [He et al. 1996] developed a system that automatically controls a camera for capturing actors in virtual environments, subtly changing the positions of virtual actors to achieve better compositions. However, these solutions were domain-specific.

We believe that ours is the first system to optimize aesthetics and composition of rendered images with simultaneous control over camera parameters, object transformations, and surface materials. We investigate this optimization problem for the novel application of image synthesis for product catalogs.

3. OVERVIEW

The core of our work is a method for optimizing 2D compositions of rendered 3D scenes by adjusting camera parameters, object transformations, and surface materials. The required input to our system is a 3D scene graph that includes *polygonal models* for all objects in the scene, a *hierarchy of support relationships* between objects (e.g., floor supports table, table supports vase), a *location and orientation* for every object, a list of possible definitions for every surface material, a list of light sources, and the *aspect ratio* of the output display device. Our system also requires a list of *focus objects* $\mathbf{O_F}$ that specify the main objects of interest in the scene. Optional inputs include a list of *context objects* $\mathbf{O_C}$ that should remain visible for context, and a list of rules that constrain the *3D spatial relationships* between objects of specific types (e.g., picture frames on walls should not be rotated) Furthermore, in some applications, users may provide possible initial values for the camera parameters. All input scene parameters can be approximate, since they will be refined by the optimization.

From this input, our system optimizes the scene description to generate a set of rendered images. In particular, our method optimizes the following scene parameters (plus other application-specific variables described in Section 6), with the degrees of freedom listed in parentheses:

—***Camera (6):*** position (3), direction (2), and field of view (1) (camera roll is constrained to be zero).
—***Object transformations (3 per object):*** position of the object centroid on its support surface (2) and rotation of the object around the normal of its support surface (1).
—***Materials (1 per material):*** choice of a material/texture definition amongst a list of possible candidates.

Our system optimizes these parameters according to an energy function that accounts for image composition, aesthetic principles, and object focus, while maintaining 3D spatial constraints (e.g., no collisions, specified spatial relationships) and 2D screen space layout constraints (e.g., preferred locations for objects of interest). In most applications (described in detail in Section 6) we envision that stylists will either use our automatically generated images directly or provide additional inputs/refinements to guide further optimization.

The following sections describe our energy function and optimization procedure in detail.

4. ENERGY FUNCTION

Our energy function estimates how effectively an image advertises the product(s) it depicts.

Of course, the quality of an advertising image depends on many complicated factors that are difficult for a computer to evaluate, including what type of room is depicted in the scene, how the layout of the room reflect who lives there, how objects near the product(s) reflect people's preferences, what season it is, etc. So, we leave those decisions to a professional stylist, who provides an initial scene layout, with constraints about which objects are on which surfaces, which object(s) are the subject of focus, which object(s) are important for context, which objects have important spatial relationships in 3D, and which objects/materials can possibly be switched with one another. Then, our system must only investigate the subspace of scenes that satisfy this high-level scene description, and the output is constrained to reflect the stylist's specifications.

Even within that subspace, there are many competing factors that affect the "quality" of the resulting image, including how well it satisfies aesthetic and composition principles, how effectively it highlights the focus objects, how well it matches the stylists specifications, etc. Our goal here is to define an energy function that reflects these factors and can be optimized efficiently.

F	The 2D image frame (viewport)
$\mathbf{O}$	The set of all objects
$\mathbf{O_F}$	The set of focus objects
$\mathbf{O_C}$	The set of context objects
$\mathbf{O_B}$	The set of background objects
O_i	An object in set $\mathbf{O}$
P_i	O_i's projection into screen space
V_i	the part of P_i visible to the camera
$\mathcal{V}(\cdot)$	Volume in scene space
$\mathcal{A}(\cdot)$	Area in screen space
$\mathcal{B}(\cdot)$	Boundary contour in screen space
$\mathcal{F}(\cdot)$	Projection onto XY plane in scene space
$\mathcal{R}(\cdot)$	Diagonal radius
$\mathcal{C}_2(\cdot)/\mathcal{C}_3(\cdot)$	Centroid in screen space/scene space
$d_2(\cdot,\cdot)/d_3(\cdot,\cdot)$	Euclidean distance in screen space/scene space
$c(\cdot,\cdot)$	Color difference in L*ab space

Table I. : *Symbols used in the energy function definition.*

Several previous papers have proposed energy functions to evaluate aesthetics and composition of images (see Section 2). Our contribution here is not the introduction of this idea, but the specialization of it to images for product advertisements.

Our approach is to interview professionals responsible for creating scenes for popular product catalogs and to devise an energy function quantifying the factors they list as important. Although the number of people suitable for such interviews is quite small (e.g., Pottery Barn might use a dozen stylists for a catalog), we were able to interview one professional stylist who works with several major companies and one technical person who works closely with stylists at a major furniture company. They first described the process used to create images for product catalogs and then explained the factors they use to create good image compositions. Summarizing briefly, the first tasks are to select products to highlight (e.g., a bedroom set), choose a space to display them (e.g., a bedroom), imagine who is using that space (e.g., a teenage boy), select contextual objects and arrange them roughly in the scene (e.g., a football on the dresser), select a rough camera viewpoint (e.g., from a person standing in the doorway), and then refine the image composition to highlight the selected products within the context of the scene. This last step is the most time consuming and the focus of our work.

During our interviews, the following factors were identified as most important for refining image compositions: 1) object placement within the 2D frame, 2) object saliency within in the 2D frame, 3) object relationships within the 3D scene, 4) camera placement, 5) image composition, and 6) consistency with the initial scene layout. We encode these factors in an energy function with terms that represent relevant compositional principles and desired object relationships:

$$\begin{aligned} E = {} & E_{rt} + E_{ce} + E_{cl} + E_{sa} + \\ & E_{sr} + E_{co} + E_{su} + E_{cv} + \\ & E_{tv} + E_{vb} + E_{cc} + E_{ir} \end{aligned} \tag{1}$$

Our system minimizes this energy function to help stylists refine rough scene layouts to produce good image compositions. The rest of this section defines the energy terms in detail. For reference, Table I. lists all of the variables used in the energy function.

4.1 Object placement within the 2D frame

For product advertisements, some of the most significant factors affecting image quality are the positions of focus objects, which we encode with the following terms.

—**Rule of thirds.** In general, focus objects should align with vertical or horizontal lines that divide the viewport into thirds and/or be centered at the intersections formed by them [Banerjee and Evans 2004][Bares 2006][Byers et al. 2004] [Datta et al. 2006][Gooch et al. 2001][Liu et al. 2010][Ward 2003]. More specifically, stylists suggest that horizontal focus objects (e.g., sofa) should align with horizontal third lines and vertical focus objects (e.g., floor lamp) should align with vertical third lines. We encode these preferences with

$$E_{rt} = \frac{w_{rt}}{\mathcal{R}(F)^2} \sum_{O_i \in \mathbf{O_F}} \left(\frac{w^2}{h^2} d_h(\mathcal{C}_2(P_i))^2 + \frac{h^2}{w^2} d_v(\mathcal{C}_2(P_i)) \right)^2$$

where w and h are the width and height of the bounding box of P_i, respectively, and d_h and d_v stand for minimum distance to the closest horizontal and vertical third lines respectively.

—**Centeredness.** For some product images (e.g., zoomed views of single objects), focus objects should appear in the center of the image [Arnheim 1988]. We encourage centeredness with the following term:

$$E_{ce} = \frac{w_{ce}}{\mathcal{R}(F)^2} \sum_{O_i \in \mathbf{O_F}} d_2\left(\mathcal{C}_2(F), \mathcal{C}_2(P_i)\right)^2$$

—**Clearance.** Since other objects should not compete with the focus objects in the composition, we introduce a term to penalize the objects that are close to focus objects.

For an object O_i, we consider its signed distance to the bounding circle of a focus object in screen space (a negative distance means O_i is inside the bounding circle), and normalize the distance by the radius of the bounding circle in order to avoid favoring small objects. We use $r(O_i)$ to denote the minimum of all the distances,

$$r(O_i) = \min_{O_j \in \mathbf{O_F}} \frac{d_2(\mathcal{C}_2(P_i), \mathcal{C}_2(P_j)) - \mathcal{R}(P_j)}{\mathcal{R}(P_j)}$$

The clearance term is then defined as,

$$E_{cl} = \frac{w_{cl}}{N} \sum_{O_i \in \mathbf{O_F}} \left(e^{-\max\{0, r(O_i)\}^2} + \max\{0, -r(O_i)\}^2 \right)$$

where $N = |\mathbf{O} \setminus \mathbf{O_F}|$. The $\max\{0, -r(O_i)\}^2$ term handles the case where O_i is inside a focus object's bounding circle.

4.2 Object saliency within the 2D frame

Visibility and object size contribute to perceived relative importance [Bares et al. 2000][Bares 2006][Bell et al. 2001][Olivier et al. 1999]. An object is perceived as less important if it is partially occluded by another object, partially clipped by the frame, or if it takes up a small amount of screen space. We encode the saliency of focus and context objects with the following terms.

—**Object size.** To quantify the effect of object size on saliency, we introduce the following term:

$$\mathscr{S}_r(O_i) = \max\left\{0, r - \frac{\mathcal{A}(V_i)}{\mathcal{A}(F)}\right\}^2$$

where $A(V_i)$ is the area of the visible part V_i of O_i, r is the minimum required size of O_i. We describe how we choose r later.

—**Visibility.** To quantify the effect of visibility on saliency, we use the following term:

$$\mathscr{V}_r(O_i) = \max\left\{0, r - \frac{\mathcal{A}(V_i)}{\mathcal{A}(P_i)}\right\}^2 + \mathscr{D}(O_i)$$

where $A(P_i)$ is the total area of the object projection P_i assuming no occlusions by other objects or clipping by the image frame. $A(V_i)$ and $A(P_i)$ can be computed efficiently using hardware occlusion queries [Govindaraju et al. 2003]. We introduce the term $\mathscr{D}(O_i)$ to encourage objects outside the frame to move towards the center of the frame:

$$\mathscr{D}(O_i) = \begin{cases} d_2(\mathcal{C}_2(F), \mathcal{C}_2(P_i))^2, & \text{if } \mathcal{A}(V_i) = 0 \\ 0, & \text{else} \end{cases}$$

Visibility and object size are both essential for focus objects. However, for context objects, we observe that there are two scenarios. If the context object is small compared to the focus object, visibility is important while its absolute size in the viewport is not (e.g. items on the dining table in Figure 4 left). On the other hand, if the context object is largely occluded, it must maintain some minimum size in the composition. To handle these cases, we compute both energies $\mathscr{V}_r$ and $\mathscr{S}_r$ for each context object and select the minimum. Thus, the complete form of our object saliency energy term is

$$E_{sa} = w_{sf} \sum_{O_i \in \mathbf{O_F}} \left(\mathscr{V}_{v_f}(O_i) + \mathscr{S}_{s_f}(O_i) \right) + \\ w_{sc} \sum_{O_i \in \mathbf{O_C}} \min\left\{ \mathscr{V}_{v_c}(O_i), \mathscr{S}_{s_c}(O_i) \right\}$$

There are four parameters in this term. In all experiments, we set $v_f = 100\%, s_f = 10\%, v_c = 80\%, s_c = 5\%$, which means a focus object is required to be fully visible *and* cover 10% of the viewport; a context object is required to be at least 80% visible *or* cover at least 5% of the viewport.

4.3 Object relationships within the 3D scene

3D spatial relationships between scene objects can influence the plausibility of a scene. For example, dining chairs usually remain near the dining table. Physical laws (e.g., collisions, gravity, etc.) also impose constraints on object positions. We introduce several terms to enforce these constraints.

—**Locked variables.** Assuming that the input scene configuration has a plausible 3D arrangement of objects, we provide a simple mechanism to allow a stylist to specify which variables can and cannot change during the optimization. For example, the stylist can specify that a picture frame cannot rotate, or that a dresser can only translate in X (along an axis-aligned wall), or that camera pitch cannot change. These are implemented as hard constraints – i.e., they reduce the degrees of freedom in the optimization.

—**Semantic relationships.** We also provide mechanisms to specify which spatial relationships between objects should be maintained from the initial scene, and which can be optimized. For example, he/she can tell the system to keep the chair in front of the desk, or keep the mouse to the right of the keyboard. These spatial relationships are specified in a short text file (with 10-15 lines on average), implemented as soft constraints using the method similar to [Bukowski and Séquin 1995; Merrell et al. 2011]:

$$E_{sr} = w_{sr} \sum_{\{O_i, O_j\} \in \mathbf{C}} \sigma_{i,j} d_3(\mathcal{C}_3(O_i), T_i^{-1}(\mathcal{C}_3(O_j)))^2$$

where $\mathbf{C}$ is a set of constrained object pairs, and T_i^{-1} is the initial transformation from the scene space into the local coordinate frame of object O_i, and $\sigma_{i,j}$ controls how much the spatial relationship can change. $\sigma_{i,j}$ can be set by the stylist empirically in practice, and we set $\sigma_{i,j} = 1$ by default in all of the results shown in this paper.

—**Collision relationships.** Object inter-penetrations should be avoided to improve the physical plausibility of the scene. In our implementation, we only penalize the collisions that are visible in the composition, and treat collisions as soft constraints with a penalty term based on the relative volume of the object intersections so that the energy function has a gradient near collision transitions:

$$E_{co} = w_{co} \sum_{\substack{O_i \in \mathbf{O} \\ \mathcal{A}(V_i) > 0}} \sum_{\substack{O_j \in \mathbf{O} \\ \mathcal{A}(V_j) > 0}} \frac{\mathcal{V}(O_i \cap O_j)}{\mathcal{V}(O_i)}$$

To compute intersections, we project every object to the XY (i.e., ground) plane and store the projection as a bitmap along with the min and max Z values of the object. We detect intersections by checking for overlaps in both the XY projections and Z ranges of objects.

—**Support relationships.** The support relationships provided by the stylist must be maintained during the optimization (e.g., object A is *on* object B or attached to the side of object B). Thus, we penalize placement of an object off its support object by measuring the fraction of its projected area outside its support surface [Fisher et al. 2012]:

$$E_{su} = w_{su} \sum_{O_i \in \mathbf{O}} \left(1 - \frac{\mathcal{A}(\mathcal{F}(O_i) \cap \mathcal{F}(S_i))}{\mathcal{A}(\mathcal{F}(O_i))}\right)^2$$

where S_i is the object supporting object O_i.

4.4 Camera placement.

Product images generally depict scenes from viewpoints that are "natural" for people. We introduce two terms that capture the notion of natural viewpoints.

—**Canonical views.** In most product images with one focus object, stylists favor canonical views of the object class. In our experiment, we manually defined a set of 1-4 canonical view directions for each object *class* [Blanz et al. 1999; Gooch et al. 2001], and then deviations from them are measured as:

$$E_{cv} = w_{cv} \min |(\theta, \phi), (\hat{\theta}_i, \hat{\phi}_i)|^2$$

where (θ, ϕ) is the view direction of the camera, and $(\hat{\theta}_i, \hat{\phi}_i)$ is the closest canonical view direction for the object class.

—**Typical views.** Product images that depict large scenes with multiple focus objects often use camera viewpoints that match how a human would typically see the scene. We encode the notion of a typical view as

$$E_{tv} = w_{ch}(h - h_0)^2 + w_{ca}\phi^2$$

where h is the height of camera off the floor, $h_0 = 5ft$ is the typical height of a human eye, and ϕ is the pitch of the camera (where 0 is horizontal). This term penalizes viewpoints that deviate from a typical human eye height and tilt the camera upwards/downwards.

4.5 Image composition

Several well-established composition guidelines are used by stylists to create aesthetically pleasing images. We have included several in our system.

—**Visual balance.** Images whose "center of mass" is close to the center of the image frame generally have better aesthetics [Arnheim 1988][Liu et al. 2010][Lok et al. 2004]. So, we add the following term, which measures the distance between the center of the frame and the center of mass of 2D object projections:

$$E_{vb} = \frac{w_{vb}}{\mathcal{R}(F)^2} d_2 \left(\mathcal{C}_2(F), \frac{\sum \mathcal{C}_2(P_i)\mathcal{A}(P_i)}{\sum \mathcal{A}(P_i)}\right)^2$$

—**Color contrast.** Greater color contrast at object contours can help a viewer understand boundaries between shapes in a scene [Kowalski et al. 2001][Wong and Low 2011]. To encourage this effect, we add the following energy term for focus objects:

$$E_{cc} = \frac{w_{cc}}{\mathcal{R}(F)^2} \sum_{O_i \in \mathbf{O_F}} \sum_{p \in \mathcal{B}(V_i)} \frac{1}{\left(\mathrm{avg}_{q \in N(p) \setminus V_i} c(p, q)\right)^2 + \epsilon}$$

where $N(p) \setminus V_i$ denotes the neighborhood of pixel p, excluding the visible pixels in V_i. To evaluate the color contrast, we extract the contour of each object in an image rendered at 1/4-th resolution, which accelerates the computation.

4.6 Inertia

Finally, we add a regularization term that encourages small changes to the scene:

$$E_{ir} = w_{ir} \sum_{O_i \in \mathbf{O}} \left(\frac{x_i^2}{\sigma_t^2} + \frac{y_i^2}{\sigma_t^2} + \frac{\theta_i^2}{\sigma_r^2}\right) + \sum_{i=0}^{5} \frac{c_i^2}{\sigma_c[i]^2}$$

where (x_i, y_i, θ_i) describe the translation and rotation of object O_i, respectively, and c_i describe the change to camera parameters, with $\sigma_t = 0.5, \sigma_r = 0.5, \sigma_c = [0.17, 0.17, 20, 20, 20, 0.17]$ controlling the flexibility of object movement and camera manipulation.

These energy terms are weighted by coefficients that adjust for scale differences and control their effects on the final results. By default, the weights are set to w_{rt} = 10000, w_{ce} = 10000, w_{cl} = 500, w_{sf} = 10000, w_{sc} = 500, w_{sr} = 100, w_{co} = 10000, w_{su} = 10000, w_{cv} = 10000, w_{ch} = 10000, w_{ca} = 10000, w_{vb} = 20000, w_{cc} = 1.0, and w_{ir} = 1.0. These weightings were determined empirically and are kept the same for all examples in this paper, except that $w_{ce} = w_{cv} = 0$ for overview images of scenes (e.g., session A in Figure 2, Figure 7, 8 and 9), and $w_{rt} = w_{ch} = w_{ca} = 0$ for zoomed-in images of specific objects (e.g., session B in Figure 2, Figure 3 and 5). *It is not expected that a user has to tweak these weights to get good results for specific scenes.*

The number of terms in our energy function reflects the inherent complexity of composing a good picture, and we found that all of the terms were useful for improving compositions in different situations. Figure 1 shows the effect of each energy term.

5. OPTIMIZATION

Our optimization procedure searches for the scene description (camera, object placements, and surface materials) that minimizes the energy function described in the previous section (Equation 1).

This is a difficult optimization problem because: 1) there are many free variables (six for the camera, three for each object transformation, one for each surface with multiple candidate materials, etc.); 2) some of the variables are continuous (camera and object transformations) while others are discrete (surface materials); and 3) the energy function is highly non-convex, with strong dependencies between multiple variables (e.g., camera and object movements). As a result, we can only hope to find a good local minimum.

Our optimization procedure interleaves optimization of discrete and continuous variables in alternate steps with an EM-style iterative algorithm. Within each iteration, it first optimizes the discrete choices of materials with camera parameters and objects transformations fixed, and then it optimizes the continuous camera parameters and object transformations with the materials fixed. The iterations terminate when neither step changes the scene significantly in the same iteration.

5.1 Discrete optimization

We use a discrete steepest-descent algorithm to optimize materials during the E-step. The input to the algorithm is a scene and a list of candidate definitions for each surface material (e.g., surfaces with material 17 may be either white painted wood, birch wood finish, etc.), and the output is ideally a selection of one candidate definition for each material that minimizes the energy function (note that the color contrast term, E_{cc}, is the only one affected by material switches).

The algorithm first builds a list of visible objects with multiple candidate materials. Then, it iteratively optimizes the materials for each such object one-by-one in decreasing order of their current contributions to E_{cc}. For each visit of an object, the algorithm selects the material switch that produces the lowest E_{cc}. The algorithm stops when no material switches are possible to lower the energy further, which usually occurs within 2-3 iterations through all objects.

5.2 Continuous optimization

We use a continuous steepest-descent algorithm to optimize camera parameters and object transformations during the M-step. The following paragraphs describe how the direction and magnitude of each step is computed.

Since the energy function contains terms whose partial derivatives are difficult to compute analytically (e.g., visibility), we compute the derivative of the energy with respect to each free variable via a centered difference approximation. Of course, a brute force implementation of centered differences for each variable would be extremely slow: a typical scene has approximately 150 free variables (3 for each of $\sim$50 object transformations plus 6 camera parameters), and thus the energy would have to be computed 300 times for each steepest descent step. Instead, we keep estimates for all partial derivatives and re-estimate only a subset after most steps. Specifically, every k steps, we estimate partial derivatives for all variables, except ones for transformations of untethered objects outside the view frustum, and make a move along the direction of steepest descent determined by all partial derivatives. We also build a list of objects that have non-zero partial derivatives T. Then, during the intervening steps, we re-estimate partial derivatives for the camera parameters and k randomly selected objects from T and make a steepest descent move based on these derivatives. We choose $k = \sqrt{|T|}$, which provides a nice trade-off between efficiency and accuracy, leveraging the fact that fewer objects have significant effect on the energy as the optimization converges.

To compute the magnitude of each steepest descent step, we conduct a line search along the direction of the estimated derivative. Specifically, we compute a minimum step length, check the energy at 10 steps increasing exponentially in length (by 1.25 times at each step), and then take the best step. To compute the minimum step length, we project the size of the minimum allowed step size in each dimension onto the direction of the derivative and take the minimum.

5.3 Timing

The full optimization procedure takes approximately 20 minutes (for 60 iterations) for the most complex examples in this paper. The discrete optimization step is usually very fast ($<$ 10 seconds), since there are relatively few ($\sim$10) candidate materials in most scenes. The continuous optimizations are slower, since there are many possible object transformations in most scenes ($\sim$ 60 objects per scene in our examples) and computing partial derivatives for each transformation variable requires rendering the scene multiple times. In our experience, computing partial derivatives takes $\sim$ 90 seconds for all variables (every k steps), but only $\sim$ 10 seconds for our randomly chosen subsets (intervening steps), at no observed accuracy difference. Performing the line search takes only 3 seconds for each step. All times are reported for a 2660 MHz Intel Core i7 processor with 8 GB of memory.

6. APPLICATIONS

In this section, we describe several applications of our scene optimization framework. These applications were chosen based on the suggestions of experts who currently create product images at large furniture companies.

6.1 Refining rough compositions

The primary application of our system is to facilitate the refinement stage of digital catalog image creation. Given a set of focus objects and a rough composition as initialization, we can apply the optimization procedure described in the previous section to automatically adjust the camera, object positions, and materials.

To evaluate whether our system can assist this application, we ran an informal experiment in which we first went through the full pro-

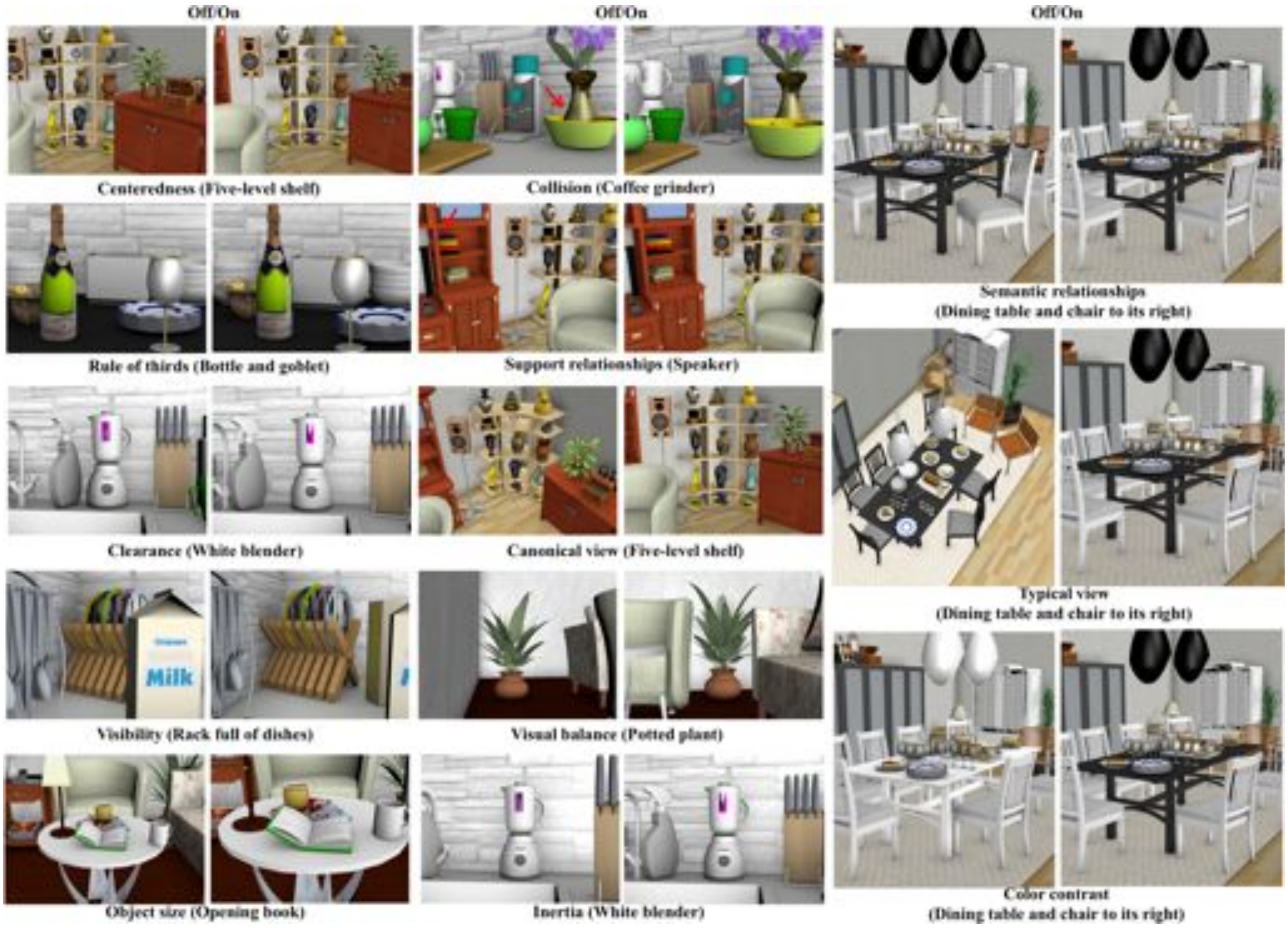

Off/On
Off/On
Off/On
Centeredness (Five-level shelf)
Collision (Coffee grinder)
Semantic relationships (Dining table and chair to its right)
Rule of thirds (Bottle and goblet)
Support relationships (Speaker)
Clearance (White blender)
Canonical view (Five-level shelf)
Typical view (Dining table and chair to its right)
Visibility (Rack full of dishes)
Visual balance (Potted plant)
Object size (Opening book)
Inertia (White blender)
Color contrast (Dining table and chair to its right)
Milk

Fig. 1: *Effects of disabling energy function terms. For each energy term, we compare the result with the term disabled (left) to our result (right). The focus object(s) is specified in the parentheses.*

cess of creating and refining a scene for a product image using an interactive modeling tool, and then we investigated how our tool could have helped during the modeling process. A trained user (a graduate student who has taken several composition classes before) was asked to create a 3D scene motivated by an image highlighting a dining room table and chair in the IKEA catalog (Figure 4 left), which she could refer to as she modeled. During the session, she started with a set of objects, candidate materials, and a random camera viewpoint (Figure 2A0), and then edited the scene interactively to achieve the final result shown in Figure 4 left. We then asked her to repeat the interactive refinement process to recreate the composition in Figure 4 right, which highlights the three goblets on the dining table. The experiment was performed exactly once with no feedback from the system regarding composition quality.

During these interactive sessions, we logged a "snapshot" scene file every 10 seconds representing the user's progress (several examples are shown in the top row of Figure 2). After the session was finished, we used the snapshot scenes to: 1) analyze whether our energy function explains changes made interactively by the user, and 2) to study at what point in the modeling process our optimization procedure could have been used to assist the user by refining the scene automatically.

The blue curve in the plot at the bottom of Figure 2 shows the value of our energy function for each snapshot of the user's interactive session. Note that for each session, the curve reveals two phases: a period of "large-scale layout" when the scene energy goes up and down ($A0 \rightarrow A47$ and $B0 \rightarrow B11$), followed by a period of "fine-scale refinements" where the energy decreases almost steadily ($A47 \rightarrow A92$ and $B11 \rightarrow B39$). This behavior suggests that the energy function correctly captures image quality differences of improvements made by the user.

The second row of images in Figure 2 shows the results of running our optimization procedure on each of the snapshot scenes shown in the top row, and the red dots in the plot below show the energy function of the optimized results (connected by a green curve). Note that the optimized results of the snapshots (bottom row) captured in the latter half of each user session (A47-A92 and B11-B39) are qualitatively similar to the final scene created by the user (top-right image), and their corresponding energy function values are comparable, or even less. These results suggest that a half of the time the

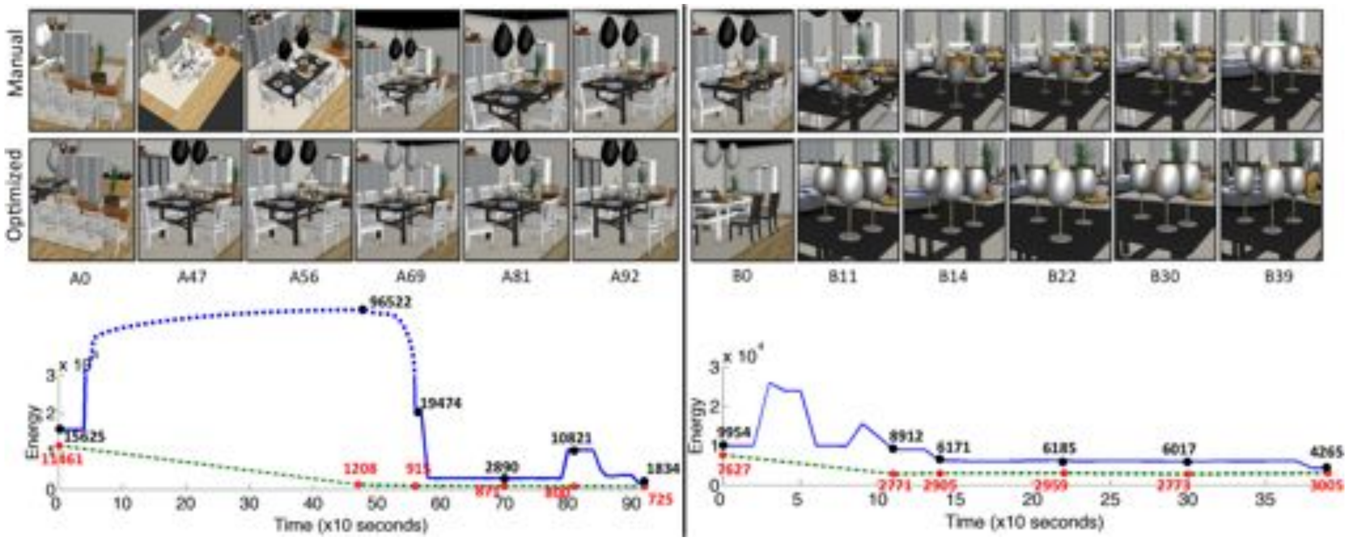
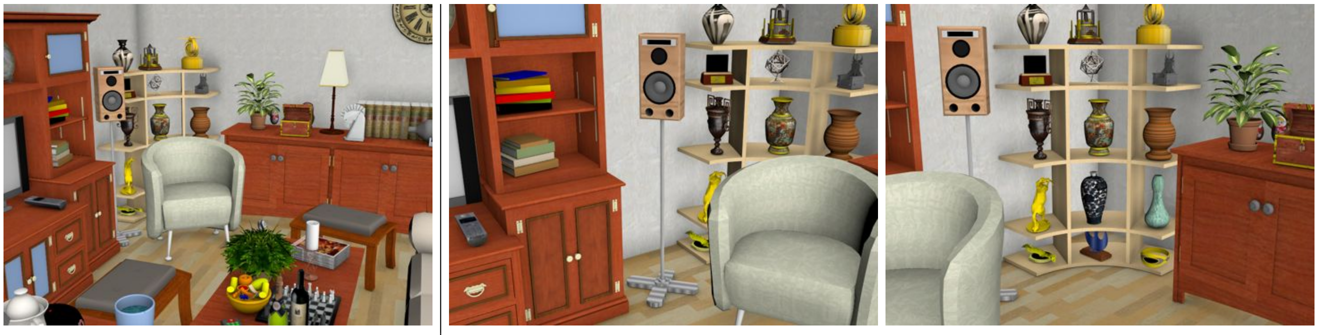


Fig. 2: *Snapshots of an interactive session (top row) and the results of refining them by our optimization tool (second row). In the first session (left), the user's goal is to achieve the composition in Figure 4 left, while in the second session (right), her goal is to achieve Figure 4 right. The plots on the bottom show the evaluation of these scenes using our energy function, with the blue representing the energy of interactive snapshots and the red points representing our optimized results. Note that the dotted section of the lefthand blue curve has been compressed to save space.*



Overview

Detail image of speaker

Detail image of shelf

Fig. 3: *Detail images generated from overview. From an overview image of a living room (a), we automatically generate detail images that highlight the speaker (b) and shelf (c). Notice how the chair moves to the right in (b) and to the left in (c) to provide an unobstructed view of the focus object (results without moving objects can be found in Figure 10).*

user spent on the scene refinement could have been off-loaded to the computer. Also, notice that the latter half of each green dotted curve is flat indicating that the optimization procedure is robust to different starting conditions created by the user.

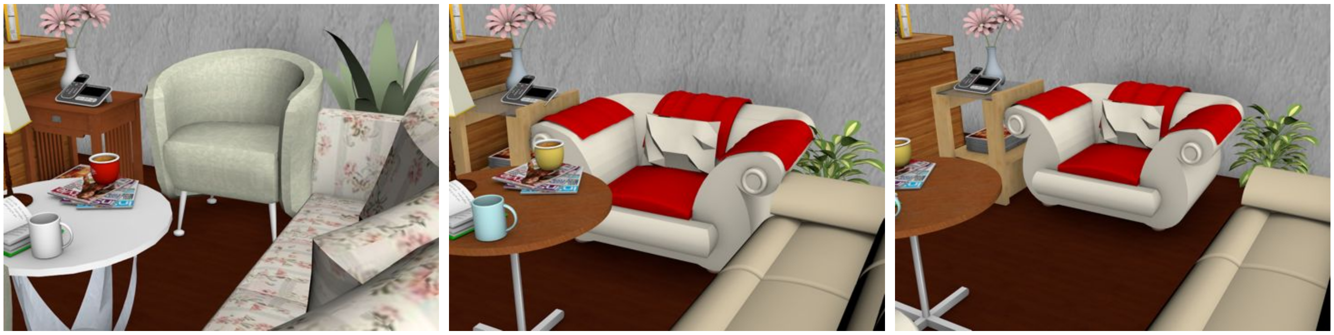
6.2 Generating detail images from an overview

In many cases, catalogs provide an overview image that shows how various objects can fit together in a room, and then one or more detail images that focus on individual products of interest. For example, the IKEA catalog image in Figure 4 includes an overview of a dining room (left) with a detail image advertising wine glasses (right).

Detail images are almost never simply cropped and zoomed-in versions of the overview image. Stylists typically choose different viewpoints and move objects slightly in order to highlight the shape and relevant features of the focus object. For example, in the right-hand detail image of Figure 4, several objects on the table have been moved to create an appropriate backdrop for the wine glasses.

To reduce this effort, stylists can use our system to automatically create detail images. For each detail object O_d, our optimization framework initializes all object positions to the arrangement in the overview image and generates a set of candidate detail images using each canonical view of the detail object as a different starting point for the camera. By default, we choose the candidate image with the lowest energy as the result.

For a given O_d and canonical view, the optimization works as follows. We set O_d as the only focus object, $\mathbf{O_F} = \{O_d\}$. To determine context objects $\mathbf{O_C}$ automatically, we set a threshold σ as 20% of the bounding box diagonal O_d, and all objects within



a) Original composition b) Composition after replacing objects c) Optimized composition

Fig. 5: *Object replacement. From the original composition (a) the chair, side table and coffee table are replaced (b). Our optimization eliminates collisions and produces a better composition for these objects (c).*

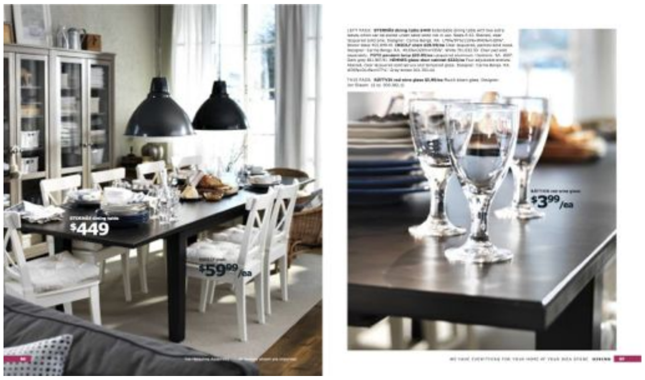


Fig. 4: *Overview and detail images in IKEA catalog. In addition to the overview image on the left, IKEA provides a detail image that advertises the glasses on the table. Note how the viewpoint and object positions are adjusted from the overview image (reprinted with permission from the 2013 IKEA catalog).*

σ from O_d are selected as context objects. We then optimize the composition with our algorithm.

As a test of our method, we generated detail images for three scenes: a kitchen, study, and living room. For each scene, we generated detail images for 20 random objects. In many cases, our optimization was able to put the camera close to canonical views only by moving objects that would otherwise occlude the detail object. For example, the gray chair is moved to different positions in Figure 3b and Figure 3c in order to reduce occlusions for different focus objects. Overall, most of our detail images produce reasonable compositions, and our user study (Section 7) indicates that our full optimization produces better results than camera-only optimizations the large majority of the time.

6.3 3D views for room planner

Home furnishing companies have recently started to provide online tools that let users create arrangements of furniture customized for their own rooms, e.g. IKEA's Home Planner (see Figure 6). After designing a room in this manner, users often want images of the room to share with others and to help them evaluate the design. Thus, another application of our system is to provide an automated solution for generating well composed images of user-designed rooms.

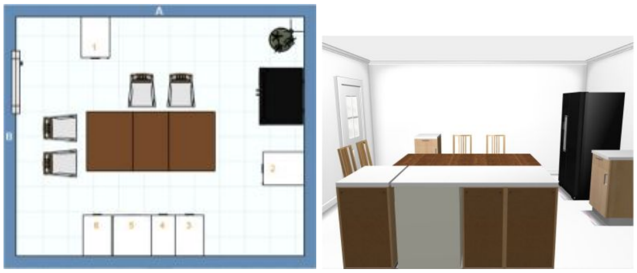


Fig. 6: *Room planner images. The IKEA Home Planner lets users create a room design in a plan view (left) and then generates a default view of the 3D scene (right).*

After generating the 3D arrangement of objects, the user selects a set of focus objects (likely the objects he is considering for purchase) and then asks our system to generate a composition. Unlike the previous two applications, we do not expect the user to provide an initial viewpoint for the scene. As a result, we modify our optimization to first search globally for the best camera parameters, which we then use as an initialization to our full optimization.

For our global camera search, we first generate a set of "plausible" initial camera parameters. We restrict the camera height to be at human eye level h_o, and we sample all of the other parameters as follows. For the other two camera position coordinates, we sample uniformly within the walls of the room at roughly 2 ft intervals; we take 20 uniformly spaced samples for azimuthal angle between 0 and 2π; to keep the camera fairly level, we take 5 uniform samples for the polar angle between $\pi/2$ (looking horizontally) and $\pi/2 + 0.28$ (looking slightly down); and we consider 4 uniformly spaced field-of-view values from 0.3 to 0.6. We then prune very poor samples by checking each camera view to see whether at least 50% of the screen space projection of every focus object bounding box is within the viewport. If not, then we discard the sample. Next, we do k-means clustering (with $k = 4$) of the pruned camera parameters. Within each cluster, we pick the camera with the lowest energy and use that as a candidate initialization. We run our optimization for each of the k candidate initializations, and pick the final composition with the lowest energy.

We used this optimization procedure to generate the results shown in Figure 7. Here, we arranged all the objects in the scene without providing an initial camera and chose the couch, coffee table and ottoman as the focus objects. In the final composition (Figure 7b), all of the focus objects are visible and the image provides a good overview of the scene from a plausible camera angle. For comparison, we show the camera-only result in Figure 7c, and notice that the initialization of it is different from that of Figure 7b. In Figure 7d, we show the camera-only result which is generated from the same initialization as in (b). With the capability of moving objects, our full optimization is able to achieve a better balance between multiple factors, and achieve a better composition.

6.4 Object replacement

Multinational furniture companies like IKEA usually customize their catalog images for different countries to match cultural preferences. This customization often involves choosing different materials or replacing objects within a scene. For example, a particular type of chair might be appropriate in the USA, while a different one is preferred in China, even though several other aspects of the scene can be shared. In many cases, the size and shape of new objects can be significantly different from the original one, which means that a stylist will have to spend a significant amount of additional effort adjusting the camera parameters and object positions to achieve a good composition for each customization. As with the previous applications, our optimization framework can automatically make these adjustments to reduce the amount of human effort required to perform these cultural customizations. Once the relevant objects and materials have been replaced, we use the current viewpoint, object positions and materials as initialization and optimize for a better composition.

Figure 5 shows an example where we replace the grey seat, side table and coffee table. When we swap in the new objects, there is a collision between the chair and plant, and in general, the composition feels cramped. When we optimize the composition, the collision is resolved and the camera pulls back to keep all the relevant objects in the frame.

6.5 Text-incorporated composition

Most catalog images have text overlays that describe the depicted scene. Such text is typically positioned over regions with nearly constant color so that it is easy to read and often appears in roughly the same location on every page (e.g., corners) so that the viewer knows where to look to find textual information. Our optimization framework can automatically position text based on all of these criteria.

In addition to a set of focus objects and an initial composition, the stylist also specifies a set of rectangles **R** where she would like overlay text to appear in the frame. We then treat each rectangle as just another object in the scene, but one that only has a 2D position and can only move within the viewport.

We apply the visibility and inertia terms to text rectangles as well. Specifically, the overlapping region of a focus object or a context object with a text rectangle is treated as occlusion. We replace E_{cc} by E_{tc} to account for the contrast between text and its background.

$$E_{tc} = w_{tv} \sum_{R_i \in \mathbf{R}} \min_{wb=w,b} \frac{1}{\mathcal{A}(R_i)} \sum_{p \in R} \frac{1}{d_l(p, wb)^2 + \epsilon}$$

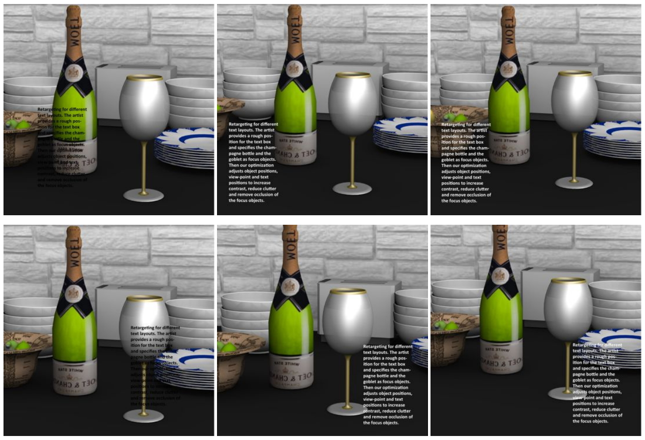
where p is a pixel in the rectangle R_i, $d_l(\cdot, \cdot)$ is the difference between the luminance of two pixels, w is white and b is black.

We also observe that to make the text with constant color stand out, it is essential to keep a low variance in luminance within each rectangle to reduce clutter behind overlaid text. We introduce E_{tv} for this reason,

$$\bar{L_{R_i}} = \frac{1}{\mathcal{A}(R_i)} \sum_{p \in R_i} L(p)$$

$$E_{tv} = w_{tv} \sum_{R_i \in \mathbf{R}} \frac{1}{\mathcal{A}(R_i)} \sum_{p \in R_i} \left(L(p) - \bar{L_{R_i}}\right)^2$$

Figure 8 presents a composition optimized with two different initial positions for the text. Notice how the objects in the scene are moved to create low contrast, low variance regions of the image where the text can be overlaid.

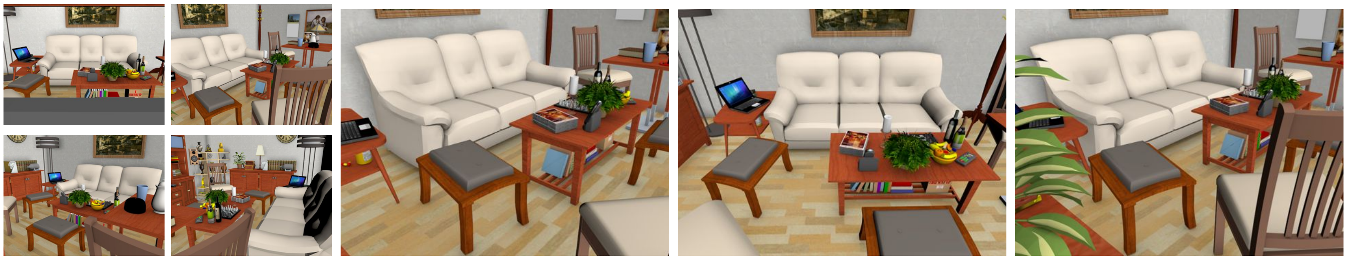


(d) *Input* **(e)** *Our method* **(f)** *Camera-only*

Fig. 8: *Retargeting for different text layouts. The artist provides a rough position for the text box and specifies the champagin bottle and the goblet as focus objects. Then our optimization adjusts object positions, viewpoint and text positions to increase contrast, reduce clutter and remove occlusion of the focus objects.*

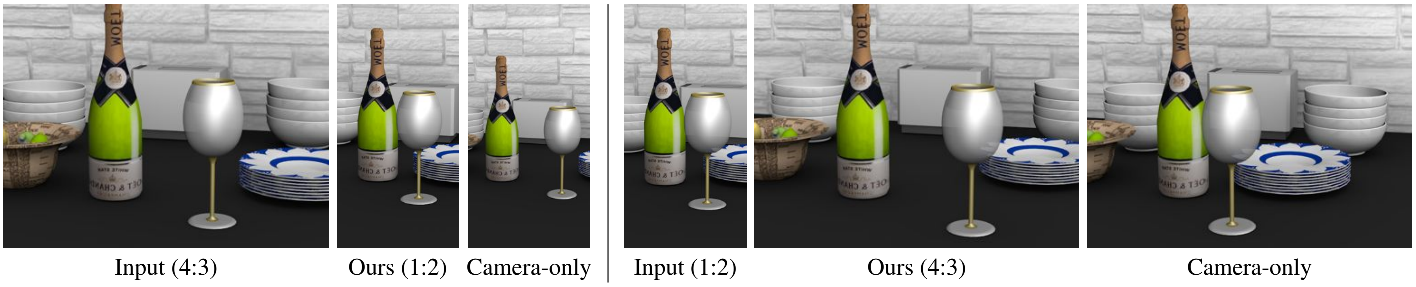
6.6 Retargeting for different aspect ratios

Our system can also automatically retarget catalog images to aspect ratios that are appropriate for different display formats. For example, a landscape image in a printed catalog may work better in portrait format for a tablet. Simple cropping is usually not sufficient to create a good retargeted composition because the relative arrangement of objects in image space remains fixed. In addition, existing 2D image retargeting methods such as Seam Carving [2007] often have trouble preserving strong structural elements (e.g., straight lines) that are prevalent in indoor scenes. In contrast, our optimization framework has the ability to adjust camera parameters and object positions to produce good compositions for different aspect ratios. For this application, we use the viewpoint and object positions from the input composition as initialization and solve for a new image with the specified dimensions.



(a) *Candidate initializations* **(b)** *Our method* **(c)** *Camera-only* **(d)** *Camera-only with the same initialization as (b)*

Fig. 7: *3D views for room planner. The focus objects are couch, coffee table and ottoman. We start by sampling plausbile camera parameters and result in 4 candidate initializations (a). We run the optimization from each of them, and select the composition with the lowest energy (b). For comparison, we show the camera-only result in (c) and the camera-only result which is generated from the same initial camera parameters as (b) in (d).*



Input (4:3) Ours (1:2) Camera-only | Input (1:2) Ours (4:3) Camera-only

Fig. 9: *Retargeting for different aspect ratios (focus objects: the champagne bottle and goblet). We start with the optimal composition in the initial aspect ratio (left in each group), and retarget it to a different one (middle). We compare our result to the one where only the camera is optimized (right).*

In Figure 9, we start with the optimal composition in one aspect ratio and then retarget it to another. For comparison, we generate images using our optimization method but without adjusting object positions (rightmost one in each group). Notably, the greater flexibility creates better retargeted images.

7. USER STUDY

A natural question to ask when considering our system is whether the additional freedom afforded by moving objects makes a positive impact on the results, or if – to the contrary – similarly good results could be obtained by performing a camera-only optimization. We investigated this question by asking people to compare 36 pairs of compositions created using our optimization procedure with object movement enabled (our method) and disabled (camera-only).

Study design: Our selection of scene compositions to compare includes all the examples shown in Section 6 of the paper, plus thirty detail images generated for different objects in three scenes (living room, study, and kitchen). For each of the three scenes, we selected – from among all the large furniture and a random subset of the smaller objects – the ten objects where the detail image generated with camera-only optimization differed most from the full-optimization.

We showed these pairs of scene compositions to study participants in randomized order, with images in each pair flipped left-right randomly. For each pair, the user selected a radio button to indicate that one composition was better at showcasing the specified focus objects (listed in the title), or that the two compositions are of the same quality (see a sample user study web page in the supplemental material).

We administered the study to two groups: *experts* who work professionally on scene layout for catalog images, and *non-experts*.

The first group was recruited through personal contacts and completed a single-page web-based survey without compensation. Given the small number of experts in this field, we were only able to administer the survey to two participants.

The second group was recruited through Amazon Mechanical Turk, and each turker was compensated 10 cents. To exclude 'lazy' turkers from our results, we tested the consistency of each turker's results. Specifically, each turker completed a multiple-page (one comparison per page) survey, where each comparsion was asked twice, with compositions swapped left-right. We excluded any input from turkers for each question where their two answers for the same pair were inconsistent, and we excluded all input from any turker whose answers were inconsistent for more than 25% of the questions. After running the study for 200 turkers, these consistency checks yielded 49 to 75 answers per comparison.

Study results: In the expert study, Expert 1 favored 'full optimization' in 22 pairs, 'camera only' in 12, and had 'no preference' in 2; Expert 2 favored 'full optimization' in 17 pairs, 'camera only' in

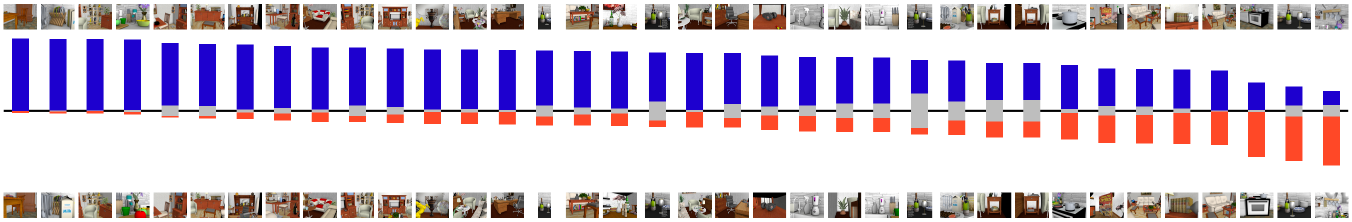


Fig. 11: *Amazon Mechanical Turk Study results. We asked participants to compare 36 pair of images generated with full optimization (top) and camera-only optimization (bottom). Bars represent the proportion of participants who favored each image (dark blue: full, red: camera-only, light grey: no preference).*

14, and had 'no preference' in 5. Results of the Amazon Mechanical Turk study are summarized in Figure 11. The top and bottom rows depict the pairs of images shown to each study participant (images in the top row show the fully optimized scenes). Upward-pointing dark blue bars reflect the fraction of study participants who favored images of the fully optimized scenes, while the downward-pointing red bars show the fraction that preferred camera-only optimizations, and the light gray bars in the middle represent cases where participants indicated no preference.

Generally, we find that full scene optimization is preferred to camera-only optimization. If the "no preference" answer is treated as half a vote for our optimization result and half a vote for camera-only, our optimization results received $\geq 75\%$ of the votes in 20 cases, 50%-75% of the votes in 13 cases, and $< 50\%$ of the votes in 3 cases in the Mechanical Turk study. In the experts study, the numbers are 16, 13 and 7 respectively. According to comments provided by participants, this is mainly because it moves objects to avoid occlusions, provide favorable contrast, and avoid awkward camera views.

8. CONCLUSION AND FUTURE WORK

In this work, we have introduced a technique for optimizing 2D compositions of 3D scenes that adjusts camera parameters, object transformations, and surface materials. Our results and informal user evaluation show the benefits of optimizing over all of these scene parameters simultaneously. In particular, the comparisons between images generated by only adjusting the camera and those generated by our full optimization clearly indicate that moving objects significantly improves the quality of compositions in many cases. We have demonstrated how our optimization framework benefits a variety of applications related to the creation of digital catalog images, from generating detail images of individual objects to rendering images of entire rooms for a home planner.

Our system has several limitations. First, the optimization procedure is currently too slow to be used in an interactive system. This is largely because speed has been sacrificed for flexibility in our implementation. We believe that the speed could be improved by orders of magnitude in a production-oriented implementation. Second, the set of objects allowed in the scene is provided as input and cannot be changed during the optimization, which limits use of our algorithm to fine-scale refinement, rather than large-scale exploration. Third, we use OpenGL rendering during our optimization, which does not account for global illumination effects in final images. Fourth, we have very primitive energy terms for controlling the 3D spatial relationships between objects. Perhaps more sophisticated probabilistic models learned from examples would be better (e.g., [Fisher et al. 2012; Yeh et al. 2012; Yu et al. 2011]). Fifth, we consider only a partial set of possible composition rules in our final experiments. Early in the project, we implemented terms for diagonal dominance, symmetry, and focusing with vanishing points, but found them less useful in our target applications – using our system to systematically investigate which energy terms are most effective for which applications would be an interesting topic of further study.

Given that companies are increasingly relying on computer-generated imagery for catalogs and other product advertisements, there are many opportunities for future work related to the automated generation of such images. For example, we imagine new advertising applications that choose furniture arrangements based on how a room will look from key viewpoints (e.g., the front door). Film, game, and real-estate companies could automatically optimize scenes for sequences of camera viewpoints (e.g, for movie shots or virtual tours). On-line advertisers could adapt product images to wide varieties (millions) of user preferences with automatically optimized aesthetics. We believe composition-aware scene modeling is a useful approach for all of these applications and as such represents a promising research direction for the computer graphics community.

REFERENCES

ABDULLAH, R., CHRISTIE, M., SCHOFIELD, G., LINO, C., AND OLIVIER, P. 2011. Advanced composition in virtual camera control. In *Smart Graphics*. Springer, 13–24.

ARNHEIM, R. 1988. *The Power of the Center*. University of California Press.

AVIDAN, S. AND SHAMIR, A. 2007. Seam carving for content-aware image resizing. *ACM Transactions on Graphics, (Proceedings SIGGRAPH 2007) 26,* 3.

BANERJEE, S. AND EVANS, B. 2004. Unsupervised automation of photographic composition rules in digital still cameras. In *SPIE Conference on Sensors, Color, Cameras, and Systems for Digital Photography*. Vol. 5301. 364–373.

BARES, W. 2006. A photographic composition assistant for intelligent virtual 3d camera systems. In *Smart Graphics*. Springer, 172–183.

BARES, W., MCDERMOTT, S., BOUDREAUX, C., AND THAINIMIT, S. 2000. Virtual 3d camera composition from frame constraints. In *Proceedings of the eighth ACM international conference on Multimedia*. ACM, 177–186.

BELL, B., FEINER, S., AND HÖLLERER, T. 2001. View management for virtual and augmented reality. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*. ACM, 101–110.

BETHERS, R. 1956. *Composition in pictures*. Pitman Pub. Corp.

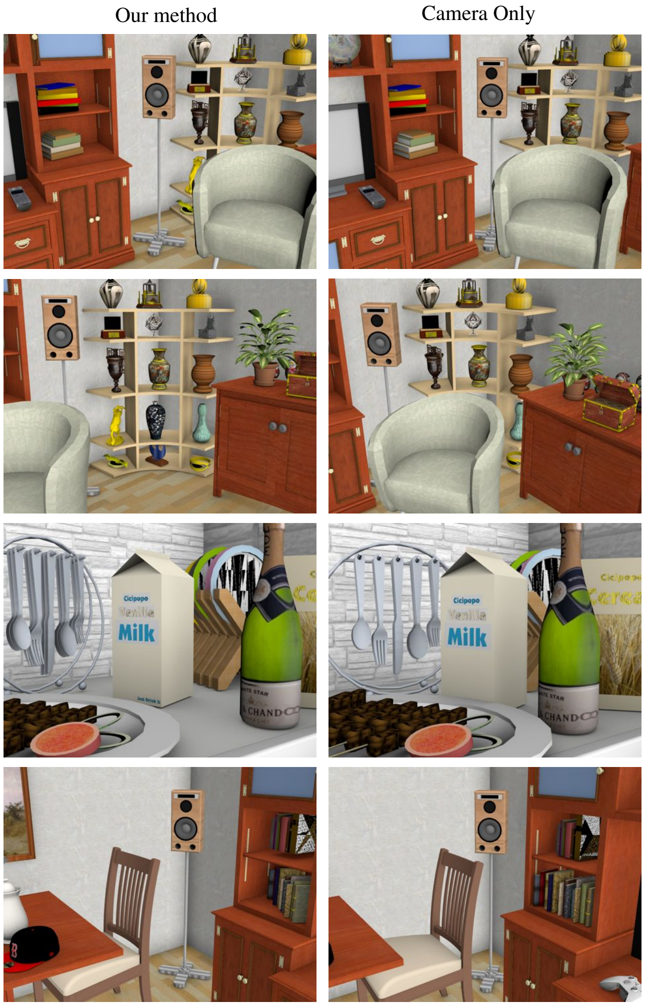


Fig. 10: *A subset of image pairs compared in our user study. Our system is able to satisfy multiple composition constraints simultanuously, which cannot be achieved by changing viewpoint only. Note visibility in row 1 and row 2, clearance for milk carton in row 3 and visual balance in row 4.*

Bhattacharya, S., Sukthankar, R., and Shah, M. 2010. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the international conference on Multimedia*. ACM, 271–280.

Blanz, V., Tarr, M., Bülthoff, H., and Vetter, T. 1999. What object attributes determine canonical views? *Perception-London 28,* 5, 575–600.

Bukowski, R. and Séquin, C. 1995. Object associations – a simple and practical approach to virtual 3d manipulation. In *ACM Symposium on Interactive 3D Graphics*.

Byers, Z., Dixon, M., Smart, W., and Grimm, C. 2004. Say cheese! experiences with a robot photographer. *AI magazine 25,* 3, 37.

Christie, M., Olivier, P., and Normand, J. 2008. Camera control in computer graphics. In *Computer Graphics Forum*. Vol. 27. Wiley Online Library, 2197–2218.

Clifton, J. 1973. *The Eye of the Artist*. North Light Publishers.

Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., and Xu, Y. 2006. Color harmonization. In *ACM Transactions on Graphics (TOG)*. Vol. 25. ACM, 624–630.

Datta, R., Joshi, D., Li, J., and Wang, J. 2006. Studying aesthetics in photographic images using a computational approach. *Computer Vision–ECCV 2006*, 288–301.

Enthed, M. 2012. 3D at IKEA. In *3D Modeling Standards*. SIGGRAPH Birds of a Feather.

Fisher, M., Ritchie, D., Savva, M., Funkhouser, T., and Hanrahan, P. 2012. Example-based synthesis of 3d object arrangements. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia) 31,* 6.

Gleicher, M. and Witkin, A. 1992. Through-the-lens camera control. In *ACM SIGGRAPH Computer Graphics*. Vol. 26. ACM, 331–340.

Gooch, B., Reinhard, E., Moulding, C., and Shirley, P. 2001. Artistic composition for image creation. In *Rendering Techniques 2001: Proceedings of the Eurographics Workshop in London, United Kingdom, June 25-27, 2001*. Springer Verlag Wien, 83.

Govindaraju, N. K., Redon, S., Lin, M. C., and Manocha, D. 2003. Cullide: Interactive collision detection between complex models in large environments using graphics hardware. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware*. Eurographics Association, 25–32.

Grill T., S. M. 1990. *Photographic Composition*. Watson-Guptill.

He, L., Cohen, M., and Salesin, D. 1996. The virtual cinematographer: a paradigm for automatic real-time camera control and directing. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 217–224.

Jin, Y., Wu, Q., and Liu, L. 2012. Aesthetic photo composition by optimal crop-and-warp. *Computers & Graphics*.

Kao, H., Ma, W., and Ming, Q. 2008. Esthetics-based quantitative analysis of photo composition. In *Pacific Graphics*.

Kowalski, M., Hughes, J., Rubin, C., and Ohya, J. 2001. User-guided composition effects for art-based rendering. In *Proceedings of the 2001 symposium on Interactive 3D graphics*. ACM, 99–102.

Krages, B. 2005. *Photography: The Art of Comopsition*. Allworth Press.

Liu, L., Chen, R., Wolf, L., and Cohen-Or, D. 2010. Optimizing photo composition. In *Computer Graphics Forum*. Vol. 29. Wiley Online Library, 469–478.

Lok, S., Feiner, S., and Ngai, G. 2004. Evaluation of visual balance for automated layout. In *Proceedings of the 9th international conference on Intelligent user interfaces*. ACM, 101–108.

Martinez, B. and Block, J. 1988. *Visual forces: an introduction to design*. Prentice Hall.

Merrell, P., Schkufza, E., Li, Z., Agrawala, M., and Koltun, V. 2011. Interactive furniture layout using interior design guidelines. In *ACM Transactions on Graphics (TOG)*. Vol. 30. ACM, 87.

Olivier, P., Halper, N., Pickering, J., and Luna, P. 1999. Visual composition as optimisation. In *AISB Symposium on AI and Creativity in Entertainment and Visual Art*. 22–30.

Souppouris, A. 2012. Ikea catalog will be 25 percent 3d renders by next year. *Wall Street Journal*.

Southern, A. 2012. Real or rendered? how 3d imagery is changing the way you shop. *Technomy*.

Taylor, E. 1938. *The How and Why of Photographic Composition*. The Galleon Publishers.

Ward, P. 2003. *Picture Composition for Film and Television*. Focal Press.

Wong, L. and Low, K. 2011. Saliency retargeting: An approach to enhance image aesthetics. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*. IEEE, 73–80.

Yeh, Y., Yang, L., Watson, M., Goodman, N., and Hanrahan, P. 2012. Synthesizing open worlds with constraints using locally annealed reversible jump mcmc. *ACM Transactions on Graphics (TOG) 31,* 4, 56.

YU, L., YEUNG, S., TANG, C., TERZOPOULOS, D., CHAN, T., AND OSHER, S. 2011. Make it home: automatic optimization of furniture arrangement. *ACM Trans. Graph 30,* 86, 1–86.