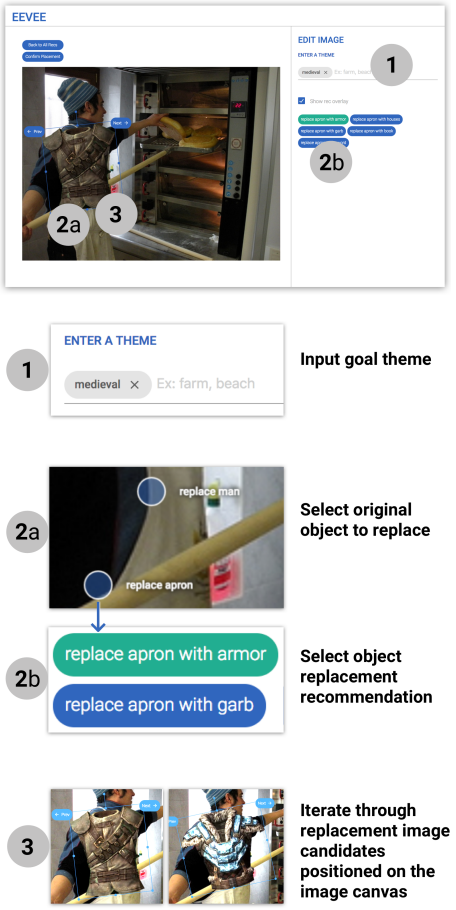

Figure 1: The Eevee system, which enables users to edit images in terms of high-level goals: (1) the user inputs goal theme (e.g. "medieval"), (2) the user selects an object to replace (e.g. "apron") and selects its object replacement recommendation (e.g. "armor"), (3) the user iterates through replacement image candidates (i.e. different pieces of armor) positioned on the image canvas.

Eevee: Transforming Images by Bridging High-level Goals and Low-level Edit Operations

Michelle S. Lam
Stanford University
Stanford, CA, USA
mlam4@cs.stanford.edu

Gracie B. Young
Stanford University
Stanford, CA, USA
gyoung@cs.stanford.edu

Catherine Y. Xu
Stanford University
Stanford, CA, USA
cxu96@cs.stanford.edu

Ranjay Krishna
Stanford University
Stanford, CA, USA
ranjaykrishna@cs.stanford.edu

Michael S. Bernstein
Stanford University
Stanford, CA, USA
msb@cs.stanford.edu

ABSTRACT

There is a significant gap between the high-level, semantic manner in which we reason about image edits and the low-level, pixel-oriented way in which we execute these edits. While existing image-editing tools provide a great deal of flexibility for professionals, they can be disorienting to novice

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3312929

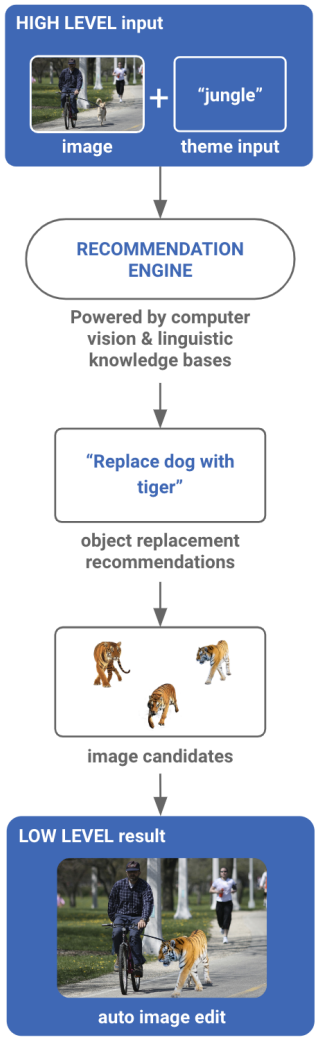

Figure 2: A summary of the user workflow with the Eevee system.

editors because of the gap between a user's goals and the unfamiliar operations needed to actualize them. We present Eevee, an image-editing system that empowers users to transform images by specifying intents in terms of high-level themes. Based on a provided theme and an understanding of the objects and relationships in the original image, we introduce an optimization function that balances semantic plausibility, visual plausibility, and theme relevance to surface possible image edits. A formative evaluation finds that we are able to guide users to meet their goals while helping them to explore novel, creative ideas for their image edit.

KEYWORDS

Image editing; natural language; object replacement; scene graphs

INTRODUCTION

We perceive images at a high level and ultimately want to craft visual stories with our images. However, current image-editing systems are centered around pixel-level adjustments to the syntax of an image rather than changes to the semantics of an image. State-of-the-art systems such as Photoshop provide a considerable degree of low-level control, but do not allow users to provide high-level semantic input (i.e. "make this image *business-like*") to reach their goals. While this approach works well for experienced editors, it may be daunting for novice image editors to be pulled into the minutiae of complex edit techniques. This approach may also limit creative experimentation during the editing process by distancing novice users from their high-level intent. Furthermore, because existing systems represent images in terms of pixels and layers and don't reason about what is *depicted in* an image, they aren't well-situated to handle semantic user input. In this paper, we aim to leverage existing technologies to bridge the gap from *high-level* goals to *low-level* image edit operations.

We present Eevee, a system that enables users to edit images in terms of high-level themes. Users provide an image and a theme for an image edit (e.g. "make this image look more like a *farm*"). High-level themes can be settings (e.g. zoo, party, office) or attributes (e.g. retro, professional, kid-friendly). Our system breaks down the provided theme into a series of low-level image-edit operations (e.g. "replace the building with a barn") based on an understanding of the scene's objects and relationships.

We introduce an optimization function that searches for optimal edit suggestions by estimating several characteristics of possible image edits: *visual plausibility* (how visually similar to the original image is the proposed edit?), *semantic plausibility* (how likely is it that the scene produced by the proposed edit would occur in the real world?), and *theme relevance* (to what extent does the proposed edit bring the image closer to the provided theme?). These metrics are calculated based on the objects and relationships in the original image as well as the prevalence of objects and relationships in visual and linguistic datasets [3, 5] and their semantic relevance to the theme [6]. By linking high-level user inputs with low-level image edit outcomes, we aim to allow novice image editors to focus on more of

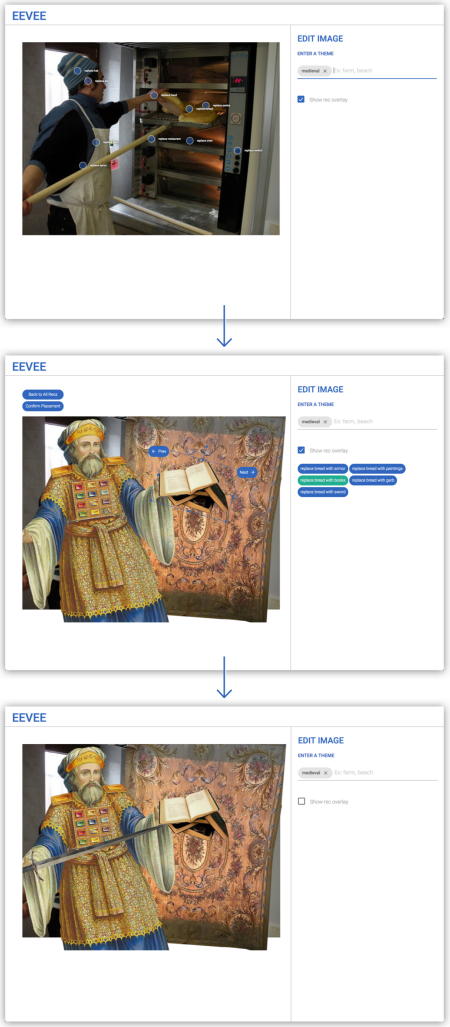

Figure 3: Samples of the Eevee editing interface: (a) User has entered in the theme "medieval"; (b) User is browsing replacement candidates for the suggestion "Replace bread with books"; (c) User has finalized their image edits to replace man with priest, oven with tapestry, bread with books, and stick with sword.

their conceptual intents around an image edit. We perform a preliminary evaluation of our system and find that users were able to transform images to match their theme and were guided to explore novel ideas in the editing process.

RELATED WORK

Prior work has explored the possibility of goal-based control in creativity tools. AttribIt focused on better surfacing system functionality to users: it eased design space exploration by smartly organizing candidate designs along high-level, interpretable axes [2]. Other systems have aimed to help the system better understand the user: PixelTone allowed users to specify desired edit operations using natural language and direct manipulation while CommandSpace mapped natural language commands to application commands to complete desired tasks [1, 4]. These works connect user commands and computer-oriented operations, but they still require users to distance themselves from their initial high-level goals to break them down into low-level, computer-understandable commands and operations (e.g. "make the shirt red"). With Eevee, we aim to enable users to interact with their images in terms of higher level goals like shifts in themes or settings.

As we see, current work in image-editing has enabled user-understandable low-level edit operations and holistic transformations, but does not enable high-level, goal-based control grounded in the contents of a particular image scene. With Eevee, we aim to account for the dynamic, emergent combinations of image components to power goal-based image edits.

EEVEE

The Eevee system is a first step towards enabling users to edit images in terms of high-level goals. We design our system to lower the threshold for novice users and aim to satisfy two design goals:

- **D1**: Enable users to make image edits that successfully move towards their high-level intent
- **D2**: Empower users to come up with new high-level ideas for their image edit

Because we are most interested in large-scale semantic changes in image edits, we focus on **object replacement edits**. These kinds of edits effectively enact significant thematic changes and are more understandable to novice users who view an image in terms of objects rather than pixels or color-regions [7]. Furthermore, working at the level of objects enables us to leverage computer vision and linguistic knowledge bases to deliver intelligent recommendations informed by an understanding of cultural norms and patterns in our physical world. Our image edits are produced in the style of **humorous memetic edits** because in this domain, the *semantics* of an edit are more important than its visual quality or fidelity, and drastic stylistic shifts and absurd elements are welcome.

$$\text{SEM\_PLAUS}(O', O_R) = \left(\sum_{i=0}^{|O_R|} 1 - \text{PLAUS}(O', O_{Ri})\right) / |O_R|$$

The **semantic plausibility cost** is the average plausibility cost of the candidate replacement object being in a relationship with all the objects that were related to the original object

$$\text{PLAUS}(O, R) = VG_{subj\_tree}(O, R) \cdot VG_{obj\_tree}(O, R)$$

The **plausibility** of an object in a relationship R is the product of its frequency as a subject with R and its frequency as an object with R

$$\text{VIS\_SIZE\_PLAUS} = \text{COSINE\_DIST}\Big(\text{SIZE\_RATIOS}(O, O_C), \text{SIZE\_RATIOS}(O', O_C)\Big)$$

The **visual size plausibility cost** is the cosine distance between the vector of size ratios for the original object and its connected objects and that of the candidate replacement object and the same connected objects

$$\text{SIZE\_RATIOS}(O, O_C) = [\text{SIZE\_RATIOS}(O, O_{C0}), \ldots, \text{SIZE\_RATIOS}(O, O_{Cn})]$$

The **size ratios** vector is the list of bounding box size ratios between an object O and all of the objects O_C that are connected to this object by a relationship

$$\text{VIS\_POS\_PLAUS} = \text{COSINE\_DIST}\Big(\text{POS\_RATIOS}(O, O_C), \text{POS\_RATIOS}(O', O_C)\Big)$$

The **visual position plausibility cost** is the cosine distance between the vector of position ratios for the original object and its connected objects and that of the candidate replacement object and the same connected objects

$$\text{POS\_RATIOS}(O, O_C) = [\text{POS\_RATIOS}(O, O_{C0}), \ldots, \text{POS\_RATIOS}(O, O_{Cn})]$$

The **position ratios** vector is the list of bounding box centroid positions (where the original object's centroid is the origin) between an object O and all of the objects O_C that are connected to this object by a relationship

$$\text{THEME\_RELEVANCE} = \text{COSINE\_DIST}(O, theme) - \text{COSINE\_DIST}(O', theme)$$

The **theme relevance cost** is the cosine dist between the original object and the theme subtracted by the cosine dist between the candidate replacement object and the theme

Figure 4: Optimization Function Equations

The Workflow

Our system is a web-based image-editing interface comprised of: (1) High-level theme specification, (2) Object replacement recommendations, (3) Image candidates for object replacement, and (4) a Canvas for final image edit staging and adjustments. Users start with an image and enter a theme word to describe their overall vision, and our system generates object replacement recommendations (e.g. "replace dog with tiger"). The user can experiment with an object replacement recommendation and iterate through the image candidates for the object replacement, which are surfaced on the image canvas at a predicted position and size. The user can make minor adjustments and can continue to experiment with other object replacement recommendations until satisfied.

Optimization Function

Our main technical contribution is an optimization function that utilizes knowledge about an image scene to surface edits that match the user's desired theme. We leverage several knowledge bases: Visual Genome (VG) [3], a dataset of images densely annotated with scene graphs (objects connected by *predicate* relationships and modified by *attributes*); WordNet [5], a lexical database that organizes English nouns, verbs, adjectives, and adverbs into synonym sets linked by semantic relations such as synonymy and hyponymy; and word vectors generated using the GloVe method for vector-space representation of words [6]. Our algorithm evaluates candidates along three dimensions: (1) **Semantic plausibility**, (2) **Visual plausibility**, and (3) **Theme relevance**, as summarized in Figure 5.

Semantic plausibility. Captures the extent to which the candidate object could plausibly take the place of the original object for all of its relationships. For an image object O (e.g. "dog"), call all of the object's relationships O_R (e.g. "man walking dog," "dog chewing bone"), and call all objects connected to object O via a relationship O_C (e.g. "man," "bone"). The semantic plausibility (SEM_PLAUS) of candidate replacement object O' is calculated as shown in Figure 4. The VG_{subj_tree} and VG_{obj_tree} scores capture the plausibility of a given subject or object and a predicate word; these scores are pregenerated across the full VG dataset as follows. Iterating over all image relationships (in *subj-predicate-obj* form), we constructed a tree, VG_{subj_tree} (or VG_{obj_tree}), by adding each subject (or object) and, iteratively, each of its hypernyms (parent-categories) until the root category. Then, for every image relationship, we incremented the count for the *subject* and *predicate* in the VG_{subj_tree} and the *object* and *predicate* in the VG_{obj_tree}. We bubbled up this increment to each parent level, exponentially scaling down the increment at each higher level. We then normalized the scores in all tree levels.

Visual plausibility. Measures the similarity of the visual properties of the candidate object and the original object in terms of relative *size* and relative *position*. For each image object O in VG, we

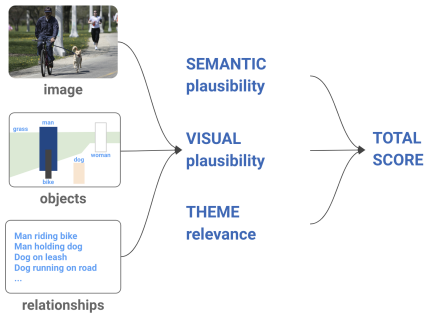

Figure 5: The Eevee system uses image objects and relationships to generate (1) Semantic plausibility, (2) Visual plausibility, and (3) Theme relevance scores, which are combined into a Total score that is used to search for edit recommendations.

calculated the relative size and centroid position of the object and its connected objects O_C to build up a size- and position-ratio vectors for each object. For the size metric (vis_size_plaus), we compare the relative size of the object O w.r.t. all related objects O_C to that of candidate object O' w.r.t. objects O_C. For the position metric (vis_pos_plaus), we compare the relative positions of the centroid of object O w.r.t. all related objects O_C to that of candidate object O' w.r.t. objects O_C.

Theme relevance. Captures how closely candidate O' matches the user-specified theme (compared to original object O). We use GloVe vectors to represent the words; then, theme_relevance is the cosine distance between O and the theme subtracted by the cosine distance between O' and the theme.

Total score. These three score metrics are calibrated to have the same range and are linearly combined to generate a holistic metric. We search for an ideal object replacement by performing a bounded search on the N objects nearest to the theme word in our word embedding space (we use $N = 100$). We calculate scores for all candidates and surface the top-ranking items as recommendations.

PRELIMINARY EVALUATION

We conducted a preliminary evaluation of Eevee with 5 college-aged participants (2 females, 3 males) from the university community. All participants had minimal photo-editing experience. Our overall goals were to (1) evaluate our **D1** and **D2** goals listed above and (2) understand how users interact with the system.

Methodology

We randomly selected 6 images from VG. Each participant received 3 of these images to edit using Eevee. For each image received, participants brainstormed a goal theme for their image. Participants were asked to use Eevee to transform each image to match their selected theme. Participants answered questions about their overall experience and used a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree) to indicate their opinions on the system and the quality of its object replacement recommendations.

Results

On average, participants made 4 object replacements per image; example edits are shown in Figure 6. Participants responded to the question "The object replacement suggestions helped me to **transform the image to match the theme**" with an average Likert scale rating of 3.9/5. In interviews, participants noted that they were pleasantly surprised by many replacement suggestions. One participant stated: *"some of the recs were surprising... like 'dynamite' to replace 'boulder' was random, but very applicable"*, while another noted: *"'navy' was a good suggestion for 'war' [theme]"*.

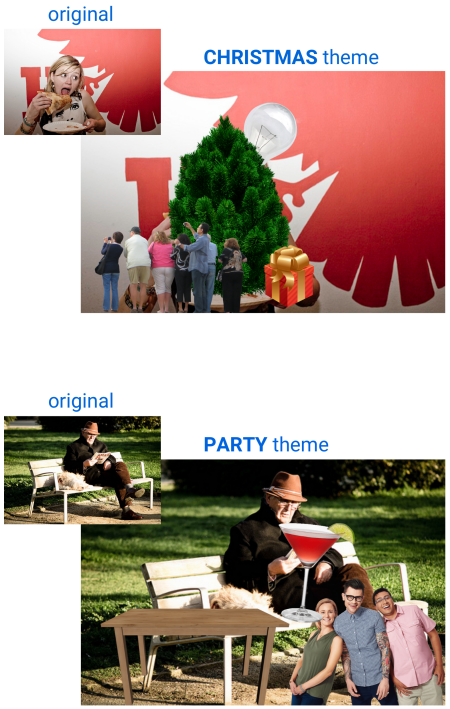

Figure 6: Image edits produced by study participants using the Eevee system. The first image is transformed to a "Christmas" theme, and the second image is transformed to a "party" theme.

Participants responded to the question "The system helped me to **think of new ideas**" with an average Likert scale rating of 4.4/5. They felt that the system helped them to explore novel ideas in the editing process: *"looking at [replacements] on the side and within each category helped me include ideas I wouldn't have thought of otherwise."* These responses indicate progress toward enabling users to make edits that move toward their high-level goal (**D1**) and helping users generate ideas in the editing process (**D2**).

FUTURE WORK & CONCLUSION

In the future, we plan to investigate greater flexibility and user input for Eevee. Users often used more than one word to describe themes and image candidates. We propose allowing users to input multiple words to describe their theme to allow specification of more refined intent. Second, participants expressed frustration with generalized image candidates *("I wanted a snow hat, not just a hat")*. We plan to refine Eevee's image candidates with attributes informed by the theme and original scene.

We have presented Eevee, a theme-based image-editing system that lowers the threshold for novice users by leveraging recent advances in image content understanding, linguistic knowledge bases, and word embeddings. Our preliminary evaluation shows promising results that Eevee can help people to generate new ideas. These results demonstrate that leveraging semantic information within an image can help users make more creative edits while guiding them to actualize their high-level goals.

REFERENCES

[1] Eytan Adar, Mira Dontcheva, and Gierad Laput. 2014. CommandSpace: modeling the relationships between tasks, descriptions and features. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 167–176.

[2] Siddhartha Chaudhuri, Evangelos Kalogerakis, Stephen Giguere, and Thomas Funkhouser. 2013. Attribit: content creation with semantic attributes. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 193–202.

[3] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[4] Gierad P Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, Jason Linder, and Eytan Adar. 2013. Pixeltone: A multimodal interface for image editing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2185–2194.

[5] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[6] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[7] Jeremy M Wolfe. 1998. Visual memory: What do you know about what you saw? *Current biology* 8, 9 (1998), R303–R304.