

BACHELOR THESIS

EXPOSÉ

Optimizing perceived aesthetics of UIs using metric guided
generative pipelines

MORITZ WÖRMANN

April 17, 2024

Variational Autoencoder Optical Character Recognition

1 MOTIVATION AND BACKGROUND

Through User Interface (UI) design tools like Figma ¹, it gets easier to create UIs for Apps and Websites for users to interact with. However, designing visually pleasing UIs still proves to be a complicated task, especially since these are highly subjective categories. [?] This challenge becomes even more significant when considering the impact initial impressions of a UI can have on the users perception and on the willingness to stay on the website or the mobile app [4]. Currently, designing a new UI is often a task for multiple teams with different professions, like graphic design and software engineering. While modern project management strategies like Scrum can alleviate the difficulties introduced by aligning and communicating stakeholder and user interests with the final product, they still rely heavily on good communication and the abundance of time and therefore dont fully alleviate the challenges associated with UI design. Automating the task of creating UIs or provding assistance via automatic algorithms is therefore a worthwhile subject as an "end-to-end" process of creating user-interfaces or at least optimising existing UIs can reduce time and effort.

Layout generation tasks describe the challenge of aligning different elements and components of user interfaces as well as controlling other parameters like font, font size and color in a visually pleasing way. While this task is evidently sufficiently difficult for humans, assigning this task to algorithms proves to be even more difficult as the challenge of defining what is considered visually pleasing for humans is not straight forward.

2 PROBLEM DESCRIPTION

The objective of this research is to develop a comprehensive methodology that can be utilized by a potential user to input an image of an existing UI and potentially provide additional instructions. This automated pipeline would be capable of segmenting the UI into its components and rearranging them in a more optimal or visually appealing manner. This segmentation process can be conceptualized as a transformation or mapping of the UI into a distinct space, which might be referred to as a latent space.

The goal of this research is to provide a potential user with a fully usable pipeline in which they can input an image of an already existing UI and potentially provide additional instructions. This automated pipeline would be able to segment the UI into its components and rearrange them in a better or visually more pleasing way. This segmentation process thus functions as

¹ <https://www.figma.com/>.

a transformation or mapping of the UI into a different space, which might even be called a latent space. An algorithm or model can operate in this space and retrieve feedback from a model, pretrained on a dataset in which users have been interrogated for their perceived aesthetics of user interfaces. It remains to be shown, if this classifier model can predict directly from this latent space or if the user interface has to be transformed out of this latent space again first (diffusion). Clearly, this transformation proves to be an additional challenge as well as deciding the size and dimension of the latent space. This research aims at exploring different approaches as to how these latent spaces can look with a focus to them being able to be used in a pipeline from the latent space to an aesthetics predictor without breaking autograd vectors in order to leverage common gradient-descent patterns for this task.

To solve these questions the following research questions will be answered:

- RQ 1 How can UIs be segmented in a way such that the layout can be optimized and reassembled?
- RQ 2 What characteristics should a suitable latent space possess in order to be utilized for the projection of user interfaces (UIs), which can then be optimized and graded by an automatic classifier?
- RQ 3 How can (accidental) adversarial attacks by the optimizer against the Aesthetic Predictor be avoided?
- RQ 4 Do Diffusion Models provide advantages, either via Pix2Pix optimization or via a different latent space which represents the UI
- RQ 5 How can the generation by all of the different approaches be constrained with user supplied input on high-level positional relationship between UI elements?
- RQ 6 How can a finished tool look which supports a potential user in a meaningful way during the layouting and design process?

In the next sections, related work is portrayed, after which a structure for the final thesis is presented.

3 STATE-OF-THE-ART

In the following, we divide the current state-of-the-art into three main parts: (1) Segmentation of UIs (2) Efforts to optimize UIs and their layout using non-diffusion based approaches (3) Efforts to optimize UIs using diffusion based approaches

3.1 *Segmentation of UIs*

Multiple different approaches to segmenting user interfaces have been explored and proven to be viable. For the training and evaluation part of this research, a pre-segmented dataset is useful. This kind of dataset has been presented in RICO [3]. Here, the segmentation is done via directly processing the semantics of the source code of UIs. Buttons, images and other elements are defined somewhere in the source of an app and through automatic agent based exploration of android apps, this segmentation is relatively straight forward. Purely optical based segmentation is also a thoroughly explored area. Approaches that leverage classic OCR (ocr) and combine it with modern object detection algorithms like YOLO [?] show promising results in segmenting user interfaces into their individual components [?].

3.2 Efforts to optimize UIs using non-diffusion based approaches

While modern generative models often leverage diffusion based approaches, a number of non-diffusion based approaches also exist. Recent research has shown large advancements while not (only) relying on diffusion based approaches like in 2022: Kong et al. [8] which shows how a layout transformer model can be used to reliably generate missing attributes from their latent space. This space is comprised of different elements, labeled with their category and their respective positioning on user interfaces. A similar approach called LayoutTransformer is presented in LayoutTransformer: Gupta et al. [5]. This approach leverages self-attention to assist with the generation and even allows for generation in the 3D Space. While these approaches may show impressive results in unguided layout generation, the topic of allowing the user to give high-level constraints such as predefined relationships between UI elements (e.g. ensuring the company logo is always on top) remains challenging and not explored in the same depth.

A different approach is the usage of a VAE (vae) like in Jiang et al. [7]. In their work, the authors propose a novel approach to segmenting the user interface into different regions. This approach involves “filling out” these regions with other user interface segments in order to combat the challenges of high-level relationships in user interfaces, which are difficult for these models to process. This research builds on Arroyo et al. [1] which initially proposed the usage of vaes for layout generation tasks. Such VAE approaches have also been explored in Xie et al. [13] and Patil et al. [11].

Still, non-VAE approaches also exist, e.g. leveraging advantages of Graph neural networks which allow for refinement of initial user controlled relationship definement like in H.-Y. Lee et al. [9].

3.3 Efforts to optimize *uis!* (*uis!*) using diffusion based approaches

Although not directly related to *uis!*, research has already been conducted in the field of metric-based optimization for Pix2Pix approaches. These approaches explore modifying images using diffusion models, effectively using an already existing image as the starting point in the latent space. Multiple research efforts have demonstrated that a simple gradient descent pipeline with a classifier at the end can optimize a prompt embedding that is passed to a stable diffusion model, as shown in their diffusion models. [2]

Another approach is to not rely on Pix2Pix models, but instead use a different autoencoder to transform the UIs into the latent space.

Deka et al. [3] already showcased an autoencoder which reduces the dimensions of a user interface layout to a 64-dimensional vector which can later be used to retrieve the layout representation again. While they did not add diffusion to their latent space, it still is a promising approach. To close the gap from this research to a finished pipeline, like it is the goal of this thesis, the component which transforms the generated layout (e.g. a picture) into a real UI (e.g. markup) is however still missing. Nevertheless, this approach could provide useful insights if be possible to use this autoencoder directly in a diffusion model.

However, all of the approaches presented that operate on user interfaces do not use a true aesthetic predictor as their metric, but rather more technical metrics that measure details such as overlap on design components.

4 PROPOSED APPROACH

As the overall goal of this research is to optimize perceived aesthetics, a clear way to measure this metric is needed. As all of the approaches will rely on the usage of a common gradient descent pipeline on one way or another, this metric needs to be measured in a way which is differentiable. For

simplicity, the same model will be used for all of the different approaches, which is the one presented in 20203: Leiva et al. [10].

The new research in this work is seeing the aesthetics predictor as part of the generation instead of using technical metrics to assess the goodness of created user interfaces after the creation has already been finished.

Table 1: Overview of the Proposed Approach

Research Questions & Related Study Phase		
RQ 1: Segmentation	RQ 2: Latent space	RQ 3: Adversarial attacks
Tasks		
<ul style="list-style-type: none"> Assess state-of-the-art Evaluate necessary changes & adaptations 	<ul style="list-style-type: none"> Comparison of common gradient-descent pipeline with naive positioning vector as latent space to SOTA approaches 	<ul style="list-style-type: none"> Evaluation if these (accidental) attacks do in fact prove to be a challenge Exploration of different mitigation approaches
Expected Results		
<ul style="list-style-type: none"> Assessment of "Fitness" of SOTA Finished Segmentation pipeline able to be differentiated in reassembling phase w.r.t. positioning vector 	<ul style="list-style-type: none"> usable decision for following quesitons 	<ul style="list-style-type: none"> Selection of specific mitigation tactic if applicable
Research Questions & Related Study Phase		
RQ 4: Introduction of diffusion	RQ 5: User Input constraints	RQ 6: Usable eze pipe
Tasks		
<ul style="list-style-type: none"> Assess SOTA layout diffusion approaches Assess Pix2Pix with common gradient descent pipeline 	<ul style="list-style-type: none"> Exploration of different constraining approaches 	<ul style="list-style-type: none"> Incorporation of prior results into usable product
Expected Results		
<ul style="list-style-type: none"> Evaluation of diffusion based approaches Pivot to explored approaches in case of superiority 	<ul style="list-style-type: none"> Incorporation in prior results if applicable 	<ul style="list-style-type: none"> Figma plugin or demo app

4.1 Research Question 1

Past Research like in [3] has shown that UI Segmentation is a task which is manageable by state of the art algorithms. It has been proven that optical segmentation into Text and Non-Text elements (by mere masking of the affected regions) can be used to train an AutoEncoder architecture which reliably reduces the dimension of the information in a user interface. As this

research is exploring a similar question (transforming a user interface into a latent space), a similar approach might provide good results for this task. It remains to be shown how significant the effects of different segmentation approaches are on the final pipeline. The maturity and reliability of models like the one presented in the mentioned paper suggests that this effect may be minimal.

4.2 *Research Question 2*

As previously described, the main challenge in this domain will be developing an entire pipeline, that includes a latent space, a function which retrieves the user interface out of this latent space and a predictor that determines the aesthetics for this retrieved and modified user interface. One such space might just be a vector of coordinates where the segments are placed on the user interface.

Once such a space and pipeline has been found, the representation of the user interface can be improved by utilizing common gradient descent patterns provided by major machine learning frameworks, provided that the whole pipeline actually converges.

4.3 *Research Question 3*

This part of the research is arguably the most important one. Keeping the pipeline from becoming too volatile or quickly “learning” how to exploit the aesthetics predictor and thus creating an adversarial attack is a complex task. These exploits might lead to undesirable results in which user interfaces might show extreme or slim changes for no apparent reason which might lead to a higher predicted aesthetic score but, are do in fact not show the same favorability during interrogation through humans. Adversarial attacks (also known as adversarial examples) have long been documented, especially in regards to humans being unable to detect them. [?]

The research area of preventing adversarial attacks has been thoroughly studied due to the ubiquitous nature of classifying neural networks. [?].

4.4 *Research Question 4*

This research question can be divided into two distinct tasks. The first task involves identifying an appropriate latent space in which the user interface can be projected. The second task assumes that satisfactory results can be achieved by solely relying on Pix2Pix approaches, such as StableDiffusion. [12]. This would undoubtedly necessitate the finetuning of the AutoEncoder in the StableDiffusion model to adapt to UIs. [?] However, this approach may prove challenging, as these models are notoriously difficult to control, which has led to the emergence of a distinct research field, prompt engineering.. [?] It is perhaps overly optimistic to suggest that Pix2Pix optimization can be used to create a user interface that is both functional and aesthetically pleasing.

Thus, the first approach, (finding a latent space that has been adapted to the UI usecase) could show improved results. For this, some research has already been done, for example by 2023: Hui et al. [6] who designed the latent space such that it only holds information about the layout of a user interface.

4.5 *Research Question 5*

Constraining the generated layouts and **uis!** has been the subject of past research [9]. While most of these efforts rely on giving the constraints at the start of the pipeline, e.g. developing relationships and going from

there on to the user interfaces, another approach could be to penalize a model/pipeline for a undesirable results which may include user defined constraints. It would then be entirely up to the model to grasp these constraints and work them into the predictions.

4.6 *Research Question 6*

As soon as a viable solution for the described problem has been found, integrating this solution in an appealing way for potential users is an additional challenge as the question arises how a finished user interface, which is still in a graphic representation at this stage can be transformed back out of the latent space.

As the main objective is to optimise existing user interfaces, the main challenge will be to transform such an existing user interface into a representation which can be used by the pipeline and, arguably more important, retrieving the user interface back from the pipeline. One such "starting point" might be rendered markup code in a tool like figma. For a segmentation and rearrangement task, this might mean the association between rendered elements and the respective code parts which produce these segments. For this task, the final extraction stage would have to have an understanding of how layering and ordering works in these kinds of markup languages.

To measure the results of this finished pipeline, already finished user interfaces could be rearranged in a way that is subjectively not aesthetic. The modified UI can be fed into the pipeline and differences between the original user interface and what the pipeline came up with can be used to grade the performance. Of course, this relies on the assumption that the original UI was already achieving a relatively high aesthetic score.

Following the described approach, the structure shown in fig. 1 is proposed for the thesis.

Figure 1: Proposed Structure

1. Introduction
 - 1.1 Motivation
 - 1.2 Problem Statement
 - 1.3 Structure of the Work
2. Background
 - 2.1 Current approaches & challenges
3. State-of-the-Art and Related Work
 - 3.1 Non-Diffusion based approaches
 - 3.2 Diffusion based approaches
 - 3.3 Current shortcomings
4. Proposed Method and Implementation
 - 4.1 Implementation of described approaches
 - 4.2
 - 4.3 Final Application
5. Evaluation
 - 5.1 Evaluation Method
 - 5.2 Comparison to prior work
 - 5.3 Use Cases
6. Results
7. Discussion
 - 7.1 Threats to Validity
 - 7.2 Future Work

REFERENCES

- [1] Diego Martin Arroyo, Janis Postels, and Federico Tombari. Variational transformer networks for layout generation, 2021.
- [2] Niklas Deckers, Julia Peters, and Martin Potthast. Manipulating embeddings of stable diffusion prompts, 2023.
- [3] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschan, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST '17, page 845–854, New York, NY, USA, 2017. Association for Computing Machinery.
- [4] Maria Douneva, Rafael Jaron, and Meinald T. Thielsch. Effects of Different Website Designs on First Impressions, Aesthetic Judgements and Memory Performance after Short Presentation. *Interacting with Computers*, 28(4):552–567, 06 2016.
- [5] Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry Davis, Vijay Mahadevan, and Abhinav Shrivastava. Layouttransformer: Layout generation and completion with self-attention, 2021.
- [6] Mude Hui, Zhizheng Zhang, Xiaoyi Zhang, Wenxuan Xie, Yuwang Wang, and Yan Lu. Unifying layout generation with a decoupled diffusion model, 2023.
- [7] Zhaoyun Jiang, Shizhao Sun, Jihua Zhu, Jian-Guang Lou, and Dongmei Zhang. Coarse-to-fine generative modeling for graphic layouts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):1096–1103, Jun. 2022.
- [8] Xiang Kong, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. Blt: Bidirectional layout transformer for controllable layout generation, 2022.
- [9] Hsin-Ying Lee, Lu Jiang, Irfan Essa, Phuong B Le, Haifeng Gong, Ming-Hsuan Yang, and Weilong Yang. Neural design network: Graphic layout generation with constraints, 2020.
- [10] Luis A. Leiva, Morteza Shiripour, and Antti Oulasvirta. Modeling how different user groups perceive webpage aesthetics. *Universal Access in the Information Society*, 22(4):1417–1424, Nov 2023.
- [11] Akshay Gadi Patil, Omri Ben-Eliezer, Or Perel, and Hadar Averbuch-Elor. Read: Recursive autoencoders for document layout generation, 2020.
- [12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- [13] Yuxi Xie, Danqing Huang, Jinpeng Wang, and Chin-Yew Lin. Canvasemb: Learning layout representation with large-scale pre-training for graphic design. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4100–4108, October 2021.