DESIGN INTENTS DISENTANGLEMENT

A Multimodal Approach for Grounding Design Attributes in Objects

MANUEL LADRON DE GUEVARA[1], ALEXANDER SCHNEIDMAN[2], DARAGH BYRNE[3] and RAMESH KRISHNAMURTI[4]
[1,2,3,4]Carnegie Mellon University, USA.
[1]manuelr@andrew.cmu.edu, 0000-0002-4585-3213
[2]amschnei@andrew.cmu.edu, 0000-0002-3788-1193
[3]daraghb@andrew.cmu.edu, 0000-0001-7193-006X
[4]ramesh@andrew.cmu.edu, 0000-0002-6327-8286

Abstract. Language is ambiguous; many terms and expressions convey the same idea. This is especially true in design fields, where conceptual ideas are generally described by high-level, qualitative attributes, called design intents. Words such as "organic", sequences like "this chair is a mixture between Japanese aesthetics and Scandinavian design" or more complex structures such as "we made the furniture layering materials like a bird weaving its nest" represent design intents. Furthermore, most design intents do not have unique visual representations, and are highly entangled within the design artifact, leading to complex relationships between language and images. This paper examines an alternative design scenario based on everyday natural language used by designers, where inputs such as a minimal and sleek looking chair are visually inferred by algorithms that have previously learned complex associations between designs and intents—vision and language, respectively. We propose a multimodal sequence-to-sequence model which takes in design images and their corresponding descriptions and outputs a probability distribution over regions of the images in which design attributes are grounded. Expectedly, our model can reason and ground objective descriptors such as black or curved. Surprisingly, our model can reason about and ground more complex subjective attributes such as rippled or free, suggesting potential regions where the design object might register such vague descriptions. Link to code: https://github.com/manuelladron/codedBert.git

Keywords. Natural Language Processing; Multimodal Machine Learning; Design Intents Disentanglement; SDG 9.

1. Introduction

Language can be ambiguous and similar ideas can be expressed in many different expressions. This is especially true in design fields, where conceptual ideas are generally described by high-level, qualitative attributes, called design intents. Even

POST-CARBON, Proceedings of the 27th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2022, Volume 1, 333-342. © 2022 and published by the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), Hong Kong.

though these descriptors are highly used in everyday language by designers—"the dining table should look more organic", "this chair is lightweight and minimal"—, they have complex visual associations due to partial subjective and conceptual components and thus, finding visual representations is a challenge. While humans might be able to identify design intents from an image of a chair with attributes such as "organic" or "minimalist" and differentiate between a "heavyweight" and a "lightweight" stand-lamp, they might also face challenges differentiating design intents such as "dynamic", "organic" or "functional". Current machine learning literature is unable to recognize these types of high-level attributes but has potential to understand them. Resolving such task would have a major impact in design communities, opening new scenarios where natural human language could directly be used in the process of design.

For computational linguistics, resolving this problem can challenge the status of theoretical understanding, problem-solving methods, and evaluation techniques (Alm, 2011). For computer vision, this presents a complex challenge of disentangling qualitative attributes—sleek, elegant, minimal—from images. Beyond its relevance in pushing machine learning research boundaries, this would significantly impact creative practice—designers, architects, and engineers. Real-time design intents understanding could open new design scenarios (e.g., voice-assisted natural language input), that reduce procedures based on intent reinterpretation as imperative commands—move, circle, radius, extrude, vertical—required by digital design engines, like AutoCAD or Rhinoceros. Such methods require a lengthy sequential process of deterministic commands that manually shape the design object.

Research on identifying high-level attributes has been done in other tasks. For instance, for selecting font types using attributes by (O'Donovan et al., 2014), or for search tasks on fashion objects with relative attribute feedback by (Kovashka et al., 2012).

In this work we aim to ground such relationships between modalities. We expand upon FashionBert (Gao et al., 2020), a framework that tackles a similar problem of disentangling design elements. The fashion descriptions from their data, however, are purely focused on the designs themselves, and most descriptions are objective rather than conceptual. In addition, we modify their image encoding strategy to make our model completely agnostic to the object itself. Finally, we adopt a modified token masking scheme to place more weight on masking adjective tokens because most keywords in design intents, such as "minimal" or "organic" are adjectives, and we hypothesize that these should provide a better training signal with respect to using the visual modality to infer language.

2. Related Work

This section firstly reviews work related to high-level attributes and design. We then review prior work done using the CODED dataset.

2.1. HIGH-LEVEL ATTRIBUTES

Research in understanding the relationships between high-level attributes and objects has not received much attention in comparison with objective or quantitative attributes. Some previous works have focused on image composition, particularly on high-level attributes of beauty, interest, and memorability and some authors described methods to predict aesthetic quality of photographs. (Datta et al., 2006) represent concepts such as colorfulness, saturation or rule-of-thirds with designed visual features, and evaluated aesthetic rating predictions on photographs. (Li & Chen, 2009) used the same approach for impressionist paintings. (Gygli et al., 2013) predicted human evaluation of image interestingness, building on work by (Isola et al., 2011), who uses various high-level features to predict human judgements of image memorability. Similarly, (Borth et al., 2013) performed sentiment analysis on images using object classifiers trained on adjective-noun pairs.

Object attributes have been explored for image search using binary attributes in works by (Kumar et al., 2011; Tao et al., 2009). Other work for searching interfaces has been done by (Parikh & Grauman, 2011), which estimates relative attributes. Whittlesearch (Kovashka et al., 2012) allows searching image collections using also relative attributes. In the following year (Kovashka & Grauman, 2013) improved their technique by using an adaptive model.

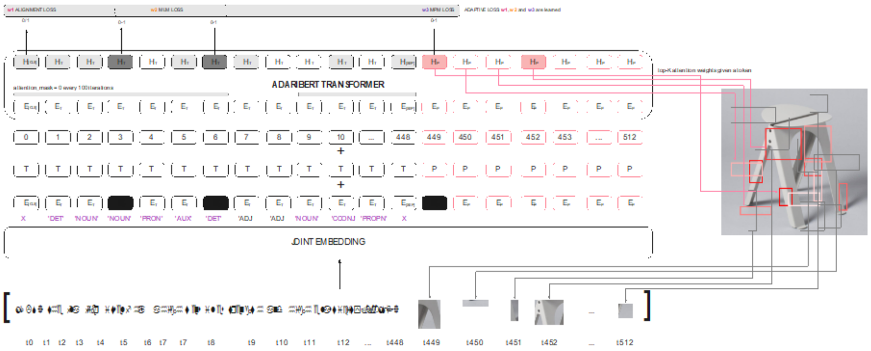

Figure 1. Overview of the CODEDBERT model. All images subjected to copyright. Used here for research purpose. Image: Dumbo, by Studio Pesi

2.2. PRIOR WORK ON THE CODED DATASET

Preliminary explorations that visually disambiguate vague terms in the context of design have been done by (Ladron de Guevara et al., 2020). The authors use a multimodal approach that combines a pretrained convolutional neural network to get the representation for images with general word indexes into a common joint subspace. A bidirectional Long Short-Term Memory (biLSTM) decoder—which models the labels co-occurrence information—learns semantic relationships between words and images.

To our knowledge, our work is the first attempt to scale work on high-level attributes on very complex unsupervised scenarios, where images do not have ground truth associating descriptors. Furthermore, CODED's language modality does not come from third party workers annotating the data, but the very own natural description from professional designers, which is a key element to integrating this knowledge and comprehending designs more naturally. While (Gao et al., 2020) explore similar

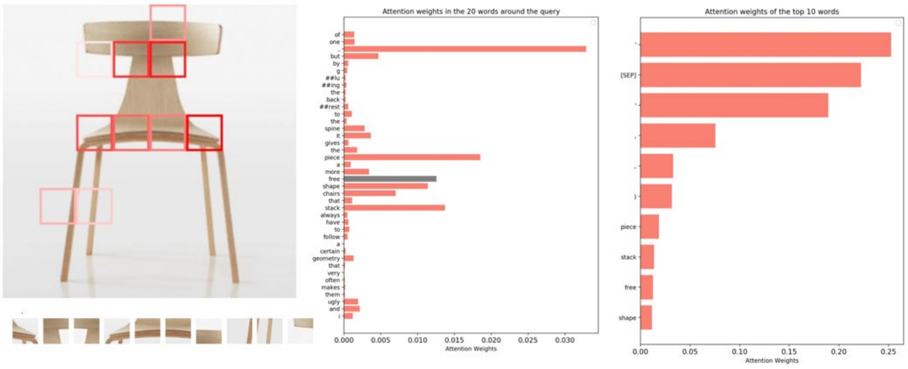

Figure 2. CODEDBERT at inference. (Left) Attention windows for the design intent "free"; higher color intensity corresponds to higher attention. (Middle) Attention weights for 20 neighboring words. (Right) Attention weights for top 10 words.

problems, their dataset only contains valid descriptors and most of them are objective attributes, our approach differs in two main aspects: unstructured language modality and high-level attributes. This motivates different masking and patching schemes.

3. Method

3.1. BERT

BERT was introduced by (Devlin et al., 2018) and uses the Transformer architecture (Vaswani et al., 2017) with a word-piece approach that divides words into tokens to be fed into the model. Every subword is projected into a set of embeddings E, and each embedding in E is computed as a sum of token embeddings particular to each subword. Each segment embedding indicates the part of the text that comes from, and a position embedding encodes the position of each token. This is fed into the multilayer BERT transformer, which generally has 12 or 24 layers, and outputs contextualized representations of each token. BERT is generally trained in two steps, pre-training and fine-tuning, where the former step is done using a combination of two objectives, mask language modeling and next sentence prediction. The latter step generally applies BERT to a particular task using different objectives according to the training task.

3.2. CODEDBERT

Our approach is to use the self-attention mechanisms of the BERT model to address two main tasks: (1) ground high-level attributes in images in an object-agnostic approach and (2) use within-language attention mechanisms as means to find relevant parts of our unstructured text and filter out those sentences that do not relate to the images, such as contextual information.

In addition to the text components in BERT, we introduce a set of visual tokens to model an image and learn joint representations of design intents and design images. The CODEDBERT model is illustrated in Figure 1. Generally, vision-and-language BERT models like VisualBert or VilBert use object-detection methods to extract objects within the image and pass the entire isolated objects through the multimodal Transfomer. Our research differs from such methods in a fundamental way: we build an object-agnostic model that disentangles design intents. For this approach, we propose two strategies to process the images. Following the FashionBert model, we resize images to 64x64 pixels and break them into 64 patches of 8x8 pixels each. We pass this sequence of patches along with a position embedding which theoretically gives the model the option to reconstruct the entire image. We called this strategy normal patching (NP). With the intention of destroying the object representation completely, our second approach takes a step further and produces random patches of positions and sizes, with a minimum of 4 and a maximum of 16 pixels in width and height. In this case, since the patches are randomly generated, there is no explicit order, and therefore, we do not use any position embeddings. We called this approach random patches (RP)—see Figure 4. The model is then tasked with grounding these patches with the design intent in the original description. We use the publicly available transformer library by HuggingFace, as a backbone for our implementation. Link to code: https://github.com/manuelladron/codedBert.git

3.2.1. Training CODEDBERT

Our training schema for CODEDBERT consists of three main objectives:

Ground Masked Language Modeling (MLM): This is a regular BERT training task in which text tokens, encoded using the wordpiece strategy, are masked with a probability of 15% and the model must minimize the negative log likelihood of predicting the original mask token, using surrounding language tokens and vision patch tokens. Given a sequence of text tokens $t_i = t_1, t_2, \dots, t_N$ the masked-out sequence is denoted by $\mathrm{t_k = t_1, t_{MASK}, \dots, t_K}$. The last layer of the transformer model output is fed into a linear classifier with vocabulary size the standard BERT model. The objective is defined as follows:

$$\mathrm{l_{MLM}(\theta) = -E_{t\sim D} \log P\,(t_i | t_k, \theta)}$$

where θ corresponds to the CODEDBERT parameters, D is the training set and $\mathrm{P(t_i | t_{/i})}$ is the probability of the masked-out token $\mathrm{t_i}$ predicted by the model given the rest of the language and vision tokens.

We propose a variant of the MLM approach that focuses on masking those words that their part of speech tag corresponds to adjectives with a probability of 13.5%, while masking non-adjectives with probability of 1.5%. We hypothesize that most design intents are captured by adjectives, and this strategy spends more training time predicting potentially more useful words for design grounding. Use within-language attention mechanisms as means to find relevant parts of our unstructured text and filter out those sentences that do not relate to the images, such as contextual information.

Text-Patches Alignment (TPA): We pre-process each sample in the dataset by pairing each image with a random text sample with a uniform distribution over the dataset. The model must predict whether any given text-image sample is paired or not. To do that, we use the pooled output from the BERT model, which is a dense representation of the [CLS] token of the entire sequence and pass it through a binary linear classifier. TPA is trained using binary cross entropy loss as follows:

$$l_{alig} - \frac{1}{n}\sum_{i=1}^{n} y \log\big(P(\hat{y})\big) + (1-y)\log\big(1-P(\hat{y})\big)$$

Masked Patch Modeling (MPM): Like the MLP task, we randomly mask out patches with probability 10%, setting the image encoder features to zero. We treat the output features as distributions over the original patches' features, and the model tries to minimize the KL divergence between the true patch features and the output masked-out features by:

$$l_{MPM}(\theta) = E_{KLp\sim D}\big(Distr.(p_i|p_k,\theta)\big|Distr.(p_i)\big)$$

3.2.2. Adaptive Loss

Following the fashionBERT model training strategy, we employ an adaptive loss algorithm to learn each of the three weights corresponding to each loss. Given the initial total loss function as

$$\mathcal{L}(\theta) = \sum_{i=1}^{L} w_i\, l_i(\theta)$$

where $L = 3$.We let the model learn the weights w_i as a new optimal problem:

$$argmin - \frac{1}{2}\sum_{i=1}^{L}\left||w_i \nabla l_i|\right|^2 + \frac{1}{2}\sum_{i,j=1}^{L}\left||w_i - w_j|\right|^2$$

$$s.t \sum_{i=1}^{L} w_i = 1 \text{ and } \exists\, w_i$$

This formulation aims to minimize the total loss while fairly treating the learning of all tasks. Considering the Karush-Kuhn-Tucher (KKT) conditions, we obtain the solution to w_i as:

$$w_i = \frac{\left(L - \nabla l_i^2\right)^{-1}}{\sum_{i=1}^{L}(L - \nabla l_i^2)^{-1}}$$

4. Experimental Setup

The CODED dataset is unique in that the text modality is not annotated by third-party workers but it is indeed leveraged from design conversations about the work in question. On the one hand this is an excellent resource for understanding design intents, as the dataset contains the original design intents rather than external labels. The downside is that the text includes rejections, negations and contextual information

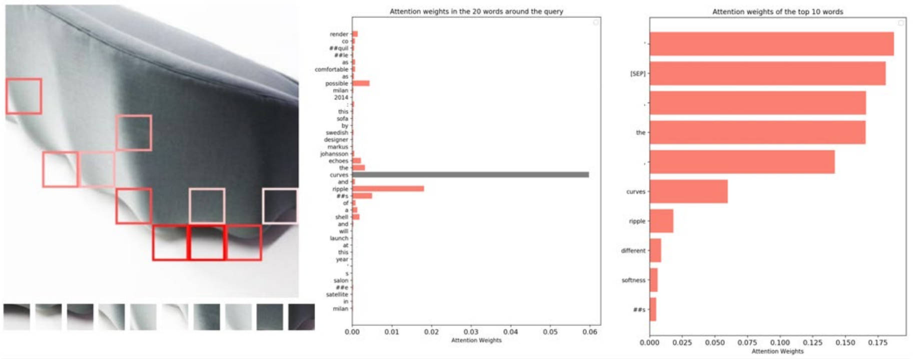

Figure 3. CODEDBERT at inference. (Left) Attention windows for the design intent "curves"; higher color intensity corresponds to higher attention. (Middle) Attention weights for 20 neighboring words. (Right) Attention weights for top 10 words.

about the designer's philosophy that can be vaguely applied to the objects we see in images. Our research examines whether it is possible to ground design intents under an object-agnostic schema, and therefore, disentangle such high-level attributes from the objects in the image modality.

4.1. DATASET AND INPUT MODALITIES

To address the issue of disentangling design intents in the context of creative practice, we use the CODED dataset, first presented in (Ladron de Guevara et al., 2020). The self-annotated CODED dataset contains a total of 33,230 samples of contemporary creative works represented by 264,028 raw sentences—provided by the original creators and by art curators—that describe 241,982 images. This dataset was assembled by collecting articles that include editorial descriptions along with associated images of the creative visual work. CODED is an organized dataset divided into seven categories: "architecture, art, design, product design, furniture, fashion" and "technology". CODED is the first dataset of pairs of images and language that, besides containing objective information of the elements in the images such as "wooden chair" or "black table", focuses on high-level attributes that correspond to design intents, such as "minimal, elegant and sleek looking chair". CODED also contains contextual information that can or cannot be indirectly applied to the images, and other more complex structures such as analogies. Note that there are ground truth labels in each image pair, but they are not indexed.

For all the experiments shown in this paper, we work with the furniture domain within the CODED dataset, and we use word modality for all our experiments. The Furniture-CODED dataset contains 17,532 images of contemporary workpieces.

4.2. MULTIMODAL BASELINE MODELS

We test 4 main experiments and provide a series of ablation studies to compare their performance. Namely, these experiments are:

Normal Patching and Normal Masking. The image is equally divided into 64 patches of 8x8 pixels, and we apply a normal common language masking strategy. The features of the patches are extracted using a pretrained ResNet50.

Random Patching and Normal Masking. The image is randomly divided into 64 patches of varying dimensions. We apply common masking and use pretrained ResNet50 as feature extractor.

Normal Patching and Adjective Masking. The images are divided by the NP approach, and we use the adjective masking approach defined in section 3. To extract the features of the patches, we fine-tuned ResNet152 trained on CODED under a contrastive learning strategy for cross-modal retrieval tasks. We also experiment with and without image-attention only by enforcing the attention mask every 100 iterations. In our model, the language modality is defined by a sequence of 448 tokens, whereas the image modality contains only 64. By looking only at the image attention weights, enforce the model to predict the masked tokens and alignment only from the visual context.

Random Patching and Adjective Masking. Like the prior experiment but using the RP approach.

4.3. IMPLEMENTATION DETAILS

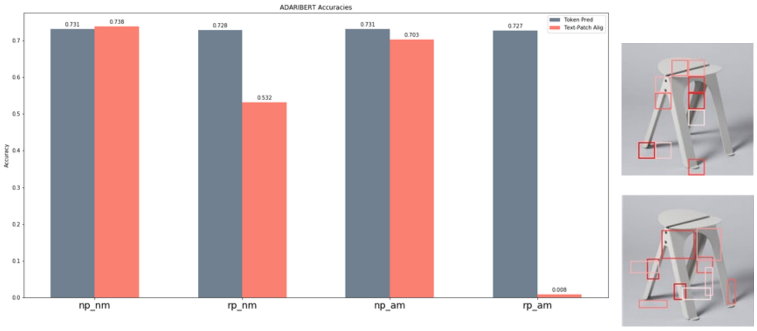

Figure 4. Accuracy between models (left). Random vs Normal patches approach (right)

We test the performance of our CODEDBERT model on two tasks, text and image alignment and masked token prediction. For text-patch alignment and token prediction, we employ accuracy as a metric to test our experiments, on a subset of 1,000 samples drawn from the test set.

We use a pre-trained 12-layers Bert-Base uncased model, which has 12 attention heads, and each hidden unit has 768D, with a total sequence length of 512. We set the text sequence to 448D, leaving the remainder 64D for the image input. For each patch, we discard the last fully connected layer of the ResNet model to have a 2,048-dimensional feature representation. We train the models for 5, 10 and 20 epochs. We discard results from the 20 epochs approach due to high overfitting issues. We set batch size to 4 due to VRAM memory constraints. We use p2.xlarge for training and g4dn.xlarge for evaluation. We use Adam optimizer with a learning rate of 2e-5, β_1 =

0.95, $\beta_2 = 0.999$, weight decay of 1e-4 and a linear schedule with warmup set to 5,000 steps. We clip gradients at 1.0.

5. Results and Discussion

We compare our four main experiments. The notation for our ablation studies is as follows: our baseline is defined by equal or normal patches and normal masking (NPNM), and it is compared against random patches and normal masking (RPNM), equal patches and adjective masking (NPAM), and random patches and adjective masking (RPAM).

We measure the performance of our models with the accuracy metric, as seen in Figure 4. For masked token prediction, slicing the image into equal patches yields slightly better results than using the random patches approach. Our best performing model in this task is our baseline (NPNM) with an accuracy of 73.15% followed very closely by NPAM with 73.10% accuracy. Likewise, the random patches approach yields almost same performance. We hypothesize that for predicting masked language tokens, the model is agnostic to the vision modality, ignoring the slicing approach. This might be due to the imbalance in the sequence length of the two modalities, or that the attention mechanism is not as strong in the vision modality as it is in the language part. We observe that the fact that skewing the masking towards adjectives and masking from a uniform distribution across the text sequence does not impact significantly in the performance of the language token prediction task.

In the alignment task, however, the slicing approach has a significant impact. We see how our baseline outperforms the rest of the ablative experiments with a 73.80% accuracy. The equal patches approach is very superior to the random patches approach. RPNM performs near to random guessing with a 53.20% accuracy, while that RPAM significantly underperforms a random classifier with a 0.8% accuracy. A reason for such low performance of the random patch scheme is that random patches do not assure covering the entire distribution of the pixels in the images, contrary to the equal patch strategy. The random patches might be scattered about a noisy background, providing a weak signal.

6. Conclusion and Future Work

We developed a transformer-based model called CODEDBERT, a joint model for image and text descriptions to address two main tasks: grounding design intents in images under an object-agnostic approach and finding the meaningful parts of the lengthy and noisy text leveraging the use of the attention mechanisms. Figure 2 and Figure 3 show how our model reasons between a chair and corresponding intent "free" and a couch and the attribute "curves", respectively. Figure 2 (left) visualizes with higher intensity those regions of the chair that are more likely to ground the design intent "free". As difficult as this task is, the model focuses the attention on the unusual kinks and convex parts of the chair. Likewise, Figure 3 correctly associates the attribute "curves" in the image.

A challenging and unique aspect of CODED is the descriptions are from long-form interviews with designers. Thus, a large portion of a given description may not be relevant to the design. We had hoped BERT could learn to distinguish which parts are relevant, but our training tasks may not have encouraged this enough. One possible direction would be to take a hierarchical approach, encoding entire sentences into single embeddings which are then passed through attention layers. This could act as a filtering mechanism to determine which sentences should be focused on.

References

Alm, C. O. (2011). Subjective natural language problems: Motivations, applications, characterizations, and implications. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 107–112.

Borth, D., Ji, R., Chen, T., Breuel, T., & Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. *Proceedings of the 21st ACM International Conference on Multimedia*, 223–232.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. *European Conference on Computer Vision*, 288–301. https://doi.org/10.1007/11744078_23

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810*.04805.

Gao, D., Jin, L., Chen, B., Qiu, M., Li, P., Wei, Y., Hu, Y., & Wang, H. (2020). Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2251–2260.

Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., & Van Gool, L. (2013). The interestingness of images. *Proceedings of the IEEE International Conference on Computer Vision*, 1633–1640.

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? *CVPR 2011*, 145–152.

Kovashka, A., & Grauman, K. (2013). Attribute adaptation for personalized image search. *Proceedings of the IEEE International Conference on Computer Vision*, 3432–3439.

Kovashka, A., Parikh, D., & Grauman, K. (2012, June). WhittleSearch: Image Search with Relative Attribute Feedback. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Kumar, N., Berg, A., Belhumeur, P. N., & Nayar, S. (2011). Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10), 1962–1977.

Ladron de Guevara, M., George, C., Gupta, A., Byrne, D., & Krishnamurti, R. (2020). *Multimodal Word Sense Disambiguation in Creative Practice.* http://arxiv.org/abs/2007.07758

Li, C., & Chen, T. (2009). Aesthetic visual quality assessment of paintings. *IEEE Journal on Selected Topics in Signal Processing*, 3(2), 236–252. https://doi.org/10.1109/JSTSP.2009.2015077

O'Donovan, P., Libeks, J., Agarwala, A., & Hertzmann, A. (2014). Exploratory Font Selection Using Crowdsourced Attributes. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 33.

Parikh, D., & Grauman, K. (2011). Relative attributes. 2*011 International Conference on Computer Vision*, 503–510.

Tao, L., Yuan, L., & Sun, J. (2009). Skyfinder: attribute-based sky image search. *ACM Transactions on Graphics (TOG)*, 28(3), 1–5.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Transformer: Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.