Generating Compositional Color Representations from Text

Paridhi Maheshwari*
Stanford University
paridhi@stanford.edu

Nihal Jain*
Carnegie Mellon University
nihalj@cs.cmu.edu

Praneetha Vaddamanu
Adobe Research
vaddaman@adobe.com

Dhananjay Raut*
Adobe
raut@adobe.com

Shraiysh Vaishay*
Advanced Micro Devices Inc.
shraiysh.vaishay@amd.com

Vishwa Vinay
Adobe Research
vinay@adobe.com

ABSTRACT

We consider the cross-modal task of producing color representations for text phrases. Motivated by the fact that a significant fraction of user queries on an image search engine follow an (attribute, object) structure, we propose a generative adversarial network that generates color profiles for such bigrams. We design our pipeline to learn composition - the ability to combine seen attributes and objects to unseen pairs. We propose a novel dataset curation pipeline from existing public sources. We describe how a set of phrases of interest can be compiled using a graph propagation technique, and then mapped to images. While this dataset is specialized for our investigations on color, the method can be extended to other visual dimensions where composition is of interest. We provide detailed ablation studies that test the behavior of our GAN architecture with loss functions from the contrastive learning literature. We show that the generative model achieves lower Fréchet Inception Distance than discriminative ones, and therefore predicts color profiles that better match those from real images. Finally, we demonstrate improved performance in image retrieval and classification, indicating the crucial role that color plays in these downstream tasks.

CCS CONCEPTS

• **Information systems** → **Query intent**; *Content ranking*; *Novelty in information retrieval*.

KEYWORDS

Color Representation, Visual Attention, Composition and Context, Contrastive Learning, Color in Ranking

*Work done when authors were at Adobe Research

1 INTRODUCTION

We consider the problem of cross-modal retrieval where textual queries are used to produce a ranked list of images. User queries can be very diverse and exhibit a range of linguistic structures. In addition, the richness and ambiguities of language make the accurate retrieval of images a challenging task. The ranker needs to incorporate multiple relevance indicators for queries as well as items. In this paper, we place constraints on both these aspects - we focus on (a) queries with an attribute-object bigram structure, and (b) the role of color as a relevance ranking feature.

Attribute-Object pairs are a commonly observed structure in search queries, e.g., 'cute dog' or 'happy child'. Since we consider the domain of image retrieval, we are interested in how these phrases manifest visually. The characteristics of the combined phrase are a composition of the visual intent of the attribute (or adjective) and the object (or noun). Modeling the constituent terms separately, and knowing how to combine them, helps us generalize to new concepts [30], e.g., 'happy dog'. Such an improved understanding of the queries enables the building of a more robust search ranker. In addition, we might be able to support higher-level intents formed by the same building blocks, such as complex composite queries like 'red bricks on a white background'.

Specifically, we focus on the impact of attribute-object compositionality on color. Our focus is motivated by the central role that color plays in image processing and retrieval [15, 17, 38]. In addition, there is an inherent color intent associated with several user queries [24]. For example, while queries such as 'raw apple' and 'deep sea' do not have explicit color mentions, they evoke specific color semantics that may be useful to infer image relevance. We take steps towards this by modeling attribute-object compositionality in user queries to generate color representations.

Modeling compositionality requires that the dataset cover the set of combinations. That is, for N_a attributes and N_o objects, the set of (attribute, object) bigrams in the dataset should ideally include the all $N_a \times N_o$ combinations with sufficient image examples. The individual terms might have varying degrees of color intent (e.g. 'dark' versus 'happy' as attributes). The corresponding bigrams might therefore also vary in how they impact color (e.g. 'dark sea' > 'dark sky' > 'happy sky'), but might also lead to unlikely pairs (e.g. 'happy sea'). To handle these scenarios, we design a novel pipeline to curate a dataset of images with their corresponding (attribute, object) pair labels. Our approach enables us to capture a diverse set of phrases with high color intent. Our dataset creation strategy could be potentially useful for other studies into compositionality along specific axes (in our case, color). Our dataset construction

includes a visual attention mechanism to ground the (attribute, object) pairs into images. This enables us to extract cleaner and more relevant color profiles from images, conditioned on the bigram.

We propose an adversarial learning approach [11] for generating color representations compositionally from textual input. Inspired by its recent success in several domains, we experiment with different loss functions from the contrastive learning paradigm for training our GAN. The generator takes word embeddings of the attribute and object as input, this aids the model in generalizing in the text modality. Since the perception of color is naturally rooted in the visual domain, we wish to provide the model with information from images even though our end-task only utilizes textual input for generating color profiles. To enable this, our model takes the image modality as input while training the discriminator. The coupled training between the generator and the discriminator, therefore, leads to the visual modality affecting the final model, while the generator utilizes only text at test time.

Our evaluation strategy is trifold: (i) assess the quality of the generated color representations for input (attribute, object) pairs; (ii) the quality of our conditional GAN architecture and training pipeline; and (iii) the effectiveness of introducing color features in downstream tasks. First, we compare the performance of our text-to-color encoder against a discriminative baseline that predicts color profiles from (attribute, object) word embeddings. We evaluate the models using $L2$ distances between the generated and ground-truth color profiles and show that our generative model outperforms the baseline. Second, we compute the Fréchet Inception Distance [14] between the predicted and real color profiles to evaluate the quality of our generative model and discuss a series of ablation experiments to study the effectiveness of components of the GAN objective. Finally, we demonstrate the usefulness of color as a feature for cross-modal retrieval and image classification.

We summarize our key contributions as follows:

(1) We present a strategy to prepare text-to-color datasets from existing public sources. The text phrases are limited to (attribute, object) bigrams to study color compositionality.

(2) We propose a generative adversarial modeling approach to produce color representations of textual phrases. This model takes visual cues from the image modality at train time but does not require these image features at test time.

(3) We perform a comparative study of loss functions adapted from contrastive learning literature for the task of text-to-color encoding using GANs.

(4) Finally, by using the generated color representations of textual queries as features for image ranking, we demonstrate that search relevance can be improved.

2 RELATED WORK

In this section, we review prior work related to the compositional structure of language and obtaining color representations from text.

2.1 Language and Color

The richness of language in being able to describe complex visual features has been a long-studied subject [41]. Various psychological studies [42, 52] have also demonstrated the strong association between colors and natural language phrases. Certain words like water and rose exemplify this association.

To study this relation between color and language, several datasets have been curated that provide a mapping between the modalities. The XKCD dataset [33] labels textual phrases with colors, and was setup via a crowd-sourced survey. [51] use this dataset to learn a probability distribution in HSV color space conditioned on the name of the color. [44] also employ the XKCD dataset along with Google's n-gram corpus [28] to prepare phrases with high color association. Google's n-gram corpus was also used by [22] to select commonly used single words for use in the task of color palette extraction. We also rely on Google's n-gram dataset for finding color related words, as described in section 3, but we describe a new mechanism to generate textual queries that not only have high color intent but also have an (attribute, object) structure.

Several research efforts have attempted to arrive at mappings between the text and color. [19] use a character-level LSTM to predict a color given a name. Other works [1, 22] focus on arriving at color palettes from textual input. In contrast, we focus on the task of generating color histograms that convey much richer information and have greater utility in a cross-modal search as an additional feature embedding. While [24] also focus on generating color histograms from text, our work focuses specifically on the compositional structure in language – which we argue is crucial for generating relevant color representations. We propose a generative model for text to color, also utilizing the image modality which contains crucial information for color intent.

2.2 Composition and Context

According to the compositionality principle observed in language, novel concepts can be constructed from primitive building blocks. Following [21, 34, 39], we model compositionality by treating attributes and objects as primitives. This intuitive principle is closely connected with the principle of contextuality which states that the behaviour of a primitive varies in the presence of others [30]. Specifically, the same attribute can affect different objects in different ways and the same object elicits different behaviours when modified by different attributes. For example, 'ripe' when used in context of 'apple' has a different visual manifestation than when 'ripe' is used in context of 'mango'. Similarly, the object 'car' evokes different intuitions when it is modified by 'sporty' and 'old'. This interaction between compositionality and contextuality has recently been a subject of study in various fields. Specifically, in recent machine learning literature [34, 35, 39, 47], these concepts have formed the basis for zero-shot learning or few-shot learning where generalization is achieved by modeling compositions of primitives which are not part of the training set.

In the present work, we explore composing attributes and objects in textual queries to extract color representations that can be used for retrieving relevant images. We further demonstrate the use of our method to compose unseen combinations of attributes and objects to derive intuitive color representations.

2.3 Contrastive Learning

Contrastive learning approaches to representation learning have recently gained traction due to their success in several domains such

as computer vision and natural language processing [6, 8, 20, 23]. The intuition behind these approaches is to bring similar pairs of data points (typically referred to as the anchor and the positive) closer to each other than dissimilar pairs (anchor and negative) in an embedding space. This is achieved through non-linear transformations learnt using objectives like the triplet loss [43, 50] or the InfoNCE objective [36] amongst others. The authors of [35, 47, 49] make use of the contrastive learning framework to model attribute-object tasks by exploiting the compositional nature of the inputs to sample positives and negatives for an anchor data point. These loss functions above operate on a triple of data points - the anchor, positive, and the negative. [49] defines a quintuplet loss, an extension of the triplet loss that introduces intuitions from compositionality into the contrastive learning setting.

Building on these approaches, we adapt the contrastive learning framework to generate color representations that respect compositionality and context. This is achieved by using the notions of similar and dissimilar pairs as described in these works to drive our sampling strategies while training our machine learning models. As an indicator of the intuition behind these methods, for a given anchor example image for the (attribute, object) bigram (A, O), positive examples are all other images tagged with the same bigram. Negative examples can also be simply obtained as all images not associated with this bigram. However, bigrams that share either the attribute or the object (i.e., (A, X) or (Y, O)) are partially related. These relative preferences amongst images associated with these classes are naturally exploited by the contrastive loss formulations.

3 DATASET

We leverage image datasets labeled with (attribute, object) pair information to develop algorithms that generate color profiles from text. It is of primary importance to ensure a rich and diverse set of (attribute, object) phrases which are not limited to trivial color mentions (such as *'red scarf'* or *'blue sky'*), but also include implicit and inherent indicators (such as *'cranberry juice'* or *'deep sea'*). While there exist public datasets on object transformations [16, 54], they attend to variations in physical state and appearance (for example, *rope* can be *thin*, *short*, *coiled*). In the current work, we focus on one particular visual aspect, i.e., color, and propose a novel approach to curate datasets that specifically capture (attribute, object) phrases with high color intent.

3.1 Curation

We start by gathering the set of commonly occurring (attribute, object) phrases from textual n-grams. The bigram corpus from Google's n-gram dataset [28] contains the list of all contiguous sequences of 2 words present in the Google corpus along with their frequency count. Based on the linguistic type of the constituent words, we extract all phrases where the first word is an adjective (attribute) and second is a noun (object). To remove non-visual concepts (such as *'old wisdom'* or *'European community'*), we restrict our vocabulary using well known lists of concrete nouns [3] and descriptive adjectives [10]. This approach results in the set of frequently occurring visual concepts in public corpora.

Given our specific focus on color, we would like to exclude phrases assumed to have no color intent (such as *'epithelial cells'* or

'electric fields'). To achieve this, we build a bipartite graph between attributes and objects and utilize a hopping logic to iteratively select pairs. Starting with the 11 Basic Color Terms [2] as attributes, we obtain the list of objects that occur most frequently (top f) with this set of seed colors. In the next step, we identify the attributes that they most commonly occur alongside. This completes one traversal, termed as a single hop of the bipartite graph. The selection process repeats with multiple hops h till the required number of (attribute, object) pairs have been selected.

For the model to learn compositionality of attributes and objects, we need to ensure sufficient occurrences of every word and we achieve this by maintaining a threshold t_a and t_o for the number of unique attributes per object and unique objects per attribute respectively. Lastly, we fetch images for every (attribute, object) pair by querying the Google Image Search engine and retrieving the top results. The statistics of the final dataset are summarized in Table 1.

Table 1: Statistics of the (attribute, object, image) datasets created using the Google Bigrams corpus.

# attributes	130
# objects	211
# pairs	1460
# images per pair	33.55
# images	48983

We essentially use the distance from Basic Color Terms [2] in the bipartite graph as a proxy for color intent. For example, starting with color terms 'red' and 'blue' as attributes gives us the frequent pairs 'red rose', 'blue sea' and so on. Now, using 'rose' and 'sea' as seed input, we fetch the pairs 'wild rose', 'deep sea' and 'stormy sea'. As the number of hops increases, the phrases become more generic and less color-centric. This technique can be generalized to other visual properties such as texture, emotions, aesthetics – by choosing an appropriate seed set of adjectives/attributes.

3.2 Color Representation for Images

Computational pipelines for color leverage well-known models and representations [53]. In this work, the downstream application (image retrieval and classification) dictates the choice of color space and distance functions. Since our task involves human perception and interpretation, we utilize the LAB space which is known to be perceptually uniform – distances in LAB space correspond to similar visually perceived changes in color.

We divide the range spanned by the 3 axes uniformly to create discrete bins. And a given pixel is mapped to one of the bins. And finally, the image is represented as a histogram over the bins such that each bar in the histogram is proportional to the fraction of pixels belonging to that bin. Note that utilizing larger bin widths leads to image level histograms that are less sparse, but with the disadvantage of having lost the detail. The choice of bin sizes, therefore, needs to trade-off the informativeness of fine-grained representations with the more robust coarse discretizations.

The discretization itself introduces some noise into the representation, this can be partially alleviated by considering multiple

bin widths. Specifically, we utilize two choices for the number of bins along (L, A, B) axes respectively – (9, 7, 8) and (10, 10, 10). This gives us two separate histograms, each of sizes $9 * 7 * 8 = 504$ and $10 * 10 * 10 = 1000$ elements. Concatenating these leads to a combined 1504-dimensional color embedding for an image. The specific choice of bin widths and number of alternative discretizations are design choices. In the current paper, we show results for a standard configuration, focussing on the central problem of interest – the building of a generative text-to-color model.

Figure 1: Final set of 1504 color bins obtained by uniformly quantizing the LAB space. Note the repeating trend of the first 504 and the last 1000 bins, a result of concatenating histograms from two different LAB space divisions.
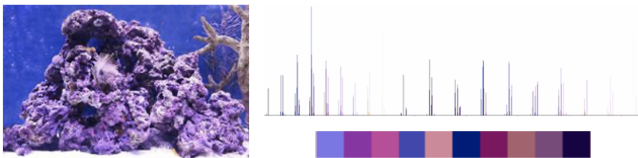

Figure 2: Sample image, its color histogram and palette. It is evident that the purple and blue bins have the highest peaks while colors like brown have fired in smaller contributions.

Figure 1 provides a visualization of the 1504 color bins and Figure 2 shows an image of *'coralline sea'* and its color histogram – the height of the bar represents the weight of the corresponding bin, and the color corresponds to its LAB value. For easier interpretation, we extract a representative palette from the histograms by clustering similar shades together and sampling from the result. This generates a diverse summary that captures the majority shades from the original histogram. We will use the palette as an intuitive visualization for the output of our models.

3.3 Modeling Visual Attention

We obtain images by querying Google's image search engine with the final set of unique (attribute, object) pairs, and utilize the technique described in the previous section to get color histograms for all images to be used for training our text-to-color models. This color representation gives uniform importance to all pixels in the image. However, conditioned on the text phrase, parts of the image may be more relevant than others, and we would like this intuition to affect the color representation appropriately. In order to identify relevant parts and extract cleaner color representations, we train a Convolutional Neural Network [18] on the classification task which internally uses visual attention to focus on parts of images. The model takes an image as input and predicts the attribute and object, while simultaneously learning an attention map over the image. We use the normalized attention weights from the trained model to give differential importance to individual pixels and create better color profiles. This is illustrated in Figure 3.
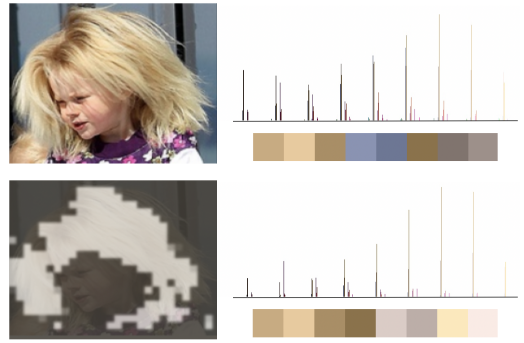

Figure 3: For the pair *'blond hair'*, we use attention map to identify relevant parts of the image and produce a less-noisy color representation with peaks towards blond and ignoring the blue from the irrelevant parts of the image.

The model architecture is summarized in Figure 4 (left). The backbone is a VGG-16 network [45] with max pooling, ReLU activation and no dense layers. Two attention modules are applied at different intermediate stages which learn a pixel-wise attention map over the image. The learnt attention weights and global features are finally average pooled to get the feature vectors. The concatenated features are passed through two different classifiers, one for predicting attribute and the other for object respectively. Each classifier is a fully connected layer that computes confidence scores for all candidate classes (set of all objects or attributes). The model is trained using cross-entropy loss on one-hot encoded labels for both attribute and object given an input image. Lastly, we extract the spatial attention map from this model and perform a pixel-wise multiplication to obtain weighted color representations.

The individual attention modules, shown in Figure 4 (right), are a function of both intermediate representations as well as global image features. After passing through separate convolution layers, the global features are upsampled using bilinear interpolation to align spatial size to that of the input image. This is followed by an element-wise addition with intermediate features to get an attention map. The output of the attention module is an attention weighted feature space, i.e., the pixel-wise product of the attention map and intermediate features.
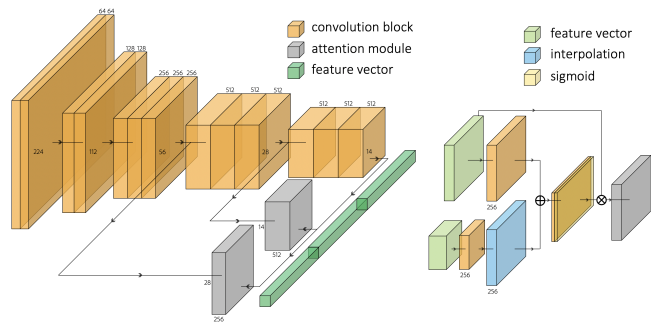

Figure 4: Overall network architecture for learning visual attention (left) and individual attention module (right).

4 GENERATIVE MODEL

Going from natural language phrases to color and being able to synthesize color profiles of unseen compositions is a challenging task. There have been prior efforts to learn this mapping from text modality alone [24, 32], but the perception of color is inherently rooted in the visual domain. To tackle this problem, we adopt a multimodal approach to learn color representations of (attribute, object) pairs, ensuring that images are required only for training, not for inference. We propose a generative model that learns compositionality and context in color space. The generator predicts plausible color profiles conditioned on the text embedding, while the discriminator attempts to distinguish between real color profiles (from images) and generator outputs. Our approach is motivated by the recent success of adversarial examples in zero-shot compositional learning for image classification [48, 49].

The generator network uses word embeddings to capture the initial context of the text, followed by fully connected layers with ReLU activation, and finally softmax to return the color embedding. It is key to note that we use different trainable embedding matrices for attributes and objects because the same word can have multiple interpretations based on its linguistic type, e.g., *sea* in *'deep sea'* is an object but plays the role of an attribute in *'sea blue'*). The image modality is only input to the discriminator network, and it provides feedback to the generator via the adversarial loss. The discriminator is another neural network that takes as input the text embeddings, pretrained image features, and a color profile; and predicts a real versus fake score between [0, 1]. The overall network architecture to map text to color profiles is shown in Figure 5, and individual components are detailed next.
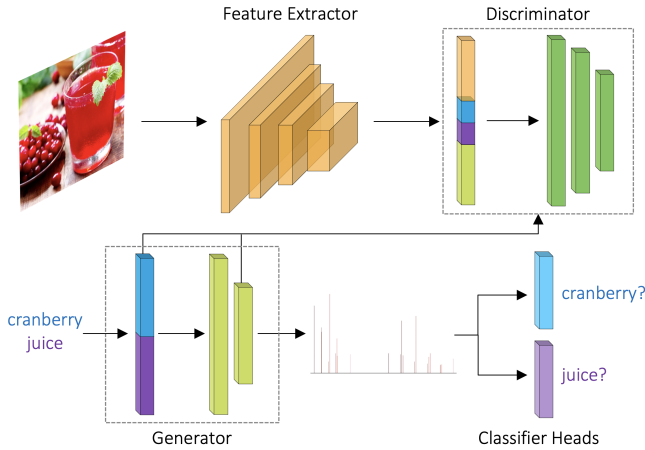

Figure 5: Generative adversarial network for learning color representations of (attribute, object) text phrases.

4.1 Training Objective

We train our GAN model using a modified Least Squares GAN objective [26]. Mathematically, the generator G and discriminator D objectives are given by

$$\mathcal{L}_G = \mathbb{E}_{\mathbf{a},\mathbf{o}}\left[\left(D\big(G(\mathbf{a},\mathbf{o})\mid \mathbf{a},\mathbf{o}\big)-1\right)^2\right] + \lambda_{\text{color}}\mathcal{L}_{\text{color}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}$$

$$\mathcal{L}_D = \mathbb{E}_{\mathbf{x}_{a,o}}\left[\left(D\big(\mathbf{x}_{a,o}\mid \mathbf{a},\mathbf{o}\big)-1\right)^2\right]$$
$$+ \mathbb{E}_{\mathbf{a},\mathbf{o}}\left[\left(D\big(G(\mathbf{a},\mathbf{o})\mid \mathbf{a},\mathbf{o}\big)\right)^2\right] + \lambda_{\text{mis}}\mathcal{L}_{\text{mis}}$$

where $D\big(\mathbf{x}_{a,o}\mid \mathbf{a},\mathbf{o}\big)$ represents the output score of the discriminator on seeing a real color profile, and $D\big(G(\mathbf{a},\mathbf{o})\mid \mathbf{a},\mathbf{o}\big)$ is for the generated color profile. Thus, the discriminator tries to minimize the score for a real color profile $\mathbf{x}_{a,o}$ which is sampled from the set of images corresponding to the composition $(\mathbf{a},\mathbf{o})$, and maximize the score of the output $G(\mathbf{a},\mathbf{o})$ by the generator. The generator, on the other hand, is trained to maximize the discriminator's score for its generated output.

Color Loss $\mathcal{L}_{\text{color}}$: This measures the distance between the predicted color representations and true ones from images, and is used to guide the training of the generator. We experiment with different contrastive losses which are described next. $\mathbf{x}_{\mathbf{a},\mathbf{o}}$ denotes the attention-weighted color profile sampled uniformly at random from the set of all images for class $(\mathbf{a},\mathbf{o})$, and $\hat{\mathbf{x}}_{a,o}$ is the model prediction $G(\mathbf{a},\mathbf{o})$.

(1) **L2 Loss**: This is a simple euclidean distance between the color profiles.

$$\mathcal{L}_{\ell 2}\left(\hat{\mathbf{x}}_{a,o}, \mathbf{x}_{a,o}\right) = \|\hat{\mathbf{x}}_{a,o} - \mathbf{x}_{a,o}\|_2$$

(2) **Triplet Loss**: Inspired from the widely-used contrastive paradigms in the vision community, triplet or margin loss [43, 50] takes a positive sample $\mathbf{x}_{a,o}$ of the same class and a negative one $\mathbf{x}_{\bar{a},\bar{o}}$, and tries to bring the anchor $\hat{\mathbf{x}}_{a,o}$ close to the positive and far from the negative.

$$\mathcal{L}_{\text{triplet}}\left(\hat{\mathbf{x}}_{a,o}, \mathbf{x}_{a,o}, \mathbf{x}_{\bar{a},\bar{o}}\right) =$$
$$\max\left(0, \mathcal{L}_{\ell 2}\left(\hat{\mathbf{x}}_{a,o}, \mathbf{x}_{a,o}\right) - \mathcal{L}_{\ell 2}\left(\hat{\mathbf{x}}_{a,o}, \mathbf{x}_{\bar{a},\bar{o}}\right) + m\right)$$

where the negative histogram $\mathbf{x}_{\bar{a},\bar{o}}$ is randomly sampled from any other class $(\bar{\mathbf{a}},\bar{\mathbf{o}})$, and m is the margin hyperparameter.

(3) **Quintuplet Loss**: This extends [49] the triplet loss by considering multiple task specific negatives. It considers one negative $\mathbf{x}_{\bar{a},\bar{o}}$ belonging to the $(\bar{\mathbf{a}},\bar{\mathbf{o}})$ class and two semi negatives $\mathbf{x}_{a,\bar{o}}$ and $\mathbf{x}_{\bar{a},o}$, which have either the same attribute $(\mathbf{a},\bar{\mathbf{o}})$ or the same object $(\bar{\mathbf{a}},\mathbf{o})$ as the anchor. The loss is a weighted sum of 3 triplet components, given by

$$\mathcal{L}_{\text{quintuplet}}\left(\hat{\mathbf{x}}_{a,o}, \mathbf{x}_{a,o}, \mathbf{x}_{\bar{a},\bar{o}}, \mathbf{x}_{a,\bar{o}}, \mathbf{x}_{\bar{a},o}\right) =$$
$$\lambda_1 \mathcal{L}_{\text{triplet}}\left(\hat{\mathbf{x}}_{a,o}, \mathbf{x}_{a,o}, \mathbf{x}_{\bar{a},\bar{o}}\right)$$
$$+ \lambda_2 \mathcal{L}_{\text{triplet}}\left(\hat{\mathbf{x}}_{a,o}, \mathbf{x}_{a,o}, \mathbf{x}_{a,\bar{o}}\right)$$
$$+ \lambda_3 \mathcal{L}_{\text{triplet}}\left(\hat{\mathbf{x}}_{a,o}, \mathbf{x}_{a,o}, \mathbf{x}_{\bar{a},o}\right)$$

where weight hyperparameters are such that $\lambda_1 > \lambda_2 = \lambda_3$.

This additional color loss in the generator's objective helps in combating mode collapse and stabilizing training.

Classification Loss $\mathcal{L}_{\text{cls}}$: The generator output is passed through two different classifier heads - one for attribute, and other object. Simultaneously, a cross-entropy loss component is added to the generator objective.

$$\mathcal{L}_{\text{cls}} = -\mathbb{E}_{\mathbf{a},\mathbf{o}}\left[\log P_{\mathbf{a}}\left(\mathbf{a}\mid \hat{\mathbf{x}}_{a,o}\right)\right] - \mathbb{E}_{\mathbf{a},\mathbf{o}}\left[\log P_{\mathbf{o}}\left(\mathbf{o}\mid \hat{\mathbf{x}}_{a,o}\right)\right]$$

where $P_{\mathbf{a}}$ and $P_{\mathbf{o}}$ denote the conditional probability of the respective classifiers making the right prediction. This is done to incorporate feedback from the closely-related task of color naming [27, 31, 32]. It has also been shown to improve the generator's ability to generalize over unseen compositions of (attribute, object) pairs.

Mismatch Loss $\mathcal{L}_{\mathbf{mis}}$: This term extends the conditional GAN loss [29] by encouraging the discriminator to classify mismatched combinations of generated color profiles and text inputs as fake [40]. Here, the discriminator minimizes the score given to the combination of the real color profile $\mathbf{x}_{a,o}$ and the mismatched composition $(\bar{\mathbf{a}}, \bar{\mathbf{o}})$, forcing the discriminator to explicitly identify class mismatch in addition to the traditional real/fake distinction.

$$\mathcal{L}_{\text{mis}} = \mathbb{E}_{\mathbf{a},\mathbf{o}} \left[\left(D\big(\mathbf{x}_{a,o} \mid \bar{\mathbf{a}}, \bar{\mathbf{o}}\big) \right)^2 \right]$$

Since our setup is a conditional GAN, the discriminator needs to evaluate the conditioning constraint of the generator's output on the input text. In turn, this feedback to the generator also ensures that the predicted color profiles are not only plausible but also correlated with the text.

Note that all loss components use the attention mechanism described before for color profiles obtained from images. We follow an alternate training strategy, wherein the generator is trained for K epochs, followed by discriminator training for K epochs, and so on. This switching is done to stabilize learning - giving both the networks sufficient iterations to train smoothly before the adversarial component drives training and further improves performance.

4.2 Results and Evaluation

We now describe our experimental setup, the baseline models, and present qualitative and quantitative evaluation of our approach.

Implementation Details: We set the hyperparameters in the dataset curation pipeline as $f = 10$, $t_a = 5$ and $t_o = 5$, and obtain the final set of (attribute, object) pairs by running the bipartite graph filtering for $h = 2$ hops. We split the set of all (attribute, object) pairs in the ratio 70 : 15 : 15 for training, validation and testing respectively. This entails that all images of a given class fall into the same set. Therefore, the compositions of the test set are never seen by the model and corresponds to zero-shot learning.

For the generator, we first embed the (attribute, object) pair into a trainable 300 dimension embedding using GloVe vectors [37]. This is followed by separate FC layers of size 400 each. The attribute and object embeddings are then concatenated and passed through another fully connected network with 1000 and 1504 hidden units respectively. We add dropout [46] to the first hidden layer, with the dropout rate set to 0.4. The embedding is finally normalized using softmax (at the resolution level, i.e., for the first 504 and last 1000 bins separately), resulting in the color representation. Both the attribute and object classifiers are a single linear layer with softmax activation to predict a probability distribution over the set of all attribute and object classes. The discriminator concatenates the 3 inputs - 800 length text conditioning, 1504 length colour representation, 2048 length image feature computed using a pretrained ResNet model [12]. It is passed through 2 hidden layers with ReLU activation of 4000 and 2000 units respectively. Lastly, a linear layer predicts a real/fake score for the conditioned input.

Figure 6: Generated color palettes. Rows 1,2 depict how an attribute alters different objects, while rows 3,4 show the same object with varying attribute types. Row 5-7 comprises of unseen combinations of (attribute, object) pairs. Row 8 illustrates complex color profiles with multiple color intents.

The end-to-end network is trained using RMSprop optimizer for 300 epochs and a batch size of 64. We use a learning rate of 10^{-4} for the generator and classifier, and 10^{-5} for the discriminator. The hyperparameters are set as follows for all involved experiments: $\lambda_{\text{cls}} = 0.05$, $\lambda_{\text{mis}} = 1$ in the overall objective; $\lambda_{\text{color}} = 0.6$ for L2 loss; $\lambda_{\text{color}} = 2$, $m = 1$ for triplet loss; $\lambda_{\text{color}} = 0.1$, $\lambda_1 = 1$, $\lambda_2 = 0.3$, $\lambda_3 = 0.3$ for quintuplet loss; and $K = 10$ for alternate training.

Baseline: We evaluate our model against a discriminative baseline that learns to compose color profiles from constituent word embeddings. The architecture of this model is the same as that of the generator, allowing for a fair comparison of the computed color representations. We refer to this baseline as *Label Embed* [34]. This network is trained using the ground-truth color profiles from images of the corresponding class, and $\mathcal{L}_{\text{color}}$ as the objective.

Evaluation: We evaluate the predicted color representations using the following metrics (i) *Macro L2* which measures the L2 distance between model predictions and average histogram across all instances of all (attribute, object) classes, (ii) *Micro L2* which measures the average L2 distance between model predictions and individual image histograms of the corresponding class and then averaged across classes, and (iii) *Fréchet Inception Distance (FID)* [14] to estimate the quality of our conditional GAN network. The FID is a comparison between statistics of the two distributions - color profiles generated by model versus real ones from images. We report these metrics at two levels - for seen (attribute, object) compositions of the training set and unseen compositions (zero-shot) of the test set. We also perform ablation experiments to study the effectiveness of different loss components while training the GAN.

The numbers are shown in Table 2 from which we can make the following observations. All variants of the generative model consistently outperform Label Embed for seen compositions. There is a drop in performance for unseen compositions, which is natural. Comparing across different $\mathcal{L}_{\text{color}}$ objectives, L2 loss leads to the lowest Macro and Micro L2 losses, primarily due to the parallels

Table 2: Evaluation of different text-to-color models on both seen and unseen compositions of (attribute, object) pairs.

Model		Seen Comp. (×10⁻³)		Unseen Comp. (×10⁻³)		FID
Network	$\mathcal{L}_{\text{color}}$	Macro L2	Micro L2	Macro L2	Micro L2	
Label Embed	$+\mathcal{L}_{\ell 2}$	0.244	0.865	0.233	0.847	0.615
	$+\mathcal{L}_{\text{triplet}}$	0.581	1.089	0.578	1.076	0.368
	$+\mathcal{L}_{\text{quintuplet}}$	0.624	1.111	0.623	1.098	0.366
Ours	$+\mathcal{L}_{\ell 2}$	0.155	0.775	0.355	0.936	0.391
	$+\mathcal{L}_{\text{triplet}}$	0.460	0.976	0.716	1.189	0.208
	$+\mathcal{L}_{\text{quintuplet}}$	0.575	1.059	0.800	1.262	0.233
Ours $-\mathcal{L}_{\text{mis}}$	$+\mathcal{L}_{\ell 2}$	0.162	0.791	0.360	0.949	0.383
	$+\mathcal{L}_{\text{triplet}}$	0.481	0.997	0.730	1.211	0.251
	$+\mathcal{L}_{\text{quintuplet}}$	0.578	1.061	0.759	1.228	0.225
Ours $-\mathcal{L}_{\text{cls}}$	$+\mathcal{L}_{\ell 2}$	0.130	0.759	0.340	0.917	0.428
	$+\mathcal{L}_{\text{triplet}}$	0.451	0.955	0.707	1.172	0.203
	$+\mathcal{L}_{\text{quintuplet}}$	0.560	1.009	0.805	1.230	0.205
Ours $-\mathcal{L}_{\text{mis}} - \mathcal{L}_{\text{cls}}$	$+\mathcal{L}_{\ell 2}$	0.137	0.763	0.335	0.915	0.445
	$+\mathcal{L}_{\text{triplet}}$	0.466	0.970	0.721	1.187	0.223
	$+\mathcal{L}_{\text{quintuplet}}$	0.557	1.023	0.801	1.238	0.229

in the training and evaluation metrics. But the importance of contrastive alternatives (Triplet and Quintuplet losses) is evident from the FID scores. The FID metric highlights the main advantage of our generative setup - the predicted color profiles are statistically much closer to real color profiles obtained from images. This is a direct consequence of our discriminator network which utilises the visual modality for enhanced training. Classification loss $\mathcal{L}_{\text{cls}}$ leads to a slight increase in the L2 metrics, but it enables the learning of realistic color profiles, noticeable from the lower FID scores.

We further provide qualitative evidence in Figure 6. Our model captures context - effect of an attribute on different objects - for abstract concepts like 'hot', and explicit color indicators like 'pink'. Our model also learns the notion of composition - how different attributes modify the same object. The color of 'young leaves' is rich in green whereas 'fallen leaves' are represented well in the brown-to-red spectrum and 'citrus leaves' are more yellowish. A similar argument follows for the object 'lemon' and modifiers such as 'fresh', 'raw' and 'pale'. It also learns meaningful colors for unseen combinations of (attribute, object) pairs. Rows 5 through 7 demonstrate effective zero-shot learning as the generated color profiles reasonably capture the semantics of the text. Another interesting behavior is its ability to highlight multiple color shades. For 'bright sun', it has learned to depict a golden yellow sun in a blue sky. Similarly, the model predicts multiple dominant color for the phrases 'coralline material' and 'orange tree'. All these example texts were obtained via the proposed dataset curation logic.

5 COLOR FOR DOWNSTREAM TASKS

In the previous section, we described our proposed generative model to go from (attribute, object) text phrases to color representations and palettes. We believe that there are multiple downstream

applications where such mapping can prove useful. Some tasks include cross-modal ranking [7, 24], language-driven image editing and manipulation [5, 13], as well as image colorization [1, 25, 55]. Our model can be used to capture users' color intent in a much more intuitive manner using natural language. For example, the phrase 'dry leaves' conveys rich semantics, while simultaneously eliminating the cumbersome process of selecting RGB values manually. While we acknowledge that we have focused on a very specific linguistic structure for the phrases, i.e adjective and noun combinations, they form a common structure of generic text inputs.

In this work, we focus on the task of cross-modal retrieval, specifically the role that color plays in it. We confine to textual queries with the (attribute, object) bigram structure. We design a relevance matching model that takes such a textual phrase and an image as input, and produces a relevance score between them. We train a standard multi-modal network [9] for shared representation learning. This network is trained using a contrastive approach where the model learns to rank relevant images higher than irrelevant ones. The overall objective is to maximise the difference of scores between the positive (relevant) and negative (not relevant). Mathematically,

$$\mathcal{L}_{\text{ranker}} = -\mathbb{E}_{\mathbf{a},\mathbf{o}}\left[\sigma\Big(R\big(\mathbf{a},\mathbf{o},\mathbf{I}_{a,o}\big) - R\big(\mathbf{a},\mathbf{o},\mathbf{I}_{\bar{a},\bar{o}}\big)\Big)\right]$$

where $\sigma(x)$ is the sigmoid activation function and R is the ranker. $R\big(\mathbf{a},\mathbf{o},\mathbf{I}_{a,o}\big)$ denotes the score predicted by the ranker for text $(\mathbf{a},\mathbf{o})$ and positive image $\mathbf{I}_{a,o}$ from the same class, and $R\big(\mathbf{a},\mathbf{o},\mathbf{I}_{\bar{a},\bar{o}}\big)$ is the ranker output for text $(\mathbf{a},\mathbf{o})$ and negative image $\mathbf{I}_{\bar{a},\bar{o}}$ chosen randomly from any other class $(\bar{a},\bar{o})$. This is inspired from the RankNet loss [4] which uses clicked-versus-not labels to improve search ranking from user behavioral data. Finally, the retrieval performance of this model is evaluated in terms of ranking the visual assets against the text queries. We also utilize and evaluate the model as a classifier, by scoring a given image against a fixed

Figure 7: Qualitative results depicting the benefits of color-centric features in cross-modal retrieval. Ranking results for two exemplar queries along with retrieval metrics. Images bound in green belong to the query class and images in red are irrelevant.

enumeration of (attribute, object) textual phrases and measuring its ability in scoring the correct one the highest.

We experiment with different combinations of input features, specifically with and without explicit color information, and attribute the gain in performance to the use of color representations. The baseline model uses only pretrained word and image embeddings (collectively termed as *Base Features*) for the text and image networks respectively. We next incrementally provide the model with color representations of images (termed as *Image Color*). Lastly, we build a model that also uses the output of our text-to-color model, i.e, color representations of text phrases (termed as *Text Color*). Here, we first wish to study the effect of text color in ranking in isolation, i.e., independent of the performance of our GAN setup. So we define a ground-truth color representation for (attribute, object) bigrams as the average over all image histograms belonging to that class. This acts as an upper limit on the ranker performance when text color is added. Then, we use the color profiles generated from our model and evaluate the ranker and GAN together. Since our goal is to evaluate a color-centric feature, we work in a controlled setup – a simple baseline that achieves reasonable accuracy. We do not use additional metadata such as image tags or captions, which are otherwise common in image retrieval systems.

Implementation Details: As before, base features are extracted from pretrained ResNet [12] and text features are a concatenation of individual GloVe embeddings [37]. All ground-truth color features are obtained from attention-weighted LAB space color histograms. The ranker comprises of 2 hidden layers with ReLU activation of 1024 and 512 units respectively, after which a linear layer returns a scalar value for the image-text relevance score.

Evaluation: Our retrieval setup is as follows - For a given a, o pair, we consider the set of all relevant images $|I_{a,o}| = n_{a,o}$ of that class and randomly sample $k * n_{a,o}$ irrelevant images from the dataset, where k is a hyperparameter. We consider a range of values of k and measure the retrieval performance of our model

using standard IR metrics – Area Under the ROC curve (AUC), Mean R-Precision (MRP) and Mean Average Precision (MAP). As k increases, the task difficulty also increases as the model now has to differentiate between relevant and irrelevant images for a text query from a much larger pool.

For image classification, we consider all the pairs in the dataset and assign relevance scores to each pair using the relevance matching model. We then calculate the top-N classification accuracy as the percentage of images for which the correct class appeared in the top N predictions made by the model. We also extend this model to attribute only and object only classification tasks. For this, we define the attribute only relevance as the average across all pairs with that attribute, and similarly for objects.

Table 3 summarizes the retrieval results at $k = 5$ and Top-20 classification accuracies. In the set of experiments in Part (1), the *Text Color* is defined as the average color profile of all images relevant for that text query. It is evident from the metrics that incorporating color of both modalities outperforms the other model variations - no color features or only color in images. It is worth noting that there is a consistent improvement in performance by adding *Text Color* to *Base Features + Image Color*, and the model is able to achieve an AUC of ~ 0.9. This indicates the significance of having a color representation for not just images, but text modality too, and further corroborates our motivation.

In Part (2), we evaluate our GAN architecture in a ranking and classification context by using the model predictions as *Text Color* features for relevance matching. While the metrics are marginally lower compared to the use of ground-truth for *Text Color*, this is a natural outcome of the use of model predictions. Even then, all variations of the GAN objective lead to better results than the baseline model which uses just *Base Features*. These results ascertain the viability of our generative model for text-to-color prediction, and the promising use of color in various downstream applications.

Table 3: Evaluation of image-text relevance matching models on cross-modal retrieval and image classification tasks. We consider different model variations in addition to base features, and they are as follows: (1a) no color-specific features (1b) only image color (1c) image color and ground-truth text color (2) image color and generated text color.

		Retrieval at $k = 5$			Top-20 Classification Accuracy		
		AUC	MRP	MAP	Pair	Attribute	Object
(1) Using Ground Truth Color Representations							
Base Features		0.837	0.510	0.534	7.521	11.403	11.200
Base Features + Image Color		0.866	0.570	0.597	13.521	16.983	15.720
Base Features + Image Color + Text Color		0.905	0.656	0.692	22.644	27.477	23.812
(2) Using Generated Color Representations for Text Color							
Ours	$+\mathcal{L}_{\ell 2}$	0.858	0.573	0.605	16.236	23.839	26.758
	$+\mathcal{L}_{\text{triplet}}$	0.862	0.577	0.607	19.576	22.753	25.210
	$+\mathcal{L}_{\text{quintuplet}}$	0.856	0.567	0.597	15.367	19.074	16.969
Ours $-\mathcal{L}_{\text{mis}}$	$+\mathcal{L}_{\ell 2}$	0.861	0.577	0.607	15.408	22.631	28.780
	$+\mathcal{L}_{\text{triplet}}$	0.862	0.576	0.604	12.367	15.381	16.915
	$+\mathcal{L}_{\text{quintuplet}}$	0.860	0.569	0.598	12.449	17.635	15.734
Ours $-\mathcal{L}_{\text{cls}}$	$+\mathcal{L}_{\ell 2}$	0.859	0.573	0.606	22.821	31.387	30.627
	$+\mathcal{L}_{\text{triplet}}$	0.854	0.568	0.594	11.987	18.666	26.622
	$+\mathcal{L}_{\text{quintuplet}}$	0.856	0.561	0.590	20.621	25.061	24.368
Ours $-\mathcal{L}_{\text{mis}} - \mathcal{L}_{\text{cls}}$	$+\mathcal{L}_{\ell 2}$	0.856	0.574	0.602	17.553	24.083	28.740
	$+\mathcal{L}_{\text{triplet}}$	0.863	0.577	0.604	19.902	24.830	26.364
	$+\mathcal{L}_{\text{quintuplet}}$	0.860	0.570	0.598	14.417	21.327	29.459

In figure 7, we provide a side-by-side comparison of image retrieval using a model that does not use explicit color information versus one that does. Consider the query 'deep sea' - the ranker with color has captured the intuition that "deep" turns the shade of water darker. Similarly, for the query 'warm sunshine', the ranker with color can retrieve more yellowish images, while the baseline ranker (which does not use color features) fetches several pictures of a blue sky. The same can also be observed from the retrieval metrics where adding color leads to improved performance.

6 CONCLUSION AND DISCUSSION

In this paper, we have considered the task of generating color representations from text. We focused our attention on textual phrases with an (attribute, object) structure, motivated by their common occurrence in user queries of an image search engine. In addition, as seen using examples throughout the paper, such phrases exhibit much visual diversity, especially on the color axis.

We have described a dataset curation strategy that we believe is a useful general workflow for studying the mapping of text to other visual axes, like texture and aesthetics. We propose a GAN architecture that when trained over the compiled dataset generates intuitive output. We have also conducted a quantitative evaluation via the use of multiple types of metrics - general notions of difference between embeddings, color-specific distance functions, as well as the Fréchet Inception Distance (FID) specifically for use with GAN outputs. While the non-generative baseline performs

equivalently on other metrics, in terms of the predicted color representations, the GAN has a better FID score indicating that the colors it produces are more realistic in general.

While the task of producing color from text has been explored in prior work, our specific focus was on the interplay between context and compositionality. When an adjective is used alongside a noun, it modifies the visual representation of the corresponding object. Encouraging compositionality (separating the network connections in the early layers of the model) aids in being able to combine previously seen primitives (i.e., attribute and object words) into novel concepts. Parts of our evaluation were directed at this zero-shot setting. In addition, we have illustrated the important role of color in the downstream task of image retrieval, and the improvements via the use of the color predictions from our model.

In future work, we wish to expand to a wider class of linguistic structures observed in phrases. It is common to have multiple attributes for a given object (e.g. 'round metallic bottle'). Modeling the compounded effect of the attributes on the object, in a manner not limited by their number, would be an interesting task. In addition, specifically focusing on the visual axis of color, some modifiers tend to be commonly observed such as 'dark' and 'deep'. Similar to the earlier point, attributes can be cascaded ('very dark blue') offering a very rich vocabulary for the text phrases. Following the pipeline described in this paper - which includes a way to enumerate a list of example phrases of that structure, obtaining ground truth derived from images, designing and training models that go from text to color - offers scope for multiple investigations.

REFERENCES

[1] Hyojin Bahng, Seungjoo Yoo, Wonwoong Cho, David Keetae Park, Ziming Wu, Xiaojuan Ma, and Jaegul Choo. 2018. Coloring with words: Guiding image colorization through text-based palette generation. In *Proceedings of the european conference on computer vision (eccv)*. 431–447.

[2] Brent Berlin and Paul Kay. 1991. *Basic color terms: Their universality and evolution*. Univ of California Press.

[3] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods* 46, 3 (2014), 904–911.

[4] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*. 89–96.

[5] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. 2018. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8721–8729.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[7] Ingemar J Cox, Matthew L Miller, Thomas P Minka, Thomas V Papathomas, and Peter N Yianilos. 2000. The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments. *IEEE transactions on image processing* 9, 1 (2000), 20–37.

[8] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766* (2020).

[9] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. (2013).

[10] Yuliya Geikhman. 2020. The Essentials of English Adjectives: 7 Key Adjective Types to Know. (2020).

[11] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *arXiv preprint arXiv:1406.2661* (2014).

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] Jeffrey Heer and Maureen Stone. 2012. Color naming models for color selection, image editing and palette design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1007–1016.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500* (2017).

[15] Wynne Hsu, ST Chua, and HH Pung. 1995. An integrated color-spatial approach to content-based image retrieval. In *Proceedings of the third ACM international conference on Multimedia*. 305–313.

[16] Phillip Isola, Joseph J Lim, and Edward H Adelson. 2015. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1383–1391.

[17] Anil K Jain and Aditya Vailaya. 1996. Image retrieval using color and shape. *Pattern recognition* 29, 8 (1996), 1233–1244.

[18] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. 2018. Learn to Pay Attention. In *International Conference on Learning Representations*.

[19] Kazuya Kawakami, Chris Dyer, Bryan R Routledge, and Noah A Smith. 2016. Character sequence models for colorfulwords. *arXiv preprint arXiv:1609.08777* (2016).

[20] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access* (2020).

[21] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. 2020. Symmetry and Group in Attribute-Object Compositions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Albrecht Lindner and Sabine Süsstrunk. 2013. Automatic color palette creation from words. In *Color and Imaging Conference*, Vol. 2013. Society for Imaging Science and Technology, 69–74.

[23] Paridhi Maheshwari, Ritwick Chaudhry, and Vishwa Vinay. 2021. Scene Graph Embeddings Using Relative Similarity Supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2328–2336.

[24] Paridhi Maheshwari, Manoj Ghuhan, and Vishwa Vinay. 2020. Learning Colour Representations of Search Queries. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[25] Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis. 2018. Learning to color from language. *arXiv preprint arXiv:1804.06026* (2018).

[26] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2794–2802.

[27] Brian McMahan and Matthew Stone. 2015. A Bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics* 3 (2015).

[28] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science* 331, 6014 (2011), 176–182.

[29] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[30] Ishan Misra, Abhinav Gupta, and Martial Hebert. 2017. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1792–1801.

[31] Will Monroe, Noah D Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. *arXiv preprint arXiv:1606.03821* (2016).

[32] Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics* (2017).

[33] Randall Munroe. 2010. Color Survey Results. https://blog.xkcd.com/2010/05/03/color-survey-results/

[34] Tushar Nagarajan and Kristen Grauman. 2018. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 169–185.

[35] Zhixiong Nan, Yang Liu, Nanning Zheng, and Song-Chun Zhu. 2019. Recognizing unseen attribute-object pair with generative model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8811–8818.

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

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

[38] Konstantinos N Plataniotis and Anastasios N Venetsanopoulos. 2013. *Color image processing and applications*. Springer Science & Business Media.

[39] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. 2019. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[40] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*. PMLR, 1060–1069.

[41] Terry Regier and Paul Kay. 2009. Language, thought, and color: Whorf was half right. *Trends in cognitive sciences* 13, 10 (2009), 439–446.

[42] Debi Roberson, Ian Davies, and Jules Davidoff. 2000. Color categories are not universal: replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General* 129, 3 (2000), 369.

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[44] Vidya Setlur and Maureen C Stone. 2015. A linguistic approach to categorical color assignment for data visualization. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 698–707.

[45] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[46] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[47] Pavel Tokmakov, Yu-Xiong Wang, and Martial Hebert. 2019. Learning compositional representations for few-shot recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6372–6381.

[48] Xin Wang, Fisher Yu, Trevor Darrell, and Joseph E Gonzalez. 2019. Task-Aware Feature Generation for Zero-Shot Compositional Learning. *arXiv preprint arXiv:1906.04854* (2019).

[49] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. 2019. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[50] Kilian Q Weinberger and Lawrence K Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research* 10, 2 (2009).

[51] Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. 2017. Learning distributions of meant color. *arXiv preprint arXiv:1709.09360* (2017).

[52] Jonathan Winawer, Nathan Witthoft, Michael C Frank, Lisa Wu, Alex R Wade, and Lera Boroditsky. 2007. Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences* 104, 19 (2007).

[53] Gunter Wyszecki and Walter Stanley Stiles. 1982. *Color science*. Vol. 8. Wiley New York.

[54] Aron Yu and Kristen Grauman. 2014. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 192–199.

[55] Changqing Zou, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu. 2019. Language-based colorization of scene sketches. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–16.