FeelingBlue: A Corpus for Understanding the Emotional Connotation of Color in Context

Amith Ananthram[1] and **Olivia Winn**[1] and **Smaranda Muresan**[1,2]
[1]Department of Computer Science, Columbia University, USA
[2]Data Science Institute, Columbia University, USA
{amith,olivia,smara}@cs.columbia.edu

Abstract

While the link between color and emotion has been widely studied, how context-based changes in color impact the intensity of perceived emotions is not well understood. In this work, we present a new multimodal dataset for exploring the emotional connotation of color as mediated by line, stroke, texture, shape, and language. Our dataset, **FeelingBlue**, is a collection of 19,788 4-tuples of abstract art ranked by annotators according to their evoked emotions and paired with rationales for those annotations. Using this corpus, we present a baseline for a new task: **Justified Affect Transformation**. Given an image I, the task is to 1) recolor I to enhance a specified emotion e and 2) provide a textual justification for the change in e. Our model is an ensemble of deep neural networks which takes I, generates an emotionally transformed color palette p conditioned on I, applies p to I, and then justifies the color transformation in text via a visual-linguistic model. Experimental results shed light on the emotional connotation of color in context, demonstrating both the promise of our approach on this challenging task and the considerable potential for future investigations enabled by our corpus.[1]

1 Introduction

Color is a powerful tool for conveying emotion across cultures, a connection apparent in both language and art (Mohr and Jonauskaite, 2022; Mohammad and Kiritchenko, 2018). Metaphoric language frequently uses color as a vehicle for emotion: Familiar English metaphors include ‘‘feeling blue’’ or ‘‘green with envy’’. Similarly, artists often pick specific colors in order to convey particular emotions in their work, while viewer perceptions of a piece of art are affected by its color palette (Sartori et al., 2015). Previous studies have mostly been categorical, focusing on confirming known links between individual colors and emotions like *blue & sadness* or *yellow & happiness* (Machajdik and Hanbury, 2010; Sartori et al., 2015; Zhang et al., 2011). However, in the wild, the emotional connotation of color is often mediated by line, stroke, texture, shape, and language. Very little work has examined these associations. Does the mere presence of blue make an image feel sad? If it is made bluer, does it feel sadder? Is it dependent on its associated form or its surrounding color context? And, if the change is reflected in an accompanying textual rationale, is it more effective?

Our work is the first to explore these questions. We present **FeelingBlue**, a new corpus of relative emotion labels for abstract art paired with English rationales for the emotion labels (see Figure 1 and Section 3). A challenge with such annotations is the extreme subjectivity inherent to emotion. In contrast to existing Likert-based corpora, we employ a Best-Worst Scaling (BWS) annotation scheme that is more consistent and replicable (Mohammad and Bravo-Marquez, 2017). Moreover, as our focus is *color in context* (colors and their form), we restrict our corpus to abstract art, a genre where color is often the focus of the experience, mitigating the effect of confounding factors like facial expressions and recognizable objects on perceived emotions (as observed in Mohammad, 2011; Sartori et al., 2015; Zhang et al., 2011; Alameda-Pineda et al., 2016).

To demonstrate **FeelingBlue**’s usefulness in explorations of the emotional connotation of color in context, we introduce a novel task, **Justified Affect Transformation**—conditional on an input image I_o and an emotion e, the task is 1) to recolor I_o to produce an image I_e that evokes e more intensely than I_o and 2) to provide justifications for why I_o evokes e less intensely and why I_e evokes e more intensely. Using **FeelingBlue**, we

[1]Our dataset, code, and models are available at https://github.com/amith-ananthram/feelingblue.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 176–190, 2023. https://doi.org/10.1162/tacl_a_00540
Action Editor: Yulan He. Submission batch: 9/2022; Revision batch: 11/2022; Published 3/2023.

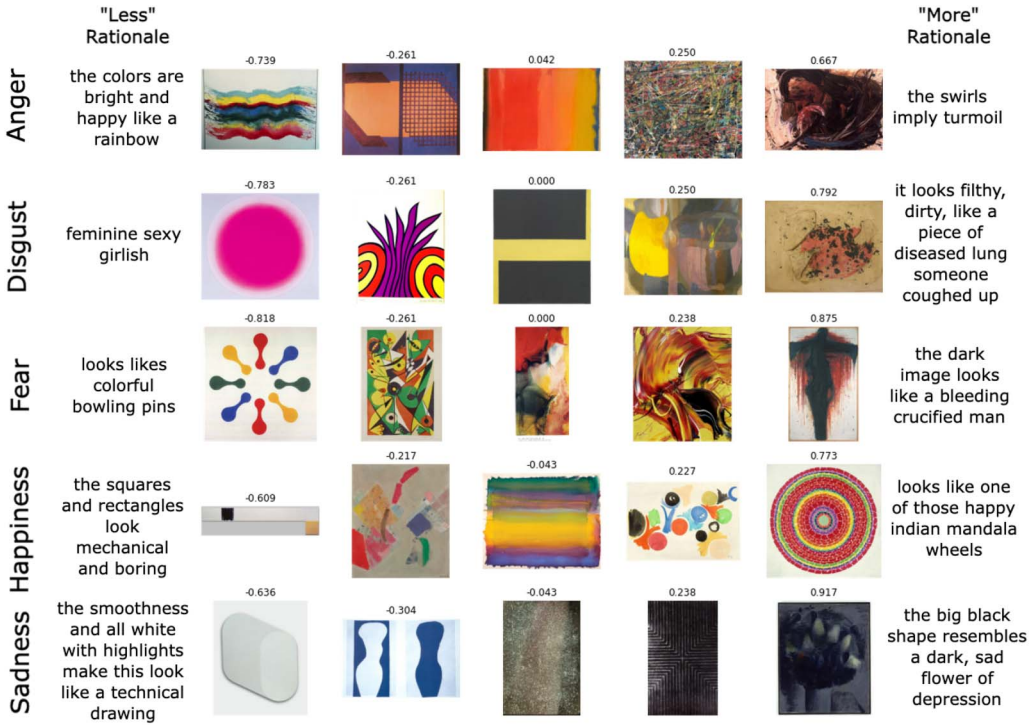

Figure 1: Representative examples spanning **FeelingBlue**'s emotion subsets. Each image in an emotion subset has a score $\in [-1, 1]$ derived from its Best-Worst Scaling annotations. Images selected as the ''least''/''most'' emotional in a 4-tuple (not shown here) have rationales explaining why they are ''less''/''more'' emotional than the rest. Information about these works can be found at `https://github.com/amith-ananthram/feelingblue`.

build a baseline system for two subtasks: image recoloring and rationale retrieval (Section 4).

We conduct a thorough human evaluation that confirms both the promise of our approach on this challenging new task and the opportunity for future investigations enabled by **FeelingBlue**. Our results reveal regularities between context-based changes in color and emotion while also demonstrating the potential of linguistic framing to mold this subjective human experience (Section 5).

Our dataset, code, and models are available at `https://github.com/amith-ananthram/feelingblue`.

2 Related Work

While the body of work studying the relationship between color and emotion is quite large, almost all of it has focused on identifying categorical relationships in text to produce association lexicons (notable works include Mohammad, 2011; Sutton and Altarriba, 2016; Mikellides, 2012).

In the domain of affective image analysis, previous work has mostly explored classifying the emotional content of images. Machajdik and Hanbury (2010), Sartori et al. (2015), Zhang et al., (2011) and Alameda-Pineda et al. (2016) focus on identifying low-level image features correlated with each perceived emotion, with the latter two examining abstract art specifically. Rao et al., (2019) employ mid- and high-level features for classification. Some work has investigated the distribution of emotions in images: Zhao et al. (2017, 2018) create probability distributions over multiple emotions and Kim et al. (2018) look at gradient values of arousal and valence, though none of the works correlate emotion values with image colors. Similarly, Xu et al. (2018) learn dense emotion representations through a text-based multi-task model but they do not explore its association with color.

Image recoloring is a very small subset of work in the field of image transformation and has mostly focused on palettes. PaletteNet recolors a source image given a target image palette (Cho et al., 2017) while Bahng et al. (2018) semantically generate palettes for coloring gray-scaled images. No previous work has examined recoloring images to change their perceived emotional content. Similarly, the field of text-to-image synthesis, which

has seen major progress recently with models like DALL-E 2 (Ramesh et al., 2022), has centered on generating images de novo or on in-painting. There has been no work that recolors an image while preserving its original structure.

Ulinski et al. (2012), Yang et al. (2019) and Achlioptas et al. (2021) (who also annotate WikiArt) have explored emotion in image captioning though their focus is much broader than color.

3 Dataset

Our dataset, **FeelingBlue**, is a collection of abstract art ranked by the emotions they most evoke with accompanying textual justifications (see Figure 1). It contains 5 overlapping subsets of images (one each for ***anger***, ***disgust***, ***fear***, ***happiness***, and ***sadness***) with continuous-value scores that measure the intensity of the respective emotion in each image compared to the other images in the subset.

While **WikiArt** (Mohammad and Kiritchenko, 2018), **DeviantArt**[2] (Sartori et al., 2015), and other emotion corpora contain images with multi-label continuous emotion scores, these scores were collected for each image in isolation without accompanying rationales. They reflect how often annotators believed that a particular emotion fit an image resulting in a measure of the presence of the emotion rather than its intensity. As such, their scores are not a suitable way to order these images by the strength of the emotion they evoke. In contrast, our annotations were collected by asking annotators to 1) rank groups of images according to a specified emotion and 2) justify their choice.

Below, we detail how we compiled this corpus.

3.1 Image Compilation

The images[3] in our dataset are drawn from both **WikiArt** and **DeviantArt**.[4] As we are most interested in the emotional connotation of color as constrained by its form, we manually removed images of photographs, statues, people, or recognizable objects. This eliminated many confounding factors like facial expressions, flowers, or skulls that might affect a person's emotional response to an image, leaving primarily color and its visual context. Our final corpus contains 2,756 images.

3.2 Annotation Collection

We began by partitioning our images into overlapping emotion subsets where each image appears twice, in both subsets of its top 2 emotions according to its original corpus (**WikiArt** or **DeviantArt**) scores. As we want meaningful continuous value scores of emotional intensity, restricting images to their top 2 emotions ensures that the scored emotion is present. Within each subset, we randomly generated 4-tuples of images such that each image appears in at least 2 4-tuples. With these 4-tuples in hand, we collected annotations via Best-Worst Scaling (BWS) (Flynn and Marley, 2014), a technique for obtaining continuous scores previously used to construct sentiment lexicons (Kiritchenko and Mohammad, 2016) and to label the emotional intensity of tweets (Mohammad and Bravo-Marquez, 2017). In fact, Mohammad and Bravo-Marquez (2017) found that BWS is a reliable method for generating consistent scaled label values. It produces more replicable results than the typical Likert scale where annotators often do not agree with their own original assessments when shown an item they have already labeled.

In BWS, annotators are presented with n options (where often $n = 4$), and asked to pick the 'best' and 'worst' option according to a given criterion. For our task, we present each annotator with a 4-tuple (i.e., 4 images) of abstract art and an emotion, and the 'best' and 'worst' options are the images that 'most' and 'least' evoke the emotion relative to the other images. In addition, we also asked each annotator to provide rationales describing the salient features of their chosen 'most' and 'least' emotional images. As is common practice with BWS, for each subset corresponding to an emotion e, we calculate continuous value scores for each image I by subtracting the number of times I was selected as 'least' evoking e from the number of times I was selected as 'most' evoking e and then dividing by the number of times I appeared in an annotated 4-tuple. We collected 3 annotations per 4-tuple task from Master Workers on Amazon Mechanical Turk (AMT) via the BWS procedure of Flynn and Marley (2014). Workers were paid consistent with the minimum wage.

[2] `www.deviantart.com`.

[3] These images are a mix of copyright protected and public domain art. We do not distribute these images. Instead, we provide URLs to where they may be downloaded.

[4] From the 283/500 images that remain available.

Emotion	# Images	# Best	# Worst	Total	Spearman-R	Maj. Agree %	# Labels
Anger	187	1,078	1,087	2,165	0.685 (0.034)	64, 68, 66	2.03, 1.92
Disgust	525	3,072	3,078	6,150	0.676 (0.019)	65, 66, 66	2.00, 1.98
Fear	396	2,352	2,315	4,667	0.691 (0.017)	65, 66, 66	1.99, 1.94
Happiness	399	2,326	2,346	4,672	0.573 (0.029)	65, 61, 63	2.00, 2.13
Sadness	183	1,084	1,050	2,134	0.672 (0.040)	68, 64, 66	1.97, 2.00
	1688	9,912	9,876	19,788			

Table 1: Summary and inter-annotator agreement statistics for our corpus. The first column contains the # of unique images in each subset. The next 3 contain the number of annotated 4-tuples of images in each subset. The final 3 columns contain measures of inter-annotator agreement: 1) the mean/stdv of the Spearman rank correlations of scores calculated from 30 random splits of each subset, 2) the % of annotations that agree with the majority annotation ('best', 'worst', 'total'), and 3) the average # of distinct labels for each annotation task ('best', 'worst')

We did not use all of the emotion classes from the original datasets in our study. Mohammad and Kiritchenko (2018) found that emotions such as arrogance, shame, love, gratitude, and trust were interpreted mostly through facial expression. We also excluded ***anticipation*** and ***trust*** as they were assumed to be related to the structural composition of the artwork rather than color choice. For each remaining emotion, 50 images from their corresponding subsets were sampled for a pilot study of the 4-tuple data collection. ***Surprise*** was removed as pilot participants exhibited poor agreement and did not reference color or terms that evoke color in its corresponding rationales. This left 5 emotions: ***anger***, ***disgust***, ***fear***, ***happiness***, and ***sadness***.

We manually filtered the collected annotations to remove uninformative rationales (such as ''it makes me feel sad''). This filtering resulted in splitting each collected BWS annotation into a 'best' 4-tuple (which contains the 4 images and the most emotional choice among them, accompanied by a rationale) and a 'worst' 4-tuple (which contains the 4 images and the least emotional choice among them, accompanied by a rationale). Our final corpus contains 19,788 annotations, nearly balanced with 9,912 'best' and 9,876 'worst' (as in some 4-tuples, either the 'best' or the 'worst' rationale was retained but not both).

Table 1 contains summary and inter-annotator agreement statistics for our corpus, broken down by emotion subset. We rely on 3 different measures to gauge the consistency of these annotations. The first captures the degree to which differences among annotations result in changes to the ranking of the images by BWS score. For each emotion, we randomly split the 3 annotations for each of its 4-tuples and then calculate BWS scores for both of the resulting random partitions. We do this 30 times, calculating the Spearman rank correlation between the pairs of scores for each partition, and present the mean and standard deviation of the resulting coefficients. The second is a measure of what percentage of all annotations agree with the majority annotation for a particular 4-tuple—cases where all annotators disagree have no majority annotation. The final is a measure of the # of distinct choices made by annotators for each 4-tuple. Given the considerable subjectivity of the annotation task, these inter-annotator agreement numbers are reasonable and consistent with those reported by Mohammad and Kiritchenko (2018) for the relatively abstract genres from which our corpus is drawn. ***Happiness***, the only emotion with a positive valence, exhibits the worst agreement.

We present images, rationales, and scores from our corpus in Figure 1.

3.3 Corpus Analysis

We explore a number of linguistic features (color, shape, texture, concreteness, emotion, and simile) in **FeelingBlue**'s rationales to better understand how they change with image emotion and color.

To measure color, we count both explicit color terms (e.g., 'red', 'green') and implicit references (e.g., 'milky', 'grass'). As the artwork is abstract and can only convey meaning through ''line, stroke, color, texture, form, and shape'' (IdeelArt, 2016), the use of adjectives and nouns with strong color correlation is a likely reference to those colors in the image. For explicit color terms, we use the base colors from the XKCD

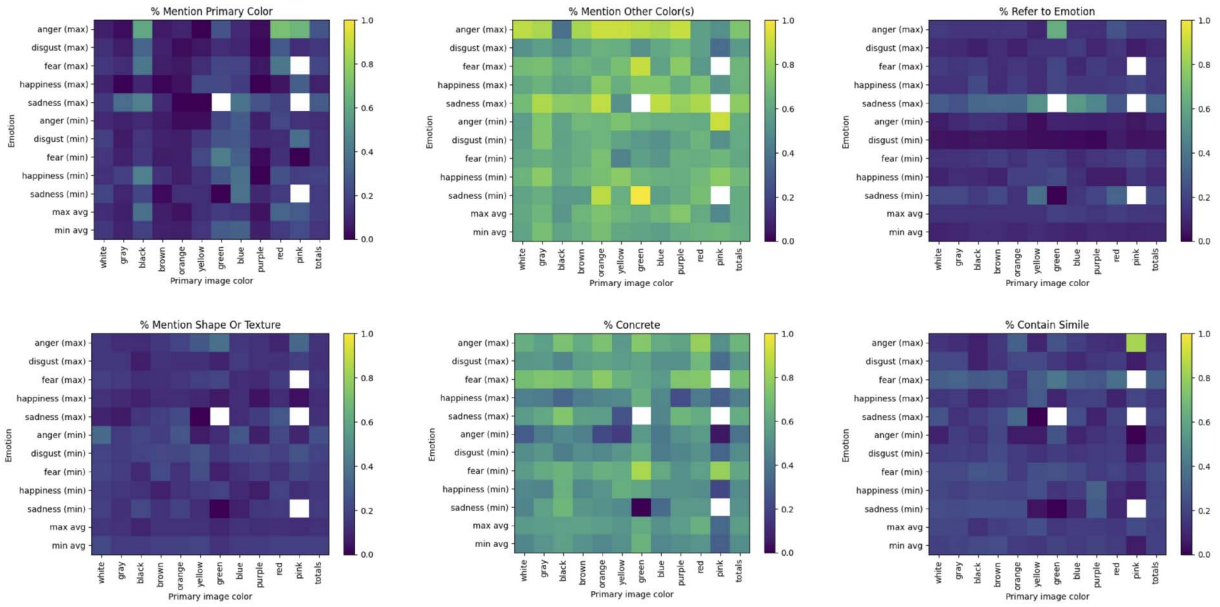

Figure 2: An analysis of the linguistic features of the rationales in **FeelingBlue**, binned by emotion, the corresponding image's primary color and whether it was 'least' or 'most' emotional. White bins contain no rationales.

dataset (Monroe, 2010). For implicit color references, we use the NRC Color Lexicon (NRC) (Mohammad, 2011), taking the words with associated colors where at least 75% of the responders agreed with the color association. In order to compare color references with the main color of each image, the primary colors of each image were binned to the 11 color terms used in the NRC according to nearest Delta-E distance. This results in an uneven number of rationales per color bin. The implicit color terms were mapped directly to the color bins; each explicit color term was mapped to its nearest.

Shape words were collected by aggregating lists of 2D shapes online and corroborating them with student volunteers. For texture words, we used the 98 textures from Bhushan et al. (1997). Concreteness was measured as a proxy for how often the rationales ground the contents of an image in actual objects and scenery, like referring to a gradient of color as a 'sunset' or a silver streak as a 'knife'. To calculate concreteness, we used the lexicon collected by Brysbaert et al. (2014), which rates 40,000 English lemmas on a scale from 0 to 5. After empirical examination, we threshold concreteness at 4 and ignore all of the explicit color terms, shapes and textures. Rationales were labeled as concrete if at least one word in the rationale was above the concreteness threshold. Similes were identified by presence of 'like' (but not 'I like') such as ''The blue looks like a monster''.

In Figure 2 we present heatmaps which break down this exploration. In total, 69.3% of the rationales refer to color, highlighting the central role color plays in **FeelingBlue**. We see that 51.2% contain explicit color references and 36.1% contain implicit references (with a 26% overlap). Surprisingly, the majority of these references were not to the primary color of each image. This suggests that viewers were drawn to colors which were either more central in the canvas or contrasted against the primary color. A notable exception is 'red' for 'anger': when the image is primarily red, it is mentioned 80% of the time in the rationale for the angriest images. Interestingly, when the image is primarily green the rationale explicitly mentions 'anger', as if the image evokes 'anger' despite the coloring. This is a surprising contrast to the sad images, where 'blue', deeply tied to 'sadness' in English, is hardly mentioned when the main color is blue, but for those same images, the rationales explicitly call the image 'sad'. It may be that blue is so intrinsically tied to 'sadness' that responders felt sad without consciously linking the two.

The happiest images are described as 'bright' and 'graceful' while the saddest are 'dark' and 'muddy'. Though the least happy are 'dark' as well, they are also 'simple' and 'dull', while the

least sad are 'simple', 'light', and 'empty'. As the language varies across these valence pairs (e.g., least happy/most sad), this suggests that the rationales reflect the full continuum of each emotion.

Shapes are referred to much less frequently, a mere 11.4%, and texture is mentioned in only 4.9% of the rationales. However, despite the images being of abstract art, 52.4% of the rationales were 'concrete' (with 17.9% containing simile), revealing the importance of grounding in rationalizations of the emotional connotation of color.

4 Justified Affect Transformation

We define a new task to serve as a vehicle for exploring the emotional connotation of color in context enabled by our corpus: **Justified Affect Transformation**. Given an image I_o and a target emotion $e \in E$, the task is:

1. change the color palette of I_o to produce an image I_e that evokes e more intensely than I_o
2. provide textual justifications, one explaining why I_o evokes e less intensely and another explaining why I_e evokes e more intensely

By focusing on changes in color (conditional on form), we can understand the affect of different palettes in different contexts. And by producing justifications for those changes, we can explore the degree to which the emotional connotation can be accurately verbalized in English.

To solve this task, we propose a two step approach: 1) an image recoloring component that takes as input an image I_o and a target emotion $e \in E$ and outputs I_e, a version of I_o recolored to better evoke e (Section 4.1); and 2) a rationale retrieval component that takes as input two images, I_o and I_e, an emotion $e \in E$, and a large set of candidate rationales R, and outputs a ranked list of rationales R_{less} that justify why I_o evokes e less than I_e and a ranked list of rationales R_{more} that justify why I_e evokes e more than I_o (Section 4.2).

4.1 Image Recoloring

Our image recoloring model takes an image I_o and an emotion e as inputs and outputs a recolored image $I_e, \forall e \in E$ that better evokes the given emotion. In an ideal scenario, this model would be trained on a large corpus that directly reflects the task of emotional recoloring: differently colored versions of the same image, ranked according to their emotion. Such a corpus is difficult to construct. Instead, we use our corpus, **FeelingBlue**, which contains 3-tuples, (I_{less}, I_{more}, e), where I_{less} and I_{more} are entirely different images and I_{less} evokes e less intensely than I_{more}.

Our image recoloring model is an ensemble of neural networks designed to accommodate this challenging training regime. It consists of two subnetworks, an emotion-guided image selector and a palette applier, each trained independently. The emotion-guided image selector takes two images and an emotion and identifies which of the two better evokes the emotion. The palette applier (PaletteNet, Cho et al. (2017)) takes an image and a c-color palette and applies the palette to the image in a context-aware manner.

To produce I_e from I_o for a specific emotion e, we begin with a randomly initialized palette p_e, apply it to I_o with the frozen palette applier to produce I_e and rank I_o against I_e with the frozen emotion-guided image selector. We update p_e via backpropagation so that the recolored I_e more intensely evokes e according to the emotion-guided image selector (see Figure 3). We avoid generating adversarial transformations by restricting the trainable parameters to the colors in the image's palette (instead of the image itself), forcing the backpropagation through the emotion-guided image selector to find a solution on the manifold of recolorizations of I_o. Additionally, we avoid local minima by optimizing 100 randomly initialized palettes for each (I_o, e) pair, allowing us to select a palette from the resulting set that balances improved expression of e against other criteria.

Our emotion-guided image selector, palette applier and palette training objectives are detailed in the Section 4.1.1, 4.1.2, and 4.1.3 respectively.

4.1.1 Emotion-Guided Image Selector

We begin by training our emotion-guided image selector. This model takes two images (I_1, I_2) and an emotion e as input and predicts which of the two images more intensely evokes e. The architecture produces dense representations from the final pooling layer of a pretrained instance of $ResNet$ (He et al., 2016), concatenates those representations and a 1-hot encoding of e and passes this through l_{ES} fully connected (FC) layers, re-concatenating the encoding of e after every

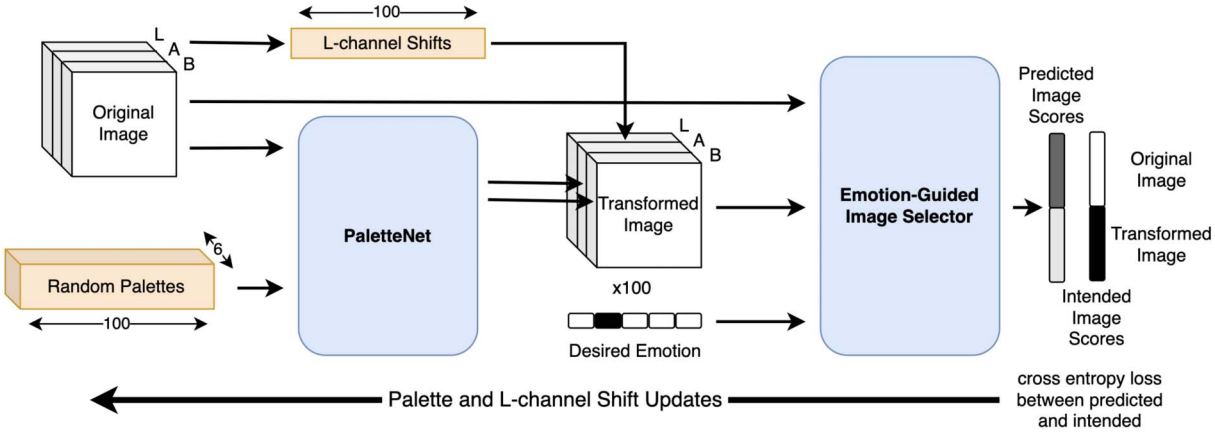

Figure 3: Our ensemble architecture for recoloring an image I_o to enhance a specified emotion e. We begin by training both PaletteNet and our Emotion-Guided Image Selector. Then, we randomly initialize 100 palettes and L-channel shifts and apply them to I_o with our frozen PaletteNet, producing 100 candidate I_e. We compare these candidate I_e to I_o with our frozen Emotion-Guided Image Selector, producing 100 predicted image scores. Then, we update only our palettes and L-channel shifts via backpropagation, minimizing a cross entropy loss between the predicted image scores and our intended selection (I_e). This process continues iteratively until convergence.

layer. We apply a dropout rate of d_{ES} to the output of the first $l_{ES}/2$ FC layers and employ leaky $ReLU$ activations to facilitate backpropagation through this network later. The model's prediction is the result of a final softmax-activated layer.

To encourage the model to be order agnostic, we expand our corpus by simply inverting each pair (producing one ''left'' example with the more intense image first and another ''right'' example with it second). We optimize a standard cross entropy loss and calculate accuracies by split (''train'' or ''valid''), side (''left'' or ''right''), and emotion. We choose the checkpoint with the best, most balanced performance across these axes.

4.1.2 Palette Applier

Our palette applier takes an image I_o and c-color palette p as input and outputs a recolored version of the image I_r using the given palette p. Its architecture is that of PaletteNet (Cho et al., 2017), a convolutional encoder-decoder model. The encoder is built with residual blocks from a randomly initialized instance of $ResNet$, while the decoder relies on blocks of convolutional layers, instance norm, leaky $ReLU$ activations, and upsampling to produce the recolored image I_r. The outputs of each convolutional layer in the encoder and the palette p are concatenated to the inputs of the corresponding layer of the decoder, ensuring that the model has the signal necessary to apply the palette properly. The palette applier outputs the A and B channels of I_r. As in Cho et al. (2017), I_o's original L channel is reused (see Figure 3).

We train this model on a corpus of recolored tuples (I_o, I_r, p) generated from our images as in Cho et al. (2017). For each image I_o, we convert it to HSV space and generate r_{PA} recolored variants I_r by shifting its H channel by fixed amounts, converting them back to LAB space and replacing their L channels with the original from I_o. We extract a c-color palette p for each of these recolored variants I_r using ''colorgram.''[5] We augment this corpus via flipping and rotation. We optimize a pixel-wise $L2$ loss in LAB space and choose the checkpoint with the best ''train'' and ''valid'' losses.

4.1.3 Palette Generation

We use frozen versions of our emotion-guided image selector ES and palette applier PA to generate a set of c-color palettes $p_e, \forall e \in E$ and L-channel shifts $b_e, \forall e \in E$ which, when applied to our original image I_o, produce recolored variants $I_e, \forall e \in E$, each evoking e more intensely than I_o.

We produce p_e and b_e iteratively, initializing p_e^0 to c random colors and b_e^0 to 0 at $t = 0$. Then, we update p_e^t and b_e^t by applying p_e^t to I_o and shifting its L channel by b_e^t, producing a recolored variant I_e^t. This recolored variant is compared against I_o by the emotion-guided image selector

[5] https://github.com/obskyr/colorgram.py.

for its specific emotion e (using both ''sides'' of ES). We optimize a cross entropy (CE) loss over these predictions with the recolored image I_e^t as the silver-label choice. We choose the final $I_e^t, \forall e \in E$ after backpropagation converges or T steps.

$$I_e^t = PA(I_o, p_e^t) + b_e^t, \forall e \in E$$
$$L = \text{CE}(ES(I_e^t, I_o, e), 0) + \text{CE}(ES(I_o, I_e^t, e), 1)$$

To avoid getting stuck in local minima, we optimize 100 randomly initialized palettes for each emotion e. Choosing the palette with the smallest loss produces similar transformations for certain emotions (such as ***fear*** and ***disgust***). One desirable property for $I_e, \forall e$ is color diversity. To prioritize this, we consider the top 50 palettes according to their loss for each emotion e and select one palette for each e such that the pairwise $L2$ distance among the resulting I_e is maximal.[6]

4.2 Rationale Retrieval

Our rationale ranking model takes as input two images, I_{less} and I_{more}, an emotion $e \in E$, and a set of candidate rationales R drawn from **FeelingBlue**. It then outputs 1) R_{less}, a ranking of rationales from R explaining why I_{less} evokes e less intensely and 2) R_{more}, a ranking of rationales from R explaining why I_{more} evokes e more intensely.

The architecture embeds I_{less} and I_{more} with $CLIP$ (Radford et al., 2021), a state-of-the-art multimodal model trained to collocate web-scale images with their natural language descriptions via a contrastive loss. We concatenate these $CLIP$ embeddings with an equally sized embedding of e and pass this through a ReLU-activated layer producing a shared representation t. We apply dropout d_{RR} before separate linear heads project t into $CLIP$'s multimodal embedding space, resulting in t_{less} and t_{more}.

Given $(I_{less}, I_{more}, r_{less}, r_{more})$, with r_{less} and $r_{more} \in R$, we optimize $CLIP$'s contrastive loss, encouraging the logit scaled cosine similarities between $t_{less|more}$ and $CLIP$ embeddings of R to be close to 1 for $r_{less|more}$ and near 0 for the rest. We weight this loss by the frequency of rationales in our corpus and reuse $CLIP$'s logit scaling factor.

[6]As this is NP-complete, we use an approximation.

4.3 Training Details

4.3.1 Corpus

We extract pairs of images ordered by the emotion they evoke from **FeelingBlue**. Each 4-tuple is ranked according to a particular emotion e resulting in a 'Least', $\boldsymbol{\ell}$, 'Most', $\boldsymbol{m}$, and two unordered middle images, $\boldsymbol{u_1}$ and $\boldsymbol{u_2}$. This provides us with 5 ordered image pairs of (*less of emotion* e, *more of emotion* e): ($\boldsymbol{\ell}$, $\boldsymbol{u_1}$), ($\boldsymbol{\ell}$, $\boldsymbol{u_2}$), ($\boldsymbol{\ell}$, $\boldsymbol{m}$), ($\boldsymbol{u_1}$, $\boldsymbol{m}$), and ($\boldsymbol{u_2}$, $\boldsymbol{m}$) which we use to train both our image recoloring and our rationale retrieval models. Note that while **FeelingBlue** restricts us to the 5 emotions for which it contains annotations, both the task and our approach could be extended to other emotions with access to similarly labeled data.

4.3.2 Preprocessing

We preprocess each image by first resizing it to 224×224, zero-padding the margins to maintain aspect ratios, converting it from RGB to LAB space, and then normalizing it to a value between -1 and 1. As LAB space attempts to represent human perception of color, it allows our model to better associate differences in perceived color with differences in perceived emotion. We note here that our emotion-guided image selector relies on fine-tuning a version of $ResNet$ pretrained on images in RGB space, not LAB space. Thus, we incur an additional domain shift cost in our fine-tuning. While this cost could be avoided by training $ResNet$ from scratch in LAB space (perhaps on a corpus of abstract art), our experimental results show that it appears to have been more than offset by the closer alignment between input representation and human visual perception.

4.3.3 Hyperparameters

For all of our models, we extract $c = 6$ color palettes. Our Emotion-Guided Image Selector uses a pre-trained $ResNet-50$ backbone and $l_{ES} = 6$ fully connected layers. It was trained with dropout $d_{ES} = 0.1$ (Srivastava et al., 2014), learning rate $lr_{ES} = 5e-5$, and a batch size of 96 for 30 epochs using Adam (Kingma and Ba, 2015). Our Palette Applier was trained for 200 epochs with a batch size of 128 using the same hyperparameters as Cho et al. (2017). To generate our palettes, we use learning rate $lr_{PG} = 0.01$ and iterate up to $T = 2000$ steps. And finally, our Rationale Retrieval model was trained with dropout $d_{RR} =$

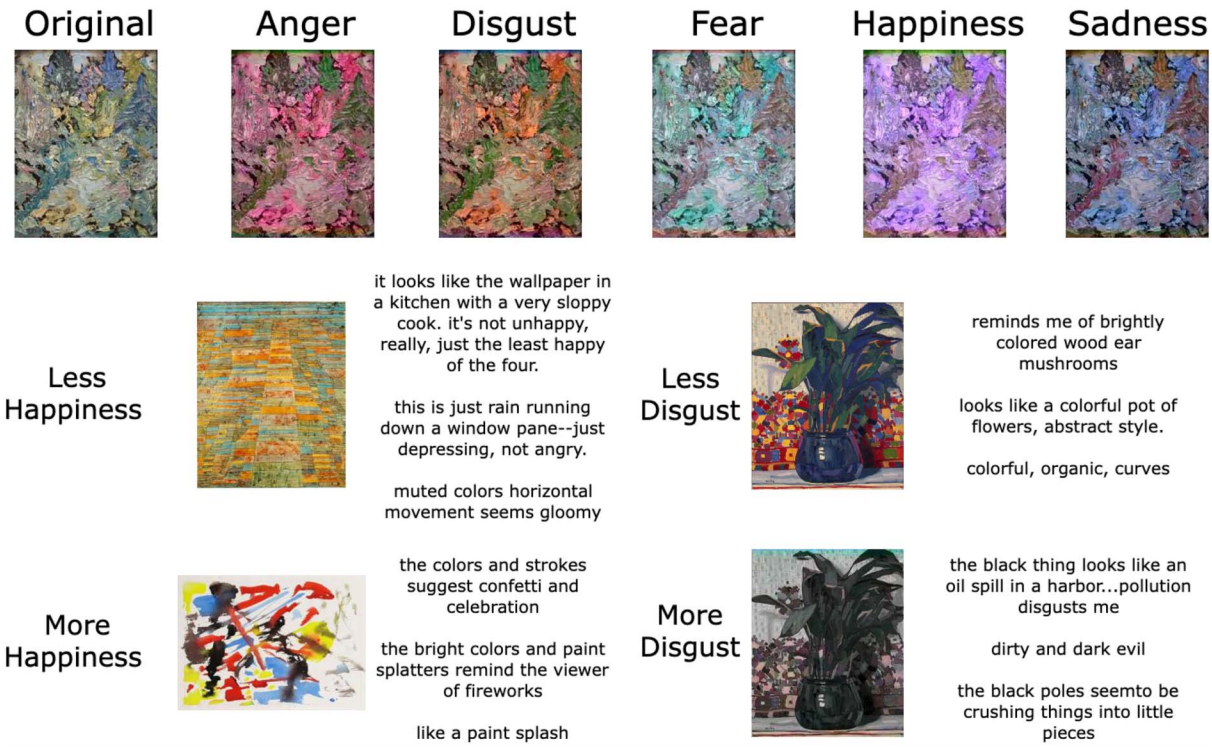

Figure 4: Samples from our model's evaluation. The top row depicts an image with all of its recolorings. The bottom left displays rationales retrieved for the depicted (I_{less}, I_{more}) distinct image pair. The bottom right displays rationales retrieved for the depicted (I_o, I_e) recolored image pair. Clockwise: *Landscape* by Arthur Beecher Carles, *Highway and Byways* by Paul Klee, *Flowers* by August Herbin, and *Farbspiele* by Ernst Wilhelm Nay.

0.4, learning rate $lr_{RR} = 1e-4$ and a batch size of 256 for 100 epochs using Adam.

5 Results and Discussion

To understand our approach's strengths, we evaluate our image recoloring model and rationale retrieval model both separately and together.

Evaluation Data for Imaging Recoloring. The evaluation set for our image recoloring model contains 100 images—30 are randomly selected from our validation set and the remaining 70 are unseen images from **WikiArt**. We generate 5 recolorized versions I_e of each image I_o corresponding to each of our 5 emotions, resulting in 1000 recolored variants for evaluation (see Section 5.1).

Evaluation Data for Rationale Retrieval. To evaluate our rationale retrieval model as a standalone module we choose an evaluation set consisting of 1000 image pairs, 200 for each emotion. For this dataset, each pair of images contains different images. Again, 30% are randomly selected from images in our **FeelingBlue** validation set and the remaining 70% consist of unseen images from **WikiArt**. To identify and order these (I_{less}, I_{more}) image pairs, we use the continuous labels produced by our BWS annotations for the former and **WikiArt**'s agreement labels as a proxy for emotional content in the latter. We retrieve and evaluate the top 5 R_{less} and R_{more} rationales for each (I_{less}, I_{more}) (see Section 5.2).

Evaluation Data for Image Recoloring+ Rationale Retrieval. Finally, to evaluate our models together, we retrieve and evaluate the top 5 R_{less} and R_{more} rationales for all 1000 recolored (I_o, I_e) pairs, which we refer to as ''recolored image rationales'' (Section 5.2).

As our domain (art recolorings) and class set (emotions) are both non-standard, automatic image generation metrics like Fréchet Inception Distance (FID) (Heusel et al., 2017) that are trained on ImageNet (Deng et al., 2009) are ill-suited to its evaluation (Kynkäänniemi et al., 2022). Thus, given the novel nature of this task, we rely more heavily on human annotation. Each evaluation task is annotated by 3 Master Workers on Amazon Mechanical Turk (AMT). Their compensation was in line with the minimum wage.

Intended Emotion e	α	≥ 2 Annotators Less	Equal	More	≥ 1 Annotators Less	Equal	More
Anger	0.063	**0.33**	0.325	0.165	**0.74**	0.72	0.56
Disgust	0.027	0.26	**0.385**	0.14	0.75	**0.79**	0.555
Fear	0.047	0.225	**0.355**	0.185	0.72	**0.795**	0.635
Happiness	0.047	0.135	0.325	**0.335**	0.555	**0.785**	*0.765*
Sadness	0.060	0.205	**0.515**	0.13	0.6	**0.86**	0.48

Annotator	1	2	3	4	5	6	7
Less	**0.436**	0.087	**0.5**	0.344	**0.492**	0.062	0.279
Equal	0.233	**0.605**	0.274	**0.43**	0.369	**0.652**	0.429
More	0.331	*0.308*	0.226	0.226	0.138	*0.286*	0.292
# Annotations	776	666	390	381	311	290	161

Figure 5: Image recoloring results. (Top left) Krippendorff's α, the fraction of recolorings with the specified majority label by intended emotion and the fraction of recolorings with the specified label (from any annotator) by intended emotion. (Bottom left) The label distributions of our top 7 most frequent annotators. (Right) The confusion matrix between our intended transformations (x-axis) and the emotion effects indicated by our annotators (y-axis). **Bold** indicates the top label per intended emotion / annotator. *Italics* indicates where ''more'' beats ''less''.

We ensure the quality of these evaluations via a control task which asks annotators to identify the colors in a separate test image. We did not restrict these evaluation tasks to native English speakers. As associations between specific colors and emotions are not universal (Philip, 2006), this may have had a negative effect on both our agreement and scores.

In total, we collected 9000 evaluation annotations which we release as **FeelingGreen**, an additional complementary corpus that could be instructive to researchers working on this task.

5.1 Image Recoloring Results

To evaluate the quality of our image recoloring, for each pair of images (I_o, I_e), we asked annotators whether the recolored image I_e, when compared to I_o, evoked *less*, an *equal* amount or *more* of all 5 of our emotions. Given that our image recolorization model is only designed to increase the specified emotion e, a task that only measures e would be trivial as the desired transformation is always *more*. Conversely, asking annotators to identify e would enforce a single label constraint for a problem that is inherently multilabel.

As is clear from the agreement scores reported in Figure 5, emotion identification is very subjective, a fact corroborated by Mohammad and Kiritchenko (2018) for the abstract genres from which our images are compiled. Therefore, in addition to reporting the percentage of tasks for each emotion with a specific majority label, we include 1) cases where at least 1 annotator selected a given label and 2) the performance of our system according to our top 7 annotators when considered individually.

The scores in Figure 5 demonstrate the difficulty of this task. More often than not, the majority label indicates that our system left the targeted emotion unchanged. In the case of *happiness*, we successfully enhanced its expression in 33.5% of tasks, the sole emotion for which *more* beats both *less* and *equal*. However, the opposite holds for *anger* where we reduced its expression in 33% of tasks while increasing it in just 16.5%. As *less* angry and *more* happy are similar in terms of valence, this suggests a bias in our approach reflected in the confusion matrix in Figure 5. Perhaps the random initialization of palettes and our preference for a diverse set of recolorings for a given I_o result in multi-colored transformations which, while satisfactory to our emotion-guided image selector, appear to most annotators as happy.

Another possibility is that distinct shape (e.g., an unambiguous circle) constrains the emotional potential of color. To test this, we calculate the max CLIP (Radford et al., 2021) similarity between each work of art and terms in our shape

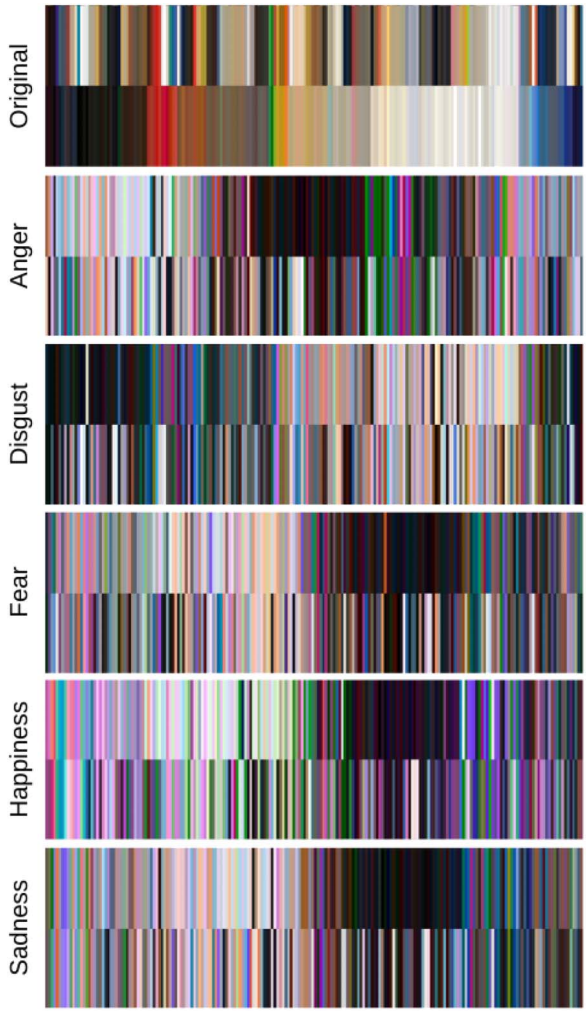

Figure 6: A summary view of our recolorings. The first row features the top 2 colors in every I_o in our evaluation set (where the bottom band displays each I_o's primary color and the top band displays its secondary color in the same position). The other 5 rows depict the top 2 colors for every recolored variant I_e in the same position as its I_o. In these 5 rows, the bands are mirrored with primary on top and secondary on the bottom.

lexicon[7] and consider the difference in scores between the 40 works in the bottom quintile and the 40 works in the top quintile (with the least distinct and most distinct shape according to CLIP). We find that on average an additional $1.5\%, 4.5\%$ (for ≥ 2 annotators and ≥ 1 annotator labeling) of our recolorings were effective (i.e., labeled as *more*) when comparing the bottom shape quintile to the top shape quintile while *less* and *equal* fell by $0\%, 2\%$ and $3\%, 1.5\%$ respectively. This lends some credence to the notion that dominant shapes restrict the breadth of emotions color can connote.

[7]We embed ‘‘an image of [SHAPE]" for each SHAPE.

When we consider our annotators individually, our recolorings were effective for at least one annotator in more than half of the tasks for each emotion. In fact, annotators 2 and 6 indicated that, when not *equal*, our system enhanced the intended emotion. This suggests an opening for emotional recolorings conditioned on the subjectivity of a particular viewer. We leave this to future work.

We display an example recoloring in Figure 4. Additionally, in Figure 6, we present a visualization of the recolorings produced by our system. When read from top (each I_o's top 2 colors) to bottom (its corresponding I_e's top 2 colors, $\forall e \in E$), some interesting properties emerge. It is clear that our diverse image selection heuristic described in Section 4.1.3 is effective, resulting in few overlapping color bands for the same image I_o across all 5 emotions. As expected, recoloring for happiness results in brighter palettes but surprisingly, when the original image begins with a light palette, our system prefers dark primary colors and bright secondary colors, that is, extreme visual contrast. While a few trends for other emotions are also identifiable, the lack of a simple relationship between emotion and generated palette or even original image color, emotion and generated palette suggests that the model is using other deeper contextual features (less prominent colors and the image's composition) to produce its recoloring.

5.2 Rationale Retrieval Results

We evaluate the rationales from R_{less} and R_{more} against two criteria: ‘Descriptive’ and ‘Justifying’. ‘Descriptive’ indicates that the rationale refers to content present in the specified image (for R_{less} this is I_o and for R_{more} this is I_e) and allows us to measure how well our rationale retrieval model correlates image features with textual content, for example, by retrieving rationales with appropriate color words. ‘Justifying’ means that the rationale is a reasonable justification for why the specified image evokes more or less of the target emotion than the other image in its pair. This allows us to measure whether the model 1) picks rationales that identify a difference between the two images and 2) more generally picks rationales that describe patterns of image differences that correspond to perceived emotional differences.

For every image and its more emotional counterpart (either the paired image or its recolored

	Distinct Image			Recolored Image		
	Descriptive	Justifying	Both	Descriptive	Justifying	Both
α	0.021	0.006	–	0.012	−0.073	–
k	@k, wi-k	@k, wi-k	@k, wi-k	@k, wi-k	@k, wi-k	@k, wi-k
1	0.716, 0.716	0.577, 0.577	0.469, 0.469	0.845, 0.845	0.761, 0.761	0.692, 0.692
2	**0.717**, 0.897	**0.587**, **0.801**	**0.475**, **0.683**	**0.842**, **0.963**	**0.753**, **0.908**	**0.682**, **0.851**
5	0.719, 0.989	0.596, 0.968	0.489, 0.908	0.845, 0.999	0.749, 0.995	0.683, 0.971
C	0.683, **0.904**	0.555, 0.779	0.441, 0.655	0.826, 0.952	0.734, 0.905	0.655, 0.839
1	0.729, 0.729	0.669, 0.669	0.545, 0.545	0.796, 0.796	0.694, 0.694	0.614, 0.614
2	**0.726**, **0.912**	**0.650**, **0.852**	**0.527**, **0.738**	0.789, 0.935	0.703, 0.877	0.613, 0.792
5	0.733, 0.990	0.644, 0.979	0.524, 0.920	0.794, 0.994	0.698, 0.979	0.613, 0.946
C	0.662, 0.866	0.580, 0.798	0.448, 0.660	**0.816**, **0.954**	**0.704**, **0.884**	**0.630**, **0.818**

	Our Model ($k = 2$)				Class-Sampled (C)			
Feature	%	Descriptive	Justifying	Both	%	Descriptive	Justifying	Both
has color	**60.3**	**0.765**	**0.665**	**0.564**	54.9	0.724	0.626	0.523
no color	39.7	0.773	**0.686**	**0.590**	**45.1**	**0.774**	0.664	0.569
is concrete	**72.7**	**0.760**	**0.665**	**0.565**	64.2	0.732	0.630	0.529
not concrete	27.3	**0.792**	**0.695**	**0.599**	**35.8**	0.773	0.667	0.569
simile	**27.8**	**0.764**	**0.655**	**0.566**	23.1	0.727	0.637	0.537
no similar	72.2	**0.770**	**0.680**	**0.578**	**76.9**	0.752	0.645	0.546

Table 2: Rationale retrieval results. (Top) Distinct image and recolored image results: Krippendorff's α, IR precisions (@k, within-k) for $k \in \{1, 2, 5\}$ and class-sampled (C) rationales explaining why I_o evokes e *less* intensely (first four rows) and why I_e evokes e *more* intensely (last four rows) across three criteria (Descriptive, Justifying and Both). (Bottom) Prevalence and IR precision (@k) for rationales grouped by ''specificity'' features: color, concreteness and simile. **Bold** indicates the higher score between our $k = 2$ predictions and C.

variant), we asked annotators to evaluate the top five rationales from the pair's R_{less} and R_{more} according to both criteria. As a strong baseline, we include 2 *class-sampled rationales* C randomly sampled from the subset of rationales in **FeelingBlue** justifying image choices for the same emotion and direction (e.g., more angry). Thus, these rationales exhibit language that is directionally correct but perhaps specific to another image. All 7 rationales were randomly ordered so annotators would not be able to identify them by position.

Table 2 reports agreement and two different metrics for each of our criteria across both the ''distinct image'' and ''recolored image'' sets: precision@k and precision-within-k, the percent of top-k rationale groups where at least one rationale satisfied the criterion. Because the validation and unseen splits had similar scores, we present only the union of both. As with our image recoloring evaluation (and emotion annotations more generally), agreement scores are again quite low (though better for 'Descriptive' than 'Justifying'). The 2 class-sampled rationales (C) are a very strong baseline for our model to beat – our model retrieves rationales by comparing combined image representations to the full set of rationales (across all emotions and for both directions), instead of drawing them from the specified emotion and direction subset as is the case for the class-sampled rationales. That precision for the class-sampled rationale is relatively high shows that people tended to gravitate towards similar features as salient to the emotional content of different images. Still, our model regularly outperforms this baseline.

One explanation for the surprising strength of the ''class-sampled'' rationales is that broader, more generally applicable rationales are overrepresented in **FeelingBlue** relative to specific rationales that only apply to certain images. To explore this, in Table 2 we also present the prevalence and scores of rationales from our model and the ''class-sampled'' baseline along three different axes of specificity: color, concrete language and simile (as identified in Section 3.3). The results show that not only was our model

more likely to prefer specific rationales, it also used them more effectively. Because specificity is more easily falsifiable than non-specificity, our model's preference for specificity depresses its aggregate scores relative to the baseline (Simpson's paradox).

Finally, it is interesting that annotators regularly found rationales for our recolored image pairs (I_o, I_e) to be 'Justifying' despite the relatively worse agreement with the intended emotion. As we ask annotators to consider a rationale 'Justifying' assuming the intended emotional difference is true, we cannot conclude that the rationales change the annotators' opinion about the recoloring. But it does show that people can recognize how others might respond emotionally to an image even if they might not agree. We include example retrievals for both variants in Figure 4.

6 Conclusion

We introduce **FeelingBlue**, a new corpus of abstract art with relative emotion labels and English rationales. Enabled by this dataset, we present a baseline system for **Justified Affect Transformation**, the novel task of 1) recoloring an image to enhance a specific emotion and 2) providing a textual rationale for the recoloring.

Our results reveal insights into the emotional connotation of color in context: its potential is constrained by its form and effective justifications of its effects can range from the general to the specific. They also suggest an interesting direction for future work—how much is our emotional response to color affected by linguistic framing? We hope that **FeelingBlue** will enable such future inquiries.

Acknowledgments

We would like to express our gratitude to our annotators for their contributions and the artists whose work they annotated for their wonderful art. Additionally, we would like to thank our reviewers and Action Editor for their thoughtful feedback.

References

2016. IdeelArt: The online gallerist.

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J. Guibas. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579. https://doi.org/10.1109/CVPR46437.2021.01140

Xavier Alameda-Pineda, Elisa Ricci, Yan Yan, and Nicu Sebe. 2016. Recognizing emotions from abstract paintings using non-linear matrix completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5240–5248. https://doi.org/10.1109/CVPR.2016.566

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)*, pages 431–447. https://doi.org/10.1007/978-3-030-01258-8_27

Nalini Bhushan, A. Ravishankar Rao, and Gerald L. Lohse. 1997. The texture lexicon: Understanding the categorization of visual texture terms and their relationship to texture images. *Cognitive Science*, 21(2):219–246. https://doi.org/10.1207/s15516709cog2102_4

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911. https://doi.org/10.3758/s13428-013-0403-5, PubMed: 24142837

Junho Cho, Sangdoo Yun, Kyoungmu Lee, and Jin Young Choi. 2017. PaletteNet: Image recolorization with given color palette. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, volume 2017-July, pages 1058–1066. https://doi.org/10.1109/CVPRW.2017.143

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE. https://doi.org/10.1109/CVPR.2009.5206848

Terry N. Flynn and Anthony A. J. Marley. 2014. Best-Worst Scaling: Theory and methods,

Handbook of Choice Modelling. Edward Elgar Publishing. https://doi.org/10.4337/9781781003152.00014

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*, pages 770–778.

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. *Advances in Neural Information Processing Systems*, 30.

Hye-Rin Kim, Yeong-Seok Kim, Seon Joo Kim, and In-Kwon Lee. 2018. Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia*, 20(11):2980–2992. https://doi.org/10.1109/TMM.2018.2827782

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations (ICLR)*, San Diego, California.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and Best-Worst Scaling. In *NAACL-HLT*, pages 811–817, San Diego, California. https://doi.org/10.18653/v1/N16-1095

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2022. The role of ImageNet classes in Fréchet Inception Distance. *arXiv preprint arXiv:2203.06026*.

Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 83–92. https://doi.org/10.1145/1873951.1873965

Byron Mikellides. 2012. Colour psychology: The emotional effects of colour perception. In *Colour Design*, Elsevier, pages 105–128. https://doi.org/10.1533/9780857095534.1.105

Saif Mohammad. 2011. Even the abstract have color: Consensus in word-colour associations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–373, Portland, Oregon, USA. Association for Computational Linguistics.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-1007

Saif Mohammad and Svetlana Kiritchenko. 2018. WikiArt Emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Christine Mohr and Domicele Jonauskaite. 2022. Why links between colors and emotions may be universal.

Randall Monroe. 2010. Color survey results.

Gill Philip. 2006. Connotative meaning in English and Italian colour-word metaphors. *Metaphorik*, 10.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.

Tianrong Rao, Xiaoxu Li, and Min Xu. 2019. Learning multi-level deep representations for image emotion classification. *Neural Processing Letters*, pages 1–19.

Andreza Sartori, Victoria Yanulevskaya, Almila Akdag Salah, Jasper Uijlings, Elia Bruni, and Nicu Sebe. 2015. Affective analysis of professional and amateur abstract paintings

using statistical analysis and art theory. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(2):1–27. https://doi.org/10.1145/2768209

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):1929–1958.

Tina M. Sutton and Jeanette Altarriba. 2016. Color associations to emotion and emotion-laden words: A collection of norms for stimulus construction and selection. *Behavior Research Methods*, 48(2):686–728. https://doi.org/10.3758/s13428-015-0598-8, PubMed: 25987304

Morgan Ulinski, Victor Soto, and Julia Hirschberg. 2012. Finding emotion in image descriptions. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pages 1–7. https://doi.org/10.1145/2346676.2346684

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2Vec: Learning generalized emotion representation by multi-task training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 292–298, Brussels, Belgium. Association for Computational Linguistics.

Jufeng Yang, Yan Sun, Jie Liang, Bo Ren, and Shang-Hong Lai. 2019. Image captioning by incorporating affective concepts learned from both visual and textual components. *Neurocomputing*, 328:56–68. https://doi.org/10.1016/j.neucom.2018.03.078

He Zhang, Eimontas Augilius, Timo Honkela, Jorma Laaksonen, Hannes Gamper, and Henok Alene. 2011. Analyzing emotional semantics of abstract art using low-level image features. In *International Symposium on Intelligent Data Analysis*, pages 413–423. Springer. https://doi.org/10.1007/978-3-642-24800-9_38

Sicheng Zhao, Guiguang Ding, Yue Gao, and Jungong Han. 2017. Approximating discrete probability distribution of image emotions by multi-modal features fusion. *Transfer*, 1000(1):4669–4675. https://doi.org/10.24963/ijcai.2017/651

Sicheng Zhao, Xin Zhao, Guiguang Ding, and Kurt Keutzer. 2018. EmotionGAN: Unsupervised domain adaptation for learning discrete probability distributions of image emotions. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 1319–1327. https://doi.org/10.1145/3240508.3240591