Cross-Modality Personalization for Retrieval

Nils Murrugarra-Llerena Adriana Kovashka
Department of Computer Science
University of Pittsburgh
{nineil, kovashka}@cs.pitt.edu

Abstract

Existing captioning and gaze prediction approaches do not consider the multiple facets of personality that affect how a viewer extracts meaning from an image. While there are methods that consider personalized captioning, they do not consider personalized perception across modalities, i.e. how a person's way of looking at an image (gaze) affects the way they describe it (captioning). In this work, we propose a model for modeling cross-modality personalized retrieval. In addition to modeling gaze and captions, we also explicitly model the personality of the users providing these samples. We incorporate constraints that encourage gaze and caption samples on the same image to be close in a learned space; we refer to this as content modeling. We also model style: we encourage samples provided by the same user to be close in a separate embedding space, regardless of the image on which they were provided. To leverage the complementary information that content and style constraints provide, we combine the embeddings from both networks. We show that our combined embeddings achieve better performance than existing approaches for cross-modal retrieval.

1. Introduction

We perceive the world through our five senses, but our perception is also affected by our experiences, personality, and bias. Thus, the meaning we attribute to the visual world is a function not only of the image pixels but of the individual way in which we look at an image. Studies show a variety of links between visual perception and the viewer's personality and emotion. For example, "open-minded" people are more likely to combine visual elements and perceive them as a unified whole [3], disorganized people or ones with low self-confidence have a high tolerance of visual blur [45], and people who believe in paranormal events are more likely to perceive actual objects in images that only contain noise [30]. Further, people's perception of physical properties is affected by emotions such as fear and joy [51].

This variance in perception due to variance in personality is important to consider when predicting what meaning viewers will extract from imagery. It is especially important to model when predicting how humans will describe images that aim to impart opinions on the viewer in subtle ways. Prior work has examined the meaning that the average human extracts from images, by learning to predict what descriptive captions are appropriate for a given image. However, not all humans will describe the image in the same manner. Further, the way they describe it depends on how they look at it. We illustrate this in Fig. 1. When shown

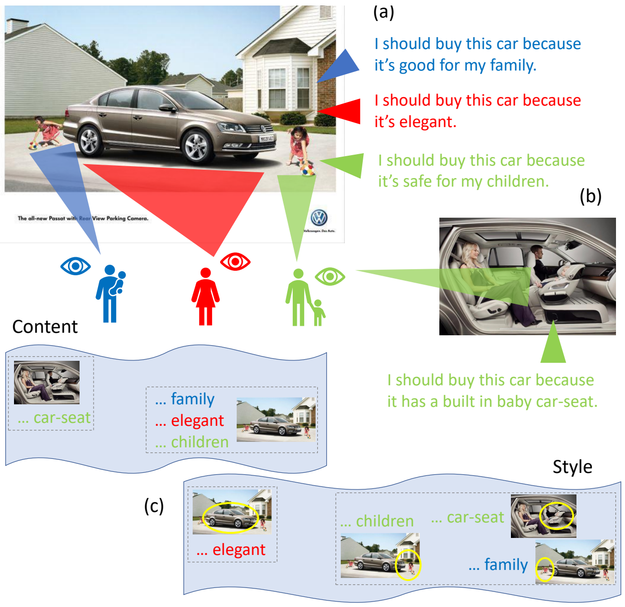

(a)
I should buy this car because it's good for my family.
I should buy this car because it's elegant.
I should buy this car because it's safe for my children.
(b)
I should buy this car because it has a built in baby car-seat.
Content
... car-seat
... family
... elegant
... children
Style
(c)
... elegant
... children
... car-seat
... family

Figure 1. People with different personalities might perceive and describe the same image differently. A social, family person might observe the children, and an artistic person might perceive the elegance of the vehicle, in this car ad (a). Further, we expect there is consistency between how the same person observes and describes images (b). To link content across modalities, but preserve different users' styles, we combine both content and style constraints (c). The former encourages samples provided on the same image to be close in a learned space, while the latter encourages samples by the same user to be close. The text in (c) denotes the full captions, and ellipses denote what part of the image the viewer focused on. Dashed boxes denote closeby items.

this car advertisement, an outgoing family man might first observe the children near the car, and interpret the message of the ad as emphasizing the safety features which are important for one's family. An artistic single woman might first fixate on the visual elegance of the car. As a result, viewers might describe the image content in different order, or even omit elements that are not interesting to them.

In this work, we study the relationship between personality, gaze and captioning. We predict how a user will caption an image, conditioned on how they looked at this image; and conversely, how they might look at the image, given how they described it. We learn a joint image-gaze-text-personality embedding space, in which we separately model content and style. Our *content* model constraints couple gaze and caption annotations that correspond to the same image, regardless of which user provided the annotation. On the other hand, our *style* constraints ensure the strongest association between gaze (or caption) samples of the same user, regardless of which image they were provided on. To leverage the complementary intuitions of content and style, we further weigh and combine the embeddings from both networks. We use these embeddings to retrieve content across modalities, in a pool of samples associated with different images and/or annotated by different users. For example, given how a person looked at an image, we learn to predict how that person might caption the image, in contrast to other users' captions on the same or different images.

We collect two cross-modality per-annotator datasets capturing gaze, captions and personality. Each annotator examines fifteen images. We record which parts of the image they examined. We also ask them to describe the meaning of each image they saw, i.e. to caption it. Finally, we ask them to respond to a ten-question personality survey. We find that when retrieving samples for each user across modalities, it is important to model the similarity in the annotations that user provided. In contrast, methods that only capture similarities in content but not personal style, produce weaker retrieval results. We also compare to a recent personality-aware method which considers single words in the form of tags, and we achieve a stronger result.

Our approach can be used in a social media context. For example, we can use it to predict how a Facebook user might caption their photos, using an estimate of their personality from their social media profile [24] and/or information about what posts they pay attention to. Alternatively, we can infer how users might construct narratives from their vacation photos, or how they might look at others' vacation photos depending on their interests.

In summary, our contributions include:

- Two datasets of caption-gaze samples for 139 and 79 unique users, respectively, and over 4000 annotations on 900 unique images, with worker identity preserved. The data is available at `http://cs.pitt.edu/~nineil/crossmod/`, and can be used by other researchers investigating personalized perception.
- A novel method that separately considers style and content, and combines them to achieve effective personality-aware retrieval across three modalities.
- An examination of the latent interdependency of these three modalities: learning all three jointly can be beneficial, even if only two are used at test time.

2. Related work

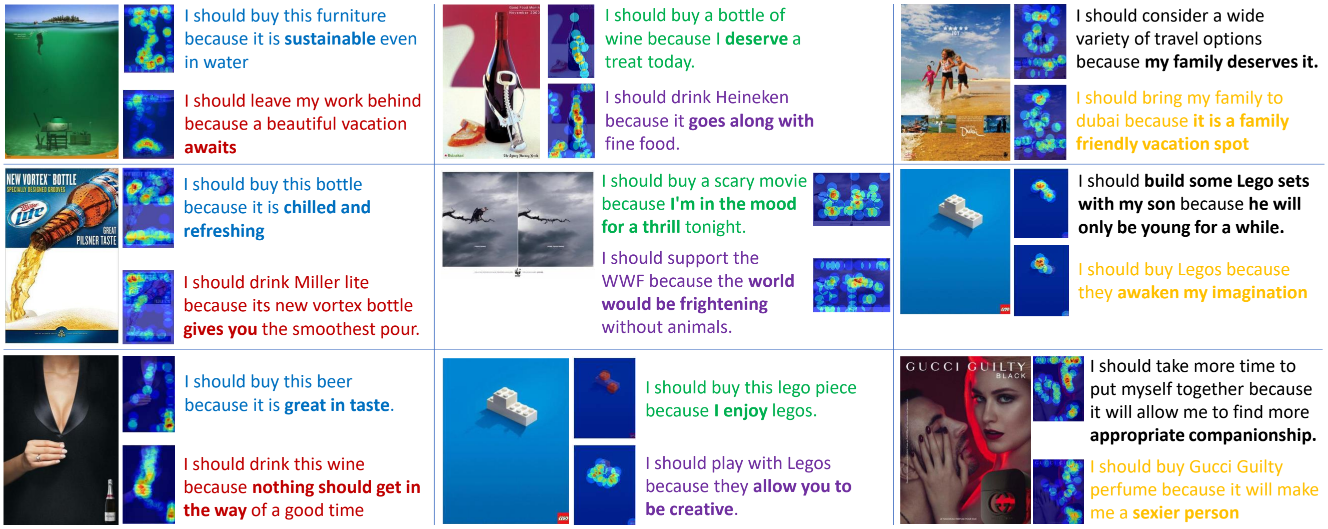
Image captioning. There is a large body of work [2, 33, 43, 49, 44, 20, 8] on automatic image captioning, or predicting a description for a given visual. Common approaches include learning a joint image-text embedding using triplet loss or by maximizing the correlation of the two modalities [10, 9, 48]; training a recurrent network that predicts a sequence of words conditioned on the image and outputs at previous timesteps [2, 44, 20, 8]; learning a template description and how to fill each position of the template with a word [27]; generative adversarial approaches [7]; etc.

Most captioning approaches assume all users would caption an image in the same way. In contrast, [6] learn individual differences in how an annotator describes an image, and [42] learn the types of hashtags a user might provide. However, none of these consider two manifestations or channels of personality as we do (i.e. gaze and captions). We show that having information from multiple modalities at training time allows us to better understand user differences.

Gaze. Saliency prediction [17, 19, 18, 29] models what humans fixate on in an image. Prior work has examined the relationship between sentiment and gaze [11] and the differences between viewers in how they look at an image [47], but none has examined the relationship between personalized *perception* and personalized *meaning*. A few authors have examined the relationship between captions and attention. For example, [50, 37, 46, 26] predict captions conditioned on an attention map (learned from human gaze or discovered from a classification loss). However, these do not consider personalized captioning or gaze as we do.

Style vs content. In our work, we aim to separate gaze/caption similarities arising due to content and style. In prior work, [52] separate content and style for handwritten Chinese characters, by training separate networks for each, and [38] use a linear model in both the content (character ID) and handwriting style. [14, 40, 12, 13, 25, 5, 39] learn domain-invariant representations for object recognition, where objects are the "content" and modalities (e.g. paintings, sketches) are the "style". We have *multiple* content modalities, and multiple styles (one per user). Also relevant is [22] which train per-user attribute models, but this work only considers one modality.

Privileged information. Our approach utilizes a type of "privileged" feature information, which is available at train-



I should buy this furniture because it is **sustainable** even in water

I should leave my work behind because a beautiful vacation **awaits**

I should buy a bottle of wine because I **deserve** a treat today.

I should drink Heineken because it **goes along with** fine food.

I should consider a wide variety of travel options because **my family deserves it.**

I should bring my family to dubai because **it is a family friendly vacation spot**

I should buy this bottle because it is **chilled and refreshing**

I should drink Miller lite because its new vortex bottle **gives you** the smoothest pour.

I should buy a scary movie because **I'm in the mood for a thrill** tonight.

I should support the WWF because the **world would be frightening** without animals.

I should **build some Lego sets with my son** because **he will only be young for a while.**

I should buy Legos because they **awaken my imagination**

I should buy this beer because it is **great in taste**.

I should drink this wine because **nothing should get in the way** of a good time

I should buy this lego piece because **I enjoy** legos.

I should play with Legos because they **allow you to be creative**.

I should take more time to put myself together because it will allow me to find more **appropriate companionship.**

I should buy Gucci Guilty perfume because it will make me a **sexier person**

Figure 2. Text and gaze samples for different users on our *Ads* data. Each column shows three images annotated by two users.

ing time only. Such information is useful to learn the structure of the space, and then utilize it at test time with only a subset of the input types. Prior work includes [41, 35, 15, 28, 4]. For example, [35] use privileged information to learn which samples are easy to learn from, and [4] regularize the parameters of one network with another learned from privileged data. In contrast, we use privileged information for caption retrieval.

3. Approach

Since no prior dataset exists that considers personalized annotations in multiple modalities, we first collect such a dataset (Sec. 3.1). We next describe the cross-modal personalized retrieval scenarios we consider (Sec. 3.2), and the cues we use to learn an embedding space using standard content (Sec. 3.3) and style (Sec. 3.4), in combination with a base network (Sec. 3.5, 3.6). We finally describe how we learn a joint space for all modalities (Sec. 3.7), and conclude with implementation details (Sec. 3.8).

3.1. Dataset

We collected two datasets. First, we collected an *Ads* dataset of 2700 annotations total, over 543 unique images (of which three were used for annotation quality validation), 3 modalities, and from 139 unique viewers (180 separate tasks, but some users completed more than one task). We used the dataset of [16] which contains 64,832 advertisements. In particular, we constructed 60 sets with 15 randomly sampled images each, from the topics: alcohol, travel, beauty, and animal rights. We showed each set to three annotators. Second, we complemented this dataset with a subset of images from COCO [23]. We selected cluttered images with many objects. Our *COCO* data contains 1350 annotations total, over 363 unique images, 3 modalities, and 79 unique viewers. For each image in the set, annotators were asked to provide the following annotations.

- **Gaze:** We simulated gaze capture, using the BubbleView interface [21]. It shows a blurred image and asks the viewer to click on parts of it, revealing clear circular regions; it is known to return data strongly correlated with gaze. This interface allows us to crowdsource the collection. We recorded the locations and order of clicks.
- **Caption text:** For ads, we asked annotators to describe the meaning of the image in the form "I should [action that the ad prompts] because [reasoning that the ad provides]." e.g. "I should buy this perfume because it will make me attractive." For COCO images, we asked annotators to describe what the image shows.
- **Personality:** Finally, we asked annotators to answer ten multiple-choice questions [32] about their personality, including characteristics such as being artistic, trusting, neurotic vs laid-back, etc. The complete personality questionnaire is provided in our supp. material.

We used Amazon Mechanical Turk (MTurk) to collect our data. To ensure quality, we restricted access to annotators with 98% approval on completed tasks, over at least 1000 submitted tasks. As a form of quality control, we incorporated validation images for which the gaze map should be simple to predict, as they contain objects in a small portion of the image and a plain background. If the acquired gaze map and the object do not intersect, the whole set of annotations are discarded and the tasks are resubmitted.

Annotation samples. In Fig. 2, we show text and gaze samples that different users provided on the same image. Each column shows the results of the same two users; the top responses are from one, and the bottom from another.

In the first column, we observe that the first user (in blue) uses more adjective words, while the second (in red) uses more verbs. For example, in the second row, the first annotator describes the drink as being "chilled and refreshing" while the second describes the ad in a more active way, i.e.

the bottle "gives you" a certain pour. From their answers to the personality questions, the second viewer is more extroverted, which aligns with energetic feelings and using verbs.

In the second column, the first user (in green) says "I deserve", "I'm in the mood for", "I enjoy", i.e. the responses come from an ego-centric perspective. The second viewer (in purple) focuses more on the state of the world and properties of products, i.e. a more analytical perception. We observe a correlation between the personality inferred from text, and the gaze maps provided. For example, the "self-centered" viewer in green has a lazier approach to examining the image, while the more analytical one is more thorough. From their personality responses, the second viewer exhibits more neuroticism (low self-esteem) than the first.

In the third column, the first viewer (in black) emphasizes his or her relationship with others (e.g. family, child, companion). The second viewer (in orange) focuses more on themselves (e.g. "awaken my imagination", "make me sexier"). Similarly, in the third image, the first viewer pays close attention to the face of the man. In contrast, the more self-centered viewer only looks at the woman (the "protagonist" of the ad). From their personality responses, the first viewer is more agreeable than the second one. Agreeableness is closely related to generosity, empathy and sympathy, which relates to making a connection with others.

Representation. For images, we extract Inception-v4 CNN features [36]. We then mask the image convolution feature with the BubbleView saliency map, by resizing the saliency map to the convolution feature size and multiplying them together. Finally, average pooling is performed to obtain a 1536-dimensional feature vector. We represent textual descriptions as their average 200-dimensional Glove embedding [31]. For personality, we use a 10-dimensional feature vector containing the scores for the personality questions in [32]. Below, we describe how we learn projections of these representations that place them in the same feature space.

3.2. Tasks and embeddings

We consider three modalities: gaze, text (captions), and personality. We consider six retrieval tasks: gaze to personality (g2p), text to personality (t2p), personality to gaze (p2g), text to gaze (t2g), gaze to text (g2t), and personality to text (p2t). In all of these, we wish to retrieve an annotation that a given user provided, upon receiving another sample from that same user on the same image, but in a different modality (e.g. retrieve the text the annotator wrote to describe the image, conditioned on how the user looked at that image).

We learn a joint embedding of images, gaze, captions, and personality. Our key hypothesis is that bridging modalities through a content loss that ensures samples on the same image, regardless of modality, project closeby, is insufficient for this task. In addition, we need to model the type of captions/gaze/personality that a user demonstrates, by also bridging samples from the same user, *regardless of the image* on which they were provided. Our approach's key intuition is summarized in Fig. 3.

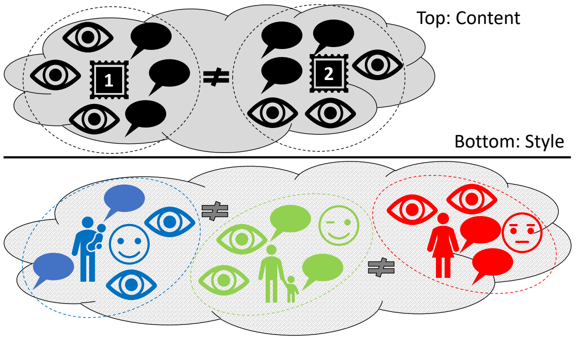


Figure 3. Standard approaches use a content-type loss for cross-modal retrieval, which ensures that samples provided for the same image map are placed in similar position in the learned space. Here these samples are gaze-masked images and captions. In contrast, we argue that a style-based loss is also necessary. In particular, we wish to ensure that samples that a particular user provided, regardless of the image on which they were provided, cluster together.

We ensure these similarities through triplet constraints. First, we project each modality to a shared 200-dimensional feature vector via a fully connected layer. Second, for every pair of modalities, $\boldsymbol{x}$ (input) and $\boldsymbol{y}$ (output), we generate the content and style constraints described below.

3.3. Content Network

We use the following constraints to learn a joint embedding that couples the representations across modalities, for data samples that correspond to the same image. Let us denote a textual description of image i provided by user a as $\mathbf{t}_i^a$, and a gaze map for the same image from the same user by $\mathbf{g}_i^a$. The image that was shown to obtain this text/gaze is denoted by $\mathbf{v}_i^a$. For compactness, we show constraints in a more general form, using $\boldsymbol{x}$ to denote one modality embedding and $\boldsymbol{y}$ to denote a different modality embedding. The original image is only used as an anchor modality; it is not part of our $\{\boldsymbol{x}, \boldsymbol{y}\}$ modality pairs, and is denoted separately.

The embeddings for the following pairs should be similar (where $*$ denotes *any user*, and i and j denote distinct images): $\{\boldsymbol{x}_i^*, \boldsymbol{y}_i^*\}$; $\{\boldsymbol{x}_i^*, \boldsymbol{x}_i^*\}$; $\{\boldsymbol{y}_i^*, \boldsymbol{y}_i^*\}$; $\{\boldsymbol{v}_i^*, \boldsymbol{x}_i^*\}$; and $\{\boldsymbol{v}_i^*, \boldsymbol{y}_i^*\}$. For example, if $\boldsymbol{x}$ refers to text and $\boldsymbol{y}$ refers to gaze, text and gaze samples provided on the same image should be similar; text samples from different users provided on the same image should be similar (and same for gaze samples); and the text and gaze samples' representations should be similar to the original image representation. The last two constraints are necessary because each image is observed by three users, and each provides a potentially different gaze map or caption. We primarily model visual content through the gaze-masked image, which we refer to as the gaze map. However, we would like to ensure the maps for the same image have similar representation.

The following representations should be dissimilar: $\{\boldsymbol{x}_i^*, \boldsymbol{y}_j^*\}$; $\{\boldsymbol{x}_i^*, \boldsymbol{x}_j^*\}$; $\{\boldsymbol{y}_i^*, \boldsymbol{y}_j^*\}$; $\{\boldsymbol{v}_i^*, \boldsymbol{x}_j^*\}$; and $\{\boldsymbol{v}_i^*, \boldsymbol{y}_j^*\}$. These are the same as before, but the subscript in the second sample in each pair is j, referring to a *different* image than the anchor. We generate triplet constraints from these, using all data in the current batch.

For content, we consider the following pairs of modalities as $\{\boldsymbol{x}, \boldsymbol{y}\}$: $\{\boldsymbol{t}, \boldsymbol{g}\}$, and $\{\boldsymbol{g}, \boldsymbol{t}\}$. We train a single network using Eq. 1 below to bridge the text and gaze modalities. It does not, however, make sense to consider the following: $\{\boldsymbol{g}, \boldsymbol{p}\}$, since the same personality matches multiple images, yet multiple different users (with different personalities) annotated the same images; nor $\{\boldsymbol{t}, \boldsymbol{p}\}$, $\{\boldsymbol{p}, \boldsymbol{g}\}$, $\{\boldsymbol{p}, \boldsymbol{t}\}$.

We would like to ensure that the distances between samples across modalities minimize the following loss:

$$\begin{aligned} L_c(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{v}; \boldsymbol{\theta}) = \sum_{i=1}^{K} \Big[& \sum_{j \in N} \left[\|\boldsymbol{x}_i^* - \boldsymbol{y}_i^*\|_2^2 - \|\boldsymbol{x}_i^* - \boldsymbol{y}_j^*\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|\boldsymbol{y}_i^* - \boldsymbol{x}_i^*\|_2^2 - \|\boldsymbol{y}_i^* - \boldsymbol{x}_j^*\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|\boldsymbol{x}_i^* - \boldsymbol{x}_i^*\|_2^2 - \|\boldsymbol{x}_i^* - \boldsymbol{x}_j^*\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|\boldsymbol{y}_i^* - \boldsymbol{y}_i^*\|_2^2 - \|\boldsymbol{y}_i^* - \boldsymbol{y}_j^*\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|\boldsymbol{v}_i^* - \boldsymbol{x}_i^*\|_2^2 - \|\boldsymbol{v}_i^* - \boldsymbol{x}_j^*\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|\boldsymbol{v}_i^* - \boldsymbol{y}_i^*\|_2^2 - \|\boldsymbol{v}_i^* - \boldsymbol{y}_j^*\|_2^2 + \alpha \right]_+ \Big] \end{aligned} \tag{1}$$

where K is batch size; N is the set of negative samples in the batch; and α is the triplet margin.

3.4. Style Network

The style network captures the similarities between different samples that the same user provided. Let a and b denote distinct users. Thus, the embeddings for the following should be similar, where $*$ denotes *any image*: $\{\boldsymbol{x}_*^a, \boldsymbol{x}_*^a\}$; $\{\boldsymbol{y}_*^a, \boldsymbol{y}_*^a\}$; and $\{\boldsymbol{x}_*^a, \boldsymbol{y}_*^a\}$. In other words, annotations provided by the same user (in the same or different modalities) should be similar, regardless of the image. The following should be dissimilar: $\{\boldsymbol{x}_*^a, \boldsymbol{x}_*^b\}$; $\{\boldsymbol{y}_*^a, \boldsymbol{y}_*^b\}$; and $\{\boldsymbol{x}_*^a, \boldsymbol{y}_*^b\}$.

We consider the following three symmetric pairs of input-output modalities $\{\boldsymbol{x}, \boldsymbol{y}\}$: $\{\boldsymbol{t}, \boldsymbol{g}\}$, $\{\boldsymbol{g}, \boldsymbol{p}\}$, $\{\boldsymbol{t}, \boldsymbol{p}\}$, We train separate networks, each bridging the corresponding two modalities. Note that when the input modality is $\boldsymbol{p}$, there can be fifteen positives (or more if an annotator completed more than one task) for text/gaze.

We seek to minimize the following expression:

$$\begin{aligned} L_s(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = \sum_{i=1}^{K} \Big[& \sum_{j \in N} \left[\|\boldsymbol{x}_*^a - \boldsymbol{y}_*^a\|_2^2 - \|\boldsymbol{x}_*^a - \boldsymbol{y}_*^b\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|\boldsymbol{y}_*^a - \boldsymbol{x}_*^a\|_2^2 - \|\boldsymbol{y}_*^a - \boldsymbol{x}_*^b\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|\boldsymbol{x}_*^a - \boldsymbol{x}_*^a\|_2^2 - \|\boldsymbol{x}_*^a - \boldsymbol{x}_*^b\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|\boldsymbol{y}_*^a - \boldsymbol{y}_*^a\|_2^2 - \|\boldsymbol{y}_*^a - \boldsymbol{y}_*^b\|_2^2 + \alpha \right]_+ \Big] \end{aligned} \tag{2}$$

3.5. Base network

We ensure these similarities through the triplet constraint losses described above, which are added on top of a base network. As our base network, we use VSE++ on Ads, which is an adaptation of VSE++ [10] on the dataset of [16], implemented in [48]. This network also employs content-type constraints. It employs the following loss:

$$\begin{aligned} L_b(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = \sum_{i=1}^{K} \Big[& \sum_{j \in N} \left[\|\boldsymbol{x}_i^a - \boldsymbol{y}_i^a\|_2^2 - \|\boldsymbol{x}_i^a - \boldsymbol{y}_j^a\|_2^2 + \alpha \right]_+ \\ & + \sum_{j \in N} \left[\|\boldsymbol{y}_i^a - \boldsymbol{x}_i^a\|_2^2 - \|\boldsymbol{y}_i^a - \boldsymbol{x}_j^a\|_2^2 + \alpha \right]_+ \Big] \end{aligned} \tag{3}$$

In other words, two samples (in different modalities) from the same user on the same image should be close by, while samples from the same user on different images should be further. However, each user only provided a single sample from each modality on a given image, so we cannot constrain samples on the same image to be close.

Note that we also experimented with ADVISE from [48] as our base network, but it performed worse due to the discrepancy between unmasked and gaze-masked images; see supp for an explanation.

3.6. Combining base, content and style

We also compute a combined embedding. We assign weights on each embedding; β_b for base, β_c for content, and β_s for style. The embedding for each modality becomes:

$$\boldsymbol{x} = \beta_b * \boldsymbol{x}^b + \beta_c * \boldsymbol{x}^c + \beta_s * \boldsymbol{x}^s \tag{4}$$

where $\boldsymbol{x}^b$ denotes the embedding obtained from Eq. 3, $\boldsymbol{x}^c$ from Eq. 1, and $\boldsymbol{x}^s$ from Eq. 2. We optimize the weights on a validation set, separately for each task, using values in the range [0, 1] with step 0.25. In the case of text-to-personality and gaze-to-personality (and vice versa), we use a subset of content constraints, only to ensure gaze/text samples on the same image are similar, and those samples are similar to the corresponding image representation.

3.7. Joint embedding and privileged information

In the above description, we create separate networks for each pair of modalities. However, we can also embed all constraints for all pairs into the same space. This means that even if our goal is to retrieve text given personality, and we do not plan to retrieve e.g. text with gaze as input, knowing about the relationship between text and gaze provides additional useful information for the main task. This can be seen as exploiting privileged information, i.e. information that is only available at training time (since at test time, we do not receive gaze as input). Thus, we add the terms from Eqs. 1, 2, 3 for any pair of modalities, into the same loss, and train a single network. We show in Sec. 4.4 that a joint embedding and privileged information improve our system's accuracy on several tasks.

3.8. Implementation details

We implemented the networks using TensorFlow [1]. We use the Adagrad optimizer, a batch size of 128, a learning rate of 2,[1] an L2 regularizer of 1e-6, 10,000 steps and $\alpha = 0.2$. Every thirty seconds, the network was evaluated on a validation set, and the network with the highest accuracy was selected for testing. For the base network, we found semi-hard negative mining [34] worked best. We selected the negative example with smallest $d(a, n)$ that satisfies $d(a, p) < d(a, n)$, where a is the anchor, p its positive annotation, n a negative example and d denotes a distance measure. If the condition was not satisfied, the negative with the largest $d(a, n)$ was selected.

4. Results

We verify the contribution of the components of our method, by comparing to the combined network. We also compare to [42]. We next show the relationship between all three modalities, using a single network trained for all tasks.

4.1. Setup and metrics

We use a test setup where one image is considered a positive; for example, if the input is a gaze sample, the one desired retrieval result is the caption the same user provided on the same image. The negatives are samples provided by other users or on other images. In other words, given a sample $\boldsymbol{x}_i^a$ (caption, gaze, personality) from user a on image i, retrieve sample $\boldsymbol{y}_i^a$ from the same user on the same image, in the presence of 14 other samples: two negatives $\boldsymbol{y}_i^b$, i.e. on the same image but from other users, and 12 negatives $\boldsymbol{y}_j^b$, where i and j are distinct images. We split the data over users, in 70% for training, 10% for validation and 20% for testing. No user is present in both the training and test sets. We run our experiments in five different shuffle splits.

We show three evaluation metrics: top-1 accuracy (is the top-retrieved result the correct one, where higher is better), top-3 accuracy (are any of the top-3 results the correct one), and rank (what is the rank of the correct result among the 15 ranked samples, where lower is better). We use top-1 accuracy to select the best network snapshot per task and per method, because retrieving the correct result at the very top of the 15 samples is the most challenging task.

	Veit [42]	Base [10]	Content	Style	**Ours**
g2p	**1.33**	1.67	5.00	3.33	3.67
t2p	4.00	2.00	5.00	2.67	**1.33**
p2g	2.67	**1.67**	5.00	3.67	2.00
t2g	4.00	2.67	2.33	5.00	**1.00**
g2t	3.33	3.67	2.00	5.00	**1.00**
p2t	4.00	3.00	5.00	2.00	**1.00**
avg	3.22	2.44	4.06	3.61	**1.67**

Table 1. Summary table for the *Ads* dataset using top-1, top-3 accuracy and rank metrics for the task-specific setup. We show the average rank (lower is better) for each method across the three metrics. The best performer per task is in **bold**.

	Veit [42]	Base [10]	Content	Style	**Ours**
g2p	2.67	**2.00**	5.00	3.00	**2.00**
t2p	3.67	3.00	5.00	2.00	**1.33**
p2g	2.33	3.67	5.00	**1.67**	2.33
t2g	4.00	3.00	2.00	5.00	**1.00**
g2t	4.00	3.00	1.67	5.00	**1.33**
p2t	3.67	2.33	5.00	3.00	**1.00**
avg	3.39	2.83	3.94	3.28	**1.50**

Table 2. Summary table for the *COCO* dataset using top-1, top-3 accuracy and rank metrics for the task-specific setup. We show the average rank (lower is better) for each method.

4.2. Methods compared

Our method is the one described in Sec. 3.6. It is composed of three constituents, each described in Sec. 3.3, 3.4 and 3.5. We compare all three components below, and their combination, and refer to these as BASE, CONTENT, STYLE, and OURS (combined). The BASE result captures the performance of VSE++ [10], which is a state of the art cross-modality embedding method but does not consider personality. We also compare to VEIT [42], which is a method that considers personality and predicts hashtags that a particular user would provide on a given image. All methods have access to the same collected data.

4.3. Benefit of combining content and style

We separately evaluate all methods according to each metric described above, and summarize the results. For each task and each metric, we rank each method from best to worst (with rank 1 being best). We then average the ranks across the three metrics, and show the result in Tab. 1 (for ads) and 2 (for COCO). We present the top-3 accuracy result in Tab. 3 and 4, and the tables for top-1 and rank are in supp. As discussed in Sec. 3.3, the CONTENT method only makes sense in the case of retrieving gaze from captions and vice versa, so it produces no result for the other

[1] In preliminary experiments, we compare 4 learning rates: 0.01, 0.05, 0.1 and 2; and select the best using a rank average of top-1, top-3 accuracy and rank. Performance was similar between 0.01, 0.05 and 2. Thus, we break ties by the best top-3 and top-1 accuracy, which is the hardest task.

	Veit [42]	Base [10]	Content	Style	**Ours**
g2p	0.2107	**0.2111**	N/A	0.206	0.2051
t2p	0.2625	**0.2894**	N/A	0.2806	0.2861
p2g	0.1671	**0.1754**	N/A	0.1643	0.1704
t2g	0.3783	0.4023	0.4384	0.2704	**0.4426**
g2t	0.3801	0.3745	0.4366	0.3074	**0.4463**
p2t	0.2556	0.2718	N/A	0.2741	**0.2768**
avg	0.2757	0.2874	0.1458	0.2505	**0.3046**

Table 3. Top-3 accuracy for the *Ads* dataset for the task-specific setup (higher is better). The best performer per task is in **bold**. N/A values were replaced with zero for average calculation.

	Veit [42]	Base [10]	Content	Style	**Ours**
g2p	0.2121	**0.2222**	N/A	0.2194	**0.2222**
t2p	0.2954	0.2926	N/A	**0.3102**	0.3074
p2g	0.1685	0.1556	N/A	**0.1759**	0.1639
t2g	0.4852	0.5371	0.6139	0.3269	**0.6250**
g2t	0.4639	0.5204	0.5972	0.3657	**0.6065**
p2t	0.2722	0.2769	N/A	0.2787	**0.2833**
avg	0.3162	0.3341	0.2019	0.2795	**0.3681**

Table 4. Top-3 accuracy for the *COCO* dataset for the task-specific setup (higher is better). The best performer per task is in **bold**. N/A values were replaced with zero for average calculation.

	Veit [42]	Base [10]	Style	**Ours**
g2p	3.33	3.33	**1.33**	2.00
t2p	4.00	3.00	1.67	**1.33**
p2g	4.00	3.00	**1.00**	2.00
t2g	3.00	2.00	4.00	**1.00**
g2t	2.67	2.33	4.00	**1.00**
p2t	3.67	3.33	**1.33**	1.67
avg	3.44	2.83	2.22	**1.50**

Table 5. Summary table showing rank of each method for the joint setup on our *Ads* data (lower is better). Content does not apply because it does not consider personality.

	Veit [42]	Base [10]	Style	**Ours**
g2p	0.2056	0.2042	**0.2083**	0.2051
t2p	0.2611	0.2852	0.3019	**0.3134**
p2g	0.1532	0.1787	**0.1815**	0.1792
t2g	0.3625	0.3843	0.2671	**0.4079**
g2t	0.3820	0.3847	0.294	**0.4120**
p2t	0.2528	0.2569	**0.2847**	0.2810
avg	0.2695	0.2823	0.2563	**0.2998**

Table 6. Top-3 accuracy for joint setup on *Ads* (higher is better).

	Per task	Joint
g2p	0.206	**0.2083**
t2p	0.2806	**0.3019**
p2g	0.1643	**0.1815**
t2g	**0.2704**	0.2671
g2t	**0.3074**	0.294
p2t	0.2741	**0.2847**
avg	0.2505	**0.2563**

Table 7. Results of Style on *Ads*; top-3 accuracy.

tasks. Here we model all tasks separately i.e. the first/third, second/sixth, and fourth/fifth rows in each table correspond to the same network.

From Tab. 1 and 2, we see our approach outperforms the rest in nine out of twelve tasks, and ranks second in two of the remaining ones. In contrast, Veit and Style are the best for one task each, and Base for two (including a tie with our method). For top-3 accuracy (Tab. 3 and 4), our approach outperforms all other methods in seven out of twelve tasks. The second and third competitors are Base and Style having the best performance in four and two tasks, respectively. We observe that our approach especially boosts the performance of t2g and g2t, where Content and Style provide complementary information. In contrast, for tasks g2p, p2g, t2p and p2t; Style can provide redundant information to the personality input/output itself, because style is very closely related to personality. We also observe that Veit is not among the top baselines. A possible reason could be the difficulty to find useful latent variables during matrix factorization. Also note that Veit requires more parameters than the other methods, which makes learning harder: due to two FC layers and matrix factorization, it has $O[(d1 + d2) * m + m^2]$ parameters where $d1, d2$ are the modality dimensions and m is the embedding dimension versus $O[(d1+d2)*m]$ for the other approaches.

Most related modalities. We observe that in terms of top-3 accuracy for the combined method, the easiest tasks (and hence the most related two modalities) are g2t/t2g on both *Ads* and *COCO*, followed by p2t/t2p, then by g2p/p2g, which is the hardest. Thus, text and gaze, and personality and text, are most tightly coupled, while the connection between gaze and personality is weaker. This finding is also confirmed by our identity classifier (Sec. 4.5). Also note that for t2g/g2t especially, prediction is much easier on *COCO* than on *Ads*, likely due to smaller variance.

4.4. Joint modeling of all tasks

We next show that all three modalities are interdependent. Even if the task is to retrieve a caption based on gaze, i.e. personality is neither input nor output, it helps to model personality jointly with text and gaze. For this experiment and the following ones, we use our *Ads* data, because retrieval on it appears more challenging (Tab. 3 and 4).

In Tab. 6, we show the top-3 accuracy result using our joint modeling of all modalities. Our joint method outperforms the baselines in three of the tasks (greatly outperforming Style for t2g and g2t) and occupies the second position for the remaining three. In Tab. 5, we show a summary result using all three performance metrics. We see that our combined method is the strongest overall.

In Tab. 7, we compare modeling all modalities jointly compared to per-task, for the style constraints only. The Joint method is trained with all three modalities at training time, and the Per task one is trained just the corresponding two modalities. Both methods receive the same inputs at test time. We see that the largest improvement (10%) between Joint and Per-task is for the personality-to-gaze task, which is the most challenging task. We also see a large gain (4-8%) between joint and per-task when the input/output pair is text-to-personality and vice versa, which we saw above is the second most challenging set of tasks.

This makes sense because joint modeling is a double-edged sword. On one hand, leaning the structure of the space from multiple modalities helps. For example, knowing about the captions a user provides helps us learn what types of users there are at training time. Thus, even if at test time we do not have their captions, we can better predict gaze or personality. On the other hand, task-specific networks are more focused, thus easier to learn the task. Thus, we expect that using a third modality at training time will only help when that third modality provides a required latent link between the input and output modalities that is otherwise missing, as in the modality pairs that are less related. The weakest performance of joint modeling is on the text-to-gaze task, since gaze and text are already tightly coupled.

4.5. In-depth look

In this section, we provide in-depth intuitions to understand the task and the performance of the methods. We first quantitatively show how distinct the samples provided by different users are; see Fig. 2 for a qualitative version. We next show the selected combination weights for our tasks. These experiments are conducted for the *Ads* dataset.

Identity classifier. If the samples from different users are very unique, it will be easy to distinguish between users. To examine how unique samples are, we train an identity classifier where the features are the samples, and the labels are the IDs of the users who provided the samples. We follow a five-fold stratified cross-validation procedure with a linear and RBF support vector machine. We select parameters for nine configurations of gamma and cost for RBF SVM and three configurations of cost for linear SVM.

In the text domain, we employ averaged 200-dimensional Glove embeddings of words in the caption. In the gaze domain, we calculate the percentage of image explored (viewed) and the max/min distance among all revealed "bubbles." These features produced the best performance for the identity classifier, but in preliminary experiments, the features used for the retrieval tasks produced similar performance. In the text space, we achieve 7% accuracy (while chance is about 1%). In the gaze space, we achieve a lower performance of 4%; and combining these two spaces, we achieve 9%. Thus, users provide reasonably different samples in all modalities, but there is more overlap in the space of gaze samples. We opt not to use percentage of exploration and bubbles distances in our retrieval task for gaze, because they do not capture any image content, hence it would be harder to find relations with text.

Content/style/base weights. Our combined approach works by combining the base, content, and style embeddings, with appropriate weights. These weights are chosen on the validation set and applied on the test set. We perform five different shuffle splits, so we obtain five sets of weights for each task. In Tab. 8, we show the average weight assigned to style, base and content. For the most content-dependent task, gaze to text and vice versa, CONTENT is most important. Then, for text to personality and vice versa, STYLE is the most important. Ads have subjectivity, thus it is required to capture well the style of the different annotators. Finally, for gaze to personality and vice versa, which is the hardest task, the weights give the same importance to the STYLE and BASE networks.

Tasks	Style	Base	Content
g2t/t2g	0.2	0.25	**0.7**
t2p/p2t	**0.7**	0.55	N/A
g2p/p2g	0.55	0.55	N/A

Table 8. Averaged weights selected for each network, on *Ads*.

5. Conclusion

We described an approach for retrieving samples capturing different perceptions of the same image input, across modalities. To understand how different viewers perceive and describe images, we use two types of constraints. One bridges samples across modalities, using images as anchors in the learned space. The other set of constraints employs viewers as anchors, i.e. samples that came from the same user should be similar, regardless of the viewed image. We combine both sets of constraints and show that the combination usually outperforms the individual sets of constraints. Further, it usually outperforms two baseline approaches. Importantly, learning about gaze, captions, and personality in the same framework improves performance over learning networks for each separate input-output pair of modalities. We validate our method on two datasets, one more subjective (ads) and another more general (COCO). We make our personality-aware captioning data publicly available.

In the future, we will investigate ways to learn more fine-grained selections over individual constraints. We will also investigate approaches for more efficient learning, by discovering groupings of viewers according to their personality and perception, and encouraging representations of gaze/text/personality for similar users to be similar (rather than just representations for the same user being similar). We will also extend our experiments to other domains beyond advertisements and image captioning.

Acknowledgements. This material is based upon work supported by the National Science Foundation under Grant No. 1566270. This research was also supported by a Google Faculty Research Award, a University of Pittsburgh Central Research Development Fund (CRDF) grant, and an NVIDIA hardware grant. We thank the reviewers and AC for their valuable suggestions. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[3] Anna Antinori, Olivia L Carter, and Luke D Smillie. Seeing it both ways: Openness to experience and binocular rivalry suppression. *Journal of Research in Personality (JRP)*, 2017.

[4] Guido Borghi, Stefano Pini, Filippo Grazioli, Roberto Vezzani, and Rita Cucchiara. Face verification from depth using privileged information. In *British Machine Vision Conference (BMVC)*. Springer, 2018.

[5] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[6] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[7] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *International Conference on Computer Vision (ICCV)*. IEEE, 2017.

[8] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[9] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[10] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: improved visual-semantic embeddings. In *British Machine Vision Conference (BMVC)*. Springer, 2018.

[11] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*. IMLS, 2015.

[13] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[14] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012.

[15] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[16] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[17] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *Transactions on pattern analysis and machine intelligence (TPAMI)*, 1998.

[18] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[19] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *International Conference on Computer Vision (ICCV)*. IEEE, 2009.

[20] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[21] Nam Wook Kim, Zoya Bylinskii, Michelle A Borkin, Krzysztof Z Gajos, Aude Oliva, Frédo Durand, and Hanspeter Pfister. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *Transactions on Computer-Human Interaction (TOCHI)*, 2017.

[22] Adriana Kovashka and Kristen Grauman. Attribute adaptation for personalized image search. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *European Conference on Computer Vision (ECCV)*. Springer, 2014.

[24] Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen Ebrahimi Moghaddam, and Lyle H Ungar. Analyzing personality through social media profile picture choice. In *International AAAI Conference on Web and Social Media (ICWSM)*. AAAI, 2016.

[25] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Neural Information Processing Systems (NIPS)*. NIPS, 2016.

[26] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[27] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[28] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Information bottleneck learning using privileged information for visual recognition. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[29] Nils Murrugarra-Llerena and Adriana Kovashka. Learning attributes from human gaze. In *Winter Conference of Computer Vision (WACV)*. IEEE, 2017.

[30] Timea R Partos, Simon J Cropper, and David Rawlings. You don't see what i see: Individual differences in the perception of meaning from visual stimuli. *PloS one*, 2016.

[31] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2014.

[32] Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality (JRP)*, 2007.

[33] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[35] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. Learning to rank using privileged information. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013.

[36] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, 2017.

[37] Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. Paying attention to descriptions generated by image captioning models. In *International Conference on Computer Vision (ICCV)*. IEEE, 2017.

[38] Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation (NC)*, 2000.

[39] Christopher Thomas and Adriana Kovashka. Artistic object recognition by unsupervised style adaptation. In *Asian Conference on Machine Learning (ACCV)*. Springer, 2018.

[40] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *International Conference on Computer Vision (ICCV)*. IEEE, 2015.

[41] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *Journal of Machine Learning Research (JMLR)*, 2015.

[42] Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. Separating self-expression and visual content in hashtag supervision. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[43] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond J Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[44] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[45] Russell L Woods, C Randall Colvin, Fuensanta A Vera-Diaz, and Eli Peli. A relationship between tolerance of blur and personality. *Investigative ophthalmology & visual science (IOVS)*, 2010.

[46] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[47] Yanyu Xu, Nianyi Li, Junru Wu, Jingyi Yu, and Shenghua Gao. Beyond universal saliency: personalized saliency prediction with multi-task cnn. In *Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI, 2017.

[48] Keren Ye and Adriana Kovashka. Advise: Symbolism and external knowledge for decoding advertisements. In *European Conference on Computer Vision (ECCV)*. Springer, 2018.

[49] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[50] Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. Supervising neural attention models for video captioning by human gaze data. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[51] Jonathan R Zadra and Gerald L Clore. Emotion and perception: The role of affective information. *Wiley interdisciplinary reviews: cognitive science (WIRCS)*, 2011.

[52] Yexun Zhang, Ya Zhang, and Wenbin Cai. Separating style and content for generalized style transfer. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.