Persuasive Faces: Generating Faces in Advertisements (Supplementary Material)

Christopher Thomas
chris@cs.pitt.edu
Adriana Kovashka
kovashka@cs.pitt.edu

Department of Computer Science
University of Pittsburgh
Pittsburgh, PA USA

We strongly encourage readers to view the supplementary qualitative results we include in the `17_way` and `5_way` folders.

In this document, we present several supplementary results to our main text. We first present an analysis of what objects appear in different types of advertisements. We find, for example, that **bottles** occur in beverage ads. We then study the visual distinctiveness of objects across ad topics. We seek to identify objects which have a unique visual presentation in a certain type of ad, which is not present in ads on the whole. We observe, for example, that **cars** in car ads appear different from cars in other ad types. Based on our object analysis, we narrowed our focus to faces because they were the most common object detected in the dataset. We performed face detection on ads and then predicted the facial attributes and expressions of those faces. We found, for example, that faces from beauty ads tended to score highly for **attractive**, **heavy makeup**, and **wearing lipstick**. We found facial expressions mattered as well, with domestic violence faces exhibiting the most **fearful** and **sad** faces. We also include three additional quantitative results from our human study. These results verify that humans did well at discriminating ad faces, show which types of ads our methods work best on, and finally rank methods by visual quality. We then discuss the qualitative results which we include with this supplementary. Finally, we include examples of real bottles we detected in ads to complement Fig. 5 in the main text.

1 Object Distributions in Ads

We wanted to understand how frequently different objects appeared in different types of ads. Overall, we find that object distributions differ across types of ads in meaningful ways. We use the Ads Dataset of [2] and select the top 10 product ad categories (alcohol, beauty, cars, etc.) and all of the seven public service announcement categories (animal rights, domestic violence, environmental conservation, etc.). We then ran a 50-layer residual RetinaNet [4] trained on the 80 COCO [3] categories on all ads in these categories. We present statistics about how frequently each object is detected in each category of ads in `object_distributions.xlsx`. Each row shows a different object and each column is a different category of ad. Because the number of ads in each category are unequal, we normalize by column so that each cell shows the frequency that a particular object appears relative to all objects detected for that category of ads. To do this, we divide the number of object detections in each cell by the total number of object detections for each topic of ads.

We also show the mean and standard deviation for each row (object) and column (ad topic). We highlight each cell by its intensity relative to other values in that row. Thus, greener cells indicate topics in which an object appears more frequently than other topics, while whiter cells indicate infrequent object appearances relative to other topics in which that object appears. On the right and bottom, we show the mean and standard deviation. We highlight these values by column (for the right side) and by row for those underneath the object table. Here, redder values indicate lower values, while greener cells indicate higher values, relative to other cells in that column or row.

We observe, for example, that **person** is the most common object by far. Many results are unsurprising. For example, **cars** are very common in car ads, **bottles** are common in alcohol and soda ads, and **cell phones** are common in electronics ads. However, there are a few noteworthy results. For example, we see that **remote** appears more often in beauty ads than in electronics ads, which is counter-intuitive. However, we we examined the detections, we found that the model frequently confused makeup palettes as remote controls, because of the "button-like" appearance of the makeup. We also observed similar confusions of the **cell phone** category with makeup powder compacts. We observe that **person** appears most frequently in domestic violence ads relative to all other topics. We believe that this is because domestic violence ads are showcasing human misery and the effects of domestic violence on individuals, while product ads, such as beauty ads, may occasionally just show the product they are selling.

2 Object Visual Distinctiveness

It is not just the number of times an object appears in a category of ad that is important; rather *how* objects are portrayed differently across types of ads provides important information about whether the object is being employed as part of a persuasive strategy, or rather, just frequently appears as a background object in that class of ads. We thus wanted to discover which objects have a distinct visual appearance in certain categories of ads when compared to their appearance across all categories of ads. To do this, we extracted SIFT features [6] from all objects we detected in ads. For each object, we then created a 100-D BoW histogram representation of the object's appearance from the extracted SIFT descriptors. In order to avoid having our metric skewed by topics with very few object instances whose appearance differs from the average simply due to having so few object detections, we only considered objects appearing in at least two topics of ads with at least 20 detected instances of that object. We computed the average intra-topic object appearance distance (the average distance between all 100-D vectors for instances of the same object class in the same topic of ads) and the average inter-topic object appearance distance (the distance between all vectors for that object excluding those that were used for the intra-topic distance calculation). We also computed the intra-inter topic distance, where we compute the average distance between all of the instances of a given object in the intra-topic ad category with all the instances of that object in other categories. We also show the difference between the intra-topic distance and the inter-intra topic distance, as well as the difference between the intra-topic distance and the inter-topic distance. Intuitively, object-topic pairs showing large positive differences correspond to that object being distinct for that topic, which might correlate with persuasiveness. Our results can be found in `visual_distinctiveness.xlsx`. We highlight each column by the magnitude of values in that column (i.e. redder values are larger). We sort the rows by the difference between the mean inter-topic object distance and the mean intra-topic object distance, i.e. objects with large inter-class difference, but low intra-class distance appear first. This should identify objects whose appearance within a topic is consistent, yet unlike the appearance of that object in other topics.

Collectively, these results allow us to identify objects whose portrayals within a given topic are unlike the portrayals of that object in other topics. This could indicate that the object is being employed in a persuasive way for that topic of ads. For example, we see that **vase** for the alcohol topic has the highest difference between the inter-intra and intra-inter-intra metrics (the two rightmost columns). This is most likely because many alcohol bottles are being incorrectly detected as vases, and thus, the "vases" detected in alcohol ads look different from vases detected in other categories. **Cars** appear distinct in car ads. This makes sense because in car ads, the purpose of the ad is to sell the car. Thus the cars appear shinier, newer, and more appealing. Other ads often feature cars in the background or for different purposes. We find a number of other interesting distinctions. For example, **bowls** and **cake** in chocolate ads, **dogs** in animal rights ads, **person** in alcohol ads, and **bottle** in alcohol and soda ads, all have significantly distinct visual appearances in their respective topics. Each of these makes sense in fairly straightforward ways; for example dogs could be shown to be abused in animal rights ads and thus unhappy or injured, therefore their appearance differs compared to other ad topics.

3 Per-Topic Face Detections

Alcohol	1167	**Animal Right**	516	**Beauty**	5299	**Cars**	875
Chocolate	1202	**Clothing**	6573	**Domestic Violence**	212	**Electronics**	1093
Environment	50	**Human Right**	151	**Restaurant**	923	**Safety**	221
Self Esteem	146	**Smoking / Alcohol Abuse**	437	**Soda**	2260	**Sports**	1137
Travel	568	**TOTAL**	22813				

Table 1: Number of faces detected in each topic of ads after removing low confidence and small detections.

In order to train a reliable generative model, sufficient data must be available. Many of the object categories identified above contain several hundred detections in total, which is insufficient to train a network on. We thus focused on the most common object detected in our dataset: people. However, because many of the person detections contained background, body occlusion, text over the body, or were simply just zoomed in faces, we narrowed our generation task to generating distinct faces. In order to further clean up the data, we trained a Faster-RCNN [9] on the Wider Face dataset [11]. We then performed detection on the same ads as above. We found that many detections were too low-resolution to be usable, so we eliminated detections smaller than 60x60 and whose detection confidence was less than 0.85. The remainder of our supplementary focuses on this dataset of per-topic ad faces. We show the per-topic count of detections above in Table 1.

4 Per-Topic Facial Attribute Analysis

We wanted to understand how facial appearance differed across topics of ads. Our results show that facial attributes significantly differ across ad topics in important ways. For this analysis, we train an Inception-v3 [10] network on the CelebA [5] dataset then apply the trained model to the ad faces we detected above to detect attributes. We present our results in `facial_attributes.xlsx`. Rows represent attributes and columns represent topics of ads. As we did with object detections, because there are an unequal number of ads in each topic, we normalize the columns, so that each cell shows the frequency that a particular attribute appears in a given ad topic relative to all other attributes for that ad topic. We also show the mean and standard deviation for each row and column.

We observe that **young** is the most common attribute detected in our dataset. This suggests that the ad categories we consider commonly target and feature younger people, which makes sense given what those topics are. We also see that beauty ads feature the most faces **wearing lipstick**, wearing **heavy makeup**, and that are **attractive**. We observe that men are commonly present in sports ads and slightly more often in environment ads, but this is probably a result of the fact that there are very few environmental conservation ads overall, so their results are probably noisy. We also notice that domestic violence ads feature the most faces with **black hair**. This makes sense because these ads are attempting to portray darker themes and colors. Overall, we observe many quantifiable semantic differences between ad topics in terms of facial attributes.

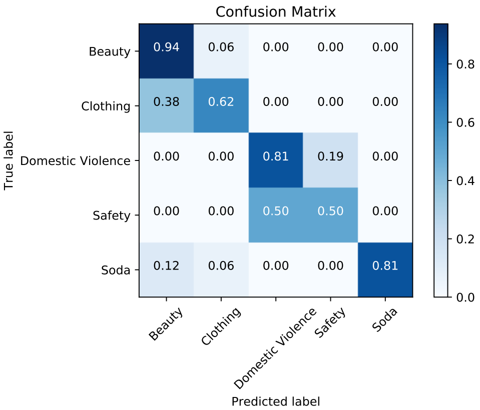
5 Per-Topic Facial Expression Analysis

Facial attributes capture a wide amount of variation in facial appearance, but do not capture facial expressions (besides smiling). We also wanted to understand how facial expressions differed in different types of ads. This makes sense because for some types of ads, like domestic violence, one might expect frightened or sad faces, while in beauty ads, one would expect happier faces. We again trained an Inception-v3 [10], this time on the AffectNet [8] dataset. The dataset contains 8 facial expressions and valence and arousal scores. Valence measures how "good" or "bad" the emotion represented by the face is, while arousal indicates how strong the expression shown on the face is [7]. We train our network to predict both valence and arousal for faces and then apply it on ad faces. We present our results in `facial_expressions.xlsx`. We normalize our results per topic as we did for facial attribute detections. For the valence and arousal rows, we show the mean valence and arousal for each ad topic.

We observe that the vast majority of ad faces are either happy or neutral. We see that travel ads have the most happy faces, followed by soda ads and alcohol ads. Collectively, we observe that beverage ad faces (alcohol and soda) tend to show happy faces drinking the beverage. Domestic violence ads show the saddest faces as well as the most fearful. Animal rights ads also feature fearful faces. For valence, we see that travel and soda ads have the highest valence faces, while domestic violence have the lowest. We observe the highest arousal in chocolate ads, followed by sports ads. This makes sense because chocolate ads show people experiencing chocolate with exaggerated expressions of happiness, while sports ads show physical activity. We conclude that facial expressions are an important signal to consider when performing facial transfer between ad topics.

6 Experimental Results

We present three additional quantitative results from our human study. The first task in our human study was designed to ensure that participants actually paid attention to the real ad



	How Well Transformed Average Rank (1-5)
Beauty	3.250
Clothing	2.375
Domestic Violence	1.875
Safety	1.875
Soda	**3.750**

Figure 1: We show two results from our human study. The first (left) shows the confusion matrix of humans on our classification task for each type of ad. The diagonal shows the accuracies for each topic. On the right, the table shows how well humans believed the best method did, on average, at transforming faces into each ad category. Higher is better.

face examples we showed them so that they would understand what faces looked like in different topics of ads. We showed participants examples of real faces from ads and then had them perform a test in which they classified rows of faces into one of five ad categories. Each row contained five faces from the same ad category and had participants classify ten rows. We show the accuracy of humans as well as which categories humans tend to confuse in Table 1 (left). We see that humans are almost always able to correctly classify beauty ads, but tend to confuse clothing and some soda ads as beauty. This makes sense because these categories show smiling women, often wearing makeup. We also notice that domestic violence and safety ads tend to be confused, most likely because both show faces with lacerations or bruising, making them hard to tell apart. We emphasize that this test was performed on rows of real, not generated, faces. Despite having five examples of real faces per row, humans still made mistakes at classifying the rows, underscoring how challenging it is to distinguish, much less generate, faces which are distinct in each topic.

At the end of our human study, we asked participants to rank the five ad topics by how well they felt the best method did on average at transforming faces into each ad topic. Participants could rate ad topics from 1-5, where 1 was the worst and 5 was the best. We also show this result in Table 1 (right). We find that humans feel that soda ads are best portrayed. This is most likely because the model is capturing the large smiles frequently present in beauty ads, as well as the vintage appearance. Beauty ads are a close second, with humans feeling that they are fairly accurately portrayed. One possible reason for this is because the models are able to capture the bright skin, makeup, lipstick, and female gender of beauty faces. Humans felt that domestic violence and safety ads were the worst portrayed. We believe this is because domestic violence and safety ads frequently tend to show faces which are damaged, bruised, or cut in some way. Our model correctly captures the dark appearance and sad expression from domestic violence ads, but is unable to recreate the facial damage seen in these ads. This is likely because domestic violence and safety faces make up a relatively small portion of our faces dataset. Thus, the model has not devoted modeling power in the latent space to representing these facial deformities, choosing instead to capture other details

	Best Visual Quality (Percent)
StarGAN (Conditional)	0.050
StarGAN (Topics)	**0.794**
Latent	0.075
Conditional	0.019
Conditional + Latent (Ours)	0.063

Table 2: We asked participants which method produced images with the best visual quality. Most participants chose **StarGAN (Topics)**. We observe that **StarGAN (Topics)** performs very little transformation to the input image, thus produces the sharpest output.

in the latent space which apply to all ad categories.

We wanted to get a sense of the visual quality of the generations produced by each method. To do this, we transformed the same face using our method and the baselines into the same five categories that humans studied the ground truth faces for. We then showed humans these generated faces, along with the original and unmodified reconstruction, for each method and asked humans which method best transferred the faces in a way that captured the unique visual appearance of each topic. We presented this result in our main text, and found that **Conditional+Latent (Ours)** performed best on this task by a large margin. However, to assess visual quality, we also asked participants which method had the best perceptual quality in each group, as our final task. Note that this metric does not measure what we are ultimately interested in in this paper, but it still useful for assessing the visual quality of the results. We show this result in Table 2.

We see that **StarGAN (Topics)** produces the images with the best visual quality by a very large margin. We believe that this is because **StarGAN (Topics)** performs very little modification of the input image. The model frequently changes skin color and adds noise to the image, but the original image (and its semantics) is unchanged. Thus, this model achieves the best visual quality because its generations are the sharpest because it simply outputs the input image with slight perturbations. **StarGAN (Conditional)** also tends to preserve the original image appearance, but does perform slightly more dramatic transformations. These transformations, however, tend to distort the face in unnatural ways. We see that of our method variations, humans felt that **Latent** produced the best perceptual quality. This makes sense because **Latent** has no conditional information to consider and only considers its learned latent representation used for constructing facial appearance. Our results show however, that this latent representation performed poorly at changing facial semantics. Thus, we conclude that conditional information is critical for the task we are interested in, but its inclusion does result in a loss of visual quality when the model is forced to change semantics.

7 Supplementary Qualitative Results

We include two types of supplementary qualitative results to our main text. In the `5_way` folder, we provide examples of faces translated into five ad topics: alcohol, beauty, clothing, domestic violence, safety, and soda. In the `17_way` folder, we show faces transformed into all seventeen topics that we study. We also show reconstructions of each face, without performing any ad topic transformation. From these results, we can see that **StarGAN (Conditional)** attempts to perform some transformation on the face, but the face's appearance becomes deformed and distorted. We also observe that this method does not tend to change facial expressions. **StarGAN (Topics)** performs only low-level transformation, for example, changing the color of the lipstick or of the skin, but does not change expressions or gender at all. **StarGAN (Topics)** also tends to add high-frequency noise into the images, most likely in order to enable the topic classifier to be able to classify them. The **Latent** method changes facial appearance, but does not tend to change facial attributes or expressions. This indicates that higher-level supervision is necessary in order to learn such semantic transformations. The **Conditional** method performs transformations and works competetitively, but does not change the overall facial appearance or color.

Our method **Conditional + Latent** changes both semantics and low-level details. We observe that alcohol, car, human right, environment, and sport ads tend to become more male, beauty ads have bright skin and bright lipstick, and clothing ads look similar to domestic violence ads, but domestic violence ads are darker and with unhappier facial expressions. We also note that safety and smoking / alcohol abuse ads tend to target men. Soda ads tend to have large smiles with a vintage skin tone. Overall, we experimentally show that our method performs best at transforming faces into different ad categories by capturing the distinctive semantics and appearances associated with each topic of ad.

8 Bottle Detections in Ads

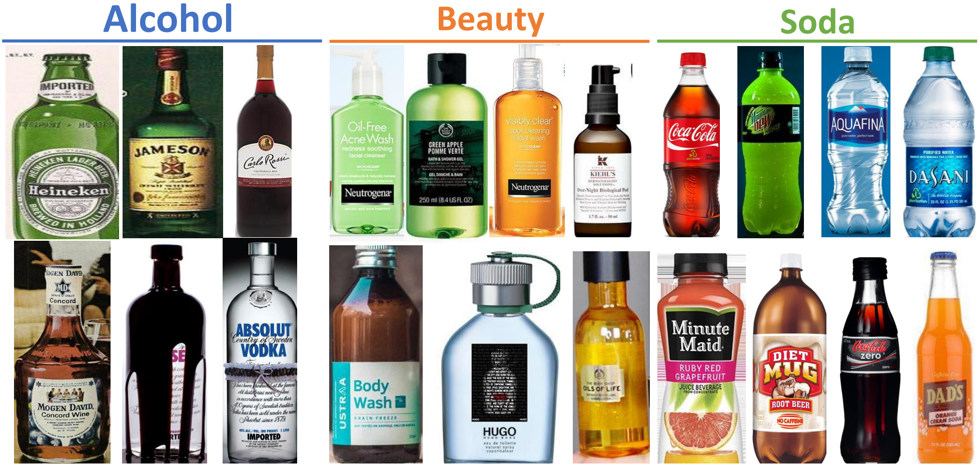


Figure 2: Bottles detected in alcohol, beauty, and soda ads. Compare to generated bottles in Fig. 5 of main text.

In our main text, we included examples of bottles generated by a conditional BEGAN [▯] model. We modified [▯]'s code to include conditional information and conditioned the model on ad topics. Based on our object distinctiveness results discussed above, we trained this model on bottles from three ad categories in which bottles appear distinctively: alcohol, beauty, and soda ads. We show examples of real bottles detected in these categories in Fig. 2. We observe that many alcohol bottles look similar, having an elongated neck, with a cap on top. Our generated alcohol bottles also seem to follow suit, having the same general shape as these ground truth bottles. Our generated beauty bottles are somewhat noisier than our alcohol generations. Compared to alcohol and soda bottles, beauty bottles tended to show the most variation in appearance. This is most likely because beauty bottles represent many different types of products (creams, lotions, makeup, etc.), while alcohol and soda bottles are almost exclusively beverage bottles, thus having less visual variation. Soda bottles tend to all have a similar shape so they are easy to hold. We observe that our generated soda bottles appear very similar to real bottles, having the same overall shape. We also note that many water bottles appear in the soda bottle category. From the generated images' similarity to these results, we conclude that our conditional model has learned meaningful ad category-specific appearance information. However, due to the small number of objects per category as well as the large variation within each category, the model's predictions are fairly noisy. Further research is needed on how conditional generative models can be trained to produce sharp images when trained on limited, diverse datasets.

References

[1] David Berthelot, Tom Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[2] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1100–1110. IEEE, 2017.

[3] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[6] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[7] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[8] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 2017.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[10] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[11] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016.