Persuasive Faces: Generating Faces in Advertisements

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

Department of Computer Science
University of Pittsburgh
Pittsburgh, PA USA

Abstract

In this paper, we examine the visual variability of objects across different ad categories, i.e. what causes an advertisement to be visually persuasive. We focus on modeling and generating *faces* which appear to come from different types of ads. For example, if faces in beauty ads tend to be women wearing lipstick, a generative model should portray this distinct visual appearance. Training generative models which capture such category-specific differences is challenging because of the highly diverse appearance of faces in ads and the relatively limited amount of available training data. To address these problems, we propose a conditional variational autoencoder which makes use of predicted semantic attributes and facial expressions as a supervisory signal when training. We show how our model can be used to produce visually distinct faces which appear to be from a fixed ad topic category. Our human studies and quantitative and qualitative experiments confirm that our method greatly outperforms a variety of baselines, including two variations of a state-of-the-art generative adversarial network, for transforming faces to be more ad-category appropriate. Finally, we show preliminary generation results for other types of objects, conditioned on an ad topic.

1 Introduction

Advertisements are persuasive tools that affect people's habits and decisions. They often advertise products and establishments, such as cosmetics and beauty, clothing, alcohol, automobiles, or restaurants. However, they can also be public service announcements that aim to educate the public about important social issues, such as domestic violence or environmental protection. Many topics advertised by ads contain distinctive objects, e.g. the most common object in car ads might be cars, bottles for alcohol ads, and faces for cosmetic ads. There is more to ads than what objects they contain, however. It is *how* objects are portrayed that makes an ad persuasive. For example, faces frequently appear in both beauty and domestic violence ads but their portrayal is vastly different.

What is it that makes a face become a beauty ad or a domestic violence prevention ad? This is what we set out to discover in this study. We first analyze the distribution of objects in common ad topics (beauty, soda, domestic violence, safety, etc.) Based on the object distributions, we select to model the appearance of faces, since faces are the most frequent object across all ad categories and have the most distinctive appearance per category. We

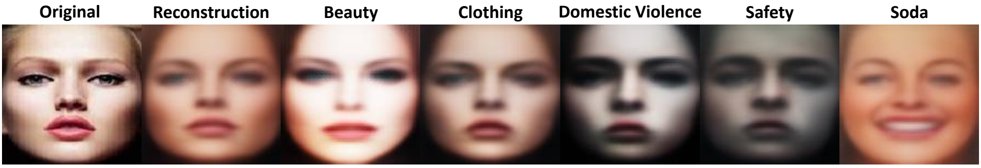

Figure 1: We transform faces so they appear more persuasive and appropriate for particular ad categories. We show an original face on the left, followed by our method's reconstruction without any transformation. We then show the face transformed according to five types of ads. Notice how the beauty face contains heavy make-up, the domestic violence face is sad and possibly bruised, the safety face is somewhat masculine, and the soda face is happy.

then learn a generative model capable of transforming faces into each ad topic. Because ads are rarer than general images, we must work with a sparser dataset than modern generative approaches usually assume. Thus, we propose a method for *transferring knowledge* from faces in other datasets, in order to mimic the variability of faces in the ads domain. We validate our approach qualitatively, by morphing the same face according to different ad categories, and quantitatively, using human judgments and classifier accuracy.

Our method works as follows. We first train facial expression and facial attribute classifiers using existing datasets. We detect faces in ads and predict their attributes and expressions. Next, we train a conditional variational autoencoder (CVAE) on our dataset of ad faces. The model learns to reconstruct an ad face from a vector comprised of a learned latent representation, facial attributes, and facial expressions. At test time, we embed all ad faces into vector space using our encoder and then compute how faces differ in that space across ad topics. Using these per-topic learned differences, we transform embeddings of other ad faces into each ad topic. Finally, we use our decoder on the transformed embeddings to generate distinct faces across ad topics. We show examples of our transformations in Fig. 1.

Note that prior work has modeled the conceptual rhetoric that ads use to convey a message [11, 39], but no work models the visual variance in the portrayal of the same object across different ad categories, nor attempts to generate such objects.

The contributions of our work are three-fold:

- We propose the problem of studying what makes an object visually persuasive and generating objects which convey appropriate visual rhetoric for a given ad topic.
- We analyze object frequency and appearance in ads, and discover objects with class-dependent appearances, which we then generate with promising quality.
- We develop a novel generative approach for modifying the appearance of faces into different ad categories, by elevating visual variance to a semantic level without the need for new semantic labels. The benefit of doing so is we can leverage semantics learned on larger datasets. Rather than directly modeling how faces in different ad categories differ on the pixel level, we model how they differ in terms of predicted attributes and facial expressions, then use these distinctions to create faces appropriate for a given ad category. Our method outperforms relevant baselines at this task.

2 Related Work

Generative models. Recently, generative adversarial networks (GANs) have produced impressive results on a number of image generation tasks [1, 2, 3, 7, 41]. Conditional GAN variants have also been proposed which generate images subject to conditional constraints. For example, recent work by [2, 20] allows conditional transformations on specific images.

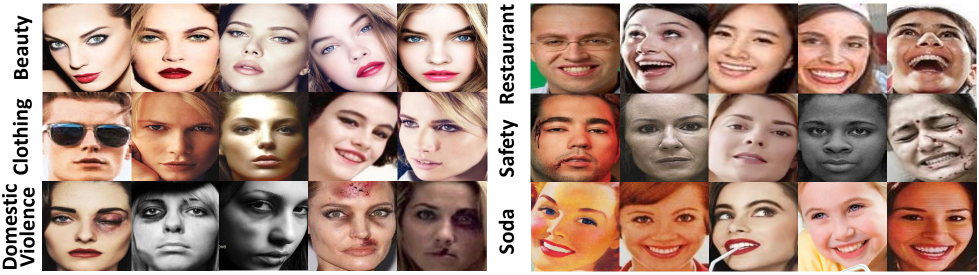

Figure 2: We show examples of real faces from different categories of ads. We notice significant differences, many of which can be captured through facial attributes and expressions.

Unfortunately, GANs are notoriously challenging to train, particularly on small and diverse datasets [7]. Our autoencoder-based method is able to contend with such a dataset, while modeling meaningful differences across ad topics. Autoencoders [9, 20, 26, 30, 36, 37] are an older type of generative model, but perform competitively when trained with recent perceptual loss functions [10]. Autoencoders learn to project an image into a learned embedding space and then to reconstruct the original image from the embedding. Our method is a conditional variational autoencoder (CVAE) [34], which adds supervised information about ad faces into the model, along with custom loss functions we found to improve result quality.

Facial expressions and attributes. We condition our model on facial expressions recognized in faces for different ad categories. Facial expression and emotion recognition is an established and popular topic [4, 15, 23, 28, 32]. Usually seven canonical expressions are recognized: happiness, sadness, surprise, fear, disgust, anger, and contempt. We also condition our generative model on facial attributes we predict on faces from ads. Attributes are semantic visual properties like "bald," "rosy cheeks," "smiling" or "attractive" [5, 19, 24].

Visual persuasion. The primary novelty of our work is to discover what makes objects in ads persuasive and then to generate such objects. While no work has been performed in this space before, researchers have studied related problems. [13] learned to predict whether a photograph portrays a politician in a positive or negative light, and [14] trained classifiers to predict the outcomes of elections based on the candidates' faces, but neither of these works creates a generative model. [11] propose a dataset of advertisements, and predict what message the ad conveys (e.g. "buy this car because it is spacious") but they do not model or generate the visual appearance of the same object across ad topics.

3 Approach

We begin by describing how we extract faces from ads. We then describe how we predict attributes and facial expressions on the detected faces. Next, we present our autoencoder architecture and then describe how we use it to transform faces across ad categories.

3.1 Ads data

We focus on the Ads Dataset of [11]. It contains ads belonging to 38 topic categories: beauty, soda, restaurants, etc. (called product ads) and domestic violence, safety, etc. (called public service announcements, or PSAs). We chose to study the ten most frequent product topics in the dataset, as well as all PSA topics, resulting in a set of 17 ad topics.

3.2 Face detection on ads

Our first step is to extract faces from ads. The remaining steps of our model work on this dataset of ad faces, rather than operating on whole ads images. This allows our model to concentrate on modeling and modifying facial appearance, without having to reconstruct the entire ad. We train Faster-RCNN [31] on the Wider Face dataset [38]. We remove face detections whose confidence is less than 0.85 or whose width or height is less than 60 pixels. We show examples of detected faces in different ad categories in Fig. 2. In total, we detected 20,532 faces. We observe, for example, that beauty ads often have brighter skin tones and feature women wearing makeup. Domestic violence faces are often darker and not smiling. Many soda faces appear vintage and smiling. Clothing ads are similar to beauty, but don't feature as bright of skin or makeup. Finally, safety ads feature more men and are not as dark as domestic violence ads. Importantly, many of the differences we observe are captured by facial attributes and expressions datasets.

3.3 Predicting facial attributes and expressions

We want our method to model the most relevant characteristics of faces in each ad topic category. As we observed in Fig. 2, the differences between faces in different ad categories can naturally be described in terms of facial attributes and expressions. Because our dataset is small and diverse, our model may not have enough signal to reliably learn to model facial attributes and expressions without explicitly being directed to do so. In other words, it may devote its modeling power to matching the precise vintage or cartoon appearance of ad faces (i.e. low-level details) without learning a high-level model of recognizable semantic differences. Thus, rather than formulating our task as modeling the unconstrained distribution of pixels from the faces in each ad group, we manually inject high-level knowledge to facilitate manipulation of specific semantic attributes and expressions across ad topics.

We use the CelebA dataset [24] of 40 facial attributes and the AffectNet dataset [28] of eight facial expressions plus valence and arousal scores. We train Inception-v3 [35] on each dataset. We train each classifier using a cross-entropy loss for classification. For the network trained on expressions, we add an additional classifier for the regression task of predicting the valence and arousal of the facial expression and also use a mean-squared error loss.

Formally, let $\mathbf{I}_t$ represent the dataset of ad faces extracted from each ad topic t (e.g. beauty faces, domestic violence faces, etc.). We use our trained attributes and expressions classifiers to predict these properties on our entire ad faces dataset. This results in an automatically labeled ads face dataset $\mathbf{I}_t = \left\{\mathbf{x}_t^i, \mathbf{y}_t^i\right\}_{i=1}^{N_t}$, where $\mathbf{x}_t^i$ represents face i from ad topic t, $\mathbf{y}_t^i$ represents the image's associated 50-dimensional vector (composed of 40 facial attributes and eight facial expressions with their accompanying valence and arousal scores), and N_t represents the total number of faces per topic. We binarize our facial attribute predictions and represent our facial expressions in a one-hot fashion. The valence and arousal scores are real numbers from $[-1,1]$. See our supplemental file for these predictions for each ad topic.

3.4 Conditional variational autoencoder

Given an image $\mathbf{x}_t^i$ and conditional vector $\widehat{\mathbf{y}_t^i}$, which may differ from the image's ground truth signature, we seek a model θ parameterizing the following transformation function:

$$f_\theta\left(\mathbf{x}_t^i, \widehat{\mathbf{y}_t^i}\right) = \widehat{\mathbf{x}_t^i} \tag{1}$$

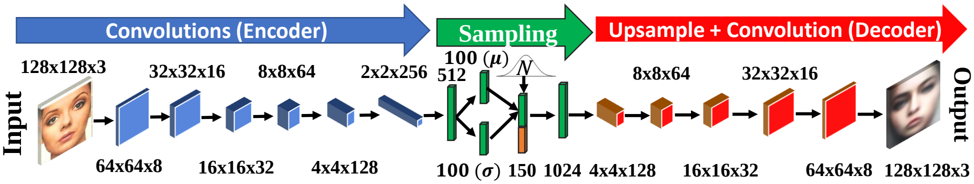

Figure 3: We show our model transforming a beauty ad face into a domestic violence face. The conditional vector (orange bar) is appended to the sampled latent vector (see Eqs. 3,4).

where $\widehat{\mathbf{x}_t^i}$ is a face retaining the overall appearance of $\mathbf{x}_t^i$, but now bearing the attributes and expressions encoded in $\widehat{\mathbf{y}_t^i}$. If $\mathbf{y}_t^i = \widehat{\mathbf{y}_t^i}$, we seek an unmodified reconstruction of $\mathbf{x}_t^i$. To modify the original appearance, we would like the reconstructed face to bear the provided set of attributes. If we denote our attribute and expression classifiers from Sec. 3.3 jointly as C, we wish to enforce the following constraint:

$$C\left(f_\theta\left(\mathbf{x}_t^i, \widehat{\mathbf{y}_t^i}\right)\right) = \widehat{\mathbf{y}_t^i} \tag{2}$$

Thus, any modifications done by our model should result in our classifiers producing the same conditional vector that was provided to the transformation model.

We also seek the capability of transforming ad topic-wise facial appearance *beyond* what is captured by our conditional vector. For example, if one topic features a predominant ethnicity, we would like our model to be capable of transforming a face into that ethnicity, even though it is not presented in our conditional vector. We thus seek a model capable of learning latent facial appearance information from our dataset. Autoencoders, which project an image into a low-dimensional space and then learn to reconstruct it from the sparse representation, are a natural choice. However, because we wish to interpolate faces across ad topics, enforcing that the learned space is smooth is important. We thus propose a custom conditional variational autoencoder, which enforces a Gaussian prior on the latent space [34].

We present our model's architecture in Fig. 3. It contains two distinct components, an encoder and decoder, which are trained end-to-end to reconstruct ad faces.

Encoder. Our encoder g_ϕ encodes any image $\mathbf{x}$ into the latent space $\mathbf{z}$ as follows:

$$\mathbf{z} = g_\phi(\mathbf{x}, \varepsilon), \varepsilon \sim \mathcal{N} \tag{3}$$

where ε represents a vector sampled at random from $\mathcal{N}$, a standard normal distribution. Specifically, g_ϕ encodes an image by predicting μ and σ for each dimension of the latent space. The latent embedding for an image is produced by combining ε with the predicted latent distribution parameters as follows: $\mathbf{z} = \mu + e^{\frac{\sigma}{2}}\varepsilon$. This mechanism of predicting the latent variable (coupled with the smoothness constraint discussed later) represents an image as a sample drawn from a Gaussian image space. Thus, the same image's latent embedding will differ each forward pass of the encoder due to random sampling of ε. This exposes our decoder network to a degree of local variation because the decoder learns that a larger space of embeddings map to the same face. This encourages smoothness in the latent space, which is important for the interpolation on latent vectors performed later.

Decoder. We concatenate each image's latent vector with its associated conditional vector (attributes and expressions) to produce the final representation given to our decoder p_ψ:

$$\mathbf{q}_t^i = \left[\widehat{\mathbf{y}_t^i}, \mathbf{z}_t^i\right] = \left[\mathbf{y}_t^i, g_\phi\left(\mathbf{x}_t^i, \varepsilon\right)\right] \tag{4}$$

During training, $\widehat{\mathbf{y}_t^i} = \mathbf{y}_t^i$. Our decoder network learns to reconstruct the original image from the embedding:

$$\widehat{\mathbf{x}_t^i} = p_\psi\left(\mathbf{q}_t^i\right) = p_\psi\left(\left[\mathbf{y}_t^i, g_\phi\left(\mathbf{x}_t^i, \varepsilon\right)\right]\right) \tag{5}$$

Learning. We train our model end-to-end to reconstruct the image provided to the encoder. However, because L2 reconstruction losses have been shown to produce blurry predictions [27], we instead use a perceptual loss similar to [10]. Rather than compute the distance between the reconstruction and original image in pixel space, we compute the distance in *feature space* of a pretrained VGG classification network following [40]. In our experiments, using a perceptual loss substantially improved the quality of reconstructions. Formally, let $\Phi\left(\mathbf{x}_t^i\right)$ and $\Phi\left(\widehat{\mathbf{x}_t^i}\right)$ represent the activations of layer `relu2_2` of a pretrained VGG-19 [33] network on the original and reconstructed images. The reconstruction loss $\mathcal{L}_r$ is given by:

$$\mathcal{L}_r = \left\|\Phi\left(\mathbf{x}_t^i\right) - \Phi\left(\widehat{\mathbf{x}_t^i}\right)\right\|_2^2 \tag{6}$$

We provide our decoder with the predicted facial attributes and expressions $\mathbf{y}_t^i$ so that we know these aspects of faces will be represented and thus modifiable across ad categories. However, the decoder might ignore less conspicuous attributes, so we force it to use the conditional information. The model should produce samples that cause our classification networks to output the same vectors provided to the decoder. If C_a and C_e represent attribute and facial expression classifiers, our conditional classification loss $\mathcal{L}_c$ is given by:

$$\mathcal{L}_c = l_{bce}\left(C_a\left(\mathbf{x}_t^i\right), C_a\left(\widehat{\mathbf{x}_t^i}\right)\right) + l_{nll}\left(C_{e_{exp}}\left(\mathbf{x}_t^i\right), C_{e_{exp}}\left(\widehat{\mathbf{x}_t^i}\right)\right) + l_2\left(C_{e_{va}}\left(\mathbf{x}_t^i\right), C_{e_{va}}\left(\widehat{\mathbf{x}_t^i}\right)\right) \tag{7}$$

where $C_{e_{exp}}$ and $C_{e_{va}}$ represent the facial expression and valence and arousal predictions from C_e respectively, l_{bce} represents the binary cross entropy loss, l_{nll} represents the negative log-likelihood loss after softmax is applied to the inputs (for multiclass classification), and l_2 represents the l_2 loss (for regression). In practice, we found our classification constraint improved reconstructions and made them more responsive to changes in the conditional vector.

To encourage smoothness in the latent space, we use a standard KL divergence term which measures the relative entropy between a spherical Gaussian distribution and the latent distribution [34]. The KL term $\mathcal{L}_{KL}$ can be analytically integrated [17] into a closed form equation as follows:

$$\mathcal{L}_{KL} = \frac{1}{2}\sum e^\sigma + \mu^2 - 1 - \sigma \tag{8}$$

We found the KL constraint critical to producing smooth faces. Our final loss is:

$$\mathcal{L} = \alpha\mathcal{L}_r + \beta\mathcal{L}_c + \gamma\mathcal{L}_{KL} \tag{9}$$

where α, β, and γ are hyperparameters weighting the contribution of each loss component.

3.5 Cross-category facial transformation

We described how to reconstruct a face, using an encoder, decoder, and fixed attributes and expressions. We now define what we input to our decoder, to translate a face to an ad class.

Notice that our model never accesses the ad topic category each face comes from. This is because the faces within topic categories are too varied for the model to make use of topic information. However, in order to transform faces so they appear to come from different topics, we first must learn how faces differ in each topic. We compute a vector for each ad topic, which, when added to an image's embedding, makes the reconstruction appear more appropriate for that topic. Specifically, we compute the *topic transformation vector* $\mathbf{v}_t$ for each topic t as follows, where the horizontal bar indicates computing the mean per dimension:

$$\mathbf{v}_t = \sum_{i}^{N_t} \overline{\mathbf{q}_t^i} - \sum_{t' \neq t} \sum_{i}^{N_{t'}} \overline{\mathbf{q}_{t'}^i} \tag{10}$$

In order to make the transformations more visible, we increase the magnitude of the vector by multiplying the conditional portion of $\mathbf{v}_t$ by 10 and the latent portion by 2.5. We found this visibly improved the distinctiveness across topic categories. To translate a face $\mathbf{x}$ into ad category t', we modify the embedding of $\mathbf{x}$ using $\mathbf{v}_{t'}$ and then reconstruct it as follows:

$$\widehat{\mathbf{x}_{t \to t'}^i} = p_\psi \left(\mathbf{q}_t^i + \mathbf{v}_{t'} \right) \tag{11}$$

3.6 Implementation details

We train our encoder and decoder end-to-end, but we do not train the VGG-19 network. We train the two classification Inception networks offline, before training our autoencoder. We train using the Adam optimizer [16] with learning rate 5.0e-4. We use minibatch size of 32 and train for 200 epochs. To ensure robustness to the highly varied ads faces dataset, we perform aggressive data augmentation. We randomly horizontally flip the training data and also randomly zoom into or out of the images. We then crop the zoomed images to 128x128. This allows our models to be less sensitive to facial alignment. We empirically found using 100 dimensions for $\mathbf{z}$ to work well. We set $\alpha = 1$ and β and γ to 0.0001; larger values caused poor reconstructions. We use Xavier initialization [6] and leaky ReLU activation [8] for inner layers with negative slope 0.01. We find using batch normalization [12] with eps 1e-4 helps stabilize training. We implement all components of our model in PyTorch [29].

4 Experimental Validation

We conduct our experiments on the image advertisement dataset of [11]. We initially sought to study general object appearance across ad topics, but our analysis below revealed faces were by far the most distinct object per topic. We thus focus primarily on modeling faces.

4.1 Objects in Ads

We ran a 50 layer residual RetinaNet [22] trained on the COCO dataset [21] on all ads in the 17 ad topics defined in Sec. 3.1. We first studied the *distributions* of objects across ad topics. We found many object-topic correlations, e.g. cars are most frequent in car ads, bottles occur frequently in alcohol and soda ads, animals are often found in animal rights and environment ads, etc. Overall, we found that people tended to occur 13 times more frequently than the second most common object (car). We next studied how objects' *appearance* differed across ad topic categories. We extracted SIFT [25] features for each object and computed BoW histograms with $k = 100$. We then analyzed the "visual distinctiveness" of objects, by measuring how each object's appearance changed within and across ad topics. We found that cars are highly visually distinct in car ads. This makes sense because cars in car ads are the *focus* of the ad, not just a background object. We also found dogs were distinct in animal rights ads, cell phones in electronics ads, cake and bowl in chocolate ads, and bottle

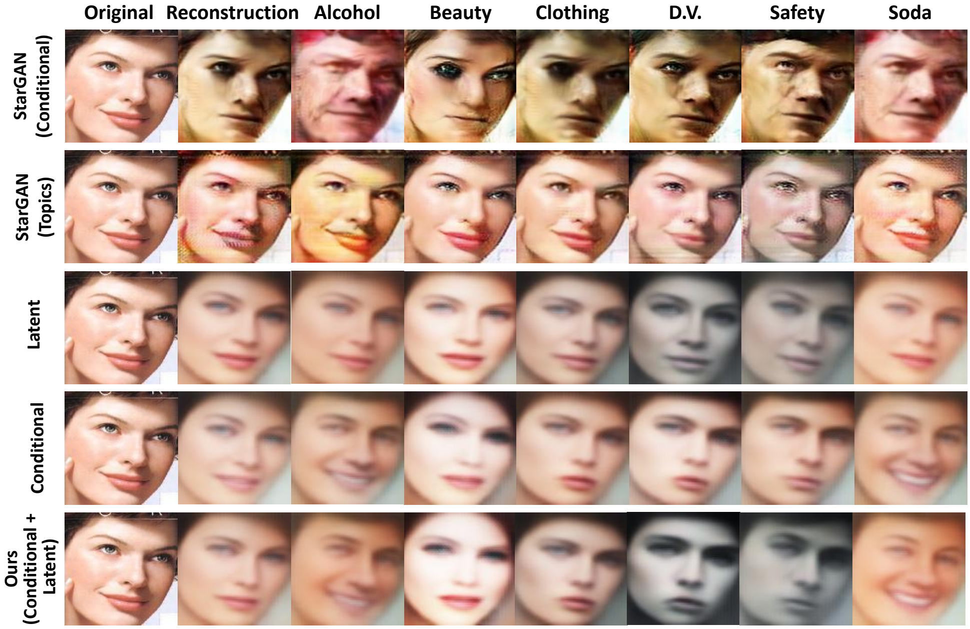

Figure 4: We show the result of transforming the same face using five different methods. Our method (bottom row) most faithfully transfers the topic-specific facial appearance as judged by our human study. We include additional qualitative results in our supplementary.

in soda ads. We include complete object distribution and visual distinctiveness tables in our supplementary file. Faces were the single object category which occurred frequently enough across topics to train a model on, thus we primarily focus on modeling faces in this paper.

4.2 Qualitative Results

We compare our method against two baselines inspired by attribute autoencoders [10, 20, 37], one of which has access to attributes and one which does not, as well as two variations of a state-of-the-art adversarial network for transforming images and attributes [2]:

- **Conditional+Latent (Ours)** - Our full model, described in Sec. 3.
- **Conditional** - Our model trained with latent and conditional information (attributes, expressions, and valence/arousal), however only the 50 conditional dimensions are changed when translating a face across topics, while the latent dimensions stay fixed.
- **Latent** - Our model *without* the conditioning on attributes and expressions.
- **StarGAN [2] (Conditional)** - We train StarGAN to modify faces to a given 50-dimensional conditional vector (facial attributes, expressions, valence/arousal).
- **StarGAN [2] (Topics)** - We train StarGAN to modify faces into a given topic. At training time, we train the model on the ground truth ad topic categories the faces are from. The model thus explicitly learns how facial appearance changes across topics.

In Fig. 4, we observe that our method **Conditional+Latent** produces the most noticeable and dramatic changes in visual appearance. We observe changes in gender, skin tone, facial expression, and facial shape. Alcohol ads feature smiling men, beauty ads tend to have light skin with lipstick, and clothing ads are similar, but with less skin brightness and less smiling. Faces in domestic violence ads are often frowning and darker, while those in safety ads tend to appear more masculine. Finally, soda ads have a vintage appearance with a large smile.

Method	**Human Judgments:** **Best At Topic Transformation**	**Classifier Topic** **Prediction Accuracy**
StarGAN (Conditional)	0.100	0.069
StarGAN (Topics)	0.113	**0.100**
Latent	0.038	0.086
Conditional	0.144	0.080
Conditional + Latent (Ours)	**0.606**	0.092

Table 1: We present two quantitative results. The first shows in what fraction of examples humans chose each method as the best, for generating visually distinct and appropriate faces in each topic. The rightmost column shows the accuracy of a classifier when trained on each row's synthetic training data and tested on real images from 17 categories.

For **Conditional**, we find that many aspects of the face change appropriately. However, the model is unable to transform other features not captured by the conditional vector: for example, making the face appear darker for domestic violence ads. For **Latent**, we find that while facial appearance overall changes, facial expressions and many facial attributes remain fixed, leaving a smiling face in inappropriate categories such as domestic violence.

We observe that both versions of StarGAN maintain the original image's appearance, but do not change the image much per topic. We notice that **StarGAN (Conditional)** tends to produce smoother skin and highlighted eyes for "beauty," but its other categories are harder to discern. **StarGAN (Topics)** adds low-level details into the generated images in order to achieve a lower topic prediction loss rather than changing the facial appearance of the image.

4.3 Quantitative Evaluation of Generated Faces

In addition to our qualitative results, we perform two quantitative experiments to assess how well our method transforms faces into each ad topic. For our first experiment, we perform a human study to assess how well humans perceive each method to do in terms of data transformation. Eight non-author participants participated in our study. We first show them examples of real faces from five ad categories: beauty, clothing, domestic violence, safety, and soda. To ensure our judges pay attention to the visual distinctions, we ask them to classify 10 rows of real faces into the correct ad topic. We then show participants the same image translated by five randomly sorted methods into the five topics, and ask them to select the method which best portrays the distinct visual appearance of faces across the five topics.

For our second experiment, we transform the same faces into all 17 ad categories. Next, we finetune a pretrained AlexNet [18] on the transformed faces to predict which topic the face is supposed to portray. Finally, we evaluate our model on *real* faces. Thus, methods which reliably transform faces in ways which capture the distinct traits of each topic will achieve higher classification accuracy. This metric assesses how well discriminative features are translated into generated ads but does not assess the visual quality of the results or the task we are ultimately interested in, namely producing visually distinct faces across topics.

We present quantitative results in Table 1. **Ours** performs best at the objective we set out to accomplish, and does competitively on the objective but less informative classifier accuracy task. In our human study, human judges found that our method best generates topic-specific faces nearly ***4 times as often*** as the next best method, **Conditional**. Interestingly, humans rarely prefer the **Latent** model, demonstrating the importance of including attributes and facial expressions. For the classification task, the classifier trained on **StarGAN (Topics)**'s data performs best, followed by our method. This makes sense because

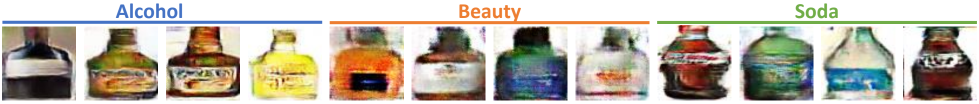

Figure 5: Alcohol, beauty, and soda bottles generated using our implementation of a conditional BEGAN [□] trained on bottles. We observe interesting differences across ad topics.

StarGAN (Topics) sees images labeled with topics at train time and learns what features are useful for topic classification. However, we see that the method only changes low-level details (e.g. color) without changing any semantics. Our method changes face semantics, never sees topic information at train time, yet performs on par with **StarGAN (Topics)** on this task, confirming our method does transfer topic-specific appearance. We observe that the accuracy for all models is similar and fairly low. This is most likely because many faces are impossible to classify since they are generic, non-persuasive background faces. Please see our supplementary file for additional results, including evaluation of visual quality and human classification performance.

4.4 Generating Other Objects

We wanted to see whether we could generate other objects besides faces as they appear in different ad categories. We conditioned BEGAN [□] on ad topics and trained on bottles from alcohol, beauty, and soda ads. We used an image size of 64x64 due to the limited amount of training data per class. We observe that the model does learn meaningful topic-wise differences in object appearance. For example, alcohol bottles look like liquor bottles with a long stem, beauty bottles are wider with a short stem (perfume), soda bottles have a soda bottle shape and label. These results show that intra-topic object appearance can be modeled, but future work is needed to address problems such as mode-collapse.

5 Conclusion

In this paper, we studied how objects appear in different categories of ads and how ads use these objects for persuasion. Based on our object analysis, we focused on faces and explored how faces could be generated across different types of ads. We proposed a conditional variational autoencoder for this task, which we augment by providing high-level facial attributes and expressions; experiments showed this auxiliary supervision was critical to achieving good results. Our experiments confirm that our method greatly outperforms a variety of baselines. We also show early results on how topic-specific objects beyond faces may be generated. In future work, we will investigate techniques for reliably generating other objects, and for creating ads complete with multiple objects and persuasive slogans.

Acknowledgements: This material is based upon work supported by the National Science Foundation under Grant Number 1566270 and a NVIDIA hardware grant. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are grateful to our human evaluation study participants and the reviewers for their time and effort. Finally, we thank Sanchayan Sarkar for preparing the AffectNet data.

References

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

[2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *To appear, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] Hui Ding, Kumar Sricharan, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[4] Irfan A. Essa and Alex Paul Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):757–763, 1997.

[5] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1785. IEEE, 2009.

[6] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[7] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 166–174, 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[9] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[10] Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1133–1141. IEEE, 2017.

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

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/ioffe15.html.

[13] Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 216–223, 2014.

[14] Jungseock Joo, Francis F Steen, and Song-Chun Zhu. Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3712–3720, 2015.

[15] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53. IEEE, 2000.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, pages 1097–1105, 2012.

[19] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(3):453–465, 2014.

[20] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic DENOYER, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in neural information processing systems (NIPS)*, pages 5963–5972, 2017.

[21] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[23] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1805–1812, 2014.

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

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

[26] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. Adversarial autoencoders. In *International Conference on Learning Representations (ICLR)*, 2016. URL http://arxiv.org/abs/1511.05644.

[27] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *International Conference on Learning Representations (ICLR)*, 2016.

[28] 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.

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in neural information processing systems - Workshops(NIPS-W)*, 2017.

[30] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems (NIPS)*, pages 2352–2360, 2016.

[31] 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 (NIPS)*, pages 91–99, 2015.

[32] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015.

[34] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3483–3491, 2015.

[35] 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 (CVPR)*, pages 2818–2826, 2016.

[36] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 835–851. Springer, 2016.

[37] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision (ECCV)*, pages 776–791. Springer, 2016.

[38] 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 (CVPR)*, pages 5525–5533, 2016.

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

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep networks as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.