UNSUPERVISED DISENTANGLING STRUCTURE AND APPEARANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

It is challenging to disentangle an object into two orthogonal spaces of structure and appearance since each can influence the visual observation in a different and unpredictable way. It is rare for one to have access to a large number of data to help separate the influences. In this paper, we present a novel framework to learn this disentangled representation in a completely *unsupervised* manner. We address this problem in a two-branch Variational Autoencoder framework. For the structure branch, we project the latent factor into a soft structured point tensor and constrain it with losses derived from prior knowledge. This encourages the branch to distill geometry information. Another branch learns the complementary appearance information. The two branches form an effective framework that can disentangle object's structure-appearance representation without any human annotation. We evaluate our approach on four image datasets, on which we demonstrate the superior disentanglement and visual analogy quality both in synthesized and real-world data. We are able to generate photo-realistic images with 256×256 resolution that are clearly disentangled in structure and appearance.

1 INTRODUCTION

Structure and appearance are the two most inherent attributes that characterize an object visually. Computer vision researchers have devoted decades of efforts to understand object structure and extract features that are invariant to geometry change (Huang et al., 2007; Thewlis et al., 2017; Rocco et al., 2018). Learning such disentangled deep representation for visual objects is an important topic in deep learning.

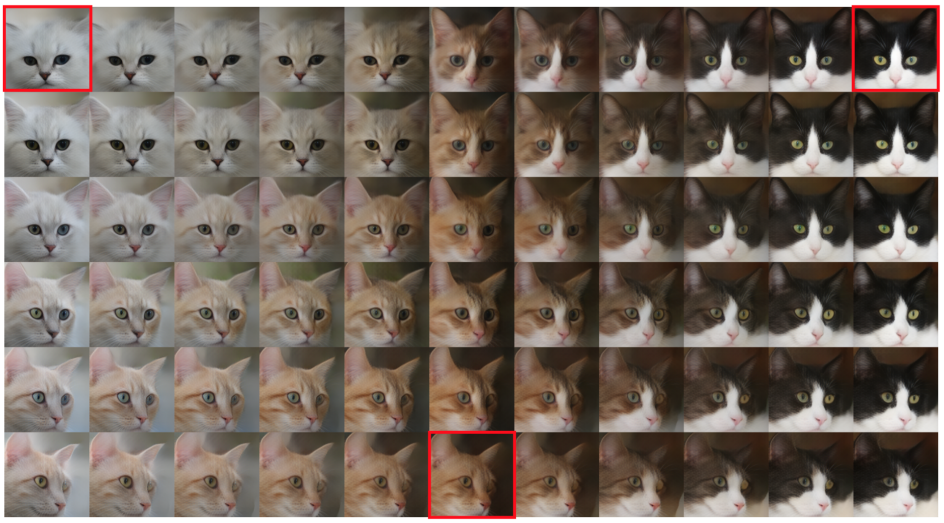

Figure 1: **Walking in the disentangled representation space:** Here we show an example learned by our algorithm. Our approach effectively disentangles the structure and appearance space. Three cat faces in the bounding box are from real data while others are interpolated through our learned representations.

The main objective of our work is to disentangle object's appearance and structure in an *unsupervised* manner. Achieving this goal is non-trivial due to three reasons: 1) Without supervision, we can hardly guarantee the separation of different representations in the latent space. 2) Although some methods like InfoGAN (Chen et al., 2016) are capable of learning several groups of independent attributes from objects, attributes from these unsupervised frameworks are uninterpretable since we cannot pinpoint which portion of the disentangled representation is related to the structure and which to the appearance. 3) Learning structure from a set of natural real-world images is difficult.

To overcome the aforementioned challenges, we propose a novel two-branch Variational Autoencoder (VAE) framework, of which the structure branch aims to discover meaningful structural points to represent the object geometry, while the other appearance branch learns the complementary appearance representation. The settings of these two branches are asymmetric. For the structure branch, we add a layer-wise softmax operator to the last layer. This can be seen as a projection of a latent structure to a soft structured point tensor space. Specifically designed prior losses are used to constrain the structured point tensors so that the discovered points have high repeatability across images yet distributed uniformly to cover different parts of the object. To encourage the framework to learn a disentangled yet complementary representation of both appearance and structure, we introduce a Kullback-Leibler (KL) divergence loss and skip-connections design to the framework.

Extensive experiments demonstrate the effectiveness of the proposed method in manipulating the structure and appearance of natural images, e.g., cat faces in Figure 1, outperform state-of-the-art algorithms (Chen et al., 2016; Higgins et al., 2017; Jakab et al., 2018). We also conduct several experiments on MNIST-Color, 3D synthesized data and real photos.

2 METHODOLOGY

In the absence of annotation on structure, we rely on prior knowledge on how object landmarks should distribute to constrain the learning and disentanglement of structural information. Our experiments show that this is possible given appropriate prior losses and learning architecture. We first formulate our loss function with a consideration on prior. Specifically, we follow the VAE framework and assume 1) the two latent variables z and y, which represent the appearance and structure, are generated from some prior distributions. 2) x follows the conditional distribution $p(x|y, z)$. We start with a Bayesian formulation and maximize the log-likelihood over all observed samples $x \in X$.

$$\begin{aligned}
\log p(x) &= \log p(y) + \log p(x|y) - \log p(y|x) \\
&\geq \log p(y) + \log \int p(x, z|y)\, \mathrm{d}z \\
&\geq \log p(y) + \mathbb{E}_q \log \frac{p(x, z|y)}{q(z|x, y)} \\
&= \log p(y) + \mathbb{E}_q \log \frac{p(x|y, z)p(z|y)}{q(z|x, y)}.
\end{aligned} \tag{1}$$

Equation 1 learns a deterministic mapping $e(\cdot; \theta)$ from x to y, which we assume y is following a Gaussian distribution over $\mathcal{N}(e(x; \omega), \Sigma)$. Term $-\log p(y|x)$ is non-negative. In the second line of the equation, we start to consider the factor z. Similar to VAE, we address the issue of intractable integral by introducing an approximate posterior $q(y, z|x; \phi)$ to estimate the integral using evidence lower bound (ELBO). By splitting the $p(x|y, z)$ from the second term of the last expression, we obtain our final loss as,

$$\mathcal{L}(x, \theta, \phi, \omega) = -\log p_\omega(y) - \mathbb{E}_{q_{\phi(z|x,y)}} \log p_\theta(x|y, z) + \mathrm{KL}(q_{\phi(z|x,y)}(z|x, y)||p_\theta(z|y)). \tag{2}$$

The first term is the prior on y. The second term describes the conditional distribution of x given all representation. Ideally, if the decoder can perfectly reconstruct the x, the second term would be a delta function over x. The third term represents the Kullback-Leibler divergence between approximate. In the rest of this paper we name these three terms respectively as prior loss L_{prior}, reconstruction loss L_{recon} and KL loss L_{KL}.

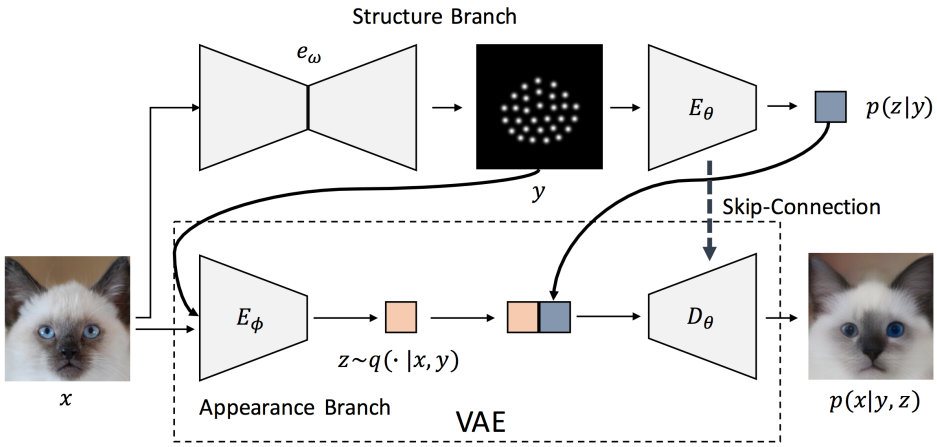

Figure 2: **Architecture:** Our framework follows an auto-encoder framework. It contains two branches: 1) the structure branch forces the representation into a Gaussian spatial probability distribution with an hourglass network e_ω. 2) the appearance branch E_ϕ learns a complementary appearance representation to the structure.

2.1 Prior Loss

Inspired by Zhang et al. (2018) and Jakab et al. (2018), we formulate our structure representation y as a soft latent structured point tensor. A re-projecting operator is applied here to force y to lie on a Gaussian spatial probability distribution space.

Following the notations from Newell et al. (2016), we denote the direct outputs of the hourglass network e_ω as landmark heatmaps h, and each channel of which represents the spatial location of a structural point. Instead of using max activations across each heatmap as landmark coordinates, we weighted average all activations across each heatmap. We then re-project landmark coordinates to spatial features with the same size as heatmaps by a fixed Gaussian-like function centered at predicted coordinates with a fixed standard deviation. As a result, we obtain a new tensor y with prior on structure representation.

Similar to the difficulty described in Zhang et al. (2018), we find that training the structure branch with general random initialization tend to locate all structural points around the mean location at the center of the image. This could lead to a local minimum from which optimizer might not escape. As such, we introduce a *Separation Loss* to encourage each heatmap to sufficiently cover the object of interest. This is achieved by the first part in Eq. 3, where we encourage each pair of ith and j^{th} heatmaps to share different activations. σ can be regarded as a normalization factor here. Another prior constraint is that we wish the structural point to behave like landmarks to encode geometry structure information. To achieve this goal, we add a *Concentration Loss* to encourage the variance of activations h to be small so that it could concentrate at a single location. This corresponds to the second term in Eq. 3.

$$L_{prior} = \sum_{i \neq j} \exp(-\frac{||h_i - h_j||^2}{2\sigma^2}) + \text{Var}(h) \tag{3}$$

It is noteworthy that some recent works have considered the prior of latent factor. Dupont (2018) proposed a Joint-β-VAE by adding different prior distribution over several latent factors so as to disentangle continuous and discrete factors from data. Our work differs in that we investigates a different prior to disentangle visual structure and appearance.

2.2 Reconstruction Loss

For the second term we optimize the reconstruction loss of whole model, which will be denoted as generator G in the following context. We assume that the decoder D_θ is able to reconstruct original input x from latent representation y and z, which is $\hat{x} = G(y, z)$. Consequently, we can design the reconstruction loss as $L_{\text{recon}} = \|x - \hat{x}\|_1$.

However, minimizing L_1 / L_2 loss at pixel-level only does not model the perceptual quality well and makes the prediction look blurry and implausible. This phenomenon has been well-observed in the literature of super-resolution (Bruna et al., 2016; Sajjadi et al., 2017). We consequently define the reconstruction loss as $L_{\text{recon}} = \|x - \hat{x}\|_1 + \sum_l \lambda_l \|\psi_l(x) - \psi_l(\hat{x})\|_1$, where ψ_l is the feature obtained from l-th layer of a VGG-19 model (Simonyan & Zisserman, 2014) pre-trained on ImageNet. It is also possible to add adversarial loss to further improve the perceptual reconstruction quality. Since the goal of this work is disentanglement rather than reconstruction, we only adopt the L_{recon} described above.

2.3 KL Loss

We model $q(z|x, y)$ as a parametric Gaussian distribution which can be estimated by the encoder network E_ϕ. Therefore, the appearance code z can be sampled from $q(z|x, y)$. Meanwhile, the prior $p(z|y)$ can be estimated by the encoder network E_θ. By using the reparametrization trick (Kingma & Welling, 2014), these networks can be trained end-to-end. In this work, only mean is estimated for the stability of learning. By modeling the two distributions as Gaussian with identity covariances, the KL Loss is simply equal to the Euclidean distance between their means. Thus, z is regularized by minimizing the KL divergence between $q(z|x, y)$ and $p(z|y)$.

Notice that with only prior and reconstruction loss. The framework only makes sure z is from x and the Decoder D_θ will recover as much information of x as possible. There is no guarantee that z will learn a complementary of y. Towards this end, we design the network as concatenating the encoded structure representation by E_θ with the inferred appearance code z. Then, the concatenated representation is decoded together by D_θ. Moreover, skip-connections between E_θ and D_θ are also used to pass multi-level structure information to the decoder. Since enough structure information can be obtained from prior, any information about structure encoded in z incurs a penalty of the likelihood $p(x|y, z)$ with no new information (*i.e.* appearance information) is captured. This design of network and the KL Loss result in a constraint to guide z to encode more information about appearance which is complementary to the structure prior.

2.4 Implementation Detail

Each of the input images x is cropped and resized to 256×256 resolution. A one-stack hourglass network (Newell et al., 2016) is used as a geometry extractor e_ω to project input image to the heatmap $y \in \mathbb{R}^{256\times256\times30}$, in which each channel represents one point-centered 2D-Gaussian map (with $\sigma = 4$). y is drawn in a single-channel map for visualization in Fig. 2. Same network (with stride-2 convolution for downsample) is use for both E_θ and E_ϕ to obtain appearance representation z and the embedded structure representation as two 128-dimension vectors. A symmetrical deconvolution network with skip connection is used as the decoder D_θ to get the reconstructed result $\hat{x}$. All of the networks are jointly trained from scratch end-to-end. We detail the architectures and hyperparameters used for our experiments in appendix A.

3 Related Work

Unsupervised Feature Disentangle: Several pioneer works focus on unsupervised disentangled representation learning. Following the propose of GANs (Goodfellow et al., 2014), Chen et al. (2016) purpose InfoGAN to learn a mapping from a group of latent variables to the data in an unsupervised manner. Many similar methods were purposed to achieve a more stable result (Higgins et al., 2017; Kumar et al., 2018). However, these works suffer to interpret, and the meaning of each learned factor is uncontrollable. There are some following works focusing on dividing latent factors into different sets to enforce better disentangling. Mathieu et al. (2016) assign one code to the specified factors of variation associated with the labels, and left the remaining as unspecified variability. Similar to Mathieu et al. (2016), Hu et al. (2018) then proposes to obtain disentanglement of feature chunks by leveraging Autoencoders, with the supervision of some same/different class pairs. Dupont (2018) divides latent variable into discrete and continuous one, and distribute them in different prior distribution. In our work, we give one branch of representation are more complicated prior, to force it to represent only the pose information for the object.

Supervised Pose Synthesis: Recently the booming of GANs research improves the capacity of pose-guided image generation. Ma et al. (2017) firstly try to synthesize pose images with U-Net-like networks. Several works soon follow this appealing topic and obtain better results on human pose or face generation. A close work to us from Esser et al. (2018) applied a conditional U-Net for shape-guided image generation. Nevertheless, existing works rely on massive annotated data, they need to treat pose of a object as input, or a strong pre-trained pose estimator.

Unsupervised Structure Learning: Unsupervised learning structure from objects is one of the essential topics in computer vision. The rudimentary works focus on keypoints detection and learning a strong descriptor to match (Thewlis et al., 2017; Rocco et al., 2018). Recent two concurrent works, from Jakab et al. (2018) and Zhang et al. (2018), show the possibility of end-to-end learning of structure in Autoencoder formulations. Our work can be seen as extending their work to learn the complementary appearance representation as well (in other words, in the loss Eq. 1, they only consider the first two terms, and ignore the factor from z).

4 Experiments

4.1 Experimental Protocol

Datasets: We evaluate our method on four datasets that cover both synthesized and real world data: 1). MNIST-Color: we extend MNIST by either colorizing the digit (MNIST-CD) or the background (MNIST-CB) with a randomly chosen color following Gonzalez-Garcia et al. (2018). We use the standard split of training (50k) and testing (10k) set. 2). 3D Chair: Aubry et al. (2014) offers rendered images of 1393 CAD chair models. We take 1343 chairs for training and the left 50 chairs for testing. For each chair, 12 rendered images with different views are selected randomly. 3). Cat & Dog Face, we collect 6k (5k for training and 1k for testing) images of cat and dog from YFCC100M (Kalkowski et al., 2015) and Standford Dog (Khosla et al., 2011) datasets respectively. All images are center cropped around the face and scaled to the same size. 4). CelebA: it supplies plenty of celebrity faces with different attributes. The training and testing sizes are 160K and 20K respectively.

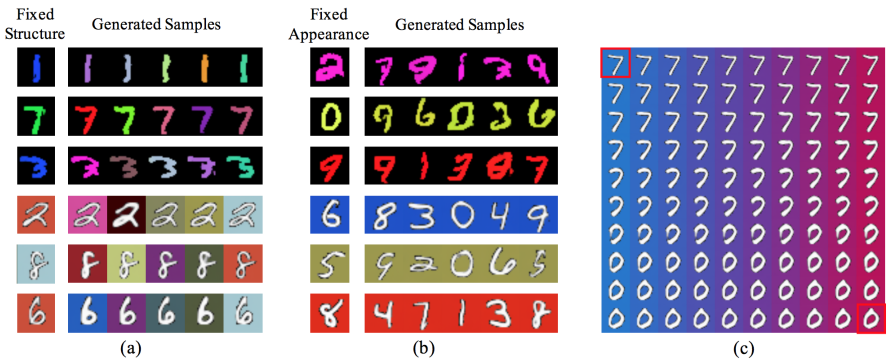

Figure 3: **Conditional generation results:**(a) Walking in the appearance space with fixed structure. (b) Walking in the structure space with fixed appearance. (c) A visualization of the disentangled space by linear interpolation. The Structure is smoothly changed in row-wise and the appearance is changed by each column.

Evaluation Metric: Less existed evaluation metric and benchmark can be utilized to evaluate the performance of disentanglement. Here we propose two forms of evaluation to study the behavior of the proposed framework: 1). **Qualitative:** we provide four kinds of qualitative results to show as many usages of the disentangled space as possible, *i.e.* conditional sampling, interpolation, retrieval, and visual analogy. 2). **Quantitative:** we apply several metrics that are widely employed in image generation (a) Structure consistency: content similarity metric (Li et al., 2017) and mean-error of landmarks (Bulat & Tzimiropoulos, 2017). (b) Appearance consistency: style similarity metric (Johnson et al., 2016) (c). Disentangled ability: retrieval recall@K (Sangkloy et al., 2016). (d). Reconstruction and generation quality: SSIM (Wang et al., 2004) and Inception Score (Salimans et al., 2016).

4.2 RESULTS ON SYNTHESIZED DATASETS

Diverse Generation. We first demonstrate the diversity of conditional generation results on MNIST-Color with the successfully disentangled structure and appearance in Fig. 3. It can be observed that, given an image as a structure condition, same digit information with different appearance can be generated by sampling the appearance condition images randomly. While given an image as appearance condition, different digits with the same color can be generated by sampling different structural conditional images. Note that the model has no prior knowledge of the digit in the image as no label is provided, it effectively learns the disentanglement spontaneously.

Interpolation. In Fig. 3, the linear interpolation results show reasonable coverage of the manifold. From left to right, the color is changed smoothly from blue to red with interpolated appearance latent space while maintaining the digit information. Analogously, the color stays stable while one digit transforms into the other smoothly from top to down.

Retrieval. To demonstrate the disentangled ability of the representation learned by the model, we perform nearest neighbor retrieval experiments following Mathieu et al. (2016) on MNIST-Color. With structure and appearance representation used, both semantic and visual retrieval can be performed respectively. The Qualitative results are shown in appendix A. Quantitatively, We use a commonly used retrieval metric Recall@K as in (Sangkloy et al., 2016; Pang et al., 2017), where for a particular query digit, Recall@K is 1 if the corresponding digit is within the top-K retrieved results and 0 otherwise. We report the most challenging Recall@1 by averaging over all queries on the test set in Table 2. It can be observed that the structure representation shows the best performance and clearly outperforms image pixel and appearance representation. In addition to the disentangled ability. This result shows that the structure representation learned by our model is useful for visual retrieval.

Visual Analogy. The task of visual analogy is that the particular attribute of a given reference image can be transformed to a query one (Reed et al., 2015). We show the visual analogy results on MNIST-Color and 3D Chair in Fig. 4. Note that even for the detail component (*e.g.* wheel and leg of 3D chair) the structure can be maintained successfully, which is a rather challenging task in previous unsupervised works (Chen et al., 2016; Higgins et al., 2017).

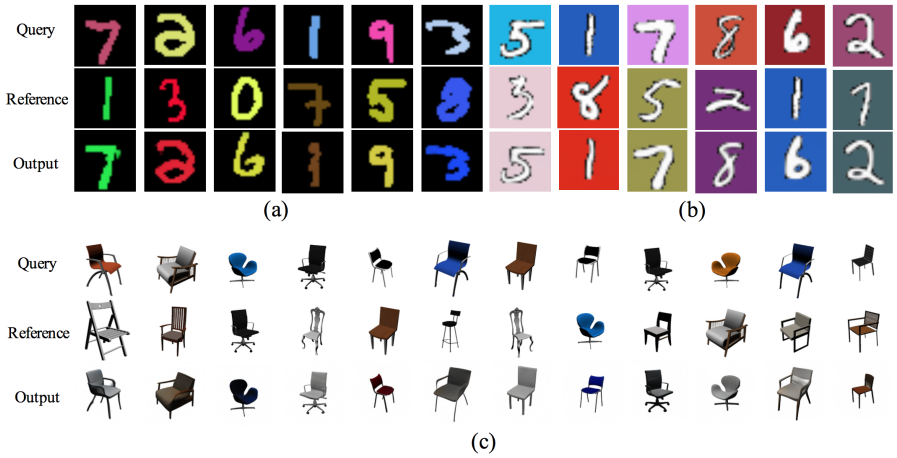

Figure 4: **Visual analogy results on synthesized datasets:** (a) MNIST-CD. (b) MNIST-CB. (c) 3D Chair. Taking the structure representation of a query image and the appearance representation of the reference one, our model can output an image which maintains the geometric shape of query image while capturing the appearance of the reference image.

4.3 RESULTS ON REAL-LIFE DATASETS

We have so far only discussed results on the synthesized benchmarks. In this section, we will demonstrate the scalable performance of our model on several real-life datasets, *i.e.*, Cat, Dog Face and CelebA. To the best knowledge of ours, there is no literature of unsupervised disentanglement before can successfully extend to photo-realistic generation with 256×256 resolution. Owing to

Method	Cat			CelebA		
	Style ($\times e^{-5}$)	Content ($\times e^{-6}$)	Landmark (%)	Style ($\times e^{-5}$)	Content ($\times e^{-6}$)	Landmark (%)
Random	7.700	1.881	0.051	5.858	1.693	0.293
Ours	5.208	1.759	0.030	3.886	1.529	0.162

Table 1: Structure and appearance consistency evaluation on Cat and CelebA dataset (lower is better).

the structural prior which accurately capture the structural information of images, our model can transform appearance information while faithfully maintain the geometry shapes.

Qualitative evaluation is performed by visually examining the perceptual quality of the generated images. In Fig. 7, the swapping results along with the learned geometry heatmaps y are illustrated on Cat dataset. In can be seen that the geometry information, *i.e.*, expression, head-pose, facial action, and appearance information *i.e.*, hair texture, can be swapped between each other arbitrarily. The learned geometry heatmaps can be shown as a map with several 2D Gaussian points, which successfully encode the geometry cues of a image by the location of its points and supply an effective prior for the VAE network. More results of visual analogy of real-life datasets on Standford Dog and CelebA dataset are illustrated in Fig. 5. We observe that the model is able to generalize to various real-life images with large variations, such as mouth-opening, eye-closing, tongue-sticking and exclusive appearance.

For quantitative measurement, there is no standard evaluation metric of the quality of the visual analogy results for real-life datasets since ground-truth targets are absent. We propose to evaluate the structure and appearance consistency of the analogy predictions respectively instead. We use content similarity metric for the evaluation of structure consistency between a condition input x_s and its guided generated images (*e.g.*, for each column of images in Fig. 7). We use style similarity metric to evaluate the appearance consistency between a condition input x_a and its guided generated images (*e.g.*, each row of images in Fig. 7). These two metrics are used widely in image generation applications as an objective for training to maintain content and texture information (Li et al., 2017; Johnson et al., 2016).

As content similarity metric is less sensitive to the small variation of images, we propose to use the mean-error of landmarks detected by a landmark detection network, which is pre-trained on manually annotated data, to evaluate the structure consistency. Since the public cat facial landmark annotations are too sparse to evaluate the structure consistency (*e.g.* 9-points (Zhang et al., 2008)), we manually annotated 10k cat face with 18-points to train a landmark detection network for evaluation purpose. As for the evaluation of celebA, a state-of-the-art model (Bulat & Tzimiropoulos, 2017) with 68-landmarks is used. The results on the testing set of the two real-life datasets are reported in Table 1. For each test image, 1k other images in the testing set are all used as the reference of structure or appearance for generating, in which mean value is calculated. In the baseline random setting, for one test image, the mean value is calculated by sampling randomly among the generated images guided by each image. The superior structure and appearance consistency of the images generated by our method can be obviously observed.

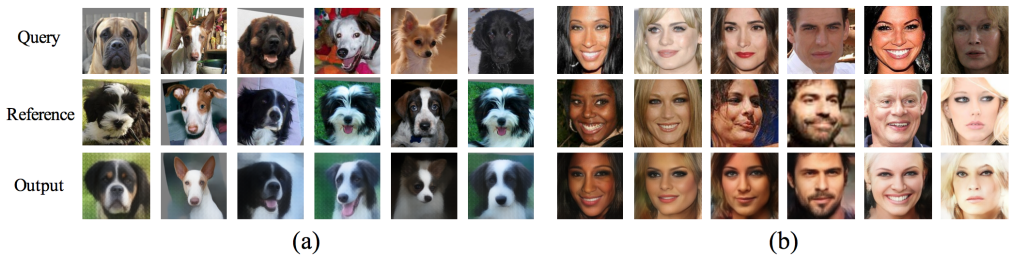

Figure 5: **Visual analogy results on real-life datasets:** (a) Standford Dog. (b) CelebA. The geometry (*e.g.* identity, head pose and expression) of query image can be faithfully maintained while the appearance (*e.g.* the color of hair, beard and illumination) of reference image can be precisely transformed. As concrete examples, the output of the dog in the third column is still tongue-sticking while the hair color is changed, and in the last column of CelebA, even the fine-grain eye make-up is successfully transformed to the query image surprisingly.

4.4 COMPARISON TO OTHER METHODS

Since there is hardly any literature before share exactly the same settings with us as discussed in the related work, we compare perceptual quality with the four most related unsupervised representation learning methods in Fig. 6, including three disentangled factor learning methods, *i.e.*, VAE (Kingma & Welling, 2014), β-VAE (Higgins et al., 2017) and InfoGAN (Chen et al., 2016), and one unsupervised structure learning method (Jakab et al., 2018). It can be observed that all of the three methods can automatically discover and learn to disentangle the factor of azimuth on 3D Chair dataset. However, it can be perceived that the geometry shape can be maintained much better in our approach than all the other methods, owing to the informative prior supplied by our structure branch. We randomly sample several query-reference pairs in testing set to compare with results reported in the paper of Jakab et al. (2018). The results of unsupervised structure learning methods have severe artifacts and look more blurred compared with ours. Nevertheless, the identity of query face can be hardly kept in the results of Jakab et al. (2018).

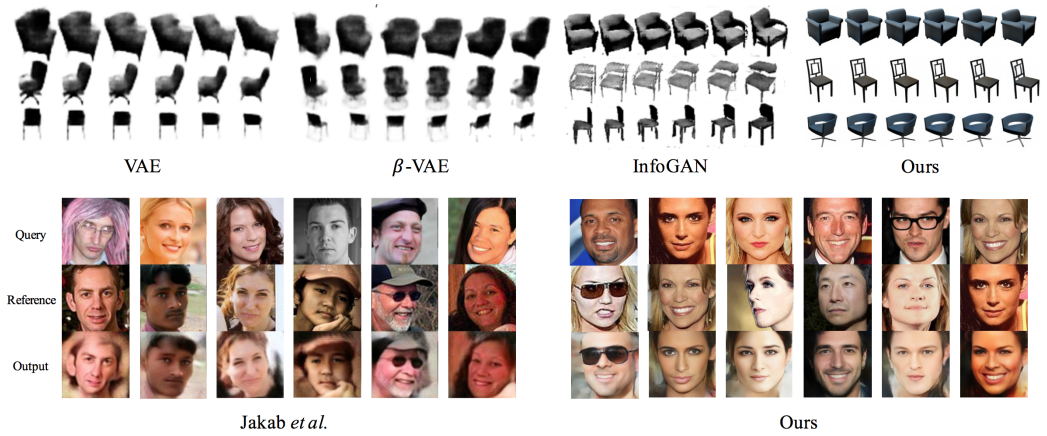

Figure 6: **Comparison to other methods.** Qualitative results of disentangling performance of VAE, β-VAE, InfoGAN and Jakab et al. (2018). We demonstrate the disentanglement of the factor of azimuth for 3D chair dataset. Visual analogy results are demonstrated for face dataset.

4.5 ABLATION STUDY

It is worth studying the effect of individual component of our method on the quality of generated images. Structural Similarities (SSIM) and Inception Scores (IS) are utilized to evaluate the reconstruction quality and the analogy quality. As reported in Table 2, without KL-Loss, the network has no incentive to learn the shape invariant appearance of representation, almost all of the metrics degraded dramatically.

Method	Color-Digit	Color-Back
Pixel	31.65	39.52
Appearance	10.25	15.32
Structure	**99.96**	**99.92**

(a)

Method	Style ($\times e^{-5}$)	Content ($\times e^{-6}$)	Landmark (%)	SSIM		Inception Score	
				mean	std	mean	std
Real Data	-	-	-	1.000	0.000	2.004	0.157
Without KL Loss	6.556	1.813	0.036	0.406	0.103	1.72	0.189
Ours	**5.208**	**1.759**	**0.030**	**0.449**	0.121	**1.968**	**0.181**

(b)

Table 2: (a) Retrieval results reported as recall@1 on MNIST-Color. (b) Ablation study.

5 CONCLUSION

We extend VAE model to disentangle object's representation by structure and appearance. Our framework is able to mine structure from a kind of objects and learn structure-invariant appearance representation simultaneously, without any annotation. Our work may also reveal several potential topics for future research: 1) Instead of relying on supervision, using strong prior to restrict the latent variables seems to be a potential and effective tool for disentangling. 2) In this work we only experiment on near-rigid objects like chairs and faces, learning on deformable objects is still an opening problem. 3) The structure-invariant appearance representation may have some potentials on recognition tasks.

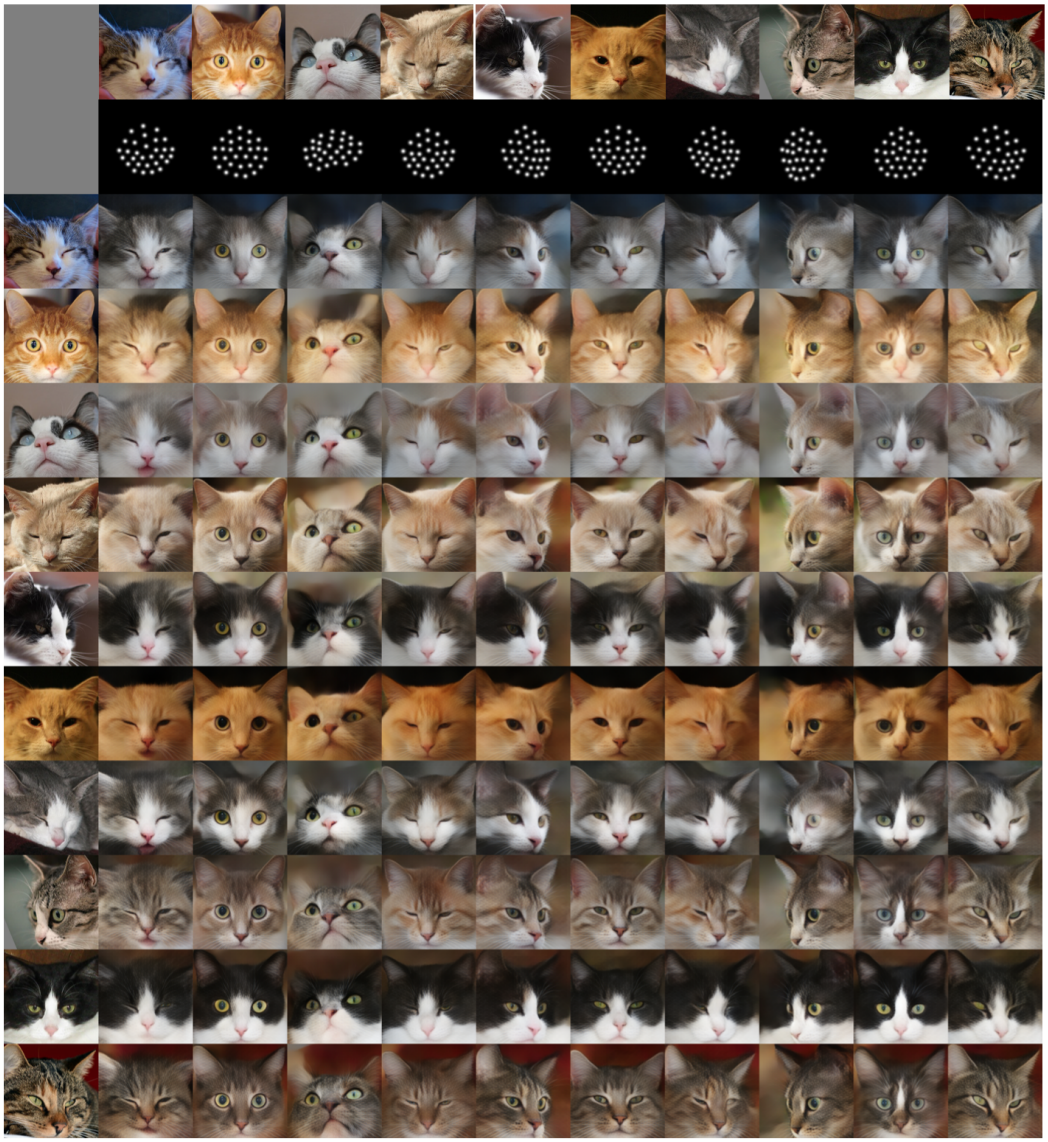

Figure 7: **A grid of structure&appearance swapping visualization.** The top row and left-most columns are random selected from the test set. In each column, the structure of the generated images are shown to be consistent with the top ones. In each row, the appearance of the generated images are shown to be consistent with the left-most ones.

REFERENCES

Mathieu Aubry, Daniel Maturana, Alexei A. Efros, Bryan C. Russell, and Josef Sivic. Seeing 3d chairs: Exemplar part-based 2d-3d alignment using a large dataset of CAD models. In *CVPR*, 2014.

Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics. In *ICLR*, 2016.

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? In *ICCV*, 2017.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

Emilien Dupont. Learning disentangled joint continuous and discrete representations. In *NIPS*, 2018.

Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018.

Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. In *NIPS*, 2018.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation by mixing them. *CVPR*, 2018.

Gary B. Huang, Vidit Jain, and Erik G. Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007.

Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Conditional image generation for learning the structure of visual objects. *NIPS*, 2018.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

Sebastian Kalkowski, Christian Schulze, Andreas Dengel, and Damian Borth. Real-time analysis and visualization of the yfcc100m dataset. In *MM Workshop*, 2015.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPR Workshop*, 2011.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *ICLR*, 2018.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *NIPS*, 2017.

Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *NIPS*, 2016.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

Kaiyue Pang, Yi-Zhe Song, Tony Xiang, and Timothy M. Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *BMVC*, 2017.

Scott E. Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *NIPS*, 2015.

Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. End-to-end weakly-supervised semantic alignment. In *CVPR*, 2018.

Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, 2017.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016.

Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. In *ACM Transactions on Graphics*, 2016.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. volume abs/1409.1556, 2014.

James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *NIPS*, 2017.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004.

Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection - how to effectively exploit shape and texture features. In *ECCV*, 2008.

Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018.

A APPENDIX

A.1 DETAILS OF ARCHITECTURE

We use Adam with parameters $\beta_1 = 0.5$ and $\beta_1 = 0.999$ to optimise the network with a mini-batch size of 8 for 160 epochs for all datasets. The initial learning rate is set to be 0.0001 and then decreasely linearly to 0 during training.

The network architecture used for our experiments is given in Table 3. We use the following abbreviation for ease of presentation: N=Neurons, K=Kernel size, S=Stride size. The transposed convolutional layer is denoted by DCONV.

	Layer	Module
Encoder (E_ϕ, E_θ)	1	CONV-(N64,K4,S2)
	2	LeaklyReLU, CONV-(N128,K4,S2), InstanceNorm
	3	LeaklyReLU, CONV-(N128,K4,S2), InstanceNorm
	4	LeaklyReLU, CONV-(N128,K4,S2), InstanceNorm
	5	LeaklyReLU, CONV-(N128,K4,S2), InstanceNorm
	6	LeaklyReLU, CONV-(N128,K4,S2), InstanceNorm
	7	LeaklyReLU, CONV-(N128,K4,S2), InstanceNorm
	8	LeaklyReLU, CONV-(N128,K4,S2), InstanceNorm
	μ	CONV-(N128,K1,S1)
Decoder (D_θ)	1	CONV-(N128,K1,S1)
	2	ReLU, DCONV-(N128,K4,S2), InstanceNorm
	3	ReLU, DCONV-(N128,K4,S2), InstanceNorm
	4	ReLU, DCONV-(N128,K4,S2), InstanceNorm
	5	ReLU, DCONV-(N128,K4,S2), InstanceNorm
	6	ReLU, DCONV-(N128,K4,S2), InstanceNorm
	7	ReLU, DCONV-(N128,K4,S2), InstanceNorm
	8	ReLU, DCONV-(N64,K4,S2), InstanceNorm
	9	ReLU, DCONV-(N3,K4,S2), Tanh

Table 3: Network architecture of encoder and decoder.

A.2 QUALITATIVE RESULTS

The qualitative retrieval results on MNIST-Color are illustrated in Fig. 8. With structure and appearance representation used, both semantic and visual retrieval can be per-formed respectively. Moreover, the interpolation results of 3D Chair with same arrangement as MNIST-Color is shown in Fig. 9.

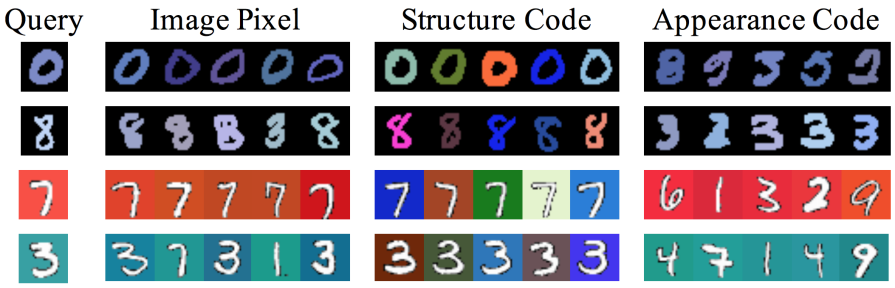

Figure 8: Random chosen 4 query images and the corresponding 5 nearest-neighbors are illustrated, which are retrieved with image pixel, structure code, appearance code respectively.

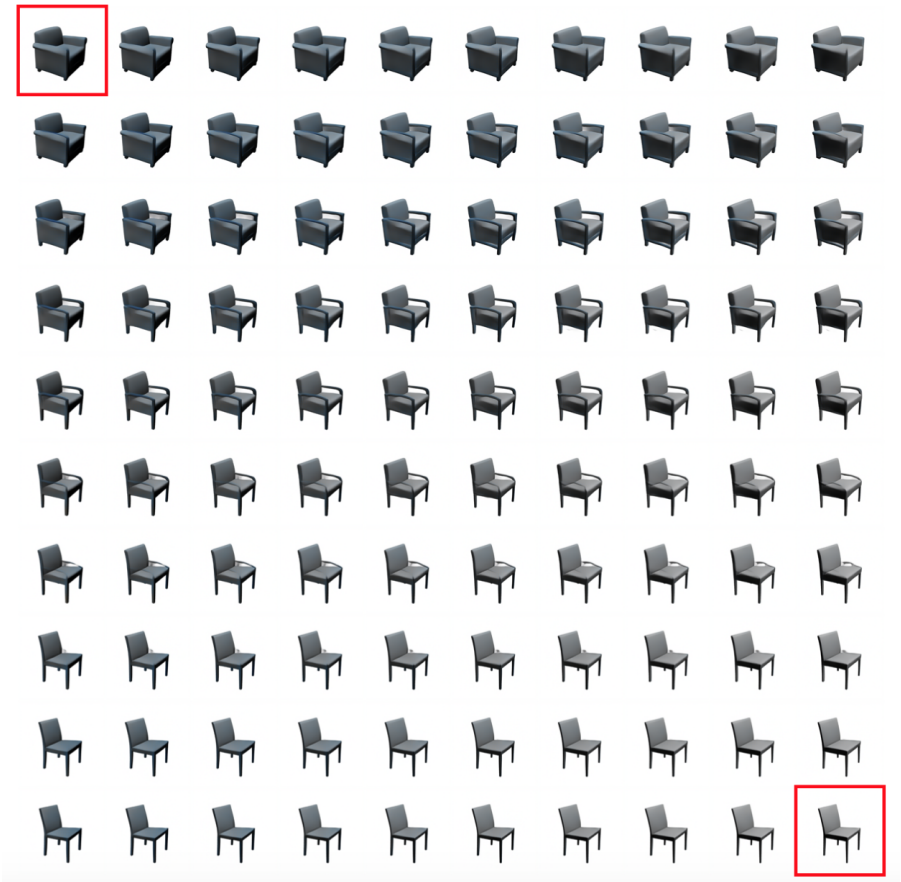

Figure 9: Interpolation results on 3D Chair.