# Unsupervised Robust Disentangling of Latent Characteristics for Image Synthesis

Patrick Esser, Johannes Haux, Björn Ommer
Heidelberg Collaboratory for Image Processing
IWR, Heidelberg University, Germany
firstname.lastname@iwr.uni-heidelberg.de

## Abstract

*Deep generative models come with the promise to learn an explainable representation for visual objects that allows image sampling, synthesis, and selective modification. The main challenge is to learn to properly model the independent latent characteristics of an object, especially its appearance and pose. We present a novel approach that learns disentangled representations of these characteristics and explains them individually. Training requires only pairs of images depicting the same object appearance, but no pose annotations. We propose an additional classifier that estimates the minimal amount of regularization required to enforce disentanglement. Thus both representations together can completely explain an image while being independent of each other. Previous methods based on adversarial approaches fail to enforce this independence, while methods based on variational approaches lead to uninformative representations. In experiments on diverse object categories, the approach successfully recombines pose and appearance to reconstruct and retarget novel synthesized images. We achieve significant improvements over state-of-the-art methods which utilize the same level of supervision, and reach performances comparable to those of pose-supervised approaches. However, we can handle the vast body of articulated object classes for which no pose models/annotations are available.*

## 1. Introduction

Supervised end-to-end training on large volumes of tediously labelled data has tremendously propelled deep learning [28]. The discriminative learning paradigm has enabled to train deep network architectures with millions of parameters to address important computer vision tasks such as image categorization [48], object detection [44], and segmentation [46]. The network architectures underlying these discriminative models have become increasingly complex [49] and have tremendously increased in depth [17] to yield great improvements in performance. However, the ability to explain these models and their decisions suffers due to this discriminative end-to-end training setup [57, 52, 42].

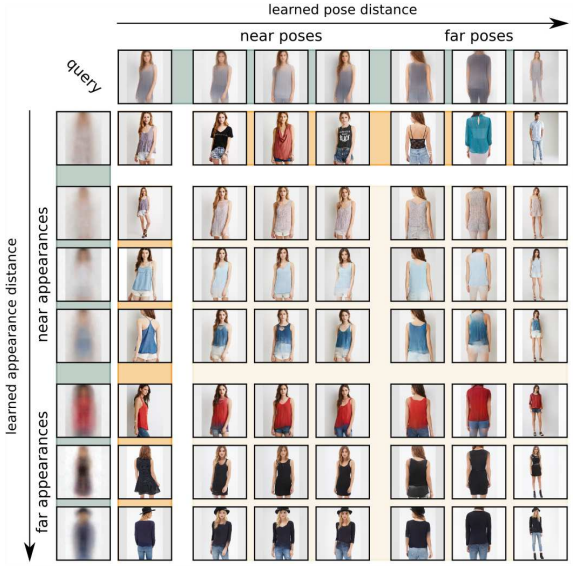



Figure 1. Without annotations about pose, we disentangle images into two independent factors: pose and appearance. Starting from a query image (top left), we can extract and visualize its pose representation (top row), and retrieve images based on their pose similarity. The visualizations (Sec. 4.2) in the first row show that pose is learned accurately and contains no information about appearance. Similarly, appearance is visualized in the first column. Because our representations are both independent and informative, we can recombine arbitrary combinations to synthesize what an appearance would look like in a specific pose. More results can be found at https://compvis.github.io/robust-disentangling.

Consequently, there has recently been a rapidly increasing interest in deep generative models [26, 45, 14, 55]. These aim for a complete description of data in terms of a joint distribution and can, in a natural way, synthesize images from a learned representation. Thus, besides explain-


Figure 2. First row: pose target. First column: appearance target. Because our method does not require keypoint-annotations or class information, it can be readily applied on video datasets [22, 13]. Besides intra-species analogies, our approach can also imagine inter-species analogies: How does a cow look like in a pose specified by a horse?

ing the joint distribution of all data, they provide a powerful tool to visualize and explain complex models [7].

However, while already simple probabilistic models may produce convincing image samples, their ability to describe the data is lacking. For instance, great progress in image synthesis and interpolation between different instances of an object category (e.g., young versus old faces) has been achieved [54, 30, 18]. But these models explain all the differences between instances as a change of appearance. Consequently, changes in posture, viewpoint, articulation and the like (subsequently simply denoted as pose) are blended with changes in color or texture.

To address the different characteristics of pose and appearance, many recent approaches started to rely on existing discriminative pose detectors. While these models show good results on disentangled image generation, their applicability is limited to domains with existing, robust pose detectors. This introduces two problems: The output of the pose detector introduces a bias into the notion of what constitutes pose, and labeling large scale datasets for each new category of objects to be explained is unfeasible. How can we learn pose and appearance without these problems?

The task naturally calls for two encoders [10, 15, 40] to extract representations of appearance and pose, and a decoder to reconstruct an image from them. To train the model, one encoder infers the pose from the image to be reconstructed; the other encoder infers the appearance from another image showing the same appearance. Without further constraints, the reconstruction task alone produces a degenerate solution [16, 51, 40]: The pose encoding contains all the information—including appearance—and the decoder ignores the appearance representation. In this case, the model collapses to an autoencoder and avoiding this is the main goal of disentanglement.

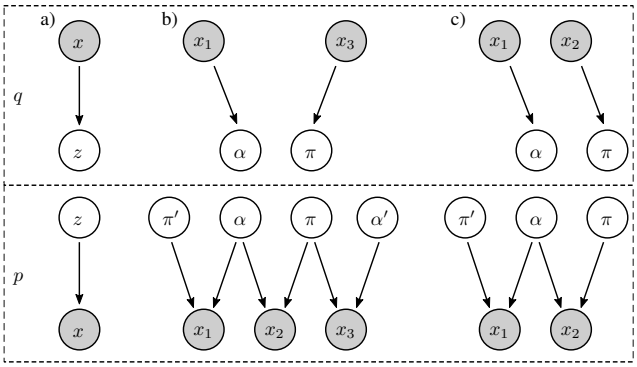
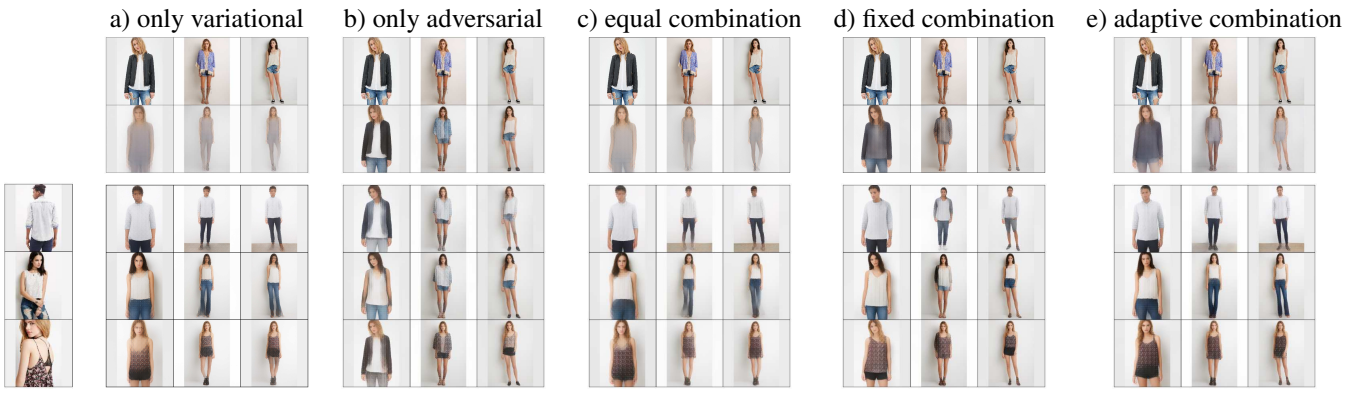



Figure 3. Graphical models describing the dependencies within a) a simple latent variable model b) a complete model of disentangled factors c) the same model with unobserved $x_3$. $p$ describes the generative process; $q$ describes a variational approximation of the inference process.

There are two principle approaches to disentanglement. Variational approaches [24, 16] utilize a stochastic representations that is regularized towards a prior distribution with the Kullback-Leibler (KL) divergence. This regularization penalizes information in the representation and—for large regularization weights—encourages disentanglement [20]. However, both entangled and disentangled content in the representation is penalized by the same amount. Disentanglement therefore comes at the price of uninformative representations: Reconstructions are blurry and the representations cannot explain the complete image.

Adversarial approaches [10, 15, 30] to disentanglement have the potential to provide a more fine-grained regularization. In these approaches a discriminator estimates entanglement and its gradients are used to guide the representations directly towards disentanglement. Therefore, they come with the promise to selectively penalize nothing but entangled content. However, we identify as the **key problem** that the encoder, having access to the discriminator's gradients, learns to produce entangled representations which are classified as disentangled. In contrast to adversarial attacks on image classifiers [53], in our case, the attack happens at the level of representations instead of images and implies that one cannot rely on adversarial approaches directly.

**Our contribution** is an approach for making adversarial approaches robust to overpowering without being affected by uninformative representations. We use a second classifier—whose gradients are never provided to the encoder—to detect overpowering: A large difference in both entanglement estimates implies that the first classifier




| | | | | | |
|---|---|---|---|---|---|
| $I_T$ | 0.1255 | 0.1254 | 0.0919 | 0.1258 | 0.1225 |
| $I_{T'}$ | 0.1239 | 0.8727 | 0.1247 | 0.4096 | 0.1201 |
| $\mathcal{L}_{\text{rec}}$ | 3.9350 | 2.9818 | 3.9615 | 3.2504 | **3.5279** |

Table 1. First row: Pose targets. Second row: Pose visualizations. First column: Appearance targets. Enforcing a MI constraint of $\epsilon = 0.125$ in eq. (5): a) leads to lossy representations. Pose is not accurately captured, leading to blurry synthesis results and high reconstruction loss. b) $T$ indicates successful disentanglement but $T'$ reveals high entanglement. Visualizations show that pose contains complete appearance information and the transfer task fails. c) Same problems as a) because disentanglement relies again on the variational approach. d) Improved compared to b) but still fails at the transfer task. e) Our adaptive combination achieves disentanglement and accurate pose representations. We obtain the lowest reconstruction error of all the methods which can enforce the disentanglement constraint (shaded green). See also Sec. 4.2.

is being tricked by the encoder. However, to achieve disentangled representations, we cannot directly utilize feedback from the second classifier, because this would reveal its gradients to the encoder and make it vulnerable to being overpowered, too. Instead, we use it indirectly to estimate the weight of a KL regularization term—we increase it when overpowering is detected and decrease it otherwise. This way, disentanglement comes from the first classifer, which is controlled by the second.

## 2. Disentanglement in probabilistic models

### 2.1. Latent variable models

Let $x$ denote an image from an unknown data distribution $p_{\mathcal{X}}(x)$. Probabilistic approaches to image synthesis approximate the unknown distribution using a model distribution $p(x)$. To fit this distribution to the data distribution, the maximum log-likelihood of data samples is maximized:

$$\max_p \mathbb{E}_{x\sim p_{\mathcal{X}}(x)} \log p(x) \tag{1}$$

The distribution $p$ can then be sampled to synthesize new images. To model $p$, latent variable models assume that images are generated due to an underlying latent variable $z$ which is not observed. The full model distribution

$$p(x, z) = p(x|z)p(z) \tag{2}$$

is then specified in terms of the factors $p(x|z)$, which are typically parameterized by a function class such as neural networks, and the factor $p(z)$ which specifies a prior on the latent variable, typically given by a simple distribution like a normal distribution.

A Generative Adversarial Network (GAN) [14] learns such a model using density ratio estimation. The training algorithm can be described as a two player game: A classifier tries to distinguish between generated and real images; a generator tries to generate images that are indistinguishable from real images. Because no inference process is learned, they cannot explain existing images.

The Variational Autoencoder (VAE) [26, 45] learns a latent variable model using variational inference and the reparameterization trick. The structure of the joint $p(x, z)$ and the corresponding encoder distribution $q(z|x)$ is shown in Fig. 3a. Variational inference involves a KL regularization of $q(z|x)$ towards the prior $p(z)$ and VAEs choose $q$ such that it can be computed efficiently, e.g. Gaussian parameterizations [26, 45] or normalizing flows [25]. To increase the flexibility of encoder distributions, [41] uses an adversarial approach that requires only samples to compute the KL regularization. Similar to GANs, it involves a two player game. This time, a classifier has to distinguish between latent codes sampled from the prior and those sampled from the encoder distribution, and the second player is the encoder.

However, images are the product of two independent factors, pose $\pi$ and appearance $\alpha$. There are both variational [20] and adversarial approaches [21] that try to discover

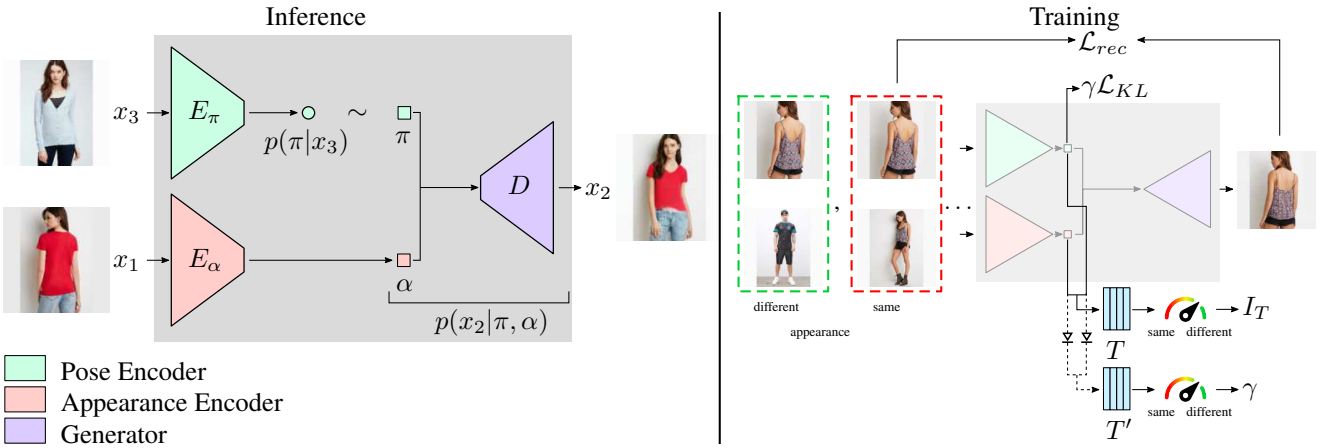



Figure 4. Left: synthesizing a new image $x_2$ that exhibits the apperance of $x_1$ and pose of $x_3$. Right: training our generative model using pairs of images with same/different appearance. $T$ and $T'$ estimate the mutual information between $\pi$ and $\alpha$. The gradients of $T$ are used to guide disentanglement, while $T'$ detects overpowering of $T$ and estimates $\gamma$ to counteract it.

such disentangled factors without any additional source of information. But simply assuming that $\pi$ and $\alpha$ are different components of $z$, i.e. $z = (\pi, \alpha)$ is problematic because the prior $p(z)$ cannot model the individual contribution of pose and appearance without inductive biases [36, 51]. The resulting models fall short in comparison to approaches that can leverage additional information [16, 27]. Thus, to learn a model in which the generative process is described by $p(x|\pi, \alpha)$ we need additional information.

## 2.2. Pose supervised disentangling

The common assumption of many recent works on disentangled image generation, e.g., [38, 39, 12, 2, 50, 11], is the observability of $\pi$, which is derived from a pretrained model for keypoint detection. While this representation of $\pi$ works quite well, it is limited to domains where robust keypoint detectors are available and sidesteps the learning task of disentangling the two latent factors $\pi, \alpha$. Instead, let us assume for a moment that we can observe samples of image triplets $(x_1, x_2, x_3)$ with the constraints that *(i)* $x_1$, $x_2$ have the same appearance, *(ii)* $x_2$, $x_3$ share the same pose, and *(iii)* $x_1$, $x_3$ have neither pose nor shape in common. Let $p_{\mathbb{T}}(x_1, x_2, x_3)$ denote this unknown joint distribution. We model each of the three images as being generated by a process of the form $p(x|\pi, \alpha)$ (which is assumed to be the same for all three images). Because $x_1, x_2$ share appearance and $x_2, x_3$ share pose, only four instead of six latent variables are required to explain how these triplets are generated. Let $\pi, \alpha$ denote the shared pose and appearance explaining $x_2$ and let $\pi', \alpha'$ denote additional realizations of pose and appearance explaining $x_1$ and $x_3$, respectively. The marginal distributions underlying $\pi$ and $\pi'$ are assumed to be the same and so are the marginal distributions of $\alpha$ and $\alpha'$. If, as assumed in [43, 29], we could observe the complete triplet $(x_1, x_2, x_3)$, a simple inference mechanism would infer $\alpha$ from $x_1$ and $\pi$ from $x_3$ as depicted in Fig. 3b. Unfortunately, the assumption that $x_2, x_3$ share the same pose but not appearance is essentially equivalent to the assumption that a keypoint estimator is implicitly available. Then pairs $x_2, x_3$ could be found by comparison of keypoints. Without this information we have to further reduce assumptions on the data and essentially train without $x_3$.

## 2.3. Disentangling without pose-annotations

Without access to $x_3$, we must rely on $x_2$ to infer $\pi$ as shown in Fig. 3c. The maximum likelihood objective for $p(x_2|\pi, \alpha)$ leads to a reconstruction loss, and without constraints on $\pi$ it encourages a degenerate solution where all information about $x_2$ is encoded in $\pi$ [40, 16], i.e. $\pi$ also encodes information about $\alpha$ instead of being independent of it.

[24] assumes the availability of labels for $\alpha$ and uses a conditional variant of the VAE, which results in a KL regularization of $q(\pi|x_2)$ towards a prior and thus a constraint on $\pi$. To improve image generations with swapped $\alpha$ and $\pi$, [40] adds an adversarial constraint on generated images. It encourages the preservation of characteristics of $\alpha$, i.e. it combines the conditional VAE with a conditional GAN similar to [3]. [51] also utilize this GAN constraint but they only require pairs $x_1, x_2$ instead of labels for $\alpha$, and instead of using the KL term to constrain $\pi$, they severely reduce its dimensionality. As pointed out by [16], these GAN constraints only encourage the decoder to ignore information about $\alpha$ in $\pi$ instead of disentangling $\alpha$ and $\pi$. [16] proposes a cycle-consistent VAE which adds a cyclic loss to the VAE objective. [37] directly models $\pi$ as keypoints. All of these methods rely on the same basic principle for disentanglement: Constraining the amount of information in $\pi$. Indeed, the VAE objective implements a variational approximation of the information bottleneck [1]. In con-

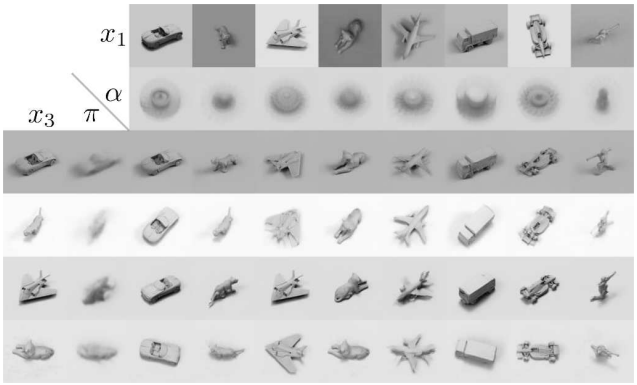



Figure 5. Generated images using examples from the norb dataset [31]. First row: images used as appearance target, second: visualizing the inferred appearance $\alpha$ marginalized over all poses. First column: images used as pose target, second: pose marginalized over all appearances. Remaining entries: decodings for different combinations of pose and appearance.

trast, we utilize this variational information bottleneck only to counteract overpowering, an issue that affects the following methods.

Similar to [41] for variational inference, [10] utilizes an adversarial approach for disentanglement: A classifier has to predict if a pair $(\pi, \pi')$ was inferred from two images of the same video sequence or different video sequences. [15] assumes that $\alpha$ is given in the form of class labels, and [30, 18] are specialized to images of faces and assume that $\alpha$ is given in the form of facial attributes but they utilize the same principle: A classifier has to predict $\alpha$ from $\pi$. Besides differences in the precise objectives used for the classifiers, all of these methods implement again an information bottleneck as in [4]. Compared to variational approximations of the bottleneck, they have the advantage that only information about $\alpha$ in $\pi$ is penalized. However, applications of adversarial approaches have been limited to synthetic datasets or facial datasets with little to no pose variations. We show that overpowering prevents their direct application to real world datasets and show how to turn them into robust methods for disentanglement. Our approach can recombine pose and appearance of any two images, while previous models for unsupervised image-to-image translation require seperate training for each appearance [23, 32] and cannot transfer to unseen appearances [8].

## 3. Approach

### 3.1. Constrained maximum-likelihood learning

We want to learn a probabilistic model of images that explains the observed image $x_2$ in terms of two disentangled representations $\pi, \alpha$. This requires a model for the decoder distribution conditioned on the two representations,

$$p(x_2|\pi, \alpha) \tag{3}$$


Figure 6. Generated images on the PKU Vehicle ID [33] dataset. First row: pose targets. First column: appearance targets.

and an encoder model $p(\pi, \alpha|x_1, x_2)$ to infer $\pi$ and $\alpha$ from the data. As shown in Fig. 3c, we estimate $\pi$ with an encoder network $E_\pi(x_2)$ from $x_2$ and $\alpha$ with an encoder network $E_\alpha(x_1)$ from $x_1$. A decoder network $D(\pi, \alpha)$ which takes $\pi$ and $\alpha$ as inputs reconstructs the image according to $p(x_2|\pi, \alpha)$.

Learning the weights of these networks depends on a constrained optimization problem. To ensure that $\pi$ and $\alpha$ describe the images well, we maximize the conditional likelihood as formulated in Eq. (4), which corresponds to a reconstruction loss. To avoid a trivial solution where $\pi$ encodes all of the information of $x_2$, we formulate the disentanglement constraint (5), such that our full optimization problem reads

$$\max_p \mathbb{E}_{x_1,x_2} \log p(x_2|\pi, \alpha) \tag{4}$$

$$\text{subject to } I(\pi, \alpha) \leq \epsilon \tag{5}$$

Here, $\epsilon$ is a small constant and $I(\pi, \alpha)$ denotes the mutual information [9] defined as

$$I(\pi, \alpha) = \mathrm{KL}(p(\pi, \alpha)|p(\pi)p(\alpha)). \tag{6}$$

Computing (6) is difficult [4] and to derive an algorithm for the solution of the optimization problem above, we must resort to approximations. Subsequently, we first derive two different estimates on the mutual information. The first one provides an upper bound, but, alas, it always overestimates it severely. A second estimate is then introduced which provides accurate estimates. However, to enforce the constraint in (5), we require gradients of the estimate and, as we will see, this enables the encoder to perform an adversarial attack on the estimate, such that it heavily underestimates the true mutual information. In Sec. 3.4, we show how to combine both estimates to obtain our method for robust maximum-likelihood learning under mutual information constraints. Thereafter, we describe the algorithm used to implement the method.

### 3.2. A variational upper bound on the mutual information

Ideally, we would like to obtain an upper bound on the MI in (6) to be able to enforce the constraint (5). Because we estimate $\pi$ from $x_2$ and $\alpha$ from $x_1$, we have the Markov-Chain $\pi \to x_2 \to \alpha$ with

$$p(\alpha, x_2, \pi) = p(\alpha|x_2)p(x_2|\pi)p(\pi) \tag{7}$$

which implies the data processing inequality [9]:

$$I(\pi, \alpha) \leq I(\pi, x_2). \tag{8}$$

The right hand side of this inequality can now be easily estimated with a variational marginal $r(\pi)$ [1]. Indeed, for any density $r$ with respect to $\pi$ we have the bound

$$I(\pi, \alpha) \leq \mathbb{E}_{x_2} \mathrm{KL}(p(\pi|x_2)|r(\pi)). \tag{9}$$

Modeling both $p(\pi|x_2)$ and $r(\pi)$ as Gaussian distributions, we can evaluate the right hand side analytically. Unfortunately, this bound is too loose for our purposes. The condition $\mathrm{KL}(p(\pi|x_2)|r(\pi)) = 0$ implies $I(\pi, x_2) = 0$ and thus $\pi$ would be completely uninformative.

### 3.3. Fine-grained estimation of mutual information

A different estimate of mutual information can be obtained with the help of density estimation [41, 4]. The KL-divergence of two densities is closely related to the associated classification problem: Let $T(\pi, \alpha)$ be a classifier that maps a pair $(\pi, \alpha)$ to a real number which represents the log probability that the pair is a sample from the joint distribution $p(\pi, \alpha)$. Denote by $\sigma(t) = (1 + e^{-t})^{-1}$ the sigmoid function. The maximum likelihood objective for this classification task reads

$$\max_T \mathbb{E}_{(\pi,\alpha)\sim p(\pi,\alpha)} \log \sigma(T(\pi, \alpha)) + \tag{10}$$

$$\mathbb{E}_{\pi\sim p(\pi),\alpha\sim p(\alpha)} \log(1 - \sigma(T(\pi, \alpha))). \tag{11}$$

The optimal solution $T^*$ of this problem satisfies

$$I(\pi, \alpha) = \mathbb{E}_{(\pi,\alpha)\sim p(\pi,\alpha)} T^*(\pi, \alpha). \tag{12}$$

When $T$ is implemented as a neural network, we obtain a differentiable estimate of $I(\pi, \alpha)$ which can be used to enforce the desired constraint during learning of $q(\pi|x_2)$. For a given classifier $T$ we write $I_T = I_T(\pi, \alpha) = \mathbb{E}_{\pi,\alpha\sim p(\pi,\alpha)} T(\pi, \alpha)$ for the resulting estimate.

### 3.4. Robust combination of variational and adversarial estimation

If we replace the constraint $I(\pi, \alpha) \leq \epsilon$ in (5) with a constraint on the estimate $I_T(\pi, \alpha) \leq \epsilon$, we observe a new type of adversarial attack: The encoder is able to overpower the classifier $T$: it can learn a distribution $p(\pi|\alpha)$ such that $T$ cannot differentiate pairs $(\pi, \alpha)$ sampled from the joint from those sampled from the marginals. However, a seperately trained classifier $T'$, whose gradients are not provided to the encoder, can still classify them (see Fig. 1). In other words, in an adversarial setting we consistently observe the situation $I_T(\pi, \alpha) \ll I(\pi, \alpha)$, i.e., we underestimate the mutual information between $\pi$ and $\alpha$. To obtain a guaranteed upper bound on the mutual information, we must utilize the variational upper bound. As we have seen before, we must be careful to enforce not too strict bounds on it. Thus, we formulate our new objective as

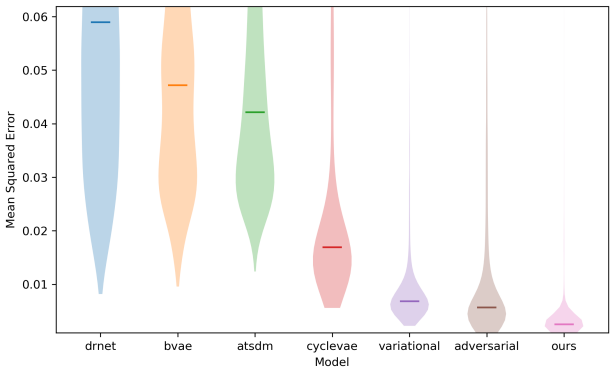



Figure 7. Distribution of average reconstruction error on the sprites dataset [43]. We evaluate against the provided ground truth.

$$\max_p \mathbb{E}_{x_1,x_2} \log p(x_2|\pi, \alpha) \tag{13}$$

$$\text{subject to } I_T(\pi, \alpha) \leq \epsilon \tag{14}$$

$$\mathrm{KL}(p(\pi|x_2)|r(\pi)) \leq C, \tag{15}$$

where $C$ has to be adaptively estimated based on the detection of adversarial attacks of the encoder against $T$. The main idea is to compare the classification performance of $T$ against an independently trained classifier. If there is a large performance gap, we cannot rely on the estimate of $T$ (it has been overpowered) and must decrease $C$. The next section describes the approach.

### 3.5. Robust disentanglement despite encoder overpowering

To obtain a training signal for our networks, we must transform problem (13) into an unconstrained problem which can be optimized by gradient ascent. Let us first consider the constraint (15) on the KL term. For a given $C$, there exists a Lagrange multiplier $\gamma \geq 0$ such that the problem can be written equivalently as

$$\max_p \mathbb{E}_{x_1,x_2} \log p(x_2|\pi, \alpha) - \gamma \mathrm{KL}(p(\pi|x_2)|r(\pi)) \tag{16}$$

$$\text{subject to } I_T(\pi, \alpha) \leq \epsilon. \tag{17}$$


Figure 8. Generated images on the BBC Pose dataset [6] dataset. First row: pose target. First column: appearance target. An animated version can be found in the supplementary.

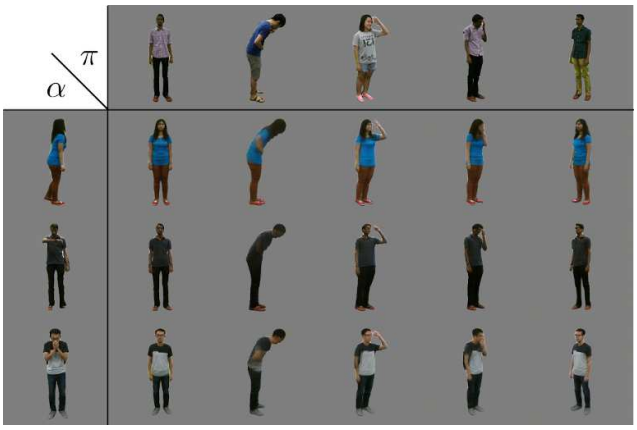



Figure 9. Generated images on the NTU dataset [47]. First row: pose target. First column: appearance target. An animated version can be found in the supplementary.

Thus, we can directly estimate $\gamma$ instead of $C$. Ideally, $\gamma$ should be very small and only active in situations were $I_T(\pi, \alpha)$ underestimates the mutual information. To achieve this, we train a second classifier $T'$ based on the same objective (10). It is crucial that its estimate $I_{T'}(\pi, \alpha)$ is never directly provided as a signal to the encoder. We merely compare the estimates of $T$ and $T'$ and if $I_T \ll I_{T'}$, we increase $\gamma$. Hence, we update $\gamma$ in each optimization step based on the proportional gain $I_{T'} - I_T$ and bias it towards zero with a small constant $b_\gamma$

$$\gamma_{t+1} = \max\{0, \gamma_t + l_\gamma(I_{T'} - I_T - b_\gamma)\}, \tag{18}$$

where $l_\gamma$ can be considered the learning rate of $\gamma$.

For the remaining constraint (17) on $I_T(\pi, \alpha)$, we utilize an Augmented Lagrangian Approach [56]. After switching from maximization to minimization, the complete unconstrained loss function $\mathcal{L}$ for training the network is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{VB}} + \mathcal{L}_{\text{MI}}, \tag{19}$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction loss given by the negative likelihood

$$\mathcal{L}_{\text{rec}} = -\mathbb{E}_{x_1,x_2} \log p(x_2|\pi, \alpha), \tag{20}$$

$\mathcal{L}_{\text{VB}}$ the penalty associated with the variational upper bound

$$\mathcal{L}_{\text{VB}} = \gamma \,\text{KL}(p(\pi|x_2)|r(\pi)), \tag{21}$$

and $\mathcal{L}_{\text{MI}}$ the loss used to enforce the the constraint (17) based on an estimated Lagrange multiplier $\lambda \geq 0$ and a penalty parameter $\mu > 0$

$$\mathcal{L}_{\text{MI}} = \begin{cases} \lambda(I_T - \epsilon) + \frac{\mu}{2}(I_T - \epsilon)^2 & \text{if } I_T - \epsilon \geq -\frac{\lambda}{\mu} \\ -\frac{\lambda^2}{2\mu} & \text{else} \end{cases}. \tag{22}$$

The update rule for $\lambda$ is

$$\lambda_{t+1} = \max\{0, \lambda_t + \mu(I_T - \epsilon)\}. \tag{23}$$

Fig. 4 outlines our network architecture during training and inference. We perform the optimization over mini-batches and alternate between the training of the classifiers $T$ and $T'$ (according to the objective defined in (10)), and the training of the generative model. The loss for the networks $D$ and $E_\alpha$ is given by $\mathcal{L}_{\text{rec}}$ and $E_\pi$ is optimized with respect to the full loss $\mathcal{L}$. After each step, $\gamma$ and $\lambda$ are updated according to (18) and (23), respectively.

## 4. Experiments

### 4.1. Comparison to state-of-the-art

In Fig. 7, we compare our method to [16] (cyclevae), the state-of-the-art among the variational approaches, and to [15] (atsdm), the state-of-the-art among the adversarial approaches. In addition to our full model (ours), we also include [10] (drnet), a version of our model that utilizes the objective of [20] (bvae), a version without $\mathcal{L}_{\text{VB}}$ (adversarial) and a version without $\mathcal{L}_{\text{MI}}$ (variational), where the update of $\gamma$ from (18) is replaced by

$$\gamma_{t+1} = \max\{0, \gamma_t + l_\gamma(I_{T'} - \epsilon)\}, \tag{24}$$

to estimate the required $\gamma$ to achieve the MI constraint (5).

Because it is difficult to obtain ground truth for triplets $(x_1, x_2, x_3)$ on real data, we resort to the synthetic sprites dataset [43] to compare the methods. It contains 672 different video game characters, each depicted in a wide variety of poses. For training, we only utilize pairs of images $(x_1, x_2)$ belonging to the same character. To measure the performance of the different approaches, we calculate the mean squared error between images $\tilde{x}_2$ generated from inputs $x_1$ and $x_3$, and the corresponding ground truth $x_2$. We

| | | Model | Market-1501 [58] | | DeepFashion [34, 35] | |
|---|---|---|---|---|---|---|
| | | | Reconstruction | Transfer | Reconstruction | Transfer |
| Appearance | | **reID** mAP [%] - *bigger is better* | | | | |
| | sup. | VUNet [12] | 25.3 | 19.9 | 21.3 | 14.6 |
| | unsup. | adversarial | 34.8 | 6.6 | 37.4 | 10.5 |
| | | ours | 30.2 | 25.4 | 52.9 | 47.1 |
| Pose | | **rePose** [% of image width] - *smaller is better* | | | | |
| | sup. | VUNet [12] | $17.9 \pm 9.4$ | $17.5 \pm 9.4$ | $1.5 \pm 3.1$ | $1.9 \pm 4.0$ |
| | unsup. | adversarial | $24.6 \pm 10.3$ | $25.3 \pm 10.3$ | $6.2 \pm 7.3$ | $7.4 \pm 8.1$ |
| | | ours | $23.9 \pm 9.9$ | $24.7 \pm 9.9$ | $5.5 \pm 6.8$ | $6.8 \pm 7.6$ |

Table 2. Evaluating how well the generated image preserves *(i)* appearance or *(ii)* pose. For *(i)* we compare input $x_1$ and output $x_2$ of our approach using a standard encoder for person reidentification [19] using retrieval performance (mAP). Conservation of pose *(ii)* is measured by comparing the results of keypoint detector [5] of $x_3$ and $x_2$. Note that [12] uses keypoint annotations.

randomly select 8000 triplets $(x_1, x_2, x_3)$ from the test set and report the error distribution in Fig. 7.

### 4.2. Visualization of encodings

To better understand the information encoded by $\pi$ and $\alpha$, we visualize these representations. Because $\pi$ does not contain information about the appearance, this corresponds to a marginalization of images depicting a given pose over all appearances. Similarly, $\alpha$ yields a marginalization for a given appearance over all poses. We show examples of these visualizations in Fig. 1, Tab. 1, and Fig. 5. This synthesis is performed independently from the training of our model with the sole purpose of interpretability and visualization. For this, a decoder network is trained to reconstruct images from only one of the factors. In Tab. 1, these visualizations demonstrate that $I_{T'}$ estimates entanglement correctly. In that figure, a) is our model without $\mathcal{L}_{\text{MI}}$ and the update of $\gamma$ replaced by (24). b) is our model without $\mathcal{L}_{\text{VB}}$, c) with $b_\gamma = 0$, d) with $\gamma = 1$ fixed and e) is our full model.

### 4.3. Shared representations across object categories

The previous dataset contained a single category of objects, namely video game characters. In this setting, a common pose representation is relatively easily defined in terms of a skeleton. It is considerably more difficult to find a representation of pose that works across different object categories which do not share a common shape. To evaluate our model in this situation, we utilize the NORB dataset [31], which contains images of 50 toys belonging to 5 different object categories. Each instance is depicted under a wide variety of camera views and lighting conditions. For this experiment, we consider camera views and lighting conditions to be represented by $\pi$. In Fig. 5 we can see that our model successfully finds two representations that can be combined *across* different object categories. Note that our model was never trained on a pair of images depicting instances of different categories.

### 4.4. Evaluation on Human Datasets

In Tab. 2, we evaluate our approach on natural images of people, which have been the subject of recent models for disentangled image generation [12]. Besides qualitative evaluations, we employ two quantitative measures to validate how much of the pose and appearance are being preserved in the generated output: (i) Since ground-truth triplets are not available for these datasets, we require a metric that captures similarity in appearance while being invariant to changes in pose. Such a measure can be obtained from a person reidentification model [19], which can identify the same person despite differences in pose. Using the evaluation protocol of [19] we report the mean average precision (mAP) of re-identifying generated images under "reID mAP" in Tab. 2. (ii) To measure how well our approach retains pose we employ Openpose [5] to obtain keypoint estimates. We extract keypoints from the pose input image $x_3$ and the output $x_2$ and compute the euclidean distance between the estimated keypoints in both images. As above, we include an ablation (*adversarial*) without $\mathcal{L}_{\text{VB}}$.

## 5. Conclusion

We have shown how an additional classifier, whose gradients are not used directly to train the encoder, prevents encoder overpowering. This enables robust learning of disentangled representations of pose and appearance without requiring a prior on pose configurations, pose annotations or keypoint detectors. Our approach can be readily applied on a wide variety of real-world datasets.

This work has been funded by the German Research Foundation (DFG) - 371923335; 421703927 and a hardware donation from NVIDIA.

## References


[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 4, 6

[2] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018. 4

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2745–2754, 2017. 4

[4] Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R Devon Hjelm, and Aaron Courville. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018. 5, 6

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 8

[6] James Charles, Tomas Pfister, Mark Everingham, and Andrew Zisserman. Automatic and efficient human pose estimation for sign language videos. *International Journal of Computer Vision*, 110(1):70–90, 2014. 7

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016. 2

[8] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 5

[9] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012. 5, 6

[10] Emily L Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4414–4423. Curran Associates, Inc., 2017. 2, 5, 7

[11] Patrick Esser, Johannes Haux, Timo Milbich, et al. Towards learning a realistic rendering of human behavior. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 4

[12] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 4, 8

[13] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. *arXiv preprint arXiv:1809.07845*, 2018. 2

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 3

[15] Naama Hadad. A two-step disentanglement method. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 772–780, 2018. 2, 5, 7

[16] Ananya Harsh Jha, Saket Anand, Maneesh Singh, and VSR Veeravasarapu. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 805–820, 2018. 2, 4, 7

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[18] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 2019. 2, 5

[19] Alexander Hermans*, Lucas Beyer*, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, 2017. 8

[20] 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. *ICLR*, 2(5):6, 2017. 2, 3, 7

[21] Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation by mixing them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3399–3407, 2018. 3

[22] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *arXiv preprint arXiv:1810.11981*, 2018. 2

[23] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018. 5

[24] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014. 2, 4

[25] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016. 3

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 3

[27] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Björn Ommer. Content and style disentanglement for artistic style transfer. In *Proceedings of the Intl. Conf. on Computer Vision (ICCV)*, 2019. 4

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q.

Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 1

[29] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, pages 2539–2547, 2015. 4

[30] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic DENOYER, and Marc Aurelio Ranzato. Fader networks:manipulating images by sliding attributes. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5967–5976. Curran Associates, Inc., 2017. 2, 5

[31] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–104. IEEE, 2004. 5, 8

[32] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018. 5

[33] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2167–2175, 2016. 5

[34] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 8

[35] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *European Conference on Computer Vision (ECCV)*, October 2016. 8

[36] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019. 4

[37] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Björn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[38] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 405–415, 2017. 4

[39] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018. 4

[40] 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 *Advances in Neural Information Processing Systems*, pages 5040–5048, 2016. 2, 4

[41] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2391–2400. JMLR. org, 2017. 3, 5, 6

[42] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization. 1

[43] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1252–1260. Curran Associates, Inc., 2015. 4, 6, 7

[44] 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. 1

[45] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32*, pages II–1278. JMLR. org, 2014. 1, 3

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[47] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 7

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[49] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv e-prints*, Sept. 2014. 1

[50] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2018. 4

[51] Attila Szabó, Qiyang Hu, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Challenges in disentangling independent factors of variation. *arXiv preprint arXiv:1711.02245*, 2017. 2, 4

[52] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[53] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus.

Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[54] Paul Upchurch, Jacob Gardner, Kavita Bala, Robert Pless, Noah Snavely, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 6090–6099, 2016. 2

[55] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4790–4798. Curran Associates, Inc., 2016. 1

[56] Stephen Wright and Jorge Nocedal. Numerical optimization. *Springer Science*, 35(67-68):7, 1999. 7

[57] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1

[58] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015. 8