Identity from here, *Pose* from there: Self-supervised Disentanglement and Generation of Objects using Unlabeled Videos

Fanyi Xiao, Haotian Liu, and Yong Jae Lee
University of California, Davis

Abstract

We propose a novel approach that disentangles the identity and pose of objects for image generation. Our model takes as input an ID image and a pose image, and generates an output image with the identity of the ID image and the pose of the pose image. Unlike most previous unsupervised work which rely on cyclic constraints, which can often be brittle, we instead propose to learn this in a self-supervised way. Specifically, we leverage unlabeled videos to automatically construct pseudo ground-truth targets to directly supervise our model. To enforce disentanglement, we propose a novel disentanglement loss, and to improve realism, we propose a pixel-verification loss in which the generated image's pixels must trace back to the ID input. We conduct extensive experiments on both synthetic and real images to demonstrate improved realism, diversity, and ID/pose disentanglement compared to existing methods.

1. Introduction

Consider the NYC street scene shown in Fig. 1 (left). As a human, it is not difficult to imagine what a red sedan would look like in place of the yellow taxis. This is likely because we have been exposed to thousands of different cars in various poses in our lifetime, and have learned how to *disentangle* a car's identity from its pose. In this paper, we propose to learn a model to perform this task – specifically, synthesizing a novel pose of an object instance conditioned on the pose of a different reference object (see Fig. 1, right), without any labels.

This task requires the model to disentangle the object's identity and pose. For example, in Fig. 1, in order to encode the ID information of the red sedan, the model needs to capture the appearance and shape that is unique to that specific car instance, *independent* of pose. Meanwhile, the model also needs to encode the pose information specified by the taxis (the pose reference image), *independent* of identity. It can then combine the identity of the red sedan with the pose of the taxis to create a new image with the desired pose.

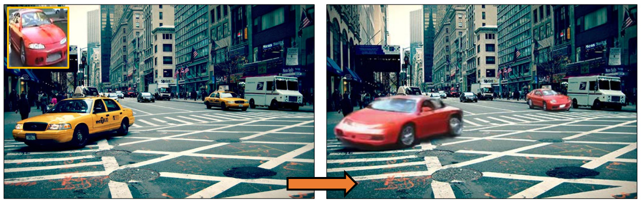

Figure 1. Our self-supervised model learns to disentangle identity (red sedan) and pose (taxis) of objects for image generation.

Disentangled representations, in which e.g., one latent subspace controls one factor of variation, can provide robustness to complex variations in the data and be useful for downstream visual recognition tasks [5]. There has been a long line of research on learning disentangled representations for images [51, 43, 55, 10, 46, 15, 20, 21, 21]. Early works like Tenenbaum and Freeman [51] operate in a fully-supervised setting in which the factors of interest (e.g., content and style) are annotated for each training image. We instead aim to solve this task with a *self-supervised* approach, without using any pose or identity annotations. Self-supervised disentanglement of identity and pose is an extremely challenging problem, since the two factors are highly intertwined. For example, shape constitutes an important part of an object's identity – to distinguish a side-view van from a side-view sedan, we need to analyze their specific shape differences. On the other hand, the difference between pose and shape is often subtle and interdependent – as the pose of the car changes, so does its perceived shape.

To tackle this, recent image generation methods either introduce cyclic constraints [35, 20, 15, 21, 25] (similar in spirit to cycleGAN [60]) or inject priors on the representation based on domain knowledge [46, 27]. Though promising, these methods typically only work well when there is no large pose change in the objects. The reason is quite intuitive: due to the lack of direct supervision (i.e., ground-truth target images), the supervisory signals provided by either the proposed constraints or the prior on the representation are often insufficient to induce disentangled representations.

We take a different approach. We utilize *unlabeled videos* to automatically construct *training triplets*, each consisting of an identity reference image, a pose reference image, and a pseudo ground-truth target, to train our model. The requirement for the pseudo ground-truth target is that it should consist of an object that has the identity of the identity reference and the pose of the pose reference. We exploit the fact that frames in a short video clip are likely to contain instances of the same object, to sample the identity reference and target image. We then find a nearest neighbor of the target image in pose space to construct the pose reference image. Though an approximation to the ground-truth, directly feeding input/output pairs to our model provides a much stronger supervisory signal than only enforcing cyclic constraints, and enables it to achieve the desired disentanglement. To supplement the direct supervision and further encourage disentanglement and realism, we propose to optimize two novel loss functions – disentanglement loss and pixel verification loss. For the disentanglement loss, we construct two explicit constraints that force the identity encoder to only capture identity information and the pose encoder to only capture pose information – the same identity feature is used to generate two different poses of the same object, while two different objects with the same pose are produced from the same pose feature. The pixel verification loss promotes realism by exploiting the fact that, a pixel in the generated image should, in most cases, be able to trace back to its root in the identity image.

Our model is a novel conditional adversarial learning framework based on Generative Adversarial Networks (GANs) [17], trained with the aforementioned loss functions to disentangle identity and pose. We conduct extensive experiments on both synthetic (3D Cars/Chairs [13, 2]) and challenging real images (YouTube-BoundingBoxes [42]) to demonstrate better realism, diversity, and ID/pose disentanglement, compared to existing unsupervised approaches.

2. Related Work

Disentangled representations Unsupervised methods for disentangling factors of variation typically employ cyclic constraints [35, 10, 20, 25, 31, 12, 23, 21, 47, 33]. A limitation with cyclic constraints is that, though necessary (they would be satisfied with perfect disentanglement), they are often insufficient for generating high-quality disentangled representations. We instead propose to employ a simple yet effective procedure to retrieve direct pseudo targets during training, to enforce a much stronger constraint. Some place disentanglement in the context of cross-domain translation [15, 21, 32], which requires a clear definition of *domains*. For example, to disentangle the identity and pose of cars, one would need to define the pose as content (according to the definition in [21]) and define one domain for each car identity, which would require one encoder-decoder pair for each identity. In contrast, our work only requires one encoder-decoder pair, and is thus much more scalable.

Others learn disentangled representations by enforcing explicit priors (e.g., a canonical appearance and a deformation field) [46] or focus on specific domains like faces/humans [41, 4, 52, 3, 34, 40]. In contrast, we avoid making strong domain-specific assumptions, and grant our model more freedom to learn directly from data. Reed et al. [44] learn a disentangled representation via a visual-analogy task, whereby a query image is transformed analogously to an example pair of reference images. Unlike visual-analogy, which takes three input images, our task only requires two (ID/pose references). Meanwhile, DDPAE [19] tackles the disentanglement problem with the motivation of simplifying future frame prediction (i.e., it is easier to predict changes based on disentangled factors). Finally, others induce disentanglement by injecting priors (e.g., maximize/minimize *factorability*, *total correlation*, *description length*, etc.) on the latent code in a variational auto-encoding framework [27, 7, 1]. However, they do not have explicit control over the semantics of the learned representation (e.g., the model does not know which dimensions correspond to "identity") whereas our approach has explicit identity and pose representations.

Novel view synthesis from a single RGB image is a highly under-determined problem that requires 3D understanding of objects. Some disentanglement work adopt novel view synthesis as their application [10, 25, 4, 52, 44]. Others tackle this problem with the help of a large stock of 3D shape models [26, 45, 56, 61], and sometimes with a large amount of human involvement [26]. [48] performs view synthesis in HOG space rather than RGB space. More recent works train CNNs to function like a graphics rendering engine [29, 55, 11] or learn appearance flow to synthesize novel views [59]. Unlike these approaches, our method does not require any 3D shape models, human intervention, or ground-truth training examples.

Conditional image-to-image translation The most successful image-to-image translation algorithms are based on Generative Adversarial Networks (GANs) [17]. Examples that learn in a supervised setting—with annotated input/output pairs—include Pix2Pix [22], Pix2PixHD [53], and GauGAN [39]. Unsupervised approaches leverage cycle-consistency [60], learn a shared latent space between domains [21, 8], or impose constraints to disentangle factors [25]. Our work leverages a large collection of *unlabeled videos* to automatically construct pseudo ground-truth targets. In this way, we can exploit the advantages of the supervised setting, without having to annotate any images.

3. Approach

Our goal is to learn a model that takes as input two images and generate a new image with one's *identity* and another's *pose*. Importantly, we do not have any identity nor pose annotations during both training and testing.

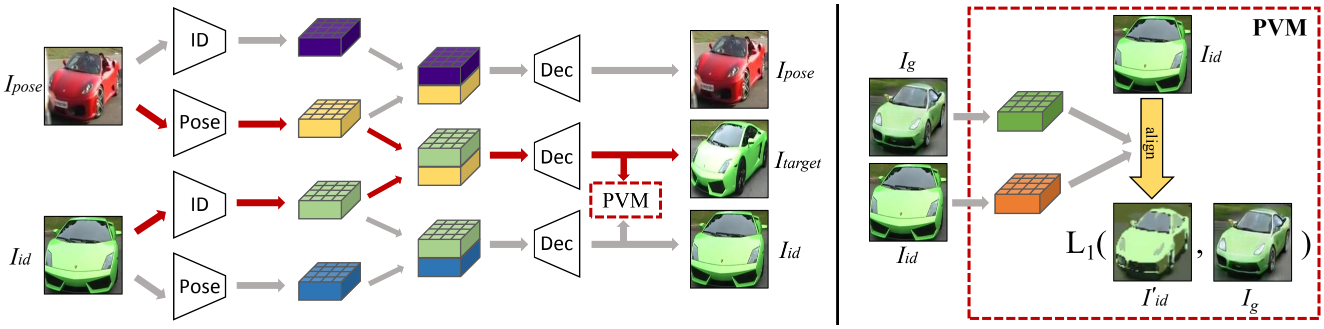

Figure 2. An illustration of the generator. Our generator takes as input both the identity reference image I_{id} and the pose reference image I_{pose}, and tries to generate an output image that matches I_{target}, which has the same identity as I_{id} but with the pose of I_{pose}. Notice how the pose encoded feature (yellow block) is used to generate both I_{target} and I_{pose}, so it cannot contain any identity information. Likewise, the identity encoded feature (green block) is used to generate both I_{target} and I_{id}, so it cannot contain any pose information. Furthermore, we propose a novel pixel verification module (PVM, details shown on the right) which computes a verifiability score between I_g and I_{id}, indicating the extent to which pixels in I_g can be traced back to I_{id}.

3.1. Network architecture

Generator To disentangle identity and pose, we use a two-branch generator network that processes the two streams of inputs separately. As shown in Fig. 2 (left, red arrows), the ID/Pose encoder processes the ID/pose reference image into a feature map that *exclusively* captures identity/pose information. The concatenated ID and pose feature maps (along the channel dimension) are fed into the decoder. Overall, our generator can be expressed as:

$$I_g = G(E_i(I_{id}), E_p(I_{pose})),$$

where I_{id} and I_{pose} denote the ID/pose reference image respectively. E_i and E_p are the ID and pose encoders and G is the decoder. The ID encoder consists of consecutive `Conv - ReLU` blocks whereas the Pose encoder consists of consecutive `Conv - Norm - ReLU` blocks. We add instance normalization (following [53]) to the Pose encoder to remove instance-specific feature means and variances which are correlated with object identity [21]. For the decoder, we follow the architecture used in [53] (from residual blocks and onwards), except we replace transposed convolutions with `Upsample - Conv` to mitigate checkerboard artifacts [36].

Discriminator For the output to preserve both *realism* and *identity*, we set up two discriminators. The first is the Real/Fake discriminator D_{real}, which takes in as input a single RGB image and classifies it as real or fake. It pushes the generated image I_g to look as real as possible, in order to fool the discriminator. The second discriminator D_{pair} focuses on preserving the object's identity in the generation and is trained to classify whether an input pair shares the same identity or not. The generator is thus trained to match the identity of the generated image to that of the input ID image. Following [53], we adopt a 2-scale discriminator, which enforces realism both locally (e.g., specific object details) and globally (e.g., overall shape).

3.2. Constructing ID-pose-target training triplets

The key difference between our work and previous unsupervised disentanglement works (e.g., [10, 20, 25, 21, 31, 15]) is that rather than relying only on indirect cyclic constraints, we instead construct a *pseudo ground-truth* target image I_{target} using *unlabeled videos* so that we can directly train the model in a supervised way, but without any labels. We demonstrate that this provides stronger supervision than cyclic constraints.

We first sample two images from the same video clip as I_{id} and I_{target}. The assumption is that these images will contain the same object instance, which is generally true for short clips (for long videos, unsupervised tracking could also be applied). We then retrieve a nearest neighbor of I_{target} from other videos (so that it's unlikely to have the same identity) using a pre-trained convnet, to serve as the pose reference image I_{pose}. Fig. 3 illustrates this process. The key insight is that retrieving objects with the same pose is much easier than retrieving objects with the same identity – objects with the same pose share a large amount of edges, which can be well-captured with an off-the-shelf feature extractor. Specifically, we use the `conv4` feature map of an AlexNet trained in a self-supervised way on ImageNet to avoid using any image labels [14]; see Fig. 4. Although the retrieved I_{target} is an approximation to the real ground-truth, we show that it is highly effective in our experiments. Finally, to ensure diversity of the sampled pairs' (I_{id}, I_{pose}) poses, we cluster all images into M different poses, and then sample a balanced number of unique pose pairs.

3.3. Loss functions

To generate images that are realistic and identity/pose-preserving, we use the following loss functions.

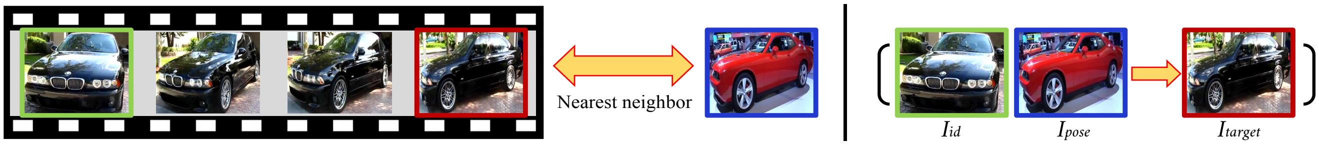

Figure 3. Constructing ID, pose, and target training triplets. With this procedure, we automatically obtain supervision to train our model.

Disentanglement loss To directly supervise our model with the pseudo ground-truth target, we minimize the L_1 difference between our model's generation and the target:

$$\mathcal{L}_{dis}^{1} = ||I_{target} - G(E_i(I_{id}), E_p(I_{pose}))||_1.$$

However, since there are many possible solutions for minimizing this loss, it alone will not necessarily enforce the desired disentanglement. To ensure that the ID/Pose encoder only encodes information about identity/pose, in addition to generating I_{target}, we also ask our model to reconstruct I_{id} and I_{pose}:

$$\begin{aligned}\mathcal{L}_{dis}^{2} =& ||I_{id} - G(E_i(I_{id}), E_p(I_{id}))||_1 \\ &+ ||I_{pose} - G(E_i(I_{pose}), E_p(I_{pose}))||_1.\end{aligned}$$

As shown in Fig. 2, this will force the ID encoder to not capture any pose information since its output is used to generate two targets with distinct poses (I_{id} and I_{target}); the same logic applies to the Pose encoder. Our final disentanglement loss is:

$$\mathcal{L}_{dis} = \mathcal{L}_{dis}^{1} + \mathcal{L}_{dis}^{2}.$$

We adopt the perceptual loss [24] as it captures the distance at a semantic level.

Pixel verification loss Recall that our final generated image should preserve the identity of the ID reference. This implies that *for (almost) every pixel in our generation, we should be able to trace it back to the ID image*. For example, for a car's front light pixel in our generation, we should be able to find the same front light pixel in the ID image, if our generation correctly preserves its identity. This will only be false when there are unobserved parts in the ID image that need to be generated. However, we can still assume that even for those unseen parts, their low-level color and texture (which are generally shared throughout an image) could still be taken from some weighted combination of pixels in the ID image.

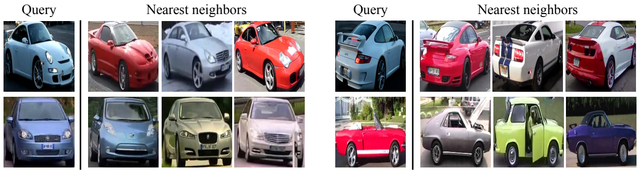

Figure 4. Retrieving nearest neighbors with a self-supervised AlexNet [14] trained on ImageNet. The nearest neighbors resemble the pose of the query well.

To this end, we propose a novel pixel verification module (PVM) that matches every pixel in the generation back to the ID image. Specifically, PVM first transforms the ID image to spatially align it to the generated image. For this, it matches each pixel in I_g to each pixel in I_{id} using their features (we use the last layer feature of our decoder, right before converting to RGB space), which results in a weight matrix $W \in \mathbb{R}^{P \times P}$, where P is the total number of pixels in both I_{id} and I_g, and W_{ij} indicates the affinity between the i-th pixel in I_g and the j-th pixel in I_{id}. To make each row of W sum to 1, we pass W through a softmax function along its rows. Then, PVM transforms the ID image by:

$$I'_{id}(i) = \sum_{j} W(i,j) \cdot I_{id}(j), \; \forall i$$

The result I'_{id} is the ID image aligned to the generation. An example is shown in Fig. 2 (right). The PVM then computes the L_1 difference between I'_{id} and I_g to compute the pixel verification loss:

$$\mathcal{L}_{pv} = ||I'_{id} - I_g||_1.$$

A low $\mathcal{L}_{pv}$ value indicates high degree of verifiability in the generation. Thus minimizing this loss ensures that every generated pixel can be traced back to the ID image. We note PVM is related to the MatchTrans module proposed in [54], however PVM does not constrain a local search window, thus allowing larger pose changes.

Adversarial loss & Auxiliary classification loss To fuel the adversarial game between the generator and the discriminators (D_{real} and D_{pair} from Section 3.1), we employ two adversarial losses $\mathcal{L}_{GAN}^{real}$ and $\mathcal{L}_{GAN}^{pair}$ to encourage realism and conditional identity-preservation, respectively. Finally, as prior research demonstrated the benefit of auxiliary classification tasks when training the discriminator [16, 37], we use the clip index as a proxy to set-up an identity classification task by assuming that cars within the same/different clip correspond to the same/different instances. This gives us $\mathcal{L}_{aux}$ as a cross-entropy classification loss.

Total loss Combining all loss functions, we form the following min-max optimization problem:

$$\min_{G} \max_{D_{real}, D_{pair}} \lambda_1 \mathcal{L}_{dis} + \lambda_2 \mathcal{L}_{pv} + \lambda_3 \mathcal{L}_{aux} + \lambda_4 \mathcal{L}_{GAN},$$

where $\mathcal{L}_{GAN} = \mathcal{L}_{GAN}^{real} + \mathcal{L}_{GAN}^{pair}$ and $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$. We alternate between fixing the generator G and training the discriminators D to maximize the losses, and fixing D and training G to minimize the losses.

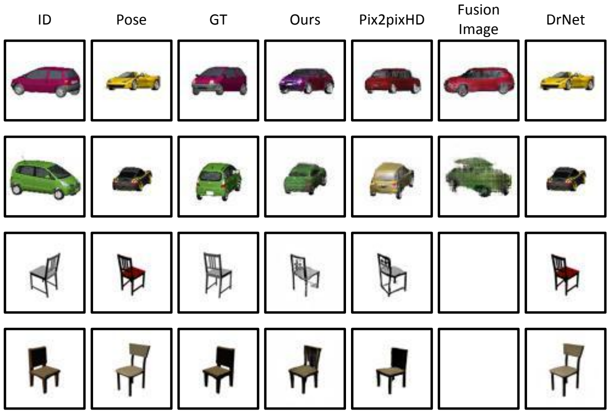

Figure 5. Comparison to baselines on 3D Cars/Chairs.

4. Experiments

In this section, we compare to state-of-the-art baselines, and perform ablative studies to demonstrate the effectiveness of our disentanglement loss and pixel-verification loss.

Datasets. We first conduct proof-of-concept experiments on two synthetic datasets: 3D Cars and 3D Chairs [13, 2], which have 183/1393 clips with varying identity and pose of cars and chairs, respectively. We then test on more challenging real images: we take 3 classes (`Car`, `Bus`, and `Truck`) from YouTube-BoundingBoxes [42] (YTBB) video dataset, which each represent unique challenges. Specifically, cars can have very different shapes (e.g., sedans, SUVs, vans), buses generally have lots of textures (e.g., logos, paints), whereas the appearance of trucks exhibits large uncertainty (it is hard to predict one view from another). Since these are real-world YouTube videos, they are quite challenging – fast motion, drastic illumination changes, compression artifacts, etc. As we will show in experiments, the challenging nature of this dataset is also demonstrated by the relatively poor quality of the results obtained by previous disentanglement methods. We use Faster-RCNN trained on MS COCO to detect instances of the object in the videos. We retain detections which have 0.9 confidence or higher, which removes inaccurate and strongly-occluded instances. This results in 2233/186, 3008/302, 1833/137 clips for training/testing on `Car`, `Bus`, and `Truck`, respectively.

Baselines. **Pix2pixHD** [53]: state-of-the-art conditional image-to-image translation approach. For the input, we directly concatenate the ID and Pose image over the channel axis (i.e., a 6-channel input). The model is trained to output the 3-channel RGB image corresponding to the pseudo ground-truth target. We use the authors' implementation.

FusionImage [25]: solely relies on cyclic constraints which, as we'll show, are not strong enough to induce the desired disentanglement due to the challenging nature of our data (e.g., drastic pose changes). For fair comparison, we adopt our generator/discriminator architectures (based on Pix2PixHD) and only change the losses to those in [25].

	3D Cars			3D Chairs		
	LPIPS	FID	ID	LPIPS	FID	ID
Ours	**0.17**	**71.33**	**0.66**	**0.19**	29.58	**0.67**
Pix2PixHD [53]	0.20	97.76	0.65	0.20	31.01	0.66
FusionImg [25]	0.28	106.96	0.57	0.60	335.39	0.51
DrNet [10]	0.27	72.01	0.57	0.21	**7.42**	0.60

Table 1. Comparison to baselines on 3D Cars/Chairs. For LPIPS and FID, the lower the better; For ID, the higher the better. Both our method and Pix2pixHD perform well on these datasets. DrNet and FusionImage perform much worse (DrNet obtains good FID score only because they incorrectly copy-paste pose images).

DrNet [10]: pits an identity classifier to classify, using pose features, whether two images are from the same video (i.e., have the same identity), and a pose encoder that tries to maximally confuse the identity classifier. This way, it can achieve disentanglement by forcing the pose encoder to not capture identity information. DrNet does not have a target image and therefore only makes use of indirect supervisory signals. We implement DrNet with our encoder and decoder architecture for fair comparison.

Evaluation metrics. We create an evaluation set of 5000 ground-truth triplets. Specifically, we sample two frames from the same video to serve as identity and target images (the same way as we construct triplets in training), whereas for pose image, we manually select an image that has the same pose as the target image.

LPIPS distance [58]: For a generated image $I_g = G(E_i(I_{id}), E_p(I_{pose}))$, we measure its LPIPS distance to the target image I_{target}. This metric essentially captures two aspects: 1) how realistic I_g is, since it has to be realistic to have a low distance to the real image I_{target}; 2) how well I_g preserves the identity of I_{id} and pose of I_{pose}, since I_{target} is a ground-truth combination of the two.

Fréchet Inception Distance (FID) [18]: measures both realism and diversity of the generated data by comparing its distribution to that of real data using the `pool3` features of the Inception-v3 network [49]. We compute FID between the set of generated images $\{I_g^1, I_g^2, ..., I_g^N\}$ and the corresponding target images $\{I_{target}^1, I_{target}^2, ..., I_{target}^N\}$.

ID and Pose preservation scores: We measure preservation of ID and Pose factors as another way to evaluate disentanglement. For the ID preservation score, we finetune an ImageNet pre-trained ResNet-50 on our data to minimize: $\max(f(x_1) \cdot f(y) - f(x_1) \cdot f(x_2) + m, 0)$, where f extracts a L_2-normalized feature from the penultimate layer of ResNet-50, x_1 and x_2 are two instances from the same video clip, and y is from another clip. This triplet loss enforces the affinity between positive pairs (frames from same clip) to be higher than that between a negative pair by a margin m. During evaluation, we average the affinity between the generated image I_g and identity image I_{id} (we sigmoid the affinity to [0, 1]) across the evaluation set as the final ID preservation score. It is harder to evaluate pose preservation

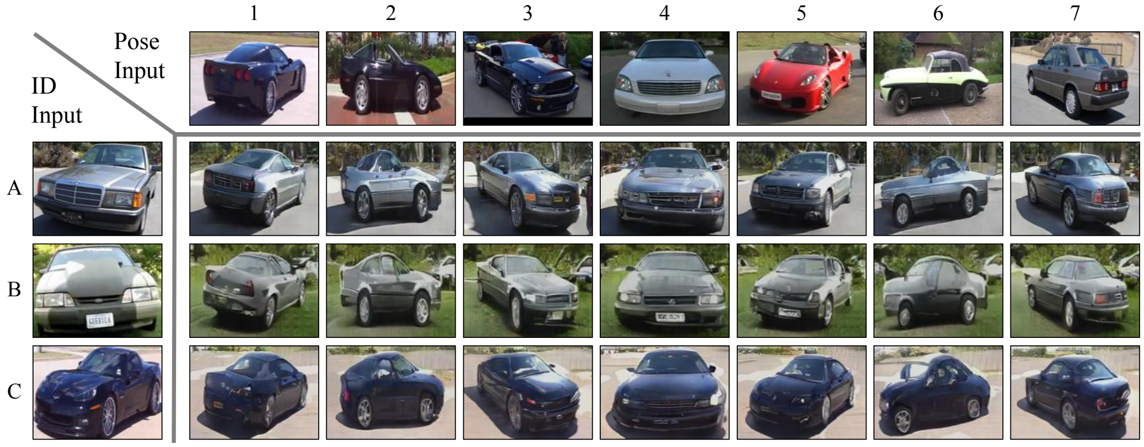

Figure 6. Our generation results for `car`. The top row shows the input pose images, while the leftmost column shows the input ID images. From these results, it is clear that our method has learned to disentangle the identity and pose; i.e., for each ID image, we can change it to different poses, while maintaining its identity.

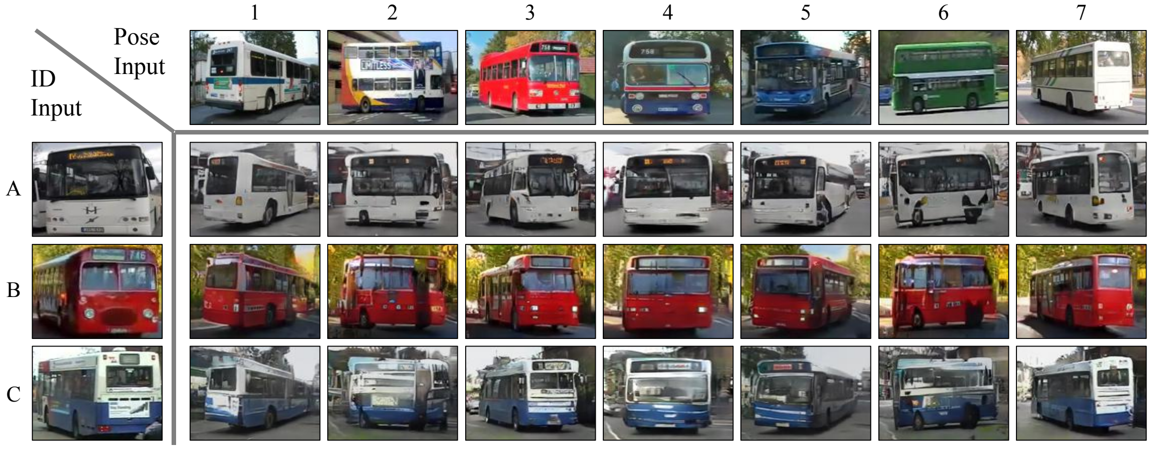

Figure 7. Our generation results for `bus`. The top row shows the input pose images, while the leftmost column shows the input ID images.

since pose annotations are lacking in general. Thus, we only evaluate on YTBB `car` by making use of the Multi-View Car Dataset [38], which has pose annotations, to train a car pose classifier and compute the pose preservation score in a similar way.

Implementation details. We train our model using Adam [28] with a learning rate of 10^{-4}. For data augmentation, we apply standard color jittering (brightness, contrast, saturation) and random cropping. To stabilize training, we perform model averaging following [57]. We generate 128x128 images for all methods (ours and baselines) on YTBB and 64x64 on 3D Cars/Chairs.

4.1. 3D Cars and Chairs datasets

We first present results on synthetic data. As shown in Fig. 5, our method learns to disentangle identity and pose for both datasets – our generation resembles the identity of the ID image and the pose of the pose image. Despite these being simple datasets, FusionImage and DrNet produce degenerate solutions and are unable to generate realistic results. Specifically, DrNet simply copies the pose image whereas FusionImage either generates a lot of artifacts (3D Cars) or generates blank images (3D Chairs). We believe this is due to the lack of supervision in their cyclic constraints when dealing with large amounts of appearance variations (from instance to instance). On the other hand, both pix2pixHD and our method work well on these simple datasets, as reflected by the quantitative results in Table 1.

4.2. YouTube-BoundingBoxes results

Qualitative results. We next present our model's results for `car`, `bus`, and `truck` in Figs. 6, 7 and 8. For each category, the leftmost column shows the input ID reference images, while the first row shows the input pose reference images. Each entry in the matrix corresponds to our model's generated image (e.g., entry C3 is result with ID image C and Pose image 3 as input).

First, it's clear that our model has learned to disentangle identity and pose, so that it can generate new images with the identity of one ID image and the pose of many dif-

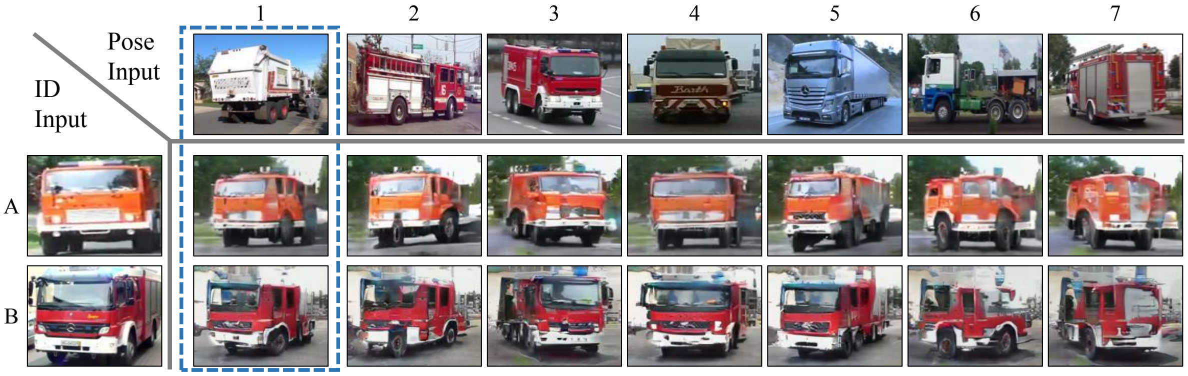

Figure 8. Our generation results for `truck`. The top row shows the input pose images, whereas the leftmost column shows the input ID images. Note how the generation in column 1 (in blue dotted box) flipped the pose by 180 degrees, exhibiting incorrect frontal views.

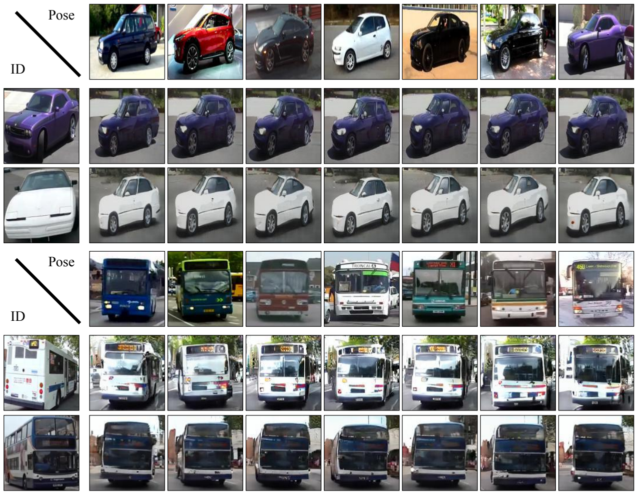

Figure 9. For each row, the pose of the input pose image is fixed while the identity is varying. Note how consistent the generation results are in each row, which suggests that our generation is invariant to the identity of the pose image.

ferent pose images (see the generated cars in Fig. 6). As mentioned before, buses usually have lots of textures (logos, paints, etc.) which makes preserving identity trickier. Still, one can see that our method preserves the fine texture details well (e.g., the blue paint on the bottom of the bus in C1 of Fig. 7). `truck` is more challenging due to the uncertainty of its appearance (e.g., it's sometimes impossible to infer a truck's side-view given only its frontal view). Still, our method is able to capture the gist of the pose while maintaining the identity. One failure mode we observe is that our model can get confused with similar-looking views (e.g., it incorrectly generates a frontal view in column 1 of Fig. 8) and this is partly because of the error from the nearest neighbor search during the triplet generation process.

Fig. 9 shows fixed pose results: for an input ID image, we vary the identity of the pose images but fix their pose. Given the consistency in generations across each row, it is clear that our model is accurately disentangling identity and pose as it is not picking up the identity of the pose image.

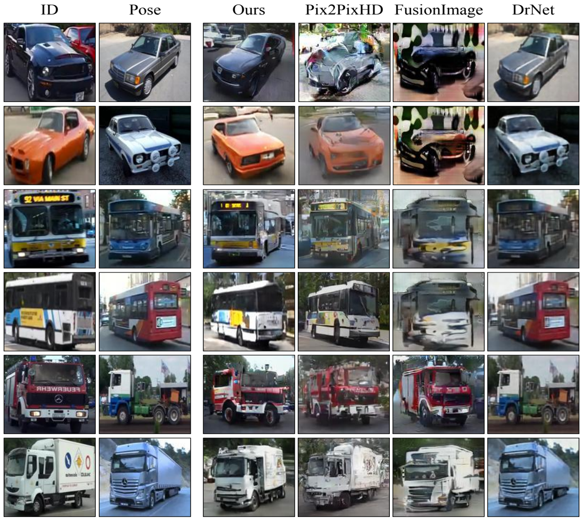

Figure 10. Comparison to baselines. The first/second column show the input ID/pose image. See text for details.

Comparison to baselines. We next show comparisons to baselines in Fig. 10. Note that these are representative examples for each method. First, FusionImage [25] experiences severe mode collapse and its output is completely independent of the pose input. DrNet [10] simply copies the content of the pose image (similar to its behavior on 3D Cars/Chairs), losing the identity information from the ID image. Pix2PixHD [53] is able to disentangle the ID and Pose factors. However, our results look more realistic (1st row) and preserves the identity/pose better (2nd and 5th row respectively). We believe the reason for the failures of FusionImage and DrNet is because the indirect cyclic constraints they optimize are not sufficient to induce disentanglement for our difficult data, and therefore lead to degenerate solutions (mode collapse/identity mapping). Unlike Pix2PixHD, our method not only optimizes the generated image to be similar to the target, but also encourages our two encoders to carry disentangled representations and thus leads to overall better generation results.

As mentioned in Sec. 2, some contemporary work learn

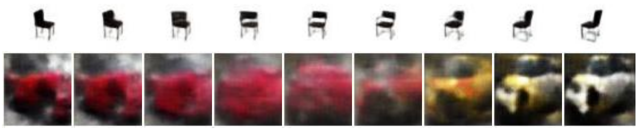

Figure 11. FactorVAE [27] on 3D Chairs and YTBB car datasets.

	car				bus			truck		
	LPIPS	FID	ID	Pose	LPIPS	FID	ID	LPIPS	FID	ID
Ours	**0.33**	**18.03**	**0.63**	0.65	**0.37**	**16.77**	**0.63**	**0.35**	**24.87**	**0.62**
Pix2PixHD [53]	0.37	25.18	0.60	0.64	0.40	32.31	0.60	0.39	75.82	0.58
FusionImg [25]	0.51	239.37	0.52	0.57	0.49	230.06	0.56	0.43	68.76	0.60
DrNet [10]	0.48	24.59	0.52	**0.69**	0.48	38.63	0.52	0.46	28.14	0.53
Ours w/o L_{pv}	0.34	18.47	0.63	0.65	0.37	19.73	0.63	0.36	26.33	0.62
w/o L_{dis}^2	0.35	19.14	0.63	0.65	0.38	24.88	0.63	0.38	37.38	0.62

Table 2. Quantitative results on YTBB car, bus, and truck. For LPIPS and FID, lower is better; for ID and Pose score, higher is better. As explained in the text, we only have pose score for car since we do not have supervised pose classifiers for bus and truck.

disentangled representations by injecting a factorability prior on the latent code in a variational auto-encoding framework. Although not directly comparable (since the model does not have direct control over the learned semantics), we present some representative results of one such model, FactorVAE [27], on 3D Chairs and YTBB car. Specifically, in Fig. 11, we display the latent code dimension that (with manual inspection) is maximally correlated with pose. As shown in the first row, on simple data like 3D Chairs, FactorVAE is able to learn a latent code that corresponds to pose. However, when applied to more challenging data like YTBB car, the latent code mixes up different factors like shape, color, and pose.

Quantitative results. We quantitatively evaluate our method's realism, diversity, and id/pose disentanglement. We also investigate the pixel verification loss, disentanglement loss, and choice of discriminator output.

How real are our generated images? Our method outperforms all baselines for all categories in FID (see Table 2), which suggests that our generated images are more realistic and diverse compared to those of the baselines.

How well do our generated images match the target image? By comparing the LPIPS distance, we can see that our results are closest to the ground-truth target.

How well does our model disentangle id and pose? Our method also outperforms the baselines on both identity and pose preservation scores (except for DrNet, which achieves a better pose preservation score since it incorrectly copy-pastes the pose image), which implies the highest degree of disentanglement between the two factors.

These results are telling in two aspects. Compared to FusionImage and DrNet, our approach clearly benefits from having a direct supervisory signal. On the other hand, the importance of explicitly enforcing disentanglement is revealed when comparing our approach to Pix2PixHD. Overall, both the qualitative and quantitative results demonstrate that our method is able to model several different object categories despite their various unique challenges.

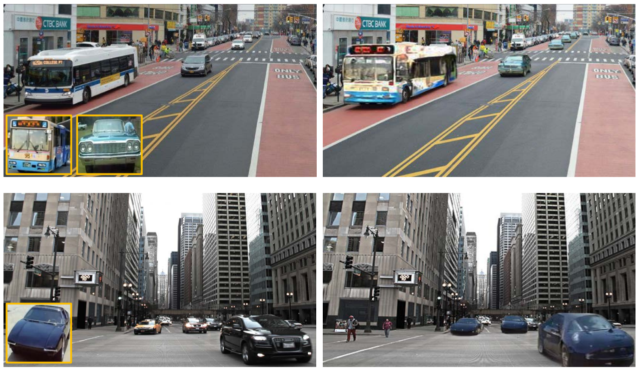

Figure 12. Image composition application. Left: original image, Right: modified image with our generations alpha-blended in.

Ablation studies. We next perform ablation studies (see Table 2 bottom). First, we remove the pixel-verification loss L_{pv}. This consistently hurts FID by a sizable margin, which suggests that pixel-verification is effective in terms of boosting the overall realism and diversity of the generation. If we also remove part of the disentanglement loss L_{dis}^2 (so we are only left with the perceptual loss L_{dis}^1), the performance further drops, both in terms of FID and LPIPS, which again demonstrates that our disentanglement loss is helping to learn a good disentangled representation.

4.3. Application: Image Composition

One potentially useful application of our approach is image composition. Standard image composition approaches [6, 30, 9, 50] require users to supply an image of the desired object pose (or a 3D CAD model matching its identity, which is even harder). For example, to replace all three cars in Fig. 12 (bottom row) with a sports car, images of the sports car facing three different directions would be needed. With our approach, we only need a single image of the desired car, *in any view*. The results in Fig. 12 are produced by alpha-blending our generation into the image.

5. Discussion

Although better than the baselines, our results are not perfect and one prominent failure mode is confusion amongst similar looking poses (e.g., frontal and rear view trucks). This is partly due to the error in nearest neighbor search for generating the training triplets. We believe this issue could potentially be mitigated with a much larger dataset, since our approach can find the nearest neighbor pose image from any image or video.

Acknowledgments. This work was supported in part by ARO YIP W911NF17-1-0410, NSF CAREER IIS-1751206, NSF IIS-1812850, AWS ML Research Award, Google Cloud Platform research credits program, XSEDE IRI180001, and GPUs donated by NVIDIA.

References

[1] Alessandro Achille, Tom Eccles, Loic Matthey, Chris Burgess, Nicholas Watters, Alexander Lerchner, and Irina Higgins. Life-long disentangled representation learning with cross-domain latent homologies. In *NeurIPS*, 2018.

[2] Mathieu Aubry, Daniel Maturana, Alexei Efros, Bryan Russell, and Josef Sivic. Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of cad models. In *CVPR*, 2014.

[3] Guha Balakrishnan, Amy Zhao, Adrian Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018.

[4] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *CVPR*, 2018.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *PAMI*, 2013.

[6] Peter Burt and Edward Adelson. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics (TOG)*, 1983.

[7] Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018.

[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 *CVPR*, 2018.

[9] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan Goldman, and Pradeep Sen. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Transactions on Graphics (TOG)*, 2012.

[10] Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *NeurIPS*, 2017.

[11] Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. In *CVPR*, 2015.

[12] Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, Narayanaswamy Siddharth, Brooks Paige, Dana Brooks, Jennifer Dy, and Jan-Willem van de Meent. Structured disentangled representations. In *AISTATS*, 2018.

[13] Sanja Fidler, Sven Dickinson, and Raquel Urtasun. 3D object detection and viewpoint estimation with a deformable 3D cuboid model. In *NeurIPS*, 2012.

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

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

[16] Ian Goodfellow. NeurIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

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

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

[19] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *NeurIPS*, 2018.

[20] Qiyang Hu, Attila Szabo, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation by mixing them. In *CVPR*, 2018.

[21] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation networks. In *ECCV*, 2018.

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[23] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *NeurIPS*, 2018.

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

[25] Donggyu Joo, Doyeon Kim, and Junmo Kim. Generating a fusion image: One's identity and another's shape. In *CVPR*, 2018.

[26] Natasha Kholgade, Tomas Simon, Alexei Efros, and Yaser Sheikh. 3D object manipulation in a single photograph using stock 3D models. *ACM Transactions on Graphics (TOG)*, 2014.

[27] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *ICML*, 2018.

[28] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[29] Tejas Kulkarni, William Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *NeurIPS*, 2015.

[30] Jean-Francois Lalonde and Alexei Efros. Using color compatibility for assessing image realism. In *ICCV*, 2007.

[31] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018.

[32] Alexander Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *NeurIPS*, 2018.

[33] 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 *ICML*, 2019.

[34] 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*, 2018.

[35] Michael Mathieu, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of

variation in deep representation using adversarial training. In *NeurIPS*, 2016.

[36] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.

[37] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *ICML*, 2017.

[38] Mustafa Ozuysal, Vincent Lepetit, and Pascal Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009.

[39] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.

[40] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017.

[41] Albert Pumarola, Antonio Agudo, Aleix Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*, 2018.

[42] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *CVPR*, 2017.

[43] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *ICML*, 2014.

[44] Scott Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *NeurIPS*, 2015.

[45] Konstantinos Rematas, Chuong Nguyen, Tobias Ritschel, Mario Fritz, and Tinne Tuytelaars. Novel views of objects from a single image. *PAMI*, 2017.

[46] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, 2018.

[47] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. FineGAN: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *CVPR*, 2019.

[48] Hao Su, Fan Wang, Li Yi, and Leonidas Guibas. 3D-assisted image feature synthesis for novel views of an object. In *ICCV*, 2015.

[49] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[50] Michael Tao, Micah Johnson, and Sylvain Paris. Error-tolerant image compositing. In *ECCV*, 2010.

[51] Joshua Tenenbaum and William Freeman. Separating style and content with bilinear models. *Neural Computation*, 2000.

[52] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017.

[53] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.

[54] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *ECCV*, 2018.

[55] Jimei Yang, Scott Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3D view synthesis. In *NeurIPS*, 2015.

[56] Shunyu Yao, Tzu-Ming Harry Hsu, Jun-Yan Zhu, Jiajun Wu, Antonio Torralba, William Freeman, and Joshua Tenenbaum. 3D-aware scene manipulation via inverse graphics. In *NeurIPS*, 2018.

[57] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1710.10916*, 2017.

[58] Richard Zhang, Phillip Isola, Alexei Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[59] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei Efros. View synthesis by appearance flow. In *ECCV*, 2016.

[60] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ECCV*, 2017.

[61] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: image generation with disentangled 3D representations. In *NeurIPS*, 2018.