Identity from here, *Pose* from there: Learning to Disentangle and Generate Objects using Unlabeled Videos

Fanyi Xiao
UC Davis

Haotian Liu
Zhejiang University

Yong Jae Lee
UC Davis

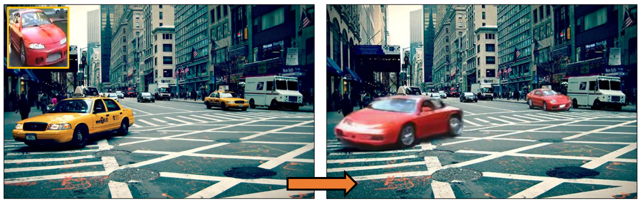

Figure 1. We propose a self-supervised method to disentangle identity (red sedan) and pose (taxis) of objects for image generation.

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.

There has been a long line of research on learning disentangled representations for objects [12, 9, 13, 1, 10, 2, 4, 5, 5]. Early works like Tenenbaum & Freeman [12] operate in a fully-supervised setting in which the factors of interest (content and style in their case) are annotated for each training image. We instead aim to solve this task in a *self-supervised* setting, without any pose or identity annotations. Self/Un-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 shape. To tackle this, recent image generation methods either introduce cyclic constraints [4, 2, 5, 6] (similar in spirit to cycleGAN [14]) or inject priors on the representation based on domain knowledge [10]. Though promising, these methods typically only work well when there is no large change of pose 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 produce visually pleasing results.

In this paper, 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 image, 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. By directly feeding input/output pairs to our model, it provides a much stronger supervisory signal than 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.

Our model is a novel conditional adversarial learning framework based on Generative Adversarial Networks (GANs) [3]. The network consists of an identity encoder, a pose encoder, and a target decoder, and is trained with aforementioned losses to disentangle identity and pose.

2. Approach

Our goal is to learn a model that generates a new image with input A's *identity* and B's *pose*. Importantly, we do not have any identity or pose annotations (i.e., the images are unlabeled) during both training and testing.

Network architecture To factorize identity and pose, we use a two-branch encoder network that processes the identity reference I_{id} and pose reference I_{pose} separately. The respective features are then combined to generate the target, through the decoder. For the output to preserve both *realism* and *identity*, we set up two discriminators for both the

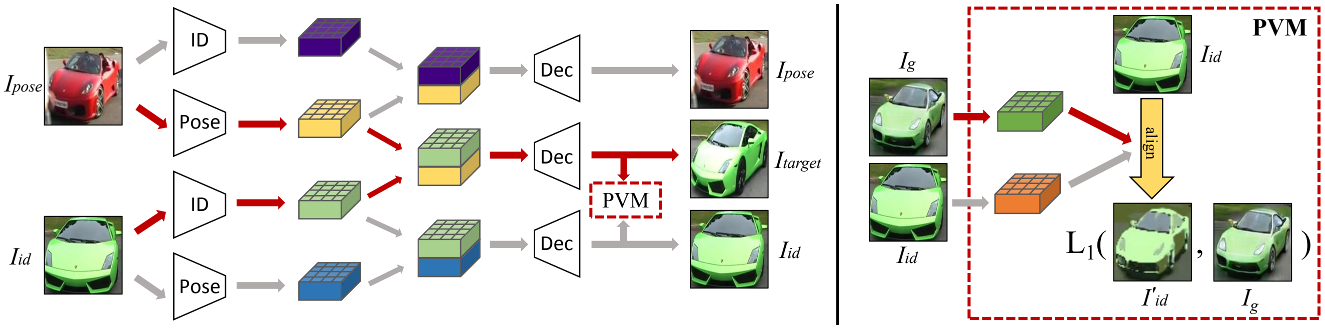

Figure 2. An illustration of the generator. Our generator takes as input both the identity reference image I_{id} and the pose reference image I_{pose}, and tries to generate an output image that matches I_{target}, which has the same identity as I_{id} but with the pose of I_{pose}. Notice how the pose encoded feature is used to generate both I_{target} and I_{pose}, so it cannot contain any identity information. Likewise, the identity encoded feature 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}.

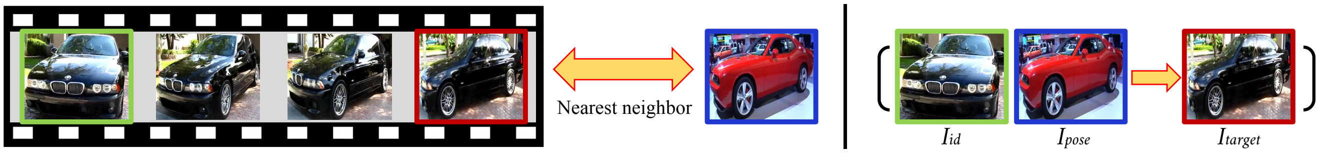

Figure 3. Constructing ID, pose, and target training triplets. With this procedure, we automatically obtain supervision to train our model.

Real/Fake task and a task of classifying whether an input pair shares the same identity or not. The overall architecture is illustrated in Fig. 2.

Constructing ID-pose-target training triplets The key difference between our work and previous self/un-supervised disentanglement works (e.g., [1, 4, 6, 5, 7, 2]) is that rather than relying only on indirect cyclic constraints, we instead construct a *pseudo ground-truth* target image I_{target} so that we can train the model in a supervised way (but without labels). Specifically, 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. We then retrieve a nearest neighbor of I_{target} from other videos/images 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 (we use the conv4 feature of VGG-16 [11]). Although an approximation to the real ground-truth, we show it is highly effective for training our model for image generation. To ensure diversity of the sampled pairs' poses, we cluster all images into M different poses, and then sample (I_{id}, I_{pose}) pairs with diversified pose configurations. Next, we will elaborate the loss functions we use to train our model.

Disentanglement loss First, 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. 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}. As shown in Fig. 2, this will force the ID encoder (the same intuition also applies to the pose encoder) to not capture any pose information since its output is used to generate two targets with distinct poses (I_{id} and I_{target}). Our disentanglement loss is the sum of these two terms.

Pixel-verification loss Recall that our generated image should preserve the identity of the ID reference image. 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 untrue 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. 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 transforms the ID image to spatially align it to the gener-

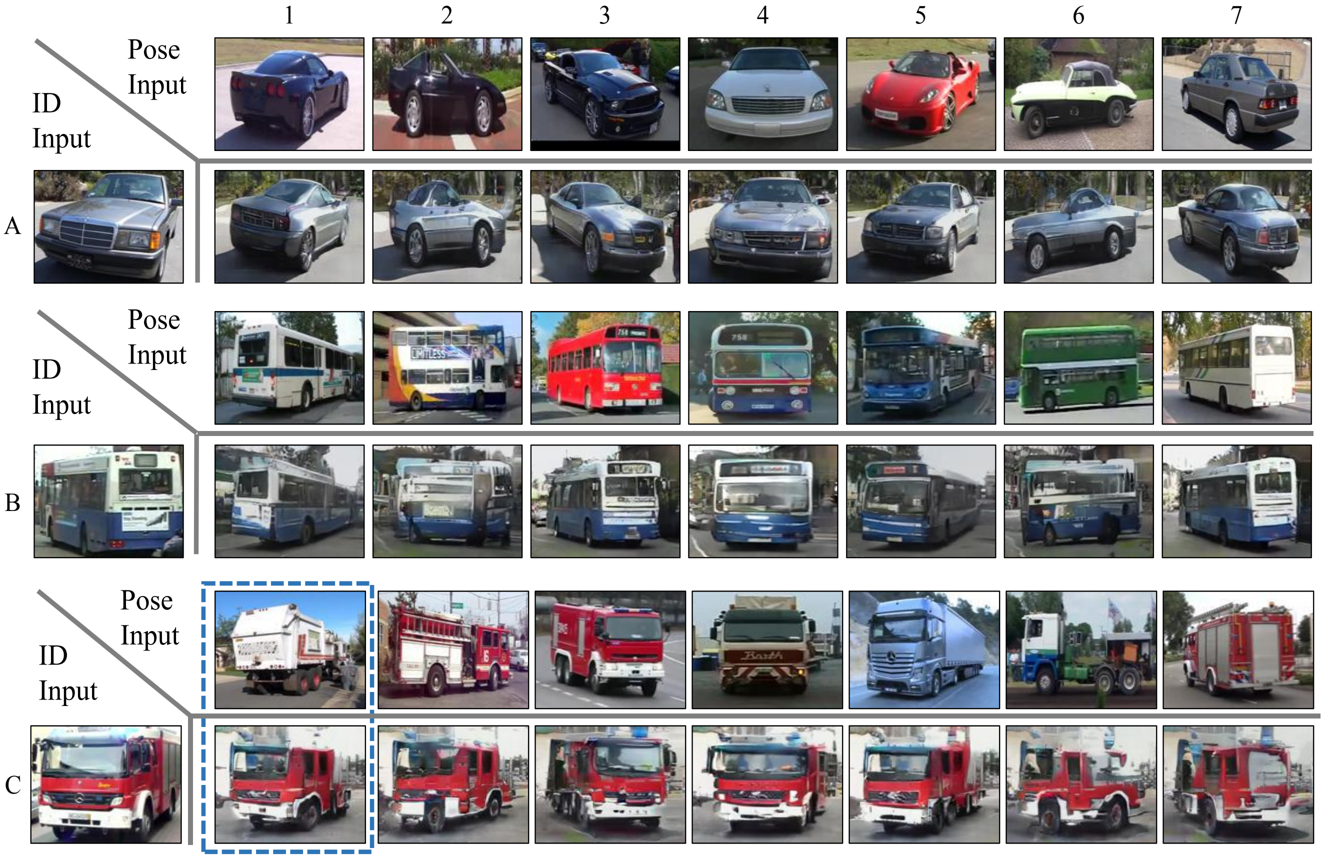

Figure 4. Our generation results. The top row shows the input pose images, while the leftmost column shows the input ID images.

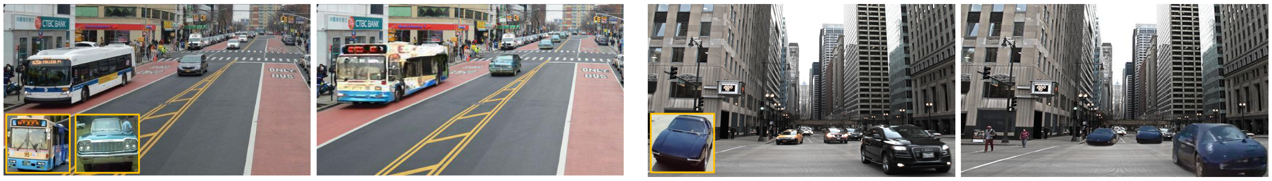

Figure 5. Image composition application. For each pair, on the left is the original image, whereas the right image is modified with our generations alpha-blended in.

ated image – it matches each pixel in I_g to each pixel in I_{id} using their features. We then align I_{id} to I_g (denoted by I'_{id}), using the matched coefficients. The PVM then computes the L_1 difference between I'_{id} and I_g to compute the pixel verification loss $\mathcal{L}_{pv}$. A low $\mathcal{L}_{pv}$ value indicates high degree of verifiability in the generation. Thus minimizing this loss acts as an extra constraint for improving realism and identity preservation. An example of the aligned image I'_{id} is shown on the right of Fig. 2.

We combine all the losses mentioned above and alternate between training the generator and the discriminators.

3. Results

Generation results We present qualitative results in Fig. 4 to demonstrate the effectiveness of our approach on 3 classes (`Car`, `Bus`, and `Truck`) from YouTube-BoundingBoxes [8] (YTBB) video dataset. We choose these classes as they represent unique challenges. Specifically, cars can have very different shapes (e.g., comparing sedans, SUVs, vans), buses generally have lots of textures (logos, paints), whereas the appearance of trucks exhibits large uncertainty (it is hard to predict one view from another). 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 generation. First, it is clear from these results that our model has learned to disentangle the identity and pose, so that it can generate new images with the identity of one ID image and the pose of many different pose images (see the generated cars in rows A). 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 B1). `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 C1) and this is partly because of the error from the nearest neighbor search during the triplet generation process.

Application: Image Composition One potentially useful application of our approach is image composition. In standard image composition approaches, users are required to find an image of the object that is in the *correct* pose (or a 3D CAD model matching its identity, which is even harder). For example, to replace all three cars in Fig. 5 (right) 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. 5 are produced by alpha-blending our generation into the image.

References

[1] E. L. Denton and V. Birodkar. Unsupervised learning of disentangled representations from video. In *NIPS*, 2017. 1, 2

[2] A. Gonzalez-Garcia, J. van de Weijer, and Y. Bengio. Image-to-image translation for cross-domain disentanglement. In *NIPS*, 2018. 1, 2

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1

[4] Q. Hu, A. Szabo, T. Portenier, P. Favaro, and M. Zwicker. Disentangling factors of variation by mixing them. In *CVPR*, 2018. 1, 2

[5] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation networks. In *ECCV*, 2018. 1, 2

[6] D. Joo, D. Kim, and J. Kim. Generating a fusion image: One's identity and another's shape. In *CVPR*, 2018. 1, 2

[7] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 2

[8] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *CVPR*, 2017. 3

[9] S. Reed, K. Sohn, Y. Zhang, and H. Lee. Learning to disentangle factors of variation with manifold interaction. In *ICML*, 2014. 1

[10] Z. Shu, M. Sahasrabudhe, R. A. Guler, D. Samaras, N. Paragios, and I. Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, 2018. 1

[11] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015. 2

[12] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 2000. 1

[13] J. Yang, S. Reed, M.-H. Yang, and H. Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *NIPS*, 2015. 1

[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ECCV*, 2017. 1