Inner Space Preserving Generative Pose Machine

Shuangjun Liu and Sarah Ostadabbas

Augmented Cognition Lab, Electrical and Computer Engineering Department,
Northeastern University, Boston, USA
{shuliu,ostadabbas}@ece.neu.edu
http://www.northeastern.edu/ostadabbas/

Abstract. Image-based generative methods, such as generative adversarial networks (GANs) have already been able to generate realistic images with much context control, specially when they are conditioned. However, most successful frameworks share a common procedure which performs an image-to-image translation with pose of figures in the image untouched. When the objective is reposing a figure in an image while preserving the rest of the image, the state-of-the-art mainly assumes a single rigid body with simple background and limited pose shift, which can hardly be extended to the images under normal settings. In this paper, we introduce an image "inner space" preserving model that assigns an interpretable low-dimensional pose descriptor (LDPD) to an articulated figure in the image. Figure reposing is then generated by passing the LDPD and the original image through multi-stage augmented hourglass networks in a conditional GAN structure, called inner space preserving generative pose machine (ISP-GPM). We evaluated ISP-GPM on reposing human figures, which are highly articulated with versatile variations. Test of a state-of-the-art pose estimator on our reposed dataset gave an accuracy over 80% on PCK0.5 metric. The results also elucidated that our ISP-GPM is able to preserve the background with high accuracy while reasonably recovering the area blocked by the figure to be reposed.

Keywords: Conditional generative adversarial networks (cGANS) · Inner space preserving · Generative pose models · Articulated bodies.

1 Introduction

Photographs are important because they seem to capture so much: in the right photograph we can almost feel the sunlight, smell the ocean breeze, and see the fluttering of the birds. And yet, none of this information is actually present in a two-dimensional image. Our human knowledge and prior experience allow us to recreate "much" of the world state (i.e. its inner space) and even fill in missing portions of occluded objects in an image since the manifold of *probable* world states has a lower dimension than the world state space.

Like humans, deep networks can use context and learned "knowledge" to fill in missing elements. But more than that, if trained properly, they can modify (repose) a portion of the inner space while preserving the rest, allowing us to significantly change portions of the image. In this paper, we present a novel

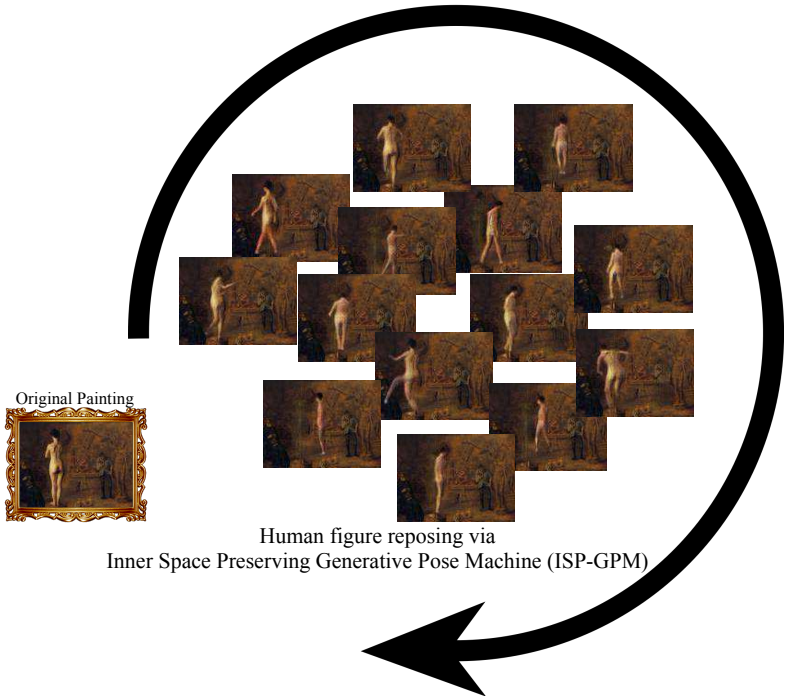


Fig. 1. Inner space preserving reposing of one of Thomas Eakins' paintings: William Rush Carving His Allegorical Figure of the Schuylkill River, 1908.

deep learning based generative model that takes an image and pose specification and creates a similar image in which a target element is reposed. In Fig. 1, we reposed a human figure a number of different ways based on a single painting by the early 20th century painter, Thomas Eakins.

In reposing a figure there are three goals: (**a**) the output image should look like a realistic image in the style of the source image, (**b**) the figure should be in the specified pose, and (**c**) the rest of the image should be as similar to the original as possible. Generative adversarial networks (GANs) [23], are the "classic" approach to solving the first goal by generating novel images that match a certain style. More recently, other approaches have been developed that merge deep learning and probabilistic models including the variational autoencoder (VAE) to generate realistic images [57, 52, 35, 16, 7, 73, 37, 48, 70].

The second goal, putting the figure in the correct pose, requires a more controlled generation approach. Much of the work in this area is based around conditional GANs (cGAN) [42] or conditional VAE (cVAE) [62, 35]. The contextual information can be supplied in a variety of ways. Many of these algorithms generate based on semantic meaning, which could be class labels, attributes, or text descriptors [22, 67, 54, 65, 47]. Others are conditioned on an image often called as image-to-image translation [70]. The success of image-to-image translation is seen in many tasks including colorization [73, 36, 26], semantic image segmentation [11, 38, 58, 24, 43, 13, 45, 19, 49, 12], texture transfer [17], outdoor

(a) (b)

Fig. 2. Generated bird figures from work presented in [56] with captions as: (a) this bird has a *black* head, a pointy orange beak, and yellow body, (b) this bird has a *red* head, a pointy orange beak, and yellow body.

photo generation with specific attributes [60, 34], scene generation with semantic layout [30], and product photo generation [72, 18].

At a superficial level, this seems to solve the reposing problem. However, these existing approaches generally either focus on preserving the image (goal **c**) or generating an entirely novel image based on the contextual image (goal **b**), but not both. For example, when transforming a photo of a face to a sketch, the result will keep the original face spatial contour unchanged [70], and when generating a map from a satellite photo, the street contours will be untouched [27]. Conversely, in attribute based generation, the whole image is generated uniquely for each description [67, 30], so even minor changes will result in completely different images. A demo case from an attribute based bird generation model from [56, 54] is demonstrated in Fig. 2, in which only changing a bird's head color from black to red will alter nearly the entire image.[1]

Recently, there have been attempts to change some elements of the inner space while preserving the remaining elements of an image. Some works successfully preserve the object graphical identities with varying poses or lighting conditions [32, 40, 33, 28, 25, 41, 15, 68]. These works include human face or office chair multi-view regeneration. Yet, all these works are conducted under simplified settings that assume a single rigid body with barren textures and no background. Another work limited the pose range to stay on the pose manifold [68]. This makes them very limited when applied on images from natural settings with versatile textures and cluttered background.

We address the problem of articulated figure reposing while preserving the image's inner space (goals **b** and **c**) via the introduction of our inner space preserving generative pose machine (ISP-GPM) that generates realistic reposed images (goal **a**). In ISP-GPM, an interpretable low-dimensional pose descriptor (LDPD) is assigned to the specified figure in the 2D image domain. Altering LDPD causes figure to be reposed. For image regeneration, we used stack of augmented hourglass networks in a cGAN framework, conditioned on both LDPD and the original image. We replaced hourglass network original downsampling mechanism by pure convolutional layers to maximize the "inner space" preservation between the original and reposed images. Furthermore, we extended the

[1] For this experiment, the random term was set to zero to rule out differences due to the input.

"pose" concept to a more general format which is no longer a simple rotation of a single rigid body, but instead the relative relationship between all the physical entities present in an image and its background. We push the boundary to an extreme case—a highly articulated object (i.e. human body) against a naturalistic background (code available at [2]). A direct outcome of ISP-GPM is that by altering the pose state in an image, we can achieve unlimited generative reinterpretation of the original world, which ultimately leads to a one-shot ISP data augmentation.

2 Related Work

Pose altering is very common in our physical world. If we take photographs of a dynamic articulated object over time, they can hardly be the same. These images share a strong similarity due to having a relatively static background with only differences caused by changes in the object's pose states. We can perceive these differences since the pose information is partially reflected in these images. However, the true "reposing" actually happens in the 3D space and the 2D mapping is just a simple projection afterwards. This fact inspired 3D rendering engines such as Blender, Maya, or 3DS Max to simulate the physical world in (semi)exact dimensions at graphical level, synthesize 3D objects in it, repose the object in 3D, and then finally render a 2D image from the reposed object using a virtual camera [37]. Following this pipeline, there are recent attempts to generate synthesized human images [51, 61, 63]. SCAPE method parameterizes the human body shapes into a generalized template using dense 3D scans of a person in multiple poses [5]. Authors in [11] mapped the photographs of clothing into SCAPE model to boost human 3D pose dataset. Physical rendering and real textures are combined in [64] to generate a synthetic human dataset. However, these methods inevitably require sophisticated 3D rendering engines and avatar data is needed either from full 3D scanning with special equipment or generated from generalized templates [39, 5], which means such data is not easily accessible or extendable to novel figures.

Image-based generative methods, such as GANs and VAEs have already been able to generate realistic images with much context control, specially when they are conditioned [27, 7, 54]. There are also works addressing pose issue of rigid (e.g. chair [14]) or single (e.g. face [68]) objects. An autoencoder structure to capture shift or rotation changes is employed in [35], which successfully regenerates images of 2D digits and 3D graphics rendered images with pose shift. Deep convolutional inverse graphics network (IGN) [33] learns interpretable representation of images including out-of-plane rotations and lighting variations to generate face and chairs from different view points. Based on IGN concept, Yang employed a recurrent network to apply out-of-plane rotations to human faces and 3D chairs to generate new images [68]. In [15], authors built a convolutional neural network (CNN) model for chair view rendering, which can interpolate between given viewpoints to generate missing ones or invent new chair styles by interpolating between chairs from the training set. By incorporating 3D mor-

phable model into a GAN structure, the authors in [71] proposed a framework which can generate face frontalization in the wild with less training data. These works as a matter of fact in a sense preserve the inner space information with the target identity unchanged. However, most are limited to a single rigid body with simple or no background, and are inadequate to deal with complex articulated objects such as human body in a realistic background setting.

In the last couple of years, there have been a few image-based generative models proposed for human body reposing. In [56] and [54], by localizing exact body parts, human figures were synthesized with provided attributes. However, though pose information is provided exactly, the appearance are randomly sampled under attribute context. Lassner and colleagues in [37] generated vivid human figures with varying poses and clothing textures by sampling from a given set of attributes. A direct result of sampling based method is a strong coupling effect between different identities in the image, in which the pose state cannot get altered without the image inner space change.

In this paper, we focus on the same pose and reposing topics but extend them to a more general format of highly articulated object with versatile background under realistic/wild settings. We are going to preserve the original inner space of the image, while altering the pose of the an specific figure in the image. Instead of applying a large domain shift on an image such as changing the day to night, or the summer to winter, we aim to model a pose shift caused by a movement in the 3D physical world, while the inner space of the world stays identical to its version before this movement. Inspired by this idea, we present our inner space preserving generative pose machine (ISP-GPM), in which rather than attribute based sampling, we focus on specific image instances.

3 World State and Inner Space of An Image

"No man ever steps in the same river twice" quoted from Heraclitus.

Our world is dynamically changing. Taking one step forward, raising hand a little bit, moving our head to the side, all these tiny motions make us visually different from a moment ago. These changes are also dependably reflected in the photographs taken from us. In most cases, for a short period of time, we can assume such changes are purely caused by pose shift instead of characteristic changes of all related entities. Let's simply call the partial world captured by an image "the world". If we model the world by a set of rigid bodies, for a single rigid body without background (the assumption in the most of the state-of-the-art), the world state can be described by appearance term $\boldsymbol{\alpha}$ and the pose state $\boldsymbol{\beta}$ of the rigid body as $W_s = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ and the reposing process is conduced by altering $\boldsymbol{\beta}$ to a target pose $\hat{\boldsymbol{\beta}}$. However, real world can hardly be described by a simple rigid body, but clustered articulated rigid bodies and background. In this case, we formulate the world state as:

$$W_s = \{\boldsymbol{\alpha_i}, \boldsymbol{\beta_i}, \phi(i,j) | i,j \in N\}. \tag{1}$$

where, N stands for the total number of rigid bodies in the world and $\phi(i,j)$ stands for the constraints between two rigid bodies. For example, a human has

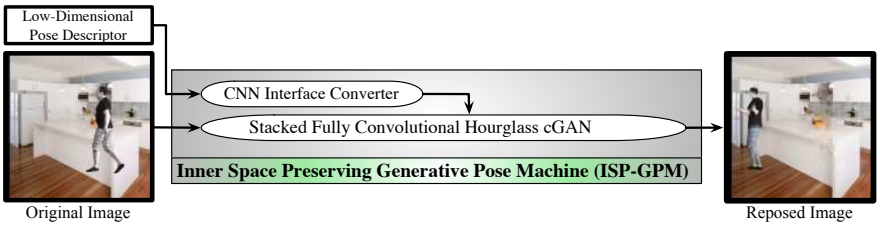


Fig. 3. An overview of the Inner Space Preserving Generative Pose Machine (ISP-GPM) framework.

N (depending on the granularity of the template that we choose) articulated limbs in which the joints between them follow the biomechanical constraints of the body. A pure reposing process in physical world should keep the $\boldsymbol{\alpha_i}$ terms unchanged. However, in imaging process, only part of the $\boldsymbol{\alpha_i}$ information is preserved as $\boldsymbol{\alpha_i^{in}}$ with $\boldsymbol{\alpha_i = \alpha_i^{in} + \alpha_i^{out}}$, where $\boldsymbol{\alpha_i^{out}}$ stands for the missing information in the image with respect to the physical world. We assume each image can partially preserved the physical world information and we call this partially preserved world state the "inner space". If $\boldsymbol{\alpha_i^{in}}$ and $\phi(i,j)$ term are preserved during figure i reposing, we call this process "inner space preserving".

Another assumption is that in the majority of cases, the foreground (F) and the background (B) should be decoupled in the image, which means if figure $i \in F$ and figure $j \in B$, the $\phi(i,j)$ is empty or vice versa. This means if a bird with black head and yellow body is the foreground, the identical bird can be in different backgrounds such as on a tree or in the sky. However, strong coupling between foreground and background is often seen in attribute-based models as shown in Fig. 2. Instead, we designed our generative pose machine to reflect: (1) inner space preserving, and (2) foreground and background decoupling.

4 ISP-GPM: Inner Space Preserving Generative Pose Machine

The ISP-GPM addresses the extensive pose transformation of articulated figures in an image through the following process: given an image with specified figure and its interpretable low-dimensional pose descriptor (LDPD), ISP-GPM outputs a reposed figure with original image inner space preserved (see Fig. 3). The key components of the ISP-GPM are: (1) a CNN interface converter to make the LDPD compatible with the first convolutional layer of the ISP-GPM interface, and (2) a generative pose machine to generate reposed figures using the regression structure of hourglass networks when stacked in a cGAN framework in order to force the pose descriptor into the regenerated images.

4.1 CNN Interface Converter

We employed an LDPD in the 2D image domain, which in the majority of the human pose dataset such as Max Planck institute informatics (MPII) [3] and Leeds sports pose (LSP) [29] is defined as the vector of 2D joint position

coordinates. To make this descriptor compatible with the convolutional layer interface of ISP-GPM, we need a CNN interface converter. The most straight forward converter could simply set the joint point in the image, similar to the work described in [56]. As human body can be represented by a connected graph [4, 8], more specifically a tree structure, in this work we further appended the edge information into our converter. Assume human pose to be represented by 2D locations of its N joints. Let's use N channel maps to hold this information as joint map, J_{Map}. For each joint i with coordinates (x_i, y_i), if joint i's parent joint exists, we are going to draw a line from (x_i, y_i) to its parent location in channel i of J_{Map}. In generating J_{Map}s, the draw operation is conducted by image libraries such as OpenCV [10].

4.2 Stacked Fully Convolutional Hourglass cGAN

Many previous works have proved the effectiveness of multi-stage estimation structure in human pose estimation, such as 2016 revolutionary work of convolutional pose machine [66]. As an inverse operation to regenerate figures of humans, we employed a similar multi-stage structure. Furthermore, human pose can be described in a multi-scale fashion, starting from simple joint description to sophisticated clothing textures on each body part, which inspired the use of an hourglass model with a stacked regression structure [44]. However, instead of pose estimation or segmentation, for human reposing problem, more detailed information needs to be preserved in both encoding and decoding phases of the hourglass network. Therefore, we replaced hourglass network's max pooling and the nearest upsampling modules by pure convolutional layers to maximize the information preservation. The skip structure of the original hourglass network is also preserved to let more original high frequency parts pass through. Original hourglass is designed for image regression purpose. In our case, we augment hourglass original design by introducing structure losses [27], which penalize the joint configuration of the output. We forced the pose into the generated image by employing a cGAN mechanism.

An overview of our stacked fully convolutional hourglass cGAN (FC-hourglass-cGAN) is shown in Fig. 4, where we employed a dual skip mechanism, a module level skip as well as the inner module level skips. Each FC-hourglass employs a encoder-decoder like structure [46, 6, 44]. Stacked FC-hourglass plays the generator role in our design, while another convolutional net plays the discriminator role. We employed an intermediate supervision mechanism similar to [44], however the supervision is conducted by both L1 loss and generator loss, as described in the following section.

4.3 Stacked Generator and Discriminator Losses

Due to the ISP-GPM stacked structure, the generator loss comes from all intermediate stages to the final one. The loss for generator is then computed as:

$$L_G(G, D) = \mathbb{E}_{u,v}[\log D(u, v)] + \sum_{i=1}^{N_{stk}} \mathbb{E}_u[\log(1 - D(u, G(u)[i])]. \quad (2)$$

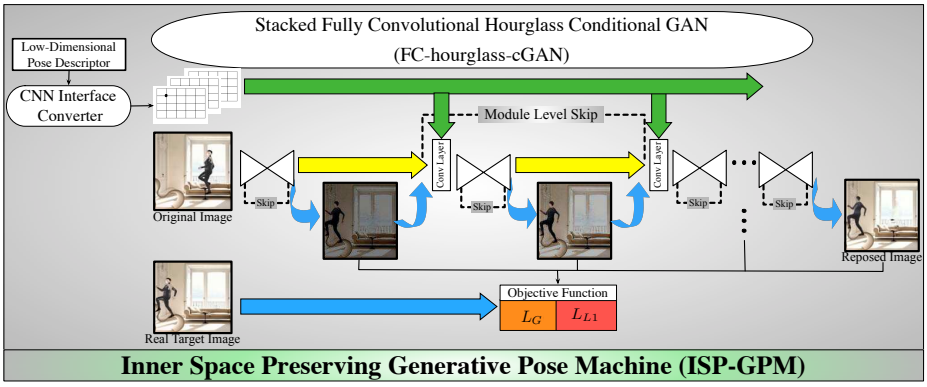


Fig. 4. Inside the stacked FC-hourglass-cGAN part of the ISP-GPM. Blue arrows stand for the image flow, yellow arrows for the hourglass feature maps, and green arrows for J_{Map} flow.

where, u stands for the combined input of J_{Map} and the original image, and v is the target reposed image. G is stacked FC-hourglass that acts as the generator role, N_{stk} stands for the total number of stacks in the generator G, and D is the discriminator part of the cGAN. Different from commonly used generator, our G gives multiple output according to the stack number. $G(u)[i]$ stands for the i-th output conditioned on u. Another difference from traditional cGAN design is that we do not include the random term z as it is common in most GAN based models [42, 62, 22, 67, 47, 23]. The particular reason to have this term in traditional GAN based model is to introduce higher variation into the sampling process. The main reason behind introducing randomness in GAN is to capture a probabilistic distribution which generates *novel* images that match a certain style. However, our ISP-GPM follows quite opposite approach, and aims to achieve a deterministic solution based on the inner space parameters, instead of generating images from a sampling process. D term is the discriminator to reveal if the input is real or fake, conditioned on our input u information.

Since our aim is regressing the figure to a target pose on its subspace manifold, low frequency components play an import role here to roughly localize the figure to the correct position. Therefore, we capture these components using a classical L1 loss:

$$L_{L1}(G) = \sum_{i=1}^{N_{stk}} \mathbb{E}_{u,v}[||v - G(u)[i]||_1]. \tag{3}$$

We used a weighted term λ to balance the importance of L1 and G losses in our target objective function:

$$L^*_{obj} = \arg\min_G \max_D L_G(G, D) + \lambda L_{L1}(G). \tag{4}$$

5 Model Evaluation

To illustrate our inner space preserving concept and the performance of the proposed ISP-GPM, we chose a specific figure as our reposing target, the hu-

man body, due to the following rationale. First and foremost, human body is a highly articulated object with over 14 components depending on the defined limb granularity. Secondly, human pose estimation and tracking is a well-studied topic [59, 20, 66, 50, 9, 53] as it is highly needed in abundant applications such as pedestrian detection, surveillance, self-driving cars, human-machine interaction, healthcare, etc. Lastly, several open-source datasets are available including MPII [3], BUFFY [21], LSP [29], FLIC [59], and SURREAL [64], which can facilitate deep learning-based model training and wide range of test samples for model evaluation.

5.1 Dataset Description

Although well-known datasets for human pose estimation [3, 29, 59] exist, few of them can satisfy our reposing purpose. As mentioned in Section 3, we aim at preserving the inner space of the original image before figure reposing. Therefor, we need pairs of images with the same $\boldsymbol{\alpha}$ term but varying $\boldsymbol{\beta}$ term, which means identical background and human. The majority of the existing datasets are collected from different people individually with no connections between images, so they have varying $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. A better option is extracting images from consecutive frames of a video. However, not many labelled video datasets from human are available. Motion capture system can facilitate auto labeling process, but they focus on the pose data without specifically augmenting the appearance $\boldsymbol{\alpha}$, such that "the same person may appear under more than one subject number" as they mentioned in [1]. The motion capture marks are also uncommon in images taken from natural settings. Another issue with daily video clips is that the background is unconstrained as it could be dynamic caused by camera motion or other independent entities in the background. Although, our framework can handle such cases by expanding world state in Eq. (1) to accommodate several dynamic figures in the scene, in this paper, we focus on a case with images from a human as the figure of interest in a static yet busy background.

Alternatively, we shift our attention to the synthesized datasets of human poses with perfect joint labeling and background control. We employed SURREAL (Synthetic hUmans foR REAL tasks) dataset of synthesized humans with various appearance textures and background [64]. All pose data are originated from the Carnegie Mellon University motion capture (mocap) dataset [1]. The total number of video clips for training is 54265 with combined different overlap settings [64]. Another group of 504 clips are used for model evaluation. One major issue of using SURREAL to suit our purpose is that the human subjects are not always shown in the video since it employs a fixed camera setting and the subjects are faithfully driven by the motion capture data. We filtered the SURREAL dataset to get rid of the frames without the human in them and also the clips with too short duration such as 1 frame clips.

5.2 ISP-GPM Implementation

Our pipeline was implemented in Torch with environment settings of CUDA8.0, CUDNN 5 with NVIDIA GeForce GTX 1080-Ti. Our implementation builds on

(a) (b)

Fig. 5. Inner space preserving human reposing with different downsampling layers: (a) downsampled with max pooling, and (b) downsampled with convolution layers. First column is the input image, second column is the ground truth image of the target pose, last column is the generated image from ISP-GPM.

the architecture of the original hourglass [44, 64]. Discriminator net follows the design in [27]. Adams optimizer with $\beta 1 = 0.5$ and learning rate of 0.0002 was employed during training [31]. We used 3 stacked hourglass with input resolution of 128×128. In each hourglass, 5 convolutions configuration is employed with lowest resolution of 4×4. There are skip layers at all scale levels.

We used the weighted sum loss during generator training with more emphasis on L1 loss to give priority to the major structure generation instead of textures. We set $\lambda = 100$ in Eq. (4) as we observed transparency in the resultant image if we give a small λ. Our input is set to $128 \times 128 \times 3$ due to the memory limitations. The pose data is 16×2 vector to indicate 16 key point positions of human body as defined in SURREAL dataset [64]. In training session, we employed a batch size of 3, epoch number of 5000, and conduct 50 epochs for each test.

5.3 ISP-GPM with Different Configurations

To compare the quality of the resultant reposed images between ISP-GPMs with different model configurations, we fixed the input image to be the first frame of each test clip and the 60th or the last frame as the target pose image.

Downsampling Strategies: We first compared the quality of the reposing when fully convolution (FC) layers vs. max pooling downsampling is used in the stacked hourglass network. To make a clear comparison, we chose same test case for different model configurations and presented the input images, ground truth and generated images in Fig. 5. Each row shows a test example. Columns from left to right stand for the input image, ground truth and generated result. With the given two examples, it is clear that the max pooling is prone to the

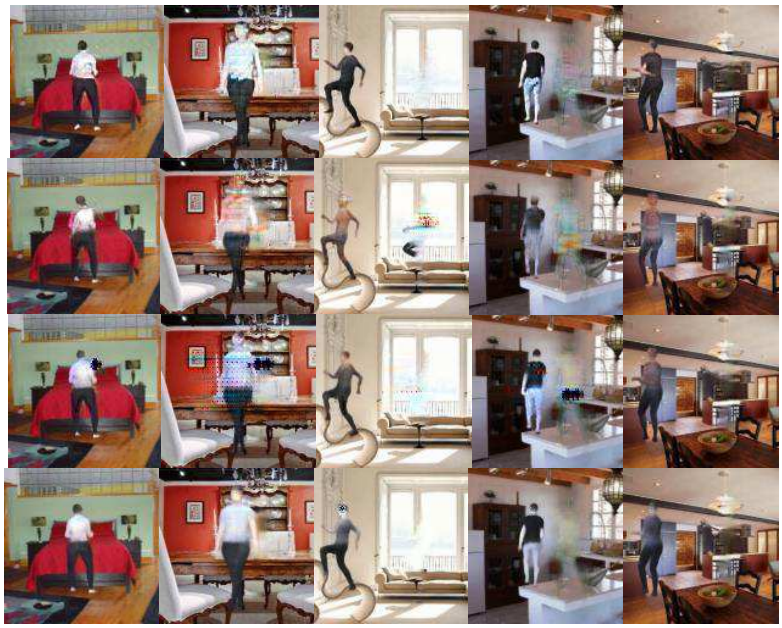
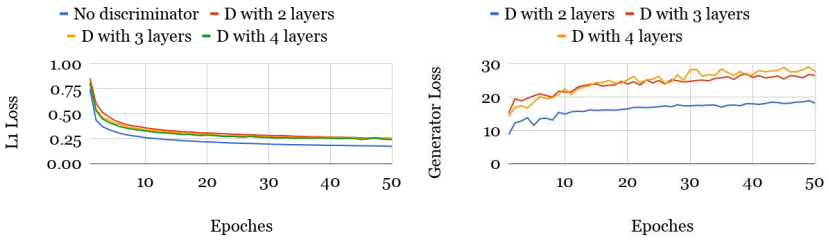


Fig. 6. Reposed human figure under different network configurations: 1st to 3rd row with two to four layers discriminator network and 4th row without discriminator but only L1 loss.

blurriness, while the FC configuration outputs more detailed textures. However, the last row of Fig. 5 uncovers that FC configuration is more likely to result in abnormal colors when compared to the max pooling configuration. This is expectable since the max pooling prefers to preserve the local information of an area.

Discriminator Layer: Inspired by [27], we employed the discriminator layer with different patch sizes to test its performance. Patch sizes can be tuned by altering the discriminator layer numbers to cover patches with different sizes. In this experiment, all the configurations we chose can effectively generate human contours at indicated position but only differs in the image quality. So we only show the outcomes by changing the discriminator layer from two to four as depicted in 1st to 3rd row of Fig. 6, respectively. The figure's last row shows the output without discriminator layer. We discover that the discriminator did help in texture generation, however larger patches in contrast will result in strong artifacts as shown in 2nd and 3rd row of Fig. 6. In the case with no discriminator and only L1 loss, the output is obviously prone to blurriness which is consistent with findings from previous works [35, 48, 27]. We believe larger patch takes higher level structure information into consideration, however the local textures on the generated human can provides better visual quality, as seen in the 1st row of Fig. 6) with two layers discriminator.

To better illustrate the discriminator's role during training session, we recorded loss of each component during training with different network configurations as



(a) (b)

Fig. 7. Losses during training for different network configurations: (a) L1 loss, (b) Generator loss. Note that model without discriminator only shows in L1 loss.

shown in Fig. 7. Model without discriminator are only shown in Fig. 7a. Though model without discriminator shows better performance on L1 metric, it does not always yield good looking images as it prefers to pick median values among possible colors to achieve better L1. There are a common trend that all G loss increase as training went on and the final G loss is even stronger than initial state. By observing the training process, we found out it is a process that the original human start fading away while the target posed human reveals itself gradually. Indeed, no matter how strong the generator is, its output cannot be as real as original one. So, at the beginning the generated image will be more likely to fool the discriminator as it keeps much of the real image information with less artifact.

5.4 Comparison with the State-of-the-art

There are few works focusing on human image generation via generative models, including Reed's [55, 56] and Lassner's [37]. We compared the outputs of our ISP-GPM model with these works as shown in Fig. 8 (excluding [55] since the code is not provided). We omitted the input images in Fig. 8 and only displayed the reposed ones to provide a direct visual comparison with other methods.

Fig. 8 shows that Lassner's [37] method preserves the best texture information in the generated images. However, there are three aspects in Lassner's that need to be noted. First of all, their generation process is more like a random sampling process from the human image manifold. Secondly, to condition this model

(a) (b) (c)

Fig. 8. Image quality comparison of the generative models for human figures presented by (a) Lassner [37], (b) Reed [56], and (c) our ISP-GPM.

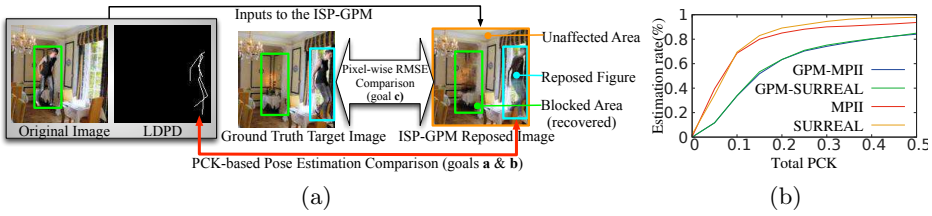


Fig. 9. (a) ISP quantitative evaluation schematic, (b) Pose estimation accuracy comparison tested on MPII, SURREAL, and our ISP-GPM datasets.

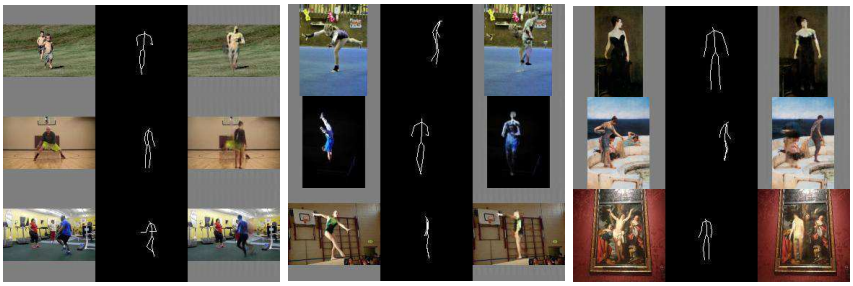
on pose, SMPL model is needed for silhouette generation, which inevitably takes advantages of a 3D engine. Thirdly, they can generate humans with vivid background, however it is like a direct mask overlapping process with fully observed background images in advance [37]. In our ISP-GPM, both human and background are generated and merged in the same pipeline. Our pose information is a low-dimensional pose descriptor that can be generated manually. Additionally, both human and background are only partially observed due to human facing direction and the occlusion caused by the human in the scene. As for [56], the work is not an ISP model, as illustrated by an example earlier in Fig. 2.

5.5 Quantitative Evaluation

To jointly evaluate goals **a** and **b**, we *hypothesized* that if the generated reposed images are realistic enough with specified pose, their pose should be recognizable by a pose recognition model trained on real-world images. We employed a high performance pose estimation model with a convolutional network architecture [44], to compare the estimated pose in the reposed synthetic image against the LDPD assigned to it in the input. We selected 100 images from both *MPII Human Pose* and *SURREAL* datasets in continuous order to avoid possible cherry picking. We selected the 20th frame of random video sequences to repose original images to form re-rendered ISP-GPM version datasets, namely MPII-GPM and SURREAL-GPM with joint labels compatible with the MPII joint definition. Please note that to synthesize the reposed images, we used ISP-GPM model with three layers discriminator and L1 loss as described in Section 5.3.

We used probability of correct keypoint (PCK) criteria for pose estimation performance evaluation, which is the measure of joint localization accuracy [69]. The average pose estimation rates (over 12 body joints) tested on MPII-GPM and SURREAL-GPM datasets are shown in Fig. 9b and compared with the the pose estimator accuracy [44] tested on 100 images from original MPII and SURREAL datasets. These results illustrate that a well-trained pose estimator model is able to recognize the pose of our reposed images with over 80% accuracy on PCK0.5 metric. Therefore, ISP-GPM not only reposes the human figure accurately, but also makes it realistic enough to fool a state-of-the-art pose detection model to take its parts as human limbs.

With respect to goal **c**, we tested the inner space preserving ability in two folds: (1) the background of the reposed image (i.e. the unaffected area) should



(a) (b) (c)

Fig. 10. ISP reposing of human figures: (a) MPII dataset [3], (b) LSP dataset [29] and (c) art works in the following order, Madame X (1884)–John Singer Sargent, Silver Favourites (1903)–Lawrence Alma-Tadema, Saint Sebastian Tended–Saint Irene and her Maid-Bernardo Strozzi.

stay as similar as possible to the original image, and (2) the blocked area by the figure in original pose should be recovered with respect to the context. To test (1), we blocked out the affected areas where the figure of interest occupies in original and target images and computed the pixel-wise mean RMSE between the unaffected area of both images (**RMSE = 0.050±0.001**). To evaluate (2), we compared the recovered blocked area with the ground truth target image (**RMSE = 0.172±0.010**). These results elucidate that our ISP-GPM is able to preserve the background with high accuracy while recovering the blocked area reasonably. Please note that the model has never seen behind the human in the original images and it attempts to reconstruct a texture compatible with the rest of the image, hence the higher RMSE.

6 ISP-GPM in Real World

To better illustrate the capability of ISP-GPM, we applied it on real world images from well-known datasets, MPII [3] and LSP [29]. As there is no ground truth to illustrate the target pose, we visualized the LDPD into a skeleton image by connecting the joints according to their kinematic relationships. ISP reposed images of MPII [3] and LSP [29] are shown in Fig. 10a and Fig. 10b, respectively. Each sample shows input image, visualized skeleton, and the generated image from left to right.

Arts are originated from real world and we believe when created, they also preserved inner space of an imagined world by the artist. So, we also applied our ISP-GPM on the arts inspired by human figures including paintings and sculptures. They are either from publicly accessible websites or art works in museums captured by a regular smartphone camera. The ISP reposing results are shown in Fig. 10c. From results of the real world images, the promising performance of ISP-GPM is apparent. However, there are still failure cases such as the residue of the original human that the network is unable to fully erased or the loss of the detailed texture and shape information.

References

1. CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/info.php (2018)
2. ISP-GPM code. http://www.northeastern.edu/ostadabbas/2018/07/23/inner-space-preserving-generative-pose-machine/ (2018)
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. Proceedings of the IEEE Conference on computer Vision and Pattern Recognition pp. 3686–3693 (2014)
4. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on pp. 1014–1021 (2009)
5. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. ACM transactions on graphics (TOG) **24**(3), 408–416 (2005)
6. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
7. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: CVAE-GAN: fine-grained image generation through asymmetric training. CoRR, abs/1703.10155 **5** (2017)
8. Bergtholdt, M., Kappes, J., Schmidt, S., Schnörr, C.: A study of parts-based object class detection using complete graphs. International journal of computer vision **87**(1-2), 93 (2010)
9. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. Computer Vision, 2009 IEEE 12th International Conference on pp. 1365–1372 (2009)
10. Bradski, G., Kaehler, A.: Opencv. Dr. Dobbs journal of software tools **3** (2000)
11. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2018)
12. Ciresan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. Advances in neural information processing systems pp. 2843–2851 (2012)
13. Dai, J., He, K., Sun, J.: Convolutional feature masking for joint object and stuff segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 3992–4000 (2015)
14. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. Advances in Neural Information Processing Systems pp. 658–666 (2016)
15. Dosovitskiy, A., Springenberg, J.T., Brox, T.: Learning to generate chairs with convolutional neural networks. Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on pp. 1538–1546 (2015)
16. Dosovitskiy, A., Springenberg, J.T., Tatarchenko, M., Brox, T.: Learning to generate chairs, tables and cars with convolutional networks. IEEE transactions on pattern analysis and machine intelligence **39**(4), 692–705 (2017)
17. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. Proceedings of the 28th annual conference on Computer graphics and interactive techniques pp. 341–346 (2001)

18. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? ACM Trans. Graph. **31**(4), 44–1 (2012)
19. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1915–1929 (2013)
20. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on pp. 1–8 (2008)
21. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on pp. 1–8 (2008)
22. Gauthier, J.: Conditional generative adversarial nets for convolutional face generation. Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester **2014**(5), 2 (2014)
23. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems pp. 2672–2680 (2014)
24. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous detection and segmentation. European Conference on Computer Vision pp. 297–312 (2014)
25. Hinton, G.E., Krizhevsky, A., Wang, S.D.: Transforming auto-encoders. International Conference on Artificial Neural Networks pp. 44–51 (2011)
26. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. ACM Transactions on Graphics (TOG) **35**(4), 110 (2016)
27. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint (2017)
28. Jampani, V., Nowozin, S., Loper, M., Gehler, P.V.: The informed sampler: A discriminative approach to bayesian inference in generative computer vision models. Computer Vision and Image Understanding **136**, 32–44 (2015)
29. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. Proceedings of the British Machine Vision Conference (2010), doi:10.5244/C.24.12
30. Karacan, L., Akata, Z., Erdem, A., Erdem, E.: Learning to generate images of outdoor scenes from attributes and semantic layouts. arXiv preprint arXiv:1612.00215 (2016)
31. Kinga, D., Adam, J.B.: A method for stochastic optimization. International Conference on Learning Representations (ICLR) **5** (2015)
32. Kulkarni, T.D., Mansinghka, V.K., Kohli, P., Tenenbaum, J.B.: Inverse graphics with probabilistic cad models. arXiv preprint arXiv:1407.1339 (2014)
33. Kulkarni, T.D., Whitney, W.F., Kohli, P., Tenenbaum, J.: Deep convolutional inverse graphics network. Advances in Neural Information Processing Systems pp. 2539–2547 (2015)
34. Laffont, P.Y., Ren, Z., Tao, X., Qian, C., Hays, J.: Transient attributes for high-level understanding and editing of outdoor scenes. ACM Transactions on Graphics (TOG) **33**(4), 149 (2014)
35. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. International Conference on Machine Learning pp. 1558–1566 (2016)
36. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. European Conference on Computer Vision pp. 577–593 (2016)

37. Lassner, C., Pons-Moll, G., Gehler, P.V.: A generative model of people in clothing. arXiv preprint arXiv:1705.04098 (2017)
38. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition pp. 3431–3440 (2015)
39. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 248 (2015)
40. Loper, M.M., Black, M.J.: Opendr: An approximate differentiable renderer. European Conference on Computer Vision pp. 154–169 (2014)
41. Michalski, V., Memisevic, R., Konda, K.: Modeling deep temporal dependencies with recurrent grammar cells. Advances in neural information processing systems pp. 1925–1933 (2014)
42. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
43. Mostajabi, M., Yadollahpour, P., Shakhnarovich, G.: Feedforward semantic segmentation with zoom-out features. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 3376–3385 (2015)
44. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. European Conference on Computer Vision pp. 483–499 (2016)
45. Ning, F., Delhomme, D., LeCun, Y., Piano, F., Bottou, L., Barbano, P.E.: Toward automatic phenotyping of developing embryos from videos. IEEE Transactions on Image Processing **14**(9), 1360–1371 (2005)
46. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. Proceedings of the IEEE International Conference on Computer Vision pp. 1520–1528 (2015)
47. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. arXiv preprint arXiv:1610.09585 (2016)
48. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 2536–2544 (2016)
49. Pinheiro, P., Collobert, R.: Recurrent convolutional neural networks for scene labeling. International conference on machine learning pp. 82–90 (2014)
50. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Strong appearance and expressive spatial models for human pose estimation. Computer Vision (ICCV), 2013 IEEE International Conference on pp. 3487–3494 (2013)
51. Pons-Moll12, G., Taylor13, J., Shotton, J., Hertzmann14, A., Fitzgibbon, A.: Metric regression forests for human pose estimation. BMVC (2013)
52. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
53. Ramanan, D.: Learning to parse images of articulated bodies. Advances in neural information processing systems pp. 1129–1136 (2007)
54. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48 pp. 1060–1069 (2016)
55. Reed, S., van den Oord, A., Kalchbrenner, N., Bapst, V., Botvinick, M., de Freitas, N.: Generating interpretable images with controllable structure. ICLR (2017)
56. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. Advances in Neural Information Processing Systems pp. 217–225 (2016)

57. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. International Conference on Machine Learning pp. 1278–1286 (2014)
58. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention pp. 234–241 (2015)
59. Sapp, B., Taskar, B.: Modec: Multimodal decomposable models for human pose estimation. Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on pp. 3674–3681 (2013)
60. Shih, Y., Paris, S., Durand, F., Freeman, W.T.: Data-driven hallucination of different times of day from a single outdoor photo. ACM Transactions on Graphics (TOG) **32**(6), 200 (2013)
61. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on pp. 1297–1304 (2011)
62. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in Neural Information Processing Systems pp. 3483–3491 (2015)
63. Taylor, J., Shotton, J., Sharp, T., Fitzgibbon, A.: The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on pp. 103–110 (2012)
64. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017) (2017)
65. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: Forecasting from static images using variational autoencoders. European Conference on Computer Vision pp. 835–851 (2016)
66. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 4724–4732 (2016)
67. Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. European Conference on Computer Vision pp. 776–791 (2016)
68. Yang, J., Reed, S.E., Yang, M.H., Lee, H.: Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. Advances in Neural Information Processing Systems pp. 1099–1107 (2015)
69. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(12), 2878–2890 (2013)
70. Yi, Z., Zhang, H.R., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. ICCV pp. 2868–2876 (2017)
71. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Towards large-pose face frontalization in the wild. arXiv preprint arXiv:1704.06244 (2017)
72. Yoo, D., Kim, N., Park, S., Paek, A.S., Kweon, I.S.: Pixel-level domain transfer. European Conference on Computer Vision pp. 517–532 (2016)
73. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. European Conference on Computer Vision pp. 649–666 (2016)