International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Appearance and Pose-Conditioned Human photograph generation using Deformable GANs

Siddharth Panda1, Shivansh Charak2, Vaibhavee Gadhave3, V. Vanlalhruaitlunga4, Prof. Noshir Tarapore5

1LY B.Tech, Computer Engineering, Science & Technology, VishwakarmaUniversity, Pune.
2LY B.Tech, Computer Engineering, Science & Technology, VishwakarmaUniversity, Pune.
3LY B.Tech, Computer Engineering, Science & Technology, VishwakarmaUniversity, Pune.
4LY B.Tech, Computer Engineering, Science & Technology, VishwakarmaUniversity, Pune.
5LY Assistant Professor of Computer Engineering, Vishwakarma University, Pune

Abstract— In this study, we address the challenge of generating images of individuals based on pose and appearance data. Specifically, we take an image, xa, of an individual and a target pose, P (xb), extracted from another image, xb. We then generate a new image of the same individual in the target pose, P (xb), while preserving the visual details from xa.

To manage pixel-to-pixel misalignments caused by pose differences between P (xa) and P (xb), we incorporate deformable skip connections in our Generative Adversarial Network's generator. Additionally, we propose a nearest-neighbour loss as an alternative to the standard L1 and L2 losses to match the texture of the generated image with the target image.

Our approach demonstrates competitive quantitative and qualitative results using standard datasets and protocols recently proposed for this task. We also carry out a comprehensive evaluation using off-the-shelf person-identification (Re-id) systems trained with person-generation-based augmented data, a key application for this task.

Our experiments reveal that our Deformable GANs can significantly boost Re-id accuracy, surpassing data-augmentation techniques specifically trained using Re-identification losses.

index Terms—Conditional GAN, Image Generation, Deformable Objects, Human Pose.

The task of generating human images based on appearance and pose conditions aims to produce an image of a person based on two specific variables: (1) the appearance of a person in a given image, and (2) the pose of the same person in another image. The generation process aims to preserve appearance details (e.g., clothing colors, texture, etc.) from the first variable while performing a deformation on the person's shape (the pose) according to the second variable.

This task can be extended to generate images where the foreground, such as a deformable object like a face or body, changes due to a perspective variation or a deformable motion. The common assumption is that the object shape can be automatically extracted using a keypoint detector.

Following the publication of the pioneering work of Ma et al., there has been a rapidly growing interest in this task, as evidenced by several recent papers on this topic. This interest is likely due to the many potential application scenarios, ranging from computer-graphics-based manipulations to data augmentation for training person re-identification (Re-id) or human pose estimation systems.

However, most of the recently proposed deep-network-based generative approaches, such as Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), do not explicitly address the problem of articulated-object generation. Common conditional techniques are also included in this category.

Fig. 1: (a) An example of a "rigid" scene era task, in which the conditioning and the output photograph local systems are nicely aligned. (b) In a deformable-item generation task, the enter and output pictures aren't spatially aligned. GANs or conditional VAEs) can synthesize pics whose ap- pearances potential rely upon a few conditioning variables (e.g., a label or any other image). for example, Isola et al. [12] proposed an "image-to-image translation" for example, Isola et al. [12] proposed an "image-to-image translation"] image y represented in any other "channel" (see Fig. 1a). but, most of those methods have problems when managing massive spatial picture. as an example, the U-net structure used by Isola et al. [12] is primarily based on skip connections which help keeping local statistics. among x and y. particularly, pass connections are used to replicate afterwhich concatenate the characteristic maps of the generator "encoder" (wherein data is downsampled the use of convolutional layers) to the generator "decoder" (containing the upconvolutional layers). but, the assumption utilized in [12] is that x and y are roughly aligned with each other and that they represent the same underlying structure. This assumption is violated whilst the foreground object in y undergoes huge spatial deformations with respect to x (seeFig.

Our research focuses on the use of a masked L1 loss to generate an intermediate image conditioned on the target pose. In the second stage, another U-net based generator is trained using an adversarial loss to generate an appearance difference map, which brings the intermediate image closer to the appearance of the conditioning image.

Unlike other methods, our U-net based approach is trained end-to-end, explicitly considering pose-related spatial deformations. We propose deformable skip connections that "move" local information according to the structural deformations represented in the conditioning variables. These layers are used in our U-net based generator.

To move information according to specific spatial deformations, we first decompose the overall deformation by a set of local affine transformations involving subsets of joints. Then, we deform the convolutional feature maps of the encoder according to these transformations and use common skip connections to transfer the transformed tensors to the decoder's fusion layers.

In addition, we propose the use of a nearest-neighbour loss as an alternative to common pixel-to-pixel losses (e.g., L1 or L2 losses) typically used in conditional generative techniques. This loss has proven beneficial in generating local details (e.g., texture) similar to the target image that are not penalized due to small spatial misalignments.

This paper extends previous work in several ways. First, we present a more detailed analysis of related work, including very recently published papers dealing with pose-conditioned human image generation. Second, we demonstrate how a variation of our method can be used to introduce a third conditioning variable: the background, represented by a third input image. Third, we provide more details about our approach.

Finally, we expand the quantitative and qualitative experiments by comparing our Deformable GANs with the latest work in this area. This comparison with the state of the art is carried out using: (1) the protocols proposed by Ma et al., and (2) Re-identification based experiments. These experiments show that Deformable GANs can significantly improve the accuracy of various Re-identification systems. Conversely, most of the other state-of-the-art techniques generate new training samples that are harmful for Re-identification systems, resulting in significantly worse performance compared to a non-augmented training dataset.

Although tested on the specific human-body problem, our approach makes few human-related assumptions and can be easily extended to other domains related to the generation of highly deformable objects. Our code and our trained models are publicly available.

1 Related work

Most common deep-learning-based methods for visual content generation can be categorized as either Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs). VAEs are based on probabilistic graphical models and are trained by maximizing a lower bound of the corresponding data likelihood. GANs consist of a generator and a discriminator, which are trained simultaneously. The generator attempts to "fool" the discriminator, and the discriminator learns to differentiate between real and fake images.Isola et al. proposed a conditional GAN framework for image-to-image translation problems, where a given scene representation is "translated" into another representation. The main assumption behind this framework is that there exists a spatial correspondence between the low-level data of the conditioning and the output image. VAEs and GANs are combined to generate realistic-looking multi-view images of clothes from a single-view input image. The target view is fed to the model using a perspective label such as front or left side, and a two-stage approach (pose integration and image refinement) is adopted.Ma et al. proposed a more general approach that allows synthesizing person images in any arbitrary pose. Similar to our proposal, the input of their model is a conditioning appearance image of the person and a target new pose defined by 18 joint locations. The target pose is defined by binary maps where small circles represent the joint locations. This work has been extended by learning disentangled representations of person images. More precisely, in the generator, the pose, the foreground, and background are separately encoded to achieve a disentangled description of the image. The input image is then reconstructed by combining the three descriptors. The main advantage of this approach is that it does not require pairs of images of the same person at training time. However, the generated images consequently suffer from a lower level of realism.

Inspired by Ma et al., several methods have been recently proposed to generate human images. In these methods, the generation process. Specifically, the first generation stage generates the body pose in the new perspective. The second and third stages generate the foreground (i.e., the person) and the background, respectively.

In addition to our proposal, Balakrishnan et al. partition the human body into different parts and separately deform each of them. Their method is based on generating a set of segmentation masks, one per body part, plus a whole-body mask which separates the human figure from the background. However, for the model to segment the human figure without relying on pixel-level annotations, training is based on pairs of conditioning images with the same background. This constraint prevents the use of this technique in applications such as Re-id data augmentation where training images are usually taken in different environments.

In contrast to these methods, we demonstrate that a single-stage approach, trained end-to-end, can be used for the same task obtaining better qualitative results and that our approach can be easily used as a useful black-box for Re-identification data augmentation.

This method relies on a dense-pose estimation system that maps pixels from images to a common surface-based coordinate framework. However, as the dense-pose estimator requires training with a large-scale ground-truth dataset with manually annotated image-to-surface correspondences, it is not directly comparable with most other works, including ours, which rely on keypoint detectors trained with less human supervision.

In another approach, a VAE is used to represent the appearance and pose with two separate encoders. The appearance and pose descriptors are then combined and passed to a decoder to generate the final image. Zanfir et al. estimate the human 3D pose using meshes, identifying the mesh regions that can be directly transferred from the input image mesh to the target mesh. The missing surfaces are then filled using a color regressor trained through Euclidean loss minimization.

Typically, U-net based architectures are often used. However, standard U-net skip connections are not well-designed for large spatial deformations as local information in the input and output images is not aligned. In contrast, we propose deformable skip connections to address this misalignment problem and "commute" local information from the encoder to the decoder driven by the specific pose difference. This allows us to simultaneously generate the overall pose and the texture-level refinement.

Landmark locations are used for other generation tasks such as face synthesis. However, as the human face is considered a more rigid object than the human body, the misalignment between the input and output images is limited and high-quality images can be obtained without feature alignment.

For discriminative tasks, different architectures have been proposed to handle spatial deformations. For example, Jaderberg et al. propose a spatial transformer layer, which learns how to transform a feature map into a "canonical" view, conditioned on the feature map itself. However, this only realizes a global, parametric transformation, while in this paper we deal with non-parametric deformations of articulated objects which cannot be defined by a single global affine transformation.

Finally, our nearest-neighbour loss is similar to the perceptual loss and to the style-transfer spatial-analogy approach. However, the perceptual loss, based on an element-by-element difference computed in the feature map of an external classifier, is different from our approach.

Deformable GANs

"In our experiment, similar to the one referenced as [1], we aim to generate an image, denoted as xˆ, that depicts a person with an appearance (such as clothing) that matches a reference image, xa. However, the body pose in the generated image is intended to resemble that in another image, xb, of the same individual. The pose P(x) is represented as a sequence of k 2D points (p1, …pk) that mark the positions of the human body joints in the image. To ensure a fair comparison with [1] and other related works, we also use the same number of joints (k = 18) and employ the same Human Pose Estimator (HPE) [9] used in [1] to extract P(). It's important to note that this HPE is utilized during both testing and training phases, which means we do not rely on manually annotated poses. Consequently, the extracted joint locations might contain some localization errors or issues such as missing detections or false positives. For the training phase, we utilize a specific dataset."

$X = \{(x^{(i)}, x^{(j)})\}$ containing

The perceptual loss is primarily based on a detailed comparison conducted within the feature map of an external classifier. This comparison involves pairs of reference and target images of the same individual in varying poses. For each image pair (xa, xb), two poses P(xa) and P(xb) are derived from the respective images.

$$H_i(p) = exp\left(-\frac{\|p - p_i\|}{\sigma^2}\right)$$

The generator G is supplied with two inputs: (1) a noise vector z, which is drawn from a noise distribution and is implicitly introduced using dropout [12], and (2) a triplet (xa, Ha, Hb). It's important to note that during testing, the target pose is known, allowing for the computation of H(P(xb)). Also, the joint locations in xa and Ha are spatially aligned by design, while they differ in Hb. This is a departure from [1] and [12], where Hb is not concatenated with the other input tensors. This is because the convolutional units in G's encoder have a small receptive field that cannot capture large spatial displacements. For instance, if there's a significant movement of a body limb in xb compared to xa, this limb is represented in different locations in xa and Hb, which may be too far apart to be captured by the convolutional units' receptive field. This is particularly pronounced in the encoder's initial layers, which represent low-level information. As a result, the convolutional filters cannot process texture-level information (from xa) and the corresponding pose information (from Hb) simultaneously. Therefore, we process xa and Ha separately from Hb in the encoder. Specifically, xa and Ha are concatenated and processed using the source stream of the encoder, while Hb is processed by the target stream, without weight sharing (Fig. 2). The feature maps of the primary stream are then fused with the layer-specific feature maps of the second stream in the decoder, following a pose-driven spatial deformation performed by our deformable skip connections (see Sec. 3.1). Our discriminator network is modeled after the conditional, fully-convolutional discriminator proposed by Isola et al. [12]. In our implementation, D accepts four tensors as input: (xa, Ha, y, Hb), where either…

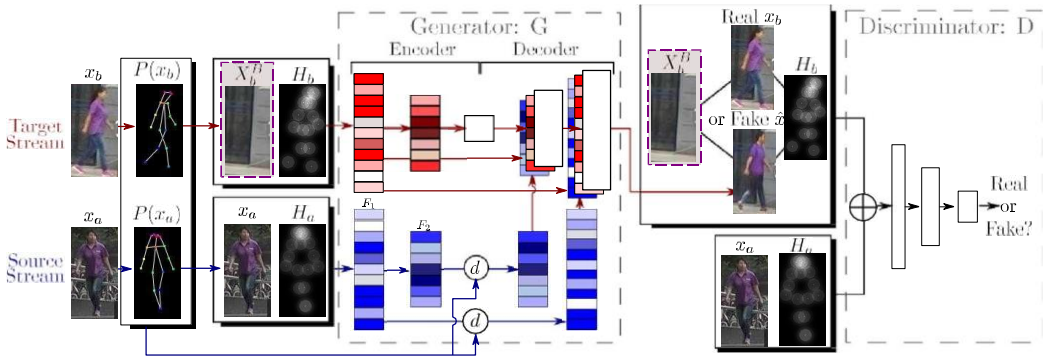

Fig. 2: A schematic representation of our network architectures

3.Training

"D and G are trained using a combination of a widespread conditional adversarial loss LcGAN and our proposed nearest-neighbor loss LNN. Specifically, in our case, LcGAN is defined as follows:

LcGAN(G,D)=E(xa,xb)∈X[logD(xa,Ha,xb,Hb)]+E(xa,xb)∈X,z∈Z [log(1−D(xa,Ha,x^,Hb))],

where

x^=G(z,xa,Ha,Hb).

Previous works on conditional GANs combine the adversarial loss with either an L2 or an L1-based loss, which is used only for G. For instance, the L1 distance computes a pixel-to-pixel difference between the generated and the real image. However, a well-known issue with using L1 and L2 is the production of blurred images. We hypothesize that this is also due to these losses' inability to tolerate minor spatial misalignments between

x^

and xb. For example, suppose that

x^

, produced by G, is visually plausible and semantically similar to xb, but the texture details on the person's clothes in the compared images are not pixel-to-pixel aligned. Both the L1 and the L2 losses will penalize this inexact pixel-level alignment."

"Despite not being semantically important from a human perspective, alignment is still considered. These misalignments do not depend on the global deformation between xa and xb, as xˆ is supposed to have the same pose as xb. To mitigate this issue, we suggest using a nearest-neighbor loss NN based on the following definition of image difference:

LNN(xˆ,xb)=p∈xˆ∑minq∈N(p)||g(xˆ(p))−g(xb(q))||,

where N(p) is a local neighborhood of point p. g(x(p)) is a vector representation of a patch around point p in image x, obtained using convolutional filters (more details below). Note that LNN () is not a metric because it is not symmetric. To effectively compute Eq. (7), we compare patches in xˆ and xb using their representation (g()) in a convolutional map of an externally trained network. In more detail, we use VGG-19 [29], trained on ImageNet and, specifically, its second convolutional layer (called conv1 2). The first convolutional maps in VGG- 19 (conv1 1 and conv1 2) are both obtained using a convolutional stride equal to 1. As a result, the feature map (Cx) of an image x in conv1 2 has the same resolution as the original image x. Leveraging this fact, we compute the nearest-neighbor field directly on conv1 2, without losing spatial precision. Subsequently, we define: g(x(p)) = Cx(p), which corresponds to the vector of all the channel values of Cx with respect to the spatial position p. Cx(p) has a receptive field of 5x5 in x, thus effectively representing a patch of dimension 5x5 using a cascade of convolutional layers interspersed by a non-linearity. Using Cx, Eq. (7) becomes:

LNN(xˆ,xb)=p∈xˆ∑minq∈N(p)||Cxˆ(p)−Cxb(q)||.

In Sec. 4.1, we demonstrate how (8) can be efficiently implemented using GPU-based parallel computing. The final LNN -based loss is:

LNN(G)=E(xa,xb)∈X,z∈ZLNN(xˆ,xb).

Combining Eq.(5) and Eq (9) we obtain our objective:

G∗=argGminDmaxLcGAN(G,D)+λLNN(G)

,

with λ = 0.01 used in all our experiments. The value of λ is small because it acts as a normalization factor in Eq. (8) with respect to the number of channels in Cx and the number of pixels in xˆ (more details in Sec.)."

4 IMPLEMENTATION DETAILS

In this stage, we provide additional technical details about our suggested approach. Firstly, we demonstrate how the proposed nearest-neighbour loss can be efficiently calculated using optimized matrix multiplications, a common technique in GPU-based programming. Secondly, we explain how the symmetry of the human body can be utilized to handle potential missing or non-detected body parts. Finally, we outline the specifics of the architectures and the training process used in our experiments.

4.1 Nearest-neighbour loss implementation

Our suggested nearest-neighbour loss is derived from LNN (xˆ, xb) as specified in Eq. (8). In this equation, for each point p in xˆ, the point q in xb that is "most similar" (in the Cx-based feature space) needs to be identified within an n x n neighborhood of p. This operation can be time-consuming if implemented using sequential computing (i.e., a "for loop"). We illustrate how this calculation can be accelerated by utilizing GPU-based parallel computing, where different tensors are processed concurrently.

Given Cxb, we compute n shifted versions of Cxb: {Cxb}, where (i, j) is a translation offset within a relative n x n neighborhood

(i, j ∈ {− n−1 , …, + n−1 }) and C(i,j) is populated with

D(i,j) = |Cxˆ − C(i,j)|

This represents the channel-by-channel absolute difference between Cxˆ(p) and Cxb (p + (i, j)). Then, for each D(i,j), we sum all of the channels where c spans all the channels and the sum is performed pointwise. S(i,j) is a matrix of scalar values, with each value representing the L1 norm of the difference between a point p in Cxˆ and a corresponding point

p + (i, j) in Cxb.

In this context, c spans all channels and the sum is computed pointwise. S(i,j) is a matrix of scalar values, each value representing the L1 norm of the difference between a point p in Cxˆ and a corresponding point p + (i, j) in Cxb:

S(i,j)(p) = ||Cxˆ(p) − Cx (p + (i, j))||1

For each point p, we can now calculate its best match in a local neighborhood of Cxb simply by using:

M (p) = min(i,j)S(i,j)(p).

Finally, Eq. (8) becomes:

LNN (xˆ, xb) =M (p). p

Since we do not normalize Eq. (12) by the number of channels nor Eq. (15) by the number of pixels, the final value LNN (xˆ, xb) is typically very high. For this reason, we use a small value λ = 0.01 in Eq. (10) when weighting LNN with respect to LcGAN.

5 EXPERIMENTS

In this section, we compare our approach with other state-of-the-art character generation methods, both qualitatively and quantitatively, and present an ablation study. Since the quantitative evaluation of generative methods remains an ongoing research challenge, we adopt various criteria, which include: (1) the evaluation protocols recommended by Ma et al., (2) human evaluations, and (3) experiments based on Re-identification training with data augmentation. It's important to note that we do not use the background conditioning information in all but the qualitative experiments shown in Section 5.6. In fact, since most of the methods we compare with do not use additional background conditioning information, we also omitted this for a fair comparison.

5.1Metrics

Evaluating generative tasks can be challenging. In our experiments, we use a variety of metrics, following the approach of 1. These include Structural Similarity (SSIM) [34], Inception Score (IS) [35], and their masked versions, mask-SSIM and mask-IS 1. The latter are obtained by masking out the image background. This is because no background information of the target image is input to G, so the network cannot predict what the target background looks like (note that we do not use background conditioning in these experiments). It's important to note that the evaluation masks we use to compute both the mask-IS and mask-SSIM values do not correspond to the mask (Mh) we use for training. The evaluation masks were constructed following the method proposed in 1 and used in that work for both training and evaluation. As a result, the mask-based metrics may be biased in favor of their method. Furthermore, we note that the IS metrics [35], based on the entropy computed over the class neurons of an external classifier [36], are not very suitable for domains with only one item class (the character class in this case). For this reason, we suggest using an additional metric that we call Detection Score (DS). In addition to the class-based metrics FCN-score, used in [12], DS is based on the detection outcome.

6.Conclusions

In this paper, we introduced a GAN-based method for generating images of individuals, conditioned on their appearance and pose. We proposed two novel concepts: deformable skip connections and nearest-neighbour loss. The former addresses common issues in U-net based generators when dealing with deformable objects, while the latter helps to mitigate the misalignment between the generated image and the ground-truth image.

Our experiments, which were based on both automated evaluation metrics and human judgments, demonstrated that our proposed method either outperforms or is on par with previous work in this task. Importantly, we showed that, unlike other popular character-generation methods, our Deformable GANs can significantly improve the accuracy of various Re-identification systems using data augmentation. The performance improvement achieved is even greater than a state-of-the-art Re-id specific data-augmentation method.

Although we tested our Deformable GANs on the specific task of human generation, we made only a few assumptions related to the human body. We believe that our concept can be easily adapted to handle other deformable-object generation tasks.

REFERENCES

[1] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Poseguided person image generation," in *NIPS*, 2017.

[2] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Videoforecasting by generating pose futures," in *ICCV*, 2017.

[3] B. Zhao, X. Wu, Z. Cheng, H. Liu, and J. Feng, "Multi-view imagegeneration from a single-view," *arXiv:1704.04886*, 2017.

[4] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditionalappearance and shape generation," in *CVPR*, 2018, pp. 8857–8866.

[5] C. Si, W. Wang, L. Wang, and T. Tan, "Multistage adversarial losses forpose-based human image synthesis," in *CVPR*, 2018, pp. 118–126.

[6] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrableperson re-identification," in *CVPR*, 2018, pp. 4099–4108.

[7] A. Siarohin, E. Sangineto, S. Lathuilie`re, and N. Sebe, "Deformable gans for pose-based human image generation," in *CVPR*, 2018.

[8] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GANimprove the person re- identification baseline in vitro," in *ICCV*, 2017.

[9] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2Dpose estimation using part affinity fields," in *CVPR*, 2017.

10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley,*ICLR*, 2014.