ELECTRONICS CLUB, IITK

3D POSE WARPING

END-TERM REPORT

INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

Acknowledgement

We are grateful to the Electronics Club and the Science and Technology Council, IITK for presenting us with this opportunity to do the project. We are grateful for all the guidance, material and help extended by them. We would like to thank our project mentors, Prakhar Maheshwari, Naiza Singla and Shivam Malhotra for their guidance.

Contents

1 Overview 1

2 Theoretical Background 2

2.1 Convolutional Neural Network 2

2.2 Latent space 3

2.3 Generative Adversarial Network 3

2.4 Our Model 4

3 Architecture 5

3.1 Generator 5

3.2 Discriminator 5

3.3 Loss Function Of GAN 6

3.4 3D warping module 6

3.5 Target Pose Encoder 6

3.6 Background inpainter 7

4 Workflow 8

4.1 Stages of Workflow 9

5 Results 10

5.1 Result 10

6 Future Prospects 12

6.1 Application in Data Augmentation 12

6.2 Application in Film-Making 12

7 Members 13

References 14

1 Overview

Being able to change a person's pose in an image gives rise to a variety of applications, from generation of large crowds or performing stunts in filmmaking to data augmentation for human-centric computer vision tasks. Some warp 2D features such that they become aligned to the target pose, which is also specified in 2D. We need to consider the actual 3D shape of the object while making changes to the object. The data is extracted into 3D voxels (a value on a grid in a 3D space). which have the all the desired information about the object and thus, these can be warped to obtain a desired pose. Warping is shuttling the voxel features to their target location. The knowledge of the 3D structure is necessary to modify the pose with accuracy. It is required to predict the depth related information from images in a volumetric representation. It is carried out by tensor reshaping operations. Once the desired pose is obtained, the data is again converted into RGB space by a decoder which result in 2D output. Mesh-based approaches fit a 3D body model to the input, infer the texture and render the mesh in the target pose. While capturing the 3D aspect, this has the downside that a specific human might not be captured well by a general model. Using only a 2D image as input, our model implicitly learns a latent volumetric representation of the input person. This representation is then warped using 3D transformations based on input and target pose to align it to the target pose. We process the warped features along with 3D target pose heatmaps with a decoder, to synthesize the reposed image.

2 Theoretical Background

We are proposing here a very novel method of reposing a given human pose into any desired pose. There have been various attempts in solving the same problem using various conventional methods. Most prior works are either based on 2D representations or require fitting and manipulating an explicit 3D body mesh. we propose to implicitly learn a dense feature volume from human images, which lends itself to simple and intuitive manipulation through explicit geometric warping. Once the latent feature volume is warped according to the desired pose change, the volume is mapped back to RGB space by a convolutional decoder

2.1 Convolutional Neural Network

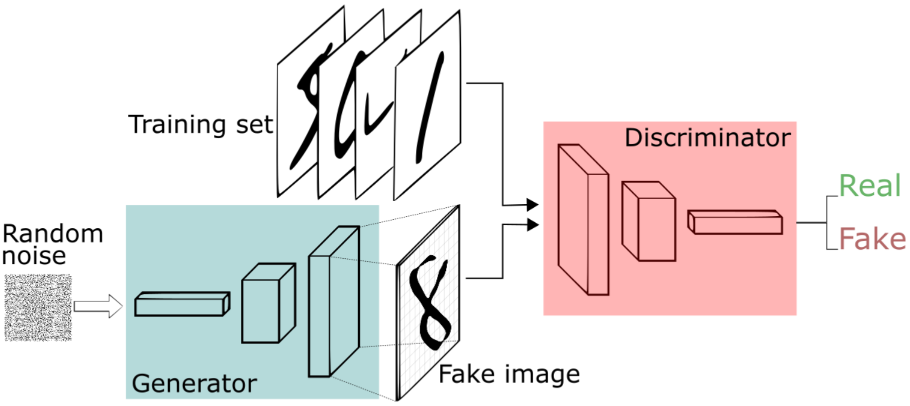
When dealing with high dimensional inputs such as images, the neurons in a layer will only be connected to a small local region of the layer before it. This mapping is done using a construct called filters. A filter is a set of trainable weights of a small dimension (e.g. 3 x 3, 5 x 5). Imagine a small filter-sized patch sliding left to right across the image from top to bottom which is elementwise multiplied with the filter, summed up, and mapped to an output image. Pooling is applied to reduce the x, y dimensions of an image by picking the maximum or average value in a window

2.2 Latent space

The latent space is simply a representation of compressed data in which similar data points are closer together in space. Latent space is useful for learning data features and for finding simpler representations of data for analysis.

2.3 Generative Adversarial Network

A generative adversarial network (GAN) is a machine learning (ML) model in which two neural networks compete with each other to become more accurate in their predictions The two neural networks that make up a GAN are referred to as the generator and the discriminator. The generator is a convolutional neural network and the discriminator is a deconvolutional neural network. The goal of the generator is to artificially manufacture outputs that could easily be mistaken for real data. The goal of the discriminator is to identify which outputs it receives have been artificially created.



2.4 Our Model

We take input image I_I of a person and a target pose P. We aim at generating an image I_T of the person in pose P. We use a two-stream generator network to tackle this problem, where the first stream reposes the person using our novel volumetric feature warping approach, while the second inpaints missing parts of the background. In the first stream, the input image is passed through **encoder** where a series of convolution layers E_{2D} apply filters with trained weights to extract different features of the input pose and along with reducing the dimension of the image. It generates $E_{2D}(I_I) \epsilon R^{H \times W \times D \cdot C}$. A reshape operation splits the channel dimension of the resulting tensor into different depth layers, yielding the feature volume $F \epsilon R^{H \times W \times D \times C}$ and 3D convolutional network (E_{3D}) is applied to yield $V \epsilon R^{H \times W \times D \times C}$.

Then it enters a latent volumetric space where our **3D warping module** reshapes it into the desired pose. Our **warping module** uses a 7-parameter Helmert transformation to achieve this task. It gets a feature volume $V \epsilon R^{H \times W \times D \times C}$, together with the 3D input and target pose $P_I, P_T \epsilon R^{(J \times 3)}$ which are given as 3D joint coordinates. The input pose P_I is used to create ten masks $M_i \epsilon \{0, 1\}^{H \times W \times D}$, one per body part, and uses in warping different body parts in volumetric space. The warped features are passed through 3D convolutions layers and then reshaped back into 2D features in **projection encoder**. The reshaped features are combined and added with the output of **background inpainter** (carries background features) to generate the required transformed pose.

3 Architecture

3.1 Generator

3.1.1 Lifting Encoder

The lifting encoder maps a 2D input image to 3D volumetric features. Bottleneck residual block is used to construct a ResNET encoder. The encoded images are then converted into 3D for the 3D warping. The images are encoded by compressing the dimension of the images dataset and changing various attributes of the image for the 3D warping. Group normalization is used instead of Batch normalization due to its better performance with batch sizes.

3.1.2 Projection Decoder

The decoder converts the 3D images back into 2D. It then decodes the images back using a ResNET model decoder consisting of residual blocks. The 3D convolutional network enhances the warped features and combines them with the output of the pose encoder. Then the output of the first network is reshaped into 2D and then a second network is applied which generates an RGB image with a mask.

3.2 Discriminator

The discriminator tries to distinguish the real data from the data created by the generator. The discriminator comprises of a convolutional and dense neural network working on the output from the generator. The discriminator connects 2 loss functions. During discriminator training the discriminator model makes a prediction and classifies images. The discriminator loss function penalizes the model misclassifying for each of the generated and real images. The weights are then updated by backpropagation through the discriminator network.

3.3 Loss Function Of GAN

The perceptual loss $\mathcal{L}_{perc}$ compares generated and target image by passing both images through an ImageNet-pretrained VGG net and computing the L_1 loss on multiple feature maps. The adversarial loss $\mathcal{L}_{adv}$ uses a discriminator net as in a classical GAN. The discriminator gets the generated or ground truth image along with the input image and the 3D target heatmap. We jointly optimize a weighted combination of these losses:

$$\mathcal{L}(\theta) = \lambda_{perc}\mathcal{L}_{perc}(\theta) + \lambda_{adv}\mathcal{L}_{adv}(\theta)$$

3.4 3D warping module

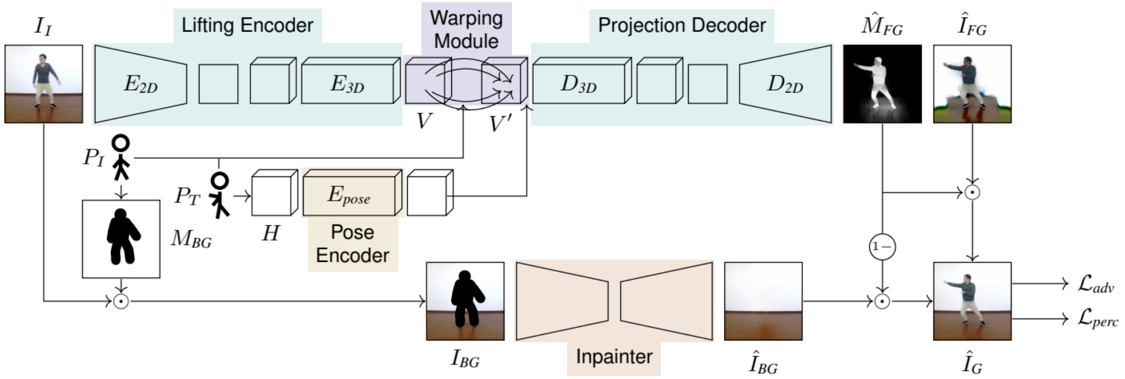
The warping module gets a 3D feature volume and a target pose as inputs. Masks are created from the input pose. The feature volume is then divided into parts and the pose is applied to each part. The module takes in as input the coordinates of the joints and of the initial mask and generates the coordinates of the masked features. Then we apply the Helmert transformation on each bodypart. Helmert transformation is a 7 parameter transformation in which the initial vector is rotated, scaled and translated along the coordinate axes to produce the transformed vector. Then warping is done on the masked bodyparts using trilinear interpolation.

3.5 Target Pose Encoder

The 3D pose encoder consists of a convolutional neural network made up of residual blocks. The encoder creates a latent space from the input pose and then voxel multiplication is applied to the 3D volume and the warped features.

3.6 Background inpainter

Background information loss occurs because the warping module only masks the bodyparts and performs transformations only to them. Background inpainter performs fills in the missing background features by using a Neural network. Any of the features that are not included in the bodypart masks are them inpainted.



4 Workflow

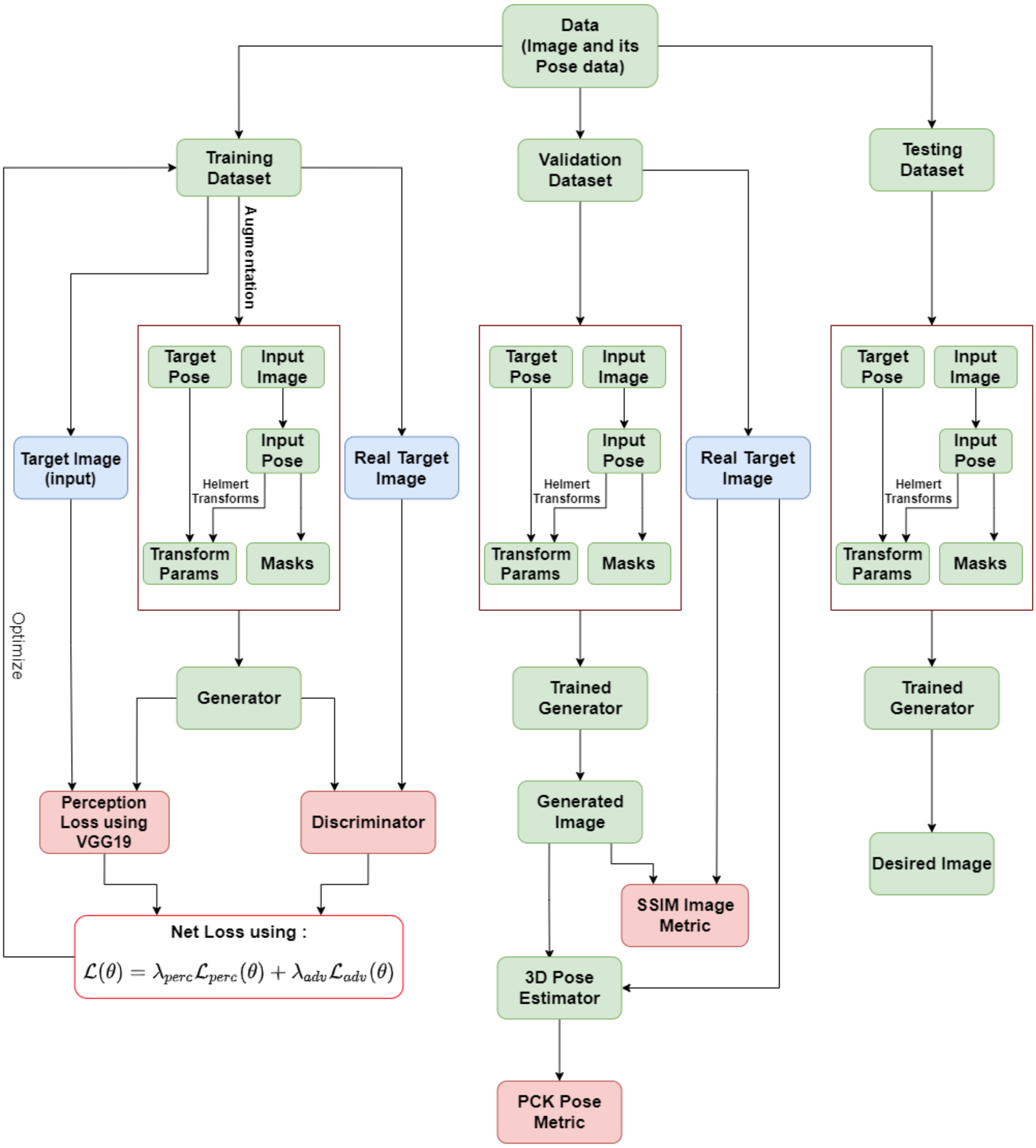


Figure 4.1: Workflow of our model

4.1 Stages of Workflow

Training Pipeline

- Create a dictionary of keys such that each image has a unique key corresponding to its path in dataset directory
- Extract the pose data of image
- We use data augmentation with rotation, scaling, translation, horizontal flip and color distortion.
- Create the masks of input data
- Estimate the helmert transform parameters for 3D warping.
- The data is then passed into the generator which generates the warped image based on its architecture.
- The perceptual loss $\mathcal{L}_{perc}$ compares generated and target image by passing both images through an ImageNet-pretrained VGG net.
- The adversarial loss $\mathcal{L}_{adv}$ is calculated using a discriminator net as in a classical GAN.
- We jointly optimize a weighted combination of these losses:

$$\mathcal{L}(\theta) = \lambda_{perc}\mathcal{L}_{perc}(\theta) + \lambda_{adv}\mathcal{L}_{adv}(\theta)$$

- The process is repeated untill the optimal result is obtained.
- We train with the Adam optimizer for 150,000 steps with batch size 2 and learning rate $\alpha = 2 \cdot 10^{-4}$. We set $\lambda_{adv} = 1, \lambda_{perc} = 3$.

5 Results

5.1 Result

We implemented a novel architecture for person reposing, which relies on 3D warping of implicitly learned volumetric features from a given 2D photograph and augmenting them to obtain the desired pose.This method is more sophisticated than augmenting 2D body features that does not require an explicit 3D human mesh model.The results indicate that volumetric representations and 3D warping are a promising way to tackle reposing and we expect that more sophisticated neural rendering techniques could further improve the results.

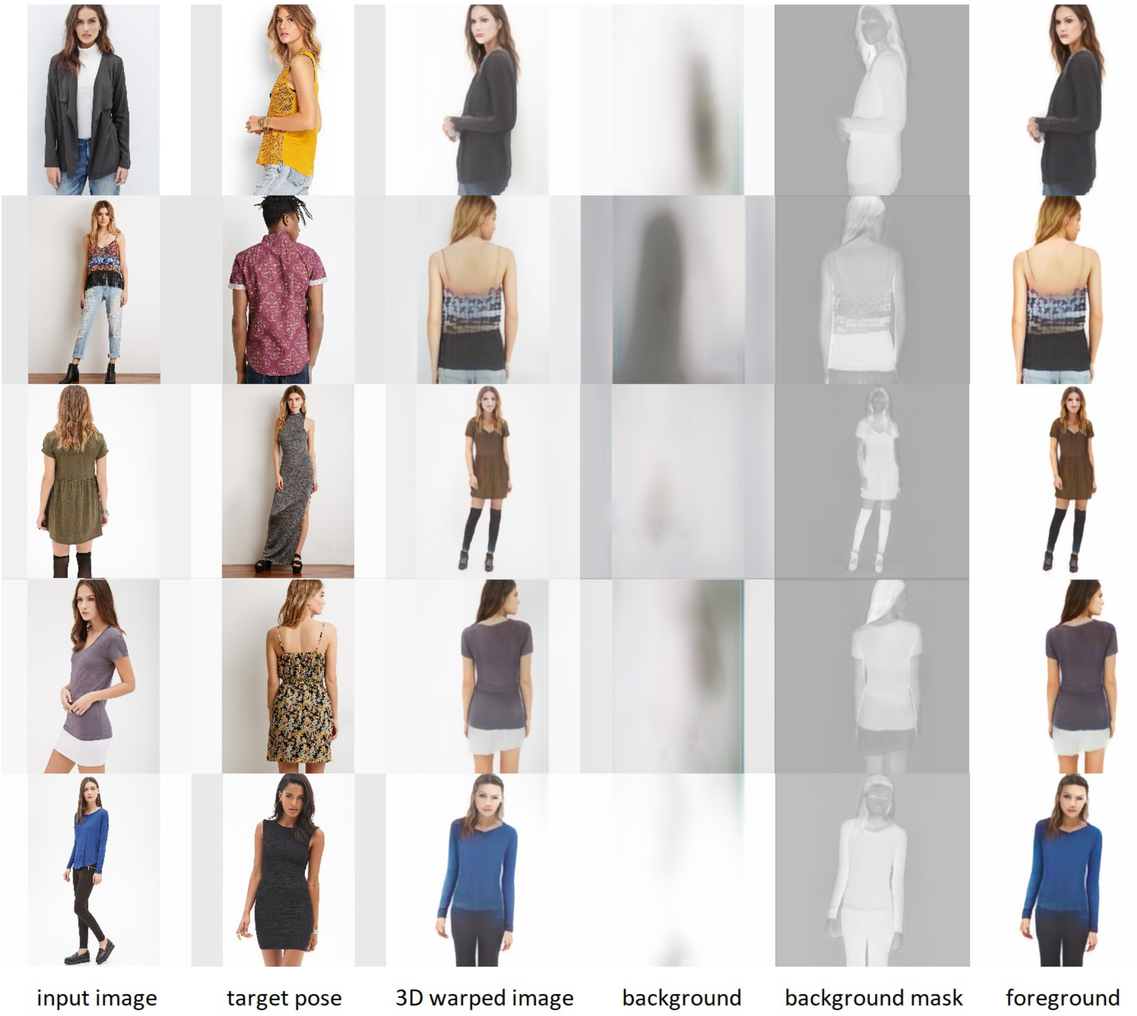


Figure 5.1: Different elements from the pipeline

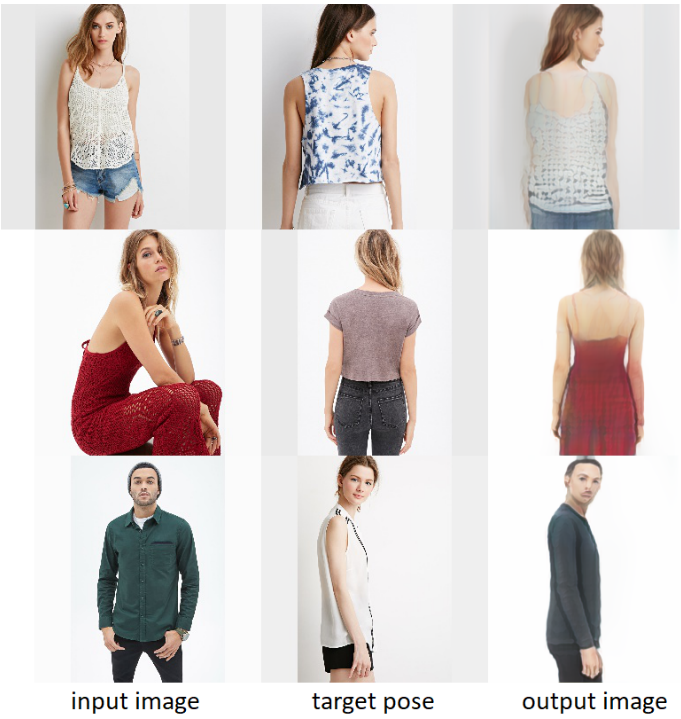


Figure 5.2: Output

Metric Scores of Our Model:

1. Computing scores on single result:
   - $Pose - loss$: 0.05913
   - $SSIM$: 0.7442
   - $SSIM_{bg}$: 0.962
   - $SSIM_{fg}$: 0.56
2. Computing scores on 100 Multiple results together:
   - $Pose - loss$: 0.03164
   - $SSIM$: 0.7293
   - $SSIM_{bg}$: 0.9346
   - $SSIM_{fg}$: 0.3464

6 Future Prospects

6.1 Application in Data Augmentation

Data Augmentation is the practice of synthesizing new data from data at hand. In all Machine Learning problems the dataset determines how well the problem can be solved. Sometimes we don't have enough data to build robust models, and what's even more common is having data with a palpable class imbalance. In cases related to lack of image dataset, 3D Pose Warping can be used to solve the problem.

6.2 Application in Film-Making

The ability to freely change a human's pose in an image opens the door for a variety of uses in the field of film making and animation. It's use may range from generating large crowds to performing difficult and complex stunts.

7 Members

1. Anant Pratap Singh
2. Atirek Aryan
3. Juee Chandrachud
4. Kaif
5. Kaushik Raj Nadar
6. Kevin Thomas Benoy
7. Maurya Jadav
8. Nikita Chauhan
9. Prateek Sogra
10. Saksham
11. Saransh Shivhare
12. Shakshi
13. Shubham Kumar
14. Vandana Basrani

References

[1] Knoche, Markus, Sárándi, István, and Leibe, Bastian. “Reposing Humans by Warping 3D Features”. In: *CVPR Workshop on Towards Human-Centric Image/Video Synthesis*. 2020.