Abstract

This is the project report for CSCI-GA.2271-001. We target human pose estimation in artistic images. For this goal, we design an end-to-end system that uses neural style transfer for pose regression. We collect a 277-style set for arbitrary style transfer and build an artistic 281-image test set. We directly run pose regression on the test set and show promising results. For pose regression, we propose a 2d-induced bone map from which pose is lifted. To help such a lifting, we additionally annotate the pseudo 3d labels of the full in-the-wild MPII dataset. Further, we append another style transfer as self supervision to improve 2d. We perform extensive ablation studies to analyze the introduced features. We also compare end-to-end with per-style training and allude to the tradeoff between style transfer and pose regression. Lastly, we generalize our model to the real-world human dataset and show its potentiality as a generic pose model. We explain the theoretical foundation in Appendix. We release code at https://github.com/strawberryfg/NAPA-NST-HPE, data, and video.

1. Introduction

The aim of human pose estimation is to locate predefined keypoints of human figures. Prior work has focused on real-world datasets. There is limited work addressing 3d pose estimation in artistic images.

Neural style transfer can effectively transfer the style of phenomenal artworks onto real images. However, to emulate the different aesthetic effects of artworks e.g. paintings, sculptures, existing style sets are not enough. Regarding our application, we build a style set covering different styles and enforce novel style transfer losses. We then empower our pose system with style transfer and attach an auxiliary style transfer in the end to further improve 2d keypoints.

To enable fine pose regression, we decouple the problem into 2d and depth using bone maps. To facilitate the learning of depth therein, we annotate pseudo 3d labels of MPII.

We are the first to showcase high-quality 2d/3d pose predictions on artwork about human figures.

We make additional contributions:

- We investigate end-to-end vs per-style training and compare stylization results. We identify the importance of instance normalization. We display more impressive stylized images using the per-style training strategy and establish a solid baseline.

- We show that our method can be extended to generic pose estimation regardless of genuineness. 1

- We perform comprehensive ablation studies to expound on all of the introduced features/components.

- We provide theoretical analysis in Appendix.

2. Related Works

Style Transfer [7] introduced neural style transfer by applying convolutional neural nets to reproduce famous painting styles on natural images. The algorithm combined the content of a natural image with the style of a famous painting by minimizing both the feature and style reconstruction losses. The algorithm produced high-quality images, but was computationally expensive as it required both forward and backward passes through a pre-trained network. [12] developed a feed-forward method of neural style transfer that is much faster than the method introduced by [7] with similar qualitative results. Our method for style transfer in this paper is based on [12].

Human Pose Estimation Early works established mathematical models e.g. exponential maps [3], shape context[22], Pfinder [29], deformable part models[5], etc. Deep learning has completely revolutionized this field[16][15][9][23], we refer to [4] for an up-to-date survey. [11] used pose in art composition for image retrieval. [19] analyzed artwork compositions via pose. The closest to our work is a very recent work [20] that used style transfer for pose estimation/image retrieval in Greek vase paintings.

1"As Apollinaire remarked, a chair will be understood as a chair from no matter what point of view it is seen if it has the essential components of a chair." [6] A human figure, whether real or artistic, has the elements to be recognized by vision algorithms.
3. Approach

We sketch the overall pipeline in Fig. 1. It consists of three parts: the neural style transfer part (green in Fig. 1), the pose regression part (red), and the self supervision part (blue). During training, the original image containing a person is presented to the image transformation network $F$, which then produces a style transferred image $O$. The following pose regression network $G$ takes only the stylized image $O$ as input and regresses the 3d keypoints $Y$. Note this is the final output we want: the 3d pose prediction. The extra self supervision part reuses the concept of neural style transfer and tries to reconstruct the stylization $O$ from the style image $S$ and the bone map $P$ which is derived from 2d keypoints $Y_{2d}$. The test stage merely maps $I$ to 3d pose $Y$ using $G$. The notation overview is in Appendix Sec. A.

Style Transfer This process stylizes the content image to emulate the artistic effect of the style target. For this stylization, we use the same image transform net $F$ as [12]. Regarding the style image $S$, we randomly select from a pool we collected (Sec. 3.1) instead of training one model per style as in [12]. We also tried the approach in [12] and detail the differences between style-specific training and arbitrary style training below in Sec. 4.4.1. Following [12] we utilize a VGG as the loss network and enforce more losses to guide the style transfer process. (Sec. 3.2, 4.4.2).

Pose Regression After transferring, we would like to generate the 3d pose (also called 3d keypoints) $Y \in \mathbb{R}^{J \times 3}$ ($J = 18$) of that human figure. The problem is decoupled into estimating 2d and lifting 2d to 3d. We extract 2d keypoints $Y_{2d}$ from $G$ [25] here. Similar to [28], we visualize the 2d in a kinematic bone map $P$ and forward it to another depth estimation network $G'$: (a simple Res50 with 1 FC regression head) to obtain depth $Y_{depth}$. Now we can piece together $Y_{2d}$ with $Y_{depth}$ to construct the 3d pose $Y$. We empirically find this decoupling and the proposed bone map benefit test cases which would otherwise cause bizarre results. (Fig. 10, Sec. 4.4.4) Losses are in Sec. 3.2.

The bone map rendering function $R$ runs as follows: For each bone in the kinematics chain [30] connecting 2 adjacent keypoints $(x_1, y_1)$ and $(x_2, y_2)$, we define the related area in the bone map as the oval (and its interior) that has $(x_2, y_2) - (x_1, y_1)$ as the minor axis $\hat{y}$, the major axis $\hat{x} \perp \hat{y}$ and bone width $d$ as the length of the major axis. $d$ effectively controls how wide we want the bone to span. The oval contains a set of loops \[ \alpha: \{ \alpha_i : [0, 1] \rightarrow R^2 \mid \hat{a}_i is the oval \} \]
\[ \left( \frac{dist((x, y), \hat{y})}{a_1^2} \right)^2 + \left( \frac{dist((x, y), \hat{x})}{a_2^2} \right)^2 = \frac{1}{2} \] for $0 \leq a_1 < a_2 < \ldots < a_i < \ldots < a_n = \frac{d}{4}$

A more formalized representation of the time-parameterized mapping between $[0, 1]$ to the bone map space $P$ is detailed in Appendix Sec. K. The bone map is finalized simply by colorizing each bone area illustrated above in a 3-channel RGB. The rendering can be better understood from the angle of loops in fundamental groups. (Appendix Sec. I.2)

Self Supervision The appended self supervision loss network corrects the predicted 2d pose by projecting it to the stylization space $O$ and measuring the difference between the reconstruction $O' \in O$ and the stylized image $O \in O$. We adopt another image transform net $F'$, which is initialized to $F$, to reconstruct the stylization given the bone map $P$ above and the style target $S$. Essentially it can be interpreted as another style transfer on the rendered bone map. We explain more details and the reason why we do not reconstruct the original image in Appendix. Sec. I.

3.1. Artistic Styles

Our pipeline supports arbitrary style transfer, which is crucial for separating pose from nuisances. Previous style transfer works only have a few most recognized paintings from most well-known painters 2, which is limited. We build a style dataset of 277 images covering different styles, which we collect from museums, galleries and books 4. We use styles ranging from Impressionism, Cubism, Fauvism, to Romanesque, Futurism, Surrealism, Expressionism, Kinetic Art etc. Some examples are in Appendix Sec. C. We also show some failure cases under this page.

3.2. Losses

Neural Style Transfer Loss

- $L_{style}(O, S)$: It is the squared Frobenius norm of the difference between the Gram matrices of the style transferred image and the style image [12].
- $L_{real}(O, I)$: The feature reconstruction loss between content and stylized image in [12].
- $L_{tv}(O)$: Total variation loss as in [12].

Note the above are used in the perceptual loss [12] paper. Besides, we propose the following:

- $L_{sent}(O, S)$: Jensen Shannon entropy loss between features from stylized image $O$ and style image $S$ at layers relu1-1, relu1-2 and relu2-1.
- $L_{rgb}(O_{rgb}, S_{rgb})$: Let $O_{rgb}, S_{rgb}$ be the rgb format of transferred image and style image respectively, the loss is the cosine similarity of these two vectors $O_{rgb}$ and $S_{rgb}$. Note losses on other color spaces [8] might further help, we leave that for future work.

---

2 "The Starry Night" by Vincent Van Gogh, "Woman with a Hat" by Henri Matisse, "Ma Jolie" by Pablo Picasso etc.
3 https://drive.google.com/drive/folders/1zNm27vXZa_lwHQzTrrSb7NIIiGw_38Xc?usp=sharing
4 MoMA, SFMOMA, Whitney Museum, Guggenheim Museum, MoCA Shanghai, NMC Beijing etc.; Lisson Gallery, Kasmin Gallery etc.; "steampunk the beginning", "Masterpieces of impressionism & post-impressionism", "Expressionism", "Gothic art", "Cubism" etc.
Figure 1. Diagram of the pipeline. During training, the neural style transfer (green) part stylizes the human image and sends the stylization to the pose regression part (red), which outputs the final 3d pose prediction as desired. The self supervision part (blue) aims to improve only the 2d keypoints and can be treated as additional supervision. During testing, the original art image is directly sent to the pose regression part which predicts the 3d pose. The style transfer and self supervision parts are completely discarded.

\[ L_{\text{hsv}}(O, S) \]: This is the HSV loss between style transferred image \( O \) and style image \( S \). Let the hue, saturation and vibrance of style image be \( h_s, s_s, v_s \) and the counterparts of the transferred image be \( h_o, s_o, v_o \). 
\[
L_{\text{hsv}} = \|h_s - h_o\|_1 + \|s_s - s_o\|_1 + \|v_s - v_o\|_1.
\]

\[ L_{\text{cos}}(O, S) \]: Defined as the cosine similarity between features from style image \( S \) and transferred image \( O \) at \text{layer relu4.2}.

\[ L_{\text{content}}(O, I) \]: The L2 loss between features from content image \( I \) and stylized image \( O \) at layer \text{relu4.2}.

\[ L_{\text{cent}}(O, I) \]: Written as the Jensen Shannon entropy loss between features from content image and stylized image at layers \text{relu5.2, relu5.3, relu5.4}.

To summarize, \( L_{\text{sent}}, L_{\text{argb}}, L_{\text{hsv}} \) and \( L_{\text{cos}} \) relate to style reconstruction loss while \( L_{\text{content}} \) and \( L_{\text{cent}} \) represent the content reconstruction loss. Practically we observe that among the additional losses apart from those in [12], the following are the most important: \( L_{\text{hsv}}, L_{\text{cos}} \) and \( L_{\text{argb}} \). We will see this afterwards in Sec. 4.4.2.

**Pose Regression Loss** As said earlier, we subscribe to the view that lifting from 2d to 3d is better than directly regressing 3d when a generic pose model is required (Sec. 4.5), and so the losses are on 2d and depth respectively.

\[ L_{2d}: \] Integral loss guides the learning of \( G \) and 2d keypoints prediction \( Y_{2d} \) (Fig. 1). It was first put forth in [25].

\[ L_{\text{depth}}: \] We use the orientation representation [18][24][27] and define an euclidean loss between predicted bone vector and ground truth bone vector. This loss forces the model to implicitly learn the depth of keypoints \( Y_{\text{depth}} \) by transforming bone vector to joint location. Bone vector is standardized (using mean and std) to facilitate the learning.

**Self Supervision Loss** The self supervision is a small “gadget” to refine the 2d by comparing the induced bone map and the stylized image in the stylization space. To initiate the comparison, we adapt the ones in neural style transfer loss above and define the counterparts:

\[ L_{\text{style-sup}}(O', O) \]: Similar to \( L_{\text{style}} \) except that the difference is between the stylized image: \( O \) and the reconstruction of the stylized image: \( O' \).

\[ L_{\text{cos-sup}}(O', O) \]: The cosine similarity between \( O \) and its reconstruction \( O' \). It is similar to \( L_{\text{cos}} \).

\[ L_{\text{feat-sup}}(O', O) \]: Now the feature to be reconstructed
is O. It is defined at the content layer relu4_2.

**Ratios** The learning is all about balancing, most importantly the balancing between the neural style transfer and the pose regression. This is critical for the end-to-end training mechanism in our pipeline, which is not needed for a separate training alternative (Sec. 4.4.1). We empirically set the ratios during training. See Appendix Sec. B.

### 3.3. Training

The entire pipeline includes a lot of prevalent network architectures. Direct end-to-end training from scratch does not work as there is a slew of losses to balance. In our exploration, we follow a 4-stage training protocol:

- **Stage 1** only learns the style transfer network $\mathcal{F}$ with VGG loss network tunable.

- **Stage 2** only targets at regressing the pose. We warm up the $\mathcal{G}(2d)$ with\footnote{https://github.com/strawberryfg/NAPA-NST-HPE/tree/main/train/per-style-training}, and $\mathcal{G}'(\text{depth})$ with ResNet50. Images are stylized using weights learned from stage 1.

- **Stage 3** fixes the VGG weight and jointly trains style transfer ($\mathcal{F}$) and pose regression ($\mathcal{G}$ & $\mathcal{G}'$).

- **Stage 4** adds self supervision loss training. Another style transfer network $\mathcal{F}'$ and another VGG are trained.

We use Human3.6M\cite{10}, MPII\cite{2} in training, same as \cite{24}\cite{25}. Concerning the pose regression part, we additionally annotate the pseudo 3d ground truth of the entire MPII \footnote{https://github.com/strawberryfg/NAPA-NST-HPE/tree/main/annotation_tools/mpii_annotator} by manually revising the model-optimized 3d while preserving projection & certain anthropometric constraints e.g. joint angle limits\cite{1}. This way, the information of various athletic 3d poses in MPII is retained and best utilized, which is underexplored in prior works. (Sec. 4.4.3, Appendix Sec. E) About the self supervision, we also make use of the vast amount of human images in COCO \cite{17}(train2017/val2017). All images are resized to $224 \times 224$. No data augmentation is enabled. Training is mostly done on an 8GB GTX 1070 or an 8GB RTX 2080 with batch size = 2, 3 or 22 depending on the stage. Learning rate is initially 0.0001, which is then decreased upon loss plateau or increment of the stage. We opt for RMSProp with a weight decay of 0.00001 and momentum set to 0.9.

### 4. Experiments

**Evaluation Dataset** We create a dataset of 281 images that contains artistic style human figures, most of which come from Museum of Modern Art, steampunk: the beginning, Gothic Art and Masterpieces of Impressionism and Post Impressionism: The Annenberg Collection. We develop an annotation tool to annotate the 2d labels. Following the standard, we report results using the PCKh metric that measures the percentage of joints for which the prediction is within a certain threshold to the ground truth. The threshold is set to 25% of the head size throughout the paper. Note this is a strict threshold that requires high-precision estimation. Regarding the 3d pose, since it’s infeasible to obtain it without wearable IMUs or marker-based MoCap systems, we only depict qualitative visualization.

#### 4.1. Baseline

We perform per-style training using the style transfer network and methodology from \cite{12} as a baseline for our neural style transfer to compare against the end-to-end training. For the per-style training, we used the MPII dataset. The images are cropped to focus on the human figures and then resized to $256 \times 256$. Training was done on K80 Tesla with a batch size of 4. We used Adam as the optimizer with a learning rate of 0.001. We train a style transfer network for each style and then the MPII dataset is fed through the trained network to generate a style transferred MPII dataset. This style transferred dataset is then used to train a pose regression model for the specific style. We will discuss the differences between the results from the per-style training and end-to-end training in Sec. 4.4.1.

#### 4.1.1 Instance Normalization

The image transformation net from \cite{12} that we used has batch normalization layers after the residual convolution layers. Instance normalization has been shown to generate better qualitative images for fast style transfer than batch normalization with improved convergence speeds. \cite{26} Instance normalization differs from batch normalization in that every image is normalized separately in contrast to in a batch for batch normalization. As we can see from Fig. 2, instance normalization provides for better style transfer than batch normalization as the batch normalization image is mostly the same color and has artifacts around the edges.

#### 4.2. Visualization

Fig. 3 and Fig. 4 demonstrate style transferred images on MPII. More details are in Sec. 4.4.1 and here \footnote{https://github.com/strawberryfg/NAPA-NST-HPE/tree/main/train/per-style-training}.

We show 2d keypoint's prediction in Fig. 5. We simultaneously present regressed 2d and 3d poses in Fig. 6, where 2d poses are plotted in the small windows. We can see that the 2d and 3d predictions are reasonable.

#### 4.3. State-of-the-art Comparison

There is no existing work tackling estimation of 3d poses in art images that include a human figure. In this part, we apply top-performing methods trained on real-world datasets to our test set. Results are visualized in Fig. 7. One can easily spot misplaced joints of these methods, which
Instance normalization improves the production quality.

may be attributed to different image domains, lack of diverse pose labels (Appendix Sec. E).

4.4. Ablation

4.4.1 End-to-end Training vs Per-style Training

Per-style training produces more pleasing stylization images than joint-training. In Fig. 3, the left 3 columns correspond to this style-specific training. However, in terms of pose accuracy, end-to-end performs better. To reflect this, we evaluate PCKh@0.25 and draw Table. 4.4.1.

4.4.2 Style Transfer Losses

Continuing the investigation on style transfer, we now move on to the extra losses introduced. As seen in Fig. 8, adding HSV loss gears the stylized image towards a warm tone, consistent with the style image. The cosine loss encourages the stylized image to be perceptually similar to the style image in high-level layers but does not necessarily require an exact match. Comparing images in another color space sRGB further boosts the stylized image quality.

4.4.3 Pseudo 3d Ground Truth

The pose regression part starts to analyze the pose after the stylization is done. As pointed out, to extend the 3d pose variety (Appendix Sec. E) we manually annotate the pseudo
3d labels of the full MPII. To signify its importance, we compare the predictions trained with and without these labels in Fig. 9. In row 1, the left arm is closer to the camera than the right one, which is correctly captured by the model trained with pseudo 3d labels of MPII. In row 2, the model preserves the lunge pose with the aid of pseudo 3d labels. Further, in row 3 the arm prediction without using pseudo 3d labels is arbitrary, whereas the one with pseudo 3d labels is only incognizant of the forward-leaning left arm.

4.4.4 Bone Map

Extending the discussion on pose regression, here we try to answer how the proposed bone map benefits the depth estimation. Fig. 10 exhibits the pose with and without bone map. It is not hard to observe the mistakes are fixed.
Figure 7. Comparison with state-of-the-art methods. I2L: [21], Integral: [25], HMR: [13], SPIN: [14]. Ours: our full pipeline.

| Joint   | w/o style | 1 style | 3 styles | 5 styles | 9 styles | 12 styles | ours |
|---------|-----------|---------|----------|----------|----------|-----------|------|
| Ankle   | 21.4      | 21.5    | 24.7     | 23.8     | 24.6     | 26.0      | 31.0 |
| Knee    | 23.0      | 33.5    | 32.4     | 31.3     | 32.0     | 31.0      | 25.3 |
| Wrist   | 30.9      | 30.1    | 31.7     | 32.4     | 32.4     | 33.4      | 37.7 |
| Elbow   | 33.1      | 33.7    | 34.0     | 35.7     | 35.2     | 35.9      | 44.4 |
| Shoulder| 41.0      | 43.0    | 43.0     | 43.7     | 42.5     | 41.2      | 55.8 |
| Head    | 52.9      | 51.2    | 55.1     | 55.3     | 55.6     | 55.6      | 58.9 |
| Hip     | 19.4      | 21.7    | 21.2     | 22.1     | 23.0     | 23.5      | 25.8 |
| Total   | 34.5      | 35.3    | 36.9     | 37.4     | 37.6     | 37.9      | 44.4 |

Table 1. PCKh@0.25 (%) (the higher, the better) of style-specific training vs arbitrary style joint training. la muse: test the model trained with only style la.muse. N styles: average the predictions of N style-specific models on the test set. ours: the full end-to-end pipeline.

Figure 8. Stylized image using different style losses. Adding more improves the production quality. Gatys: [7].

with the usage of bone map.

4.4.5 Self Supervision

The self supervision explicitly improves the 2d pose, which we measure by comparing the PCKh@0.25 of the model Figure 9. Visualization of 3d pose predictions with and without the pseudo 3d ground truth of MPII.
with and without self supervision. Looking at Table 2, it is most effective for improving upper body joints i.e. elbow, wrist, shoulder and head, which is understandable since upper body joints are generally easier to estimate. Be aware of that PCKh@0.25 is a stricter threshold than the commonly used PCKh@0.5, but the following is true: self supervision leads to non-trivial improvements. Appendix Sec. I.2 reflects the property of the bone map space $P$ and expresses the bone map from fundamental groups.

| Joint    | w/o self_sup | w/ self_sup
|----------|--------------|------------|
| Ankle    | 28.9         | 29.4       |
| Knee     | 30.7         | 31.1       |
| Wrist    | 34.7         | 38.7       |
| Elbow    | 41.3         | 44.4       |
| Shoulder | 48.1         | 49.5       |
| Head     | 58.3         | 60.2       |
| Hip      | 27.0         | 27.3       |
| Total    | 41.7         | 43.0       |

Table 2. PCKh@0.25 (%) between the model trained with and without self supervision. The higher, the better.

4.5. Cross-dataset Generalization

In this section we will take an excursion into generalizing to another domain: real-world images. The main focus of the project is to evaluate on artistic style human images, yet we would also like to see how it will adapt to real human figures. For this purpose, we apply the model trained exclusively on styled images to H36M. The MPJPE (w/ Procrustes Alignment) is 82.91 mm. It is certainly much worse than state-of-the-art works, but do note that H36M dataset is not used anywhere for this setting. Recall in Sec. 4.3 we apply methods trained on real images to our artistic test set. One interesting question to ask then would be, what would happen if we include both real images and artistic images in training? We adjust the ratio between real and art images in training pose network $G$, which means a random portion of $O$ will be replaced with the original RGB $I$. We load weights of the main system and fine tune from there. (Appendix Sec. H) In Table 4.5, we see naturally using more real images leads to better accuracy on H36M. Yet we find the result on our test set is only slightly improved. Since the depth is estimated from the bone map by decoupling, we do not see a significant boost. However, having said that, including images from both real and artistic domain would be conducive if we demand a generic pose model that works on any images containing a human. (See Appendix Sec. G)

5. Conclusion

We present the first method to estimate 3d human poses in artworks. We show considerably well results in varied cases. Key to our approach is the customization of neural style transfer and the exploration of pose regression. We exploit neural style transfer to (a) stylize real images for pose regression (b) self supervise 2d pose predictions. We create a solid baseline using per-style training with instance normalization and exhibit notable stylized image collections. Notice this already gives strong pose results. For further improvement, we build an end-to-end system featuring novel style losses, self supervision, bone map. We infuse the system with a self-collected style set and self-annotated pseudo 3d labels of MPII. We build a test set for evaluation and ablation. Along the way, we elevate the system to a generic pose model with real/art balancing. Theory comes up next.
References

[1] Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1446–1455, 2015. 4

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3686–3693, 2014. 4

[3] Christoph Bregler and Jitendra Malik. Tracking people with twists and exponential maps. In Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231), pages 8–15. IEEE, 1998. 1

[4] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. Computer Vision and Image Understanding, 192:102897, 2020. 1

[5] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In 2008 IEEE conference on computer vision and pattern recognition, pages 1–8. IEEE, 2008. 1

[6] E.F. Fry. Cubism. World of art. Oxford University Press, 1978. 1

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2414–2423, 2016. 1, 7

[8] Shreyank N. Gowda and Chun Yuan. Colornet: Investigating the importance of color spaces for image classification. CoRR, abs/1902.00267, 2019. 2

[9] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10905–10914, 2019. 1

[10] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7):1325–1339, jul 2014. 4

[11] Tomas Jenicek and Ondrej Chum. Linking art through human poses. In 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 1338–1345. IEEE, 2019. 1

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision, pages 694–711. Springer, 2016. 1, 2, 3, 4

[13] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7122–7131, 2018. 7

[14] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In Proceedings of the IEEE International Conference on Computer Vision, pages 2252–2261, 2019. 7

[15] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6152–6162, 2020. 1

[16] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6173–6183, 2020. 1

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014. 4

[18] Chenzhu Luo, Xiao Chu, and Alan Yuille. Orinet: A fully convolutional network for 3d human pose estimation. arXiv preprint arXiv:1811.04989, 2018. 3

[19] Prathmesh Madhu, Tilman Marquart, Ronak Kosti, Peter Bell, Andreas Maier, and Vincent Christlein. Understanding compositional structures in art historical images using pose and gaze priors. arXiv preprint arXiv:2009.03807, 2020. 1

[20] Prathmesh Madhu, Angel Villar-Corrales, Ronak Kosti, Torsten Bendschus, Corinna Reinhardt, Peter Bell, Andreas Maier, and Vincent Christlein. Enhancing human pose estimation in ancient vase paintings via perceptually-grounded style transfer learning. 2020. 1

[21] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-voxel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. arXiv preprint arXiv:2008.03713, 2020. 7

[22] Greg Mori and Jitendra Malik. Estimating human body configurations using shape context matching. In European conference on computer vision, pages 666–680. Springer, 2002. 1

[23] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In 2018 international conference on 3d vision (3DV), pages 484–494. IEEE, 2018. 1

[24] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In Proceedings of the IEEE International Conference on Computer Vision, pages 2602–2611, 2017. 3, 4

[25] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In Proceedings of the European Conference on Computer Vision (ECCV), pages 529–545, 2018. 2, 3, 4, 7

[26] Dmitriy Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022, 2016. 4

[27] Qingfu Wan, Weichao Qiu, and Alan L Yuille. Patch-based 3d human pose refinement. In CVPRW, 2019. 3
[28] Qingfu Wan, Wei Zhang, and Xiangyang Xue. Deepskeleton: Skeleton map for 3d human pose regression. \textit{arXiv preprint arXiv:1711.10796}, 2017. 2

[29] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Pfinger: Real-time tracking of the human body. \textit{IEEE Transactions on pattern analysis and machine intelligence}, 19(7):780–785, 1997. 1

[30] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In \textit{European Conference on Computer Vision}, pages 186–201. Springer, 2016. 2
Appendix

A. Notations

Just to clarify the notations used in the main paper, in Table 1 we list them. We define $I$ as the original RGB image. ($I \in I$, $I$ is the space that contains $I$). $S \in S$ as the style image. $O \in O$ as the stylized output of the image transform network. During training, the $O$ is forwarded to the pose network to generate a 3D pose prediction output $Y \in Y$. (the corresponding pose label is referred to as $\hat{Y}$). For the inference phase, $I$ is passed to generate $Y$. More generally, the pose network $G : X \rightarrow Y$. ($X = I \cup O$).

| Symbol | Meaning |
|--------|---------|
| $I$    | the original RGB |
| $S$    | the style target |
| $O$    | the style transferred image |
| $O'$   | the reconstruction of $O$ via self supervision part |
| $Y$    | the ground truth 3d pose |
| $Y_{2d}$ | the predicted 2d pose |
| $Y_{2d}$ | the ground truth 2d pose |
| $P$    | the rendered kinematic bone map from $Y_{2d}$ |
| $Y_{depth}$ | the predicted joint depth |
| $I$    | the space where $I$ sits inside |
| $S$    | the space that contains $S$ |
| $O$    | the space of $O$ |
| $\mathcal{X}$ | the space of both real and artistic images |
| $\mathcal{Y}$ | the space of poses |
| $P$    | the space encapsulating $P$ |
| $F$    | the image transform net to stylize $I$ |
| $G$    | the pose network that obtains $Y_{2d}$ |
| $R$    | the parameter-free function to render $P$ |
| $G'$   | the pose network to deduce depth from $P$ |
| $F'$   | the image transform net to reconstruct $O$ |

Table 1. Notation table

B. Loss Ratios

To train such a large network, balancing the losses is of great essence so that not a single one is dominating. We list the ratios in Table. 2.

| Loss Type                   | Loss Name                  | Ratio  |
|-----------------------------|----------------------------|--------|
| Neural Style Transfer Loss  | $\mathcal{L}_{style}(O, S)$ | 1.0    |
|                            | $\mathcal{L}_{tv}(O)$      | 1.0    |
|                            | $\mathcal{L}_{sent}(O, S)$ | $10^9$ |
|                            | $\mathcal{L}_{srgb}(O_{srgb}, S_{srgb})$ | 200    |
|                            | $\mathcal{L}_{hae}(O, S)$  | 300    |
|                            | $\mathcal{L}_{cos}(O, S)$  | 1000   |
|                            | $\mathcal{L}_{content}(O, I)$ | 1.0    |
|                            | $\mathcal{L}_{cent}(O, I)$ | $10^6$ |
| Pose Regression Loss        | $\mathcal{L}_{2d}$         | 1.0    |
|                            | $\mathcal{L}_{depth}$      | 1000   |
| Self Supervision Loss       | $\mathcal{L}_{style, sup}(O', O)$ | 0.0035 |
|                            | $\mathcal{L}_{cos, sup}(O', O)$ | 0.0035 |
|                            | $\mathcal{L}_{feat, sup}(O', O)$ | 0.0035 |

Table 2. Loss ratio table.

C. Styles

We show some sample images for different styles in Fig. 1. Please refer to Google Drive for a full view.

D. Preliminaries

Before analyzing the test cases, we borrow some preliminaries from topology.

**Definition D.1** A topological space is an ordered pair $(\mathcal{X}, T_{\mathcal{X}})$ where $\mathcal{X}$ is a set (of points), $T_{\mathcal{X}}$ is a collection (family) of subsets satisfying the three following:

1. $\emptyset, \mathcal{X}$ in $T_{\mathcal{X}}$ are open.
2. Any (infinite) union of set in $T_{\mathcal{X}}$ is in $T_{\mathcal{X}}$: If $\{U_{\alpha} | \alpha \in \mathcal{I} \}$ are open sets, then $\cup_{\alpha} U_{\alpha}$ is open.
3. Any intersection of 2 sets (finite intersection) of open sets are open: If $U, V$ are open in $T_{\mathcal{X}}$, then $U \cap V$ is open.

$T_{\mathcal{X}}$ is the topology on $\mathcal{X}$, and with $\mathcal{X}$ form the topological space. Usually $T_{\mathcal{X}}$ is dropped and we will refer to $\mathcal{X}$ as the topological space.
**Definition D.2** \( G : \mathcal{X} \rightarrow \mathcal{Y} \) is continuous if and only if \( G^{-1}(\text{any open subset in } \mathcal{Y}) \) is also open in \( \mathcal{X} \).

**Definition D.3** The relative topology is defined as follows: if \( \mathcal{X} \) has a topological space, and \( \mathcal{A} \subset \mathcal{X} \). \( \mathcal{X} \) has the topology \( \mathcal{T}_x \) (a collection of open sets), then \( \mathcal{A} \) has the relative topology \( \mathcal{T}_A = \{ \mathcal{A} \cap U | U \in \mathcal{T}_x \} \).

**Definition D.4** A subset \( \mathcal{A} \subset \mathcal{X} \) of a topological space is called compact if whenever \( \mathcal{A} \subset \cup_{\alpha} U_\alpha \) with \( U_\alpha \) open in \( \mathcal{X} \), then \( \mathcal{A} \subset \cup_{\alpha=1}^{N} U_\alpha \) (finite subcover property).

**Definition D.5** A path \( \alpha : [0, 1] \rightarrow \mathcal{Y} \) is a time-parameterized mapping (a function; a "parameterized curve") between the interval \( I = [0, 1] \) (note this \( I \) is different from the original RGB I) and the space \( \mathcal{Y} \).

**Definition D.6** A path \( \alpha : [0, 1] \rightarrow \mathcal{Y} \) is said to be a loop if \( \alpha(0) = \alpha(1) \).

**Definition D.7** A path \( \alpha : [0, 1] \rightarrow \mathcal{Y} \) is homotopic to \( \beta : [0, 1] \rightarrow \mathcal{Y} \) written as \( \alpha \sim_h \beta \) if they have the same end points: \( \alpha(0) = \beta(0) \) and \( \alpha(1) = \beta(1) \).

**E. Test Case Analysis 1: Generalize to Artistic Style Images**

When we say \( \text{Img}_1, \text{Img}_2 \) are similar in a space \( \mathcal{X} \) within metric space \( (\mathcal{X}, d) \) for some unknown metric \( d \), it is formalized as \( d(\text{Img}_1, \text{Img}_2) < r \) for some \( r \) (pre-defined by as the acceptable range radius). This leads us to the open ball concept in metric space \( B_r(\text{Img}_1) = \{ \text{Img}_a \in \mathcal{X} | d(\text{Img}_1, \text{Img}_a) < r \} \) is the open ball of radius \( r \) about \( \text{Img}_1 \) in \( \mathcal{X} \) for some unknown distance metric \( d \) (euclidean, cosine similarity, maximum mean discrepancy etc.). All the images (as individual points inside \( \mathcal{X} \) in \( B_r(\text{Img}_1) \) is \( \text{Img}_1 \)'s neighborhood. Let an image \( \text{Img}_1 \) with its corresponding pose output \( Y_1 \), and another image \( \text{Img}_2 : \text{Img}_1 + \Delta \) with its pose output \( Y_2 \). If \( \Delta \) is relatively small, namely \( \text{Img}_2 \) is \( \text{Img}_1 \)'s neighbor \( \text{Img}_2 \in B_r(\text{Img}_1) \), \( Y_2 \) would also be close to \( Y_1 \): \( Y_2 \in B_r(Y_1) \).

However, it introduces 2 unknown variables: \( r \) and distance metric \( d \).

This can be abstracted in topology as follows: write \( N(\text{Img}_1) \in \mathcal{T} \) (\( \mathcal{T} \) is the topology) as its neighborhood. \( N(\text{Img}_1) \) is the open set containing image \( \text{Img}_1 \). The description bypasses the need to use \( r \) and \( d \), which would later be convenient.

Consider the case where the model performs well on the pose prediction. The ground truth 3D pose (unknown) is \( Y_1 \). Let \( \mathcal{A} \) be an open subset in \( \mathcal{Y} \) containing \( Y_1 \); the neighborhood of the unknown pose label \( Y_1 \). \( \mathcal{A} \) can be conceptually interpreted as the acceptable (reasonable) range without rigorous examination of the accuracy. The fact pose prediction is good tells us \( G(I_1) \subset \mathcal{A} \), \( I_1 \subset G^{-1}(\mathcal{A}) \). \( G^{-1}(\mathcal{A}) \)

![Figure 1. Style examples.](image-url)
is the open set covering the images that produce reasonable pose predictions within the acceptable open set “range” \( A \). \( G^{-1}(A) \) is open in \( X \) be definition of \( G : X \rightarrow Y \) being continuous. Intuitively this means \( I_1 \) is of some distance \( \epsilon > 0 \) to a point in \( G^{-1}(A) \). (Note we have not yet formally defined what the distance is, but it does not affect the illustration).

We write the following 2 conditions:

(a) The pose of \( I_1 \) is of some distance \( \epsilon_{\text{pose}} > 0 \) to a pose in \( G^{-1}(A) \).

(b) The style of \( I_1 \) is of some distance \( \epsilon_{\text{style}} > 0 \) to a style in \( G^{-1}(A) \).

Meeting the condition

When both (a) and (b) are met, it naturally follows the feature of \( I_1 \) is of some distance \( \epsilon > 0 \) to the feature of a point in \( G^{-1}(A) \). Following this, \( I_1 \) is close to some point in \( G^{-1}(A) \).

To meet (a), the literature generally resorts to the following:

1. Jointly train 2D outdoor dataset with only 2D annotation and indoor 3D dataset with limited pose variety compared to MPII.
2. Use synthetic images which also don’t exhibit the pose variety of 2D dataset e.g. MPII.
3. Use IMU in the wild but the poses are limited compared to athletic poses presented in MPII e.g. diving, skiing, swimming, surfing etc.

Models trained using either way fail on some extreme and rare 3D pose scenarios of an in-the-wild 2D image, though they have achieved relatively high overall accuracy on most images. Since we annotate the pseudo ground truth of MPII, (a) is not an issue.

If condition (b) is not satisfied, when we restrict the discussion of close in the space \( X \), it cannot be said that \( I_1 \) is in the neighborhood of \( G^{-1}(A) \). Extending the discussion range to deeper layers of the conv net, if the feature of \( I_1 \) is of some \( \epsilon > 0 \) to the feature of a point in \( G^{-1}(A) \) starting from layer index 1, we say \( I_1 \) is close to some point in \( G^{-1}(A) \).

Corollary E.0.1 This means the convolution network is capable of completely separate the human pose content from the style (background) arguably.

F. Test Case Analysis 2: Generalize to the Original Training Images

Note the training phase refrains from making use of the original training images. (See Sec. A). In this scenario, training is done on artistic style images while testing is on the original RGB input of H36M subject S9 and S11, following standard practice.

Recall in Sec. A \( X = I \cup O \). \( T_X \) is the fully topology on set \( X \). \( T_X = \{(u) | u \in X \} \).

During training, only stylized image \( O \) is forwarded to the pose component, \( O \) has a relative topology \( T_O = \{ O \cap U | U \in T_X \} \) (check Sec. D)

During inference on the original H36M, \( I \) has a relative topology \( T_I = \{ I \cap U | U \in T_X \} \).

Recall in test case 1, when the model performs well, \( G(I_1) \subset A, I_1 \subset G^{-1}(A) \).

From the perspective of metric space, for point \( I_1 \in \mathcal{I} \), for all \( \epsilon > 0 \), \( \exists \) finite set of points \( O_1, O_2, \ldots, O_k \in \mathcal{O} \) (since we have finite styles and infinite images) such that \( I_1 \) is of distance \( \epsilon \) of at least 1 of these \( O_j \) \((j = 1..k) \) points. Equivalently, \( \exists \) finite collection of open balls \( B_{\epsilon}(O_j) \) which together cover \( \{ I_1 \} \). Then \( (X, d) \) is precompact. Unfortunately, \( \mathbb{R}^n \) is not precompact because it is not closed. Given the original image \( I_1 \), style \( S_1, S_2, \ldots, S_k, \ldots \) a sequence \( \{ O_1, O_2, \ldots, O_k, \ldots \} \) does not have a limit in itself.

Through the lens of topology, given a cover of \( \{ I_1 \} \) by open sets in \( X = I \cup O \), \( \{ U_j \}_{j \in J} \), we would like to see finitely many of these \( U_j \) (open sets in \( X \)) cover \( \{ I_1 \} \). Since \( I_1 \subset G^{-1}(A) \) (remember it delivers a good pose prediction), \( \{ I_1 \} \subset \bigcup_{j=1}^{k} U_j \) (\( k \) is the number of open sets). Because during training, we restrict ourselves to the relative topology \( T_O \) (remember the pose network only takes the stylized \( O \) as input), the cover is actually done in the open sets of \( O \).

The finite subcover property as discussed in the last paragraph gives an important topological property called compactness.

Based on the fact

1. \( X, Y \) are Hausdorff spaces
2. \( \mathcal{I} \) is compact

\( G(\mathcal{I}) \) is also compact, given \( G : X \rightarrow Y \) is a 1-1 correspondence, we derive that \( G \) is a homeomorphism.

The indication is that: given the original image \( I_1 \) and its stylized variants which together cover it (finite subcover) \( O_1, O_2, \ldots, O_k \) \((k \) styles), as well as the pose label \( Y \), it follows then \( G(I_1) \) is the neighbor to \( G(O_1), G(O_2), \ldots, G(O_k) \). (in the same open set).

Corollary F.0.1 They (\( I_1 \) and \( O_1, \ldots, O_k \)) are similar in visual content. Again, as in test case 1, restricting ourselves to the space \( X \) does not guarantee that. Because this would imply the stylized image is too close to the real-world images, which nullifies the stylization process. Contradiction. The only way to get around this is to extend the space where closeness is analyzed to deeper feature layers of the conv network.
G. Summary on 2 Test Cases

If the model outputs reasonably well pose predictions on both I₁ and O₁ with the same pose label Y₁, it means either of the following holds:

1. I₁ is too close to O₁ in space \( \mathcal{X} \) (recall the corollary part of test case 2). Namely, the style transfer procedure fails completely and degenerates to have the same (or very similar) output as input.

2. The pose network G arguably yields similar features at some layer l₁ (and of course layers onwards) for I₁ and O₁. If we extend the discussion as did in test case 1 and 2; extend the map \( G \) from \( \mathcal{X} \rightarrow \mathcal{Y} \) to \( H : \mathcal{X} \rightarrow \mathcal{Y} \) and \( M : \mathcal{V} \rightarrow \mathcal{Y} \). (\( G = M \circ H \) for some intermediate feature space \( \mathcal{V} \)). Then \( (\mathcal{V}, d) \) is precompact from metric point of view, and also \( M : \mathcal{V} \rightarrow \mathcal{Y} \) is now a homeomorphism \( \Rightarrow U \) open in \( \mathcal{V} \Leftrightarrow M(U) \) open in \( \mathcal{Y} \).

Meeting the condition To achieve condition 2, there are several ways:

1. Jointly train the original image with the stylized counterpart within a batch, namely each batch is a mixture of both original and stylized images. This is what we did for the cross-dataset generalization experiment in the main body.

2. Use a siamese network, and pass both O and its related I as input.

3. Pass both I and O to two separate streams and enforce similarity losses at certain layers.

4. Decouple the problem into 2D plus depth. 2D is first estimated (which contains a certain degree of depth ambiguity), from which depth is deduced.

We subscribe to the fourth solution in the pipeline.

H. Joint Training of Real-world Images and Stylized Images

Homotopic paths Use the above notation, I₁ is the original H36M rgb image, O₁ is some stylized image of I₁, while Y₁ is the ground truth pose. If the input to the pose network is the artistic style transferred image, the training goes through the path \( \alpha \) in Fig.: \( I₁ \rightarrow O₁ \rightarrow Y₁ \). If otherwise, the input is the original rgb image, then training flows through the path \( \beta : I₁ \rightarrow Y₁ \). If both paths produce the same final outcome: the pose prediction for the same original image, \( \text{the original image is not path-connected} \), we say path \( \alpha \) is homotopic to path \( \beta \), i.e., \( \alpha \simeq β \).

The homotopy \( H \) is

1. \( H(0, s) = I₁, 0 < s < 1 \)
2. \( H(1, s) = Y₁, 0 < s < 1 \)
3. \( H(t₀, 1) = O₁, t₀ \) is the moment the pose part starts (to take input images), \( 0 < t₀ < 1 \)
4. \( H(t₀ + \Delta t, 1) = \text{features of layer index } \frac{\Delta t}{t₀} \text{ total number of layers, } 0 \leq \Delta t < 1 - t₀ \)
5. \( H(t, 0) = \text{features of layer index } t \text{ number of total layers, } 0 < t < 1 \).

We observe that the path from \( O₁ \rightarrow Y₁ \) is moving at a faster speed than \( I₁ \rightarrow Y₁ \).

This poses a challenge when jointly training them, since one can imagine in the path \( \alpha \) the original image I₁ is first diverted dramatically to a distant image O₁ (otherwise the style transfer does not take effect), then the pose component tries to pull it back to the apropos track so that it has the same output as the original image I₁. Issues about gradient magnitude may arise. In the cross-dataset generalization section of the main paper, we finetune the weights trained exclusively on stylized images to initiate the "joint training" of both real and stylized images.

I. Self Supervision

Motivation The motivation follows the analysis-by-synthesis mechanism where the output usually is mapped to the original image space for self supervision. In our scenario, the output (either stylized O or joint prediction Y is projected to the input image space \( \mathcal{I} \), from which losses can be enforced.)

Quotient space Given space \( \mathcal{O} \) (Sec. A) (stylized) and equivalence relationship \( \sim \) among the points in \( \mathcal{O} \), the equivalence relationship is defined as \( O₁ \sim O₂ \) if \( O₁ \) and \( O₂ \) are stylized using the same image content I₁ \( \mid O \) is the equivalence class of \( O \) in the set of equivalence classes of points of \( \mathcal{O} \).

Path-connectedness We would like the propagated gradients to flow smoothly through the input space \( \text{wherever the image resides for the self supervision part} \). In the following, the first attempt does not satisfy this while the second one meets this requirement.

1.1. Reconstructing I from O

This effectively means \( O \xrightarrow{q} O/ \sim \xrightarrow{r} I \): we first map \( O \) to the equivalence class where each image is derived from the same original image content. However, it is practically difficult to map \( r \).

Remark The space \( \mathcal{X} = \mathcal{I} \cup \mathcal{O} \) is not path-connected because path-connectedness implies \( \forall a, b \in \mathcal{X}, \exists \text{ a path } \omega : [0, 1] \rightarrow \mathcal{X} \text{ from } a \text{ to } b \). In this case let \( a = O \) and \( b = I \), it’s hard to find a path from \( O \rightarrow I \).
Measuring the difference in the original image space is difficult due to geometric and photometric transformations. We instead opt to implicitly optimize the stylization (via network) again and gauge the discrepancy between stylization $O$ and re-optimized stylization $O'$.

I.2. Reconstructing $O$ from $S$ and $Y_{2d}$

Another approach is to render a 2D bone map $P$ from the 2D pose prediction $Y_{2d}$, together with the style image $S$, an image transform network $F'$ is used to reconstruct the stylized output $O'$, which we did in the main algorithm.

Suppose the ground truth pose is $Y_{2d}$.

1. If $Y_{2d}$ is not in $Y_{2d}$'s neighborhood (the open set that contains $Y_{2d}$: $Y_{2d} \in A \subset Y_{2d}$; $Y_{2d} \notin A$.

Then the reconstructed output $O'$ has the pose content astray from the label $Y_{2d}$. (The spatial correspondence between $O'$ and $P$ is roughly preserved by FCN)

This imposes a large $L_{cos,sup}$ which could flow back to $Y_{2d}$

2. If $Y \in A$, note the space $P$ of the rendered 2D bone map has a nice property: it’s locally path-connected. That is, for each $p \in P$, $\exists$ an open set that contains $p$. Further, it is path-connected, for $\forall p, q \in P$, $\exists$ a path $\omega : [0, 1] \rightarrow P$ from $p$ to $q$.

Intuitively this can be explained by the fact we can deduce 2D from such a clean bone map and perform 2D deformation thereafter.

Specifically, consider 2 adjacent bones in the kinematic chain (for instance pelvis $\rightarrow$ left_hip and left_hip $\rightarrow$ left_knee) connecting $(x_1, y_1) \rightarrow (x_2, y_2)$ and $(x_2, y_2) \rightarrow (x_3, y_3)$. The path from $(x_1, y_1)$ to $(x_2, y_2)$ induces a homeomorphism $f_x$ that maps the loop (the first bone) $\hat{\alpha}_{n,1}^0$ (connecting $(x_1, y_1)$ to $(x_2, y_2)$ and back to $(x_1, y_1)$) in the fundamental group $\pi_1(P, (x_1, y_1))$ to the loop (the second bone) $\hat{\alpha}_{n,2}^2 ((x_2, y_2) \rightarrow (x_3, y_3) \rightarrow (x_2, y_2))$ in $\pi_1(P, (x_2, y_2))$:

$f_x(\hat{\alpha}_{n,1}^0) = \hat{\alpha}_{n,2}^2$, which basically means the first bone is mapped to the second bone. Continue this process, bones further down in the kinematic chain can be mapped by the induced homeomorphism in the fundamental group.

The fundamental group $\pi_1(P, (x_1, y_1))$ is isomorphic ($\cong$) to the fundamental group $\pi_1(P, (x_2, y_2))$. In particular, the fundamental group $\pi_1(P)$ does not depend upon the choice of base point. Which means there is a loop that starts and ends at any pixel inside the bone map (interior or on edge of the oval areas pertaining to each bone): the trivial loop $e_x$.

At a high level, the 2D bone map is nothing but a loop based at $x_0$ (since the base point doesn’t matter, let’s assume for now $x_0 = pelvis$) in the fundamental group $\pi_1(P)$. The idea of continuously deforming the 2D bone map according to backward gradient signal is just:

- Using induced homeomorphism by paths connecting 2 points $(x_1, y_1)$ and $(x_2, y_2)$ to do mapping between loops (bones) in the fundamental group.
- Using path-connectedness of space $P$ and the homeomorphisms to construct the new bone map bone by bone.

Then, it follows that

(a) If $Y = \hat{Y}$, 2D bone map overlaps entirely with the stylized image $O$. For the background pixel of the rendered $P$, the gradient has no impact on the pose prediction $Y$. For the overlapped part, the style input $S$ will guide the reconstruction to have the style as $S$, and so won’t incur a large pixel difference loss.

(b) If $Y \neq \hat{Y}$ but $Y \in A$ (in $\hat{Y}$’s neighborhood), the loss would be small. As described earlier, the style is not an issue, and the background pixel of rendered 2D bone map $P$ does not affect $Y$. Since $Y$ is in $\hat{Y}$’s neighborhood, $\exists$ a path $\omega$ from $Y$ to $\hat{Y}$. For the overlapped pixels, each pixel is in some loop determined by $\hat{\alpha}_t^i(t)$ for some $t$. And the pixel value is parameterized by the 2 consecutive keypoints in the kinematic chain. The backpropagation ensures $Y$ is moving towards $\hat{Y}$ along the path as iterations increase. (recall the bone map space $P$ is path-connected.)

We empirically found in our experiments that with the help of neural style transfer, the 2D prediction $Y_{2d}$ is already very close to the ground truth $Y_{2d}$, meaning (b) is met. The self supervision loss slightly refines the 2D output by virtue of

- The path-connectedness of space $P$.
- The stylization loss. (It does not require an exact per-pixel match between $O$ and $O'$)

J. Per-style Training

The ablation study about per-style training trains parallel independent style transfer networks, upon which we condition independent pose networks to deliver the final pose output. It is possible to use boosting algorithm like Adaboost to fix previous mistakes made by previous styles, meaning the training of later styles depends on earlier styles.
K. Bone Map

We continue the definition of bone map here. With regards to the loop, go to Preliminaries Definition D.5/D.6.

For the loop $\tilde{\alpha}_i$, the starting point $\tilde{\alpha}_i(0)= (x_1, y_1)$, the halfway midpoint $\tilde{\alpha}_i(\frac{1}{2}) = (x_2, y_2)$. For other $t$, the mapped point in space $\mathbb{R}^2$: $\tilde{\alpha}_i(t)$ is determined such that it satisfies one of the following case:

$0 \leq t \leq \frac{1}{4}$: $t = \frac{\text{dist}(\mathbf{(x,y)}, \mathbf{x})}{4a_i}$

$\frac{1}{4} < t \leq \frac{1}{2}$: $t = \frac{1}{4} + \left(\frac{1}{4} \left(1 - \frac{\text{dist}(\mathbf{(x,y)}, \mathbf{x})}{a_i}\right)\right)$

$\frac{1}{2} < t \leq \frac{3}{4}$: $t = \frac{1}{2} + \frac{\text{dist}(\mathbf{(x,y)}, \mathbf{x})}{4a_i}$

$\frac{3}{4} < t \leq 1$: $t = \frac{3}{4} + \left(\frac{1}{4} \left(1 - \frac{\text{dist}(\mathbf{(x,y)}, \mathbf{x})}{a_i}\right)\right)$

L. Non-exhaustive standalone notes about artistic human figure images

• At the beginning of the 14-th century, proportions of human body were lengthened.

• The 15-th century was a time of change, the rise in realism became noticeable in people’s attitude towards the human condition.

• Impressionism concentrates on the landscape rather than on the human figure. People become part of nature.

• Expressionism uses bolder and brighter colors than impressionism and focuses more on human emotions and expressions. Subject and style are highly simplified. Colors become brighter and more arbitrary, and are arranged to look more powerful.

• According to Virginia Woolf, on or about Dec 1910 human nature changed. Scientists had disproved many of the ideas on which nature’s laws appeared to have been based and there was a degree of questioning and hypothesizing in all areas of human activities unparalleled in history. The new view of the world had begun to be reflected in the visual arts. Many of the modernist styles had already been created. Fauvism, Cubism, and Expressionism all put forward their own versions of reality. Each in its own way had questioned the tenets of Impressionism.

• Cubists like Picasso broke away from two central characteristics of European painting since the Renaissance, one of which being the classical norm for the human figure.

• Some painters like Renoir never wandered in the sense of structure and topography, on the other hand, some painters particularly pronounce the shapelessness of their paintings. And so some human figures are more amorphous than others with anatomical constraints completely forgone e.g. Franc Marc.

• As time goes, some think abstract art is the only form of modern art while some are in search of pure colors. A new attitude towards the expressive potentialities of the human figure is not on gesture but on complete freedom, meaning unexpressed states of minds are unleashed, which results in surrealism.

• Some paintings are softer/sweeter colors, some are darker/freer/richer.