Improving Pose Estimation on Art Collections with Style Transfer

Tristan Verheecke
Ghent University
Ghent, Belgium
tristan.verheecke@ugent.be

Dieter De Witte
IDLab, Ghent University
Ghent, Belgium
dieter.dewitte@ugent.be

Steven Verstockt
IDLab, Ghent University
Ghent, Belgium
steven.verstockt@ugent.be

Kenzo Milleville
IDLab, Ghent University
Ghent, Belgium
kenzo.milleville@ugent.be

Ravi Kharti
IDLab, Ghent University
Ghent, Belgium
ravi.khatri@ugent.be

***Abstract*—Through digitalization, museums are given the ability to more efficiently analyze their art collections. Important connections between artworks can be uncovered this way, which can be useful for classification or retrieval. Museums put a great amount of effort in this process, but it can be very labor intensive doing this manually. To eliminate this issue, they've sought to automate these tasks using computer vision methods. In computer vision, there's a rich volume of research in image classification, semantic segmentation, object detection and 2D/3D human pose estimation (HPE). It turns out however, that these algorithms aren't suitable for tasks on art collections as they were trained on photographs. This paper will deal with the HPE problem and what methods can be used to improve performance on art collections. Two shortcomings can be identified: incomplete keypoint prediction and wrong pose association. To solve this problem, this paper proposes a method which fine-tunes state-of-the-art (SOTA) HPE models with a combination of stylized COCO datasets. Three datasets were created from the WikiArt dataset representing baroque, renaissance and impressionism. From those genres a selection of figurative paintings is made using content-based image retrieval. Then for each style transfer model, first, a mixture of genres is used and, second, one with only impressionism to create a stylized COCO dataset. This is done for CycleGAN and AdaIN. Then, the SWAHR and ViTPose pose estimation models are fine-tuned on the COCO dataset in combination with the stylized COCO dataset and with only the stylized COCO dataset. This makes a total of 16 models that are evaluated and in which a consistent improvement in pose estimation prediction was found.**

***Index Terms*—Cultural Heritage, Computer Vision, Deep Learning, 2D Human Pose Estimation, Style Transfer, CycleGAN, AdaIN, SWAHR, ViTPose**

I. Introduction

Part of the modern age is the digitalization of information. Digitization makes information more accessible to a broader audience and allows it to be processed more efficiently. Museums have put huge efforts in digitalizing their catalogue as well, as this can help in Iconography; this is the branch of art history that concerns itself with the themes and motifs of artworks. Through the analysis of artworks, different connections between different artworks can be established, which can be useful for classification or retrieval. However, art collections don't contain much metadata and it is time-consuming to enhance them manually. Museums want to utilize computer vision to automate this process, but the algorithms that were developed over the last few decades, are mainly for photography and it turns out that art collections (paintings, statues, drawing, etc) are less interpretable by these algorithms.

Computer vision can perform a wide range of tasks, including image classification, semantic segmentation, object detection, and 2D/3D human pose estimation (HPE). For this paper, the focus will be on 2D HPE. A database can be created with the different poses found in the artworks which can be used to discover similar themes and categorize them. There is an extensive amount of research based around HPE that can be useful for this.

To make the vast quantity of research around HPE available to art collections, the following proposed solution will be explored: If the pre-trained models can't be used, it's still an option to retrain one with an augmented dataset. With style transfer the images of existing datasets can be stylized and added to the datasets. This will increase the size and variance of the dataset, making it better to train on. This can increase performance on art collections as stylized images are also being trained on but can also potentially increase the performance on photographs.

II. Related Work

A. Generative Adversarial Network

Proposed by Goodfellow et al. [1], a Generative Adversarial Network (GAN) is a framework where a generative model competes against a discriminative model. It consists of a generator G which takes in random noise and attempts to output a sample from a specific distribution, while the discriminator D will try to distinguish between the sample from the generator and a real sample from that distribution. A popular variation of GAN is the conditional GAN, this architecture feeds an extra label to the generator and discriminator, so that it can be conditioned to generate certain images based on the label.

B. Human Pose Estimation

Human Pose Estimation (HPE) aims to detect human features from input data such as images and videos. It's an elementary part of computer vision with many applications among which are: human action recognition (sign language), human tracking (surveillance), and human-computer interaction (video games). This is an extensively researched area with a diverse range of different techniques. The human body has a high degree-of-freedom due to all the limbs, self-similar parts and body types, which may cause self-occlusion or rare/complex poses. The variations in configuration are made even larger due to clothing, lighting, foreground occlusion, as well as viewing angles and truncation, among others. This makes HPE one of the most difficult tasks in computer vision.

1) Single-Person Human Pose Estimation: **Heatmap and Detection-based Methods** predict the individual body parts using heatmaps. This method results in an easier optimization and a more robust generalization [2]. Most of the latest HPE methods use heatmaps because of this. After the joints are found, they are assembled to fit a human skeleton, as shown in Figure 1.

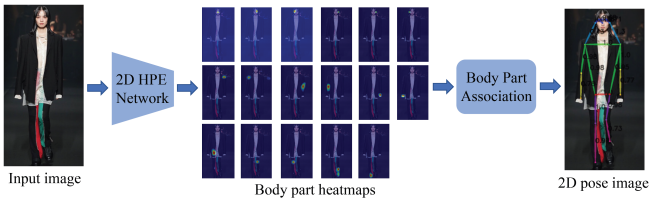
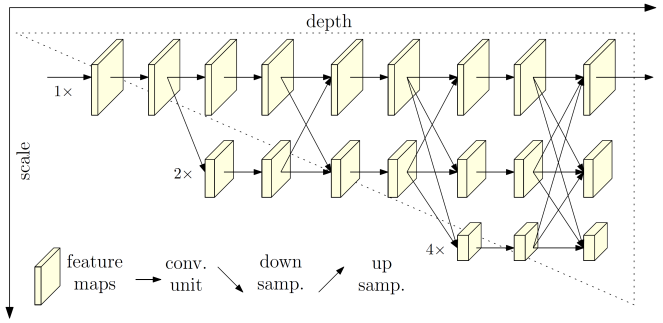
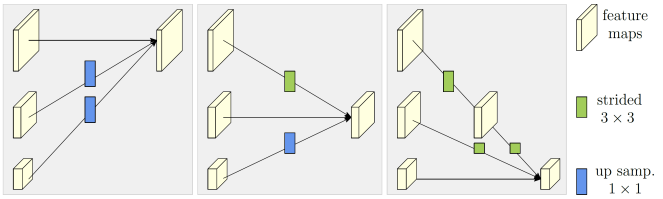


Fig. 1: Heatmap-based Methods as presented in [3]

A fundamental work written by Wei et al. [4] combines convolution networks with Pose Machines [5]. Pose Machines is an iterative architecture which consists of two models: The first is used for stage one where it predicts potential heatmaps for the joints. The second model is used for subsequent stages where the result of the previous stage is fed in together with the results of its own convolution network on the input image. This gradually refines the predictions for the joints and their positioning. Another influential work was being written at the same time by Newell et al. [6]. Similar to CPMs, this is also an iterative architecture. They suggest what they call a "stacked hourglass" network, where "hourglass" modules are repeated. In an "hourglass" module, first, the features are downsampled and, afterwards, upsampled again. This network captures different spatial relationships between joints at different resolutions. Both use intermediate supervision to tackle the problem of vanishing gradients. This still doesn't build a deep sub-network for feature extraction which limits the predictions. This has become less of a problem with the emergence of the ResNet [7] which allows better back-propagation at deeper levels through shortcuts. A more recent work by Sun et al. [8] maintains high-resolution representations instead of working with the high-resolution from the low-to-high sub-network. After a first high-resolution sub-network, it gradually adds high-to-low sub-networks in parallel to predict multi-resolution features. Before each branch, they apply multi-scale fusion, which joins the predicted features from each scale on each other scale (Figure 2).



(a) HRNet



(b) Multi-Scale Fusion

Fig. 2: The architecture of the High-Resolution network and how it applies multi-scale fusion [8].

2) Multi-Person Human Pose Estimation: **Top-Down Human Pose Estimation** will first try to detect all persons in the image with a human detector. Each person is cropped by the bounding box and single-person HPE predicts a pose. Iqbal et al. [9] use Faster R-CNN [10] to detect the human boundaries, after which it applies integer linear programming on each person's fully connected graph to obtain the final pose estimates. The use of a human detector comes with its own set of problems, which Fang et al. [11] try to remedy with Regional Multi-person Pose Estimation. They try to tackle inaccurate bounding boxes with a Symmetric Spatial Transformer Network and redundant detections with Parametric Pose Non-Maximum-Suppresion. Papandreou et al. [12] use a two stage pipeline, where they, first, employ the Faster R-CNN detector and, second, estimate the pose in each found bounding box using their own network. It predicts heatmaps using a fully convolutional ResNet and then uses a heatmap-offset aggregation procedure. Afterwards, they do post-processing using keypoint-based Non-Maximum-Suppression. A continuous effort is taken by Chen et al. [13] to deal with occlusion and truncation. They suggest a two stage architecture, where, the "simple" keypoints are captured with GlobalNet, a feature pyramid network based on [14], and the "hard" keypoints are handled by their RefineNet. It integrates the information via upsampling and concatenating of HyperNet [15], using an adapted stacked hourglass. In more recent research, a new method was become competitive with CNNs. Based on work in language modeling, attention mechanisms, an optimization of recurrent networks, allow the modeling of dependencies

without regard of the distance in the input or output sequences. The Transformer, introduced by Vaswani et al. [16], eliminates recurrence and relies solely on attention mechanisms. This enables it to work better in parallel, while it still maintains state-of-the-art performance. Based on this new architecture, Dosovitskiy et al. [17] created a new model that can work with images; the Visual Transformer (ViT). Xu et al. [18] use ViT to apply it to the HPE task. It works by splitting the input image into fixed-size patches which are linearly embedded and then fed into the transformer blocks. The output of this is then processed by different decoders to form the heatmaps.

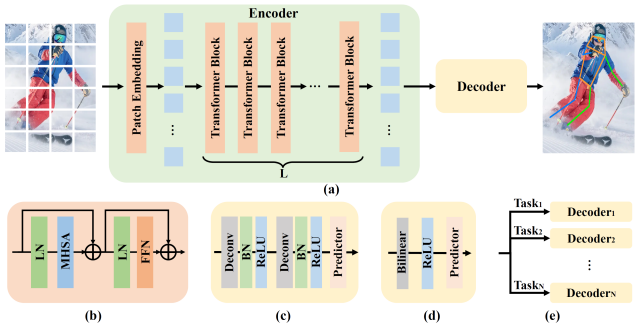


Fig. 3: (a) The framework of ViTPose. (b) The transformer block. (c) The classic decoder. (d) The simple decoder. (e) The decoders for multiple datasets. [18]

Bottom-Up Human Pose Estimation first locates all joints in the image and then afterwards assemble them in potential poses. DeepCut by Pishchulin et al. [19], uses Fast R-CNN [10], it detects the body parts and labels each. With the joints found, it then uses Integer Linear Programming to assemble them. However, this method is very computationally expensive. Insafutdinov et al. [20] therefor introduce a stronger part detector and better optimization strategy with DeeperCut. Convolutional Pose Machines make a return with OpenPose by Cao et al. [21]. They're used to predict the joints with heatmaps and Part Affinity Fields, which also encodes the position and orientation of the limb which makes the assembly more reliable. Kreiss et al. [22] continue on the idea of fields and introduce the Part Association Fields (PIF) and Part Intensity Fields (PAF). First, they predict the location of the different joints with PIF. Afterwards, they use PAF to find the inter-joint relationships. They are able to outperform any previous OpenPose-based proposals on low-resolution and occlusions. Newell et al. [23] introduce associative embedding. They make use of the stacked hourglass network from [6] with some small modifications, and produce joint heatmaps and associative embedding tags. Continuing on the idea of associative embedding, Cheng et al. [24] use HRNet [8] as backbone for their HigherHRNet (Figure 4). Their method focuses on the scale-variance problem, so it can localize keypoints for small persons better. Lou et al. [25] introduce Scale-adaptive Heatmap Regression and Weight-adaptive Heatmap Regression to the scale-variance problem. SAHR adaptively adjusts the standard deviation of each heatmap corresponding with the scale of the person. WAHR rebalances the foreground and background samples, so SAHR can work to its fullest extent.

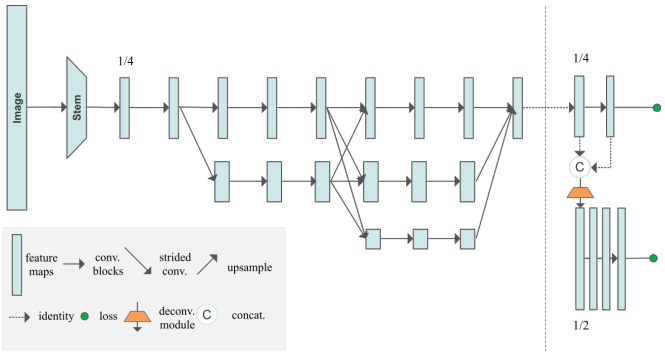


Fig. 4: The architecture of HigherHRNet. It uses HRNet as backbone. [24]

C. Image Style Transfer

Image Style Transfer is the technique of applying the style of one image to the content of another. Traditionally, this is a problem reserved for only artists, but more recently this has also interested computer scientists. There are several different ideas on how this can be achieved, ranging from how to separate the style from the content to how well an algorithm can generalize.

1) Optimization-based Networks: Gatys et al. [26] introduce deep neural networks to image style transfer. Using a modified VGG-network [27], they extract the features from the higher layers of an image, which they argue represents the content, and then reconstruct it on a white noise image. They also extract the style representation of another image by using the Gram matrix and then reconstructs it on the same white noise image. The Gram matrix is the vector product of two sets of vectorized feature maps. The resolution affects the performance of the algorithm and is thus restricted to low resolutions.

2) Feed-forward Generation Networks: To improve the performance, Ulyanov et al. [28] suggest using a feed-forward generation network instead of reconstruction. To train such a network, they use a pre-trained network for image classification, and calculate a texture and content loss by extracting the features similar to [26]. Johnson et al. [29] propose an almost identical framework independently. The work of Ulyanov et al. suggest further improvements to their network [30]. First, they replace Batch Normalization (BN) [31] with Instance Normalization (IN) which alone has a significant impact on quality. Second, they teach the generator to sample from the Julesz ensemble [32] which improves variation in the outputs. Until now, style transfer was only possible on styles that were seen during training. Huang et al. [33] try to remedy this by introducing an Adaptive Instance Normalization (AdaIN) layer. Unlike the other normalization techniques, (AdaIN) does not have affine parameters, and will adaptively compute these from the style image. Figure 5 shows that their network first

extracts features from the content and style image using a fixed VGG-19 network as their encoder. The AdaIN layer then performs style transfer in the feature space and with the results the decoder constructs a new image. During training, the content loss and style loss are calculated by extracting the features using the same VGG encoder.

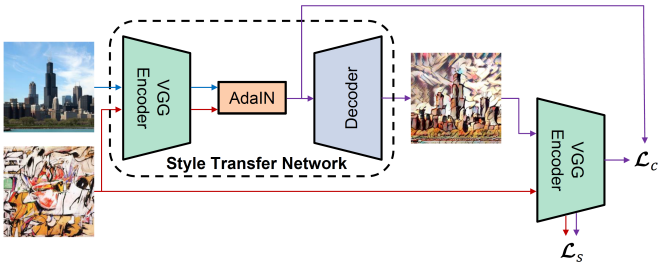


Fig. 5: Adaptive Instance Normalization network by Huang et al. [33].

3) Generative Adversarial Networks: Among the first to use GAN are Isola et al. [34], who use cGAN. However, their method still requires paired training samples. Meanwhile, Taigman et al. [35] are doing research in unsupervised domain transfer. Their network uses an autoencoder as the generator and they assume that the encoder is fixed between domains. The discriminator has a ternary output and distinguishes between real, fake and reconstruction. They add several new loss functions which check the consistency between the two domains (consistency loss) and whether G performs perfect reconstruction (reconstruction loss). For the encoder, they use a pre-trained network that is trained on paired samples though. In order to make the network completely unsupervised, Yi et al.[36] propose DualGAN, Kim et al. [37] DiscoGAN and Zhu et al. [38] CycleGAN, which are all three essentially the same proposal. The entire model consists of two cycle-consistent networks where each translates from one domain to the other. A cycle-consistent network will first translate the input to a target domain and then back to the original domain. Each domain has a discriminator which compares the real input from one network with the fake from the other (adversarial loss). In addition to this there's a cycle-consistency loss, which is the MSE between the input and the reconstructed image as you can see in Figure 6. The goal is to minimize the adversarial and cycle-consistency loss, while maximizing the discriminators' accuracy.

D. Content-Based Image Retrieval

Content-Based Image Retrieval (CBIR), a long-established research area, is the task of finding semantically matching or similar content images for a specified query image. This has become increasingly relevant with the exponential growth of image and video data and the need to effectively search these image collections. CBIR has been used specifically for person re-identification, remote sensing, medical image search, and shopping recommendations in online marketplaces, among many others [39]. Image retrieval can be categorized into two

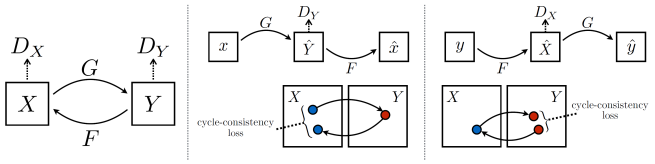


Fig. 6: The cycle-consistent network by Zhu et al. [38]. Unsupervised image-to-image translation between domains X and Y is established by training the generators G, F and discriminators D_X, D_Y. During training cycle-consistency loss is calculated under the assumption that $F(G(x)) \stackrel{!}{=} x$ and $G(F(y)) \stackrel{!}{=} y$.

different groups: Category Image Retrieval (CIR) and Instance Image Retrieval (IIR). CIR's goal is to find images within the same category as the query, while IIR tries to find images with a particular instance given in the query image. For IIR, Radenovic et al. [40] train a VGG network from reconstructed 3D models obtained by retrieval and structure-from-motion methods. This allows them to use the geometry and camera positions to enhance the feature extraction along with several other optimization techniques. During training, they make use of the contrastive loss. Contrastive loss is minimized when similar image pairs are close to each other in embedding space and different pairs are far away.

E. Deep Learning in the Art domain

Through digitization, analysis of art collections has become more efficient. Artists are constantly inspiring and being inspired, and in order to correctly analyze the relation between paintings and artists, it's beneficial to have a method that finds their inspirations. One way that can be achieved is by using image retrieval, which gives good results, but only when the works are visually similar. In other cases, the inspiration is drawn from themes, which involve composition, lighting and poses. Jenicek et al. [41] propose finding these relations by analyzing the similarity between poses. From a database of images the poses are estimated and normalized. They then employ a two step process: With a query image and fast matching, they generate a shortlist of possible hits. Afterwards, geometric validation filters out impossible alignments with the query image. Their experiments show significant improvements over previous methods. They also note some failure cases where pose estimation falls short, like failing to find keypoints or making associations with wrong poses. To improve these shortcomings on Greek vases, Madhu et al. [42] apply style transfer to the COCO dataset with AdaIN in the style of those vases and use this for fine-tuning. They use a top-down architecture with Faster R-CNN as detector and HRNet for pose estimation. They also created their own small dataset to evaluate their improvements (Figure 7). Kadish et al. [43] have the same idea and also use AdaIN to stylize the COCO dataset. They randomly sample artworks from the Painter by Numbers dataset from Kaggle [44] for the style images and using it to fine-tune the Faster R-CNN network

for object detection. Both papers found an improvement in the performance of the networks on art collections.

(a) Original (b) OpenPose (c) Madhu et al.

Fig. 7: Improvements to the state-of-the-art by Madhu et al. [42].

III. Datasets

A. Human Pose Estimation

1) Common Object in Context (COCO) Dataset [45]: COCO is a large-scale dataset for a wide range of computer vision algorithms. For HPE, the set contains more than 200,000 images in which 250,000 persons are annotated. Each person has 17 keypoints, a bounding-box and visibility labels. This dataset has become the most popular for benchmarking.

2) Human-Art Dataset [46]: Human-Art dataset bridges the gap between natural and artificial images. The set contains 50,000 high-quality images with 123,000 annotated humans. Each person has 17 keypoints, bounding boxes, self-contact points, and text information.

B. Image Style Transfer

1) Arbitrary Style Transfer Image Quality Assessment Database (AST-IQAD) [47]: AST-IQAD is a set specifically made to measure style transfer. It constructs the set around several inter-subjective characteristics and categories. This means that these criteria of subjective evaluation are mostly agreed upon across a group of people. Among those are: color tone, brush stroke, distribution of objects, and contents. It also declares a set of style images.

2) WikiArt Dataset [48]: WikiArt consists of 80,000 fine-art paintings. All are annotated for 27 styles, 60,000 are annotated for 20 genres and 20,000 for 23 artists.

IV. Methodology

The goal is to improve pose estimation on art collections. SWAHR and ViTPose will be fine-tuned on an COCO dataset which is augmented with synthetic COCO images. This needs a style transfer model that is trained to do a transformation between an art movement and realistic images. For this, AdaIN and CycleGAN will be analyzed. The style transfer models will be trained on three datasets of different art movements.

A. Style Transfer

1) Datasets: None of the selected models has any pre-trained weights for certain art movements, so new models need to be made. The WikiArt dataset will be used to assemble different datasets and specifically on the Baroque, Renaissance and Impressionism styles. The impressionist style is chosen because it is more colorful and abstract than the others. Baroque and Renaissance are both very dark and very similar in style, but renaissance artworks are just a bit more stylized. For each art movement, a subset needs to be created with images that contain full body poses as well as crowded images, as this is what the pose estimation models are trained on. While there's a high variation of genres in the WikiArt dataset, they do not adequately subdivide the dataset for this problem. It is important then to have a consistent style in each dataset which is not possible by just splitting the genres provided by WikiArt. To achieve this, the Content Based Image Retrieval (CBIR) model by Radenovic et al. [40] is used. Using a query image it can find similar looking images. This results in four datasets: The photograph dataset which consists of 825 images, the baroque dataset with 518 images, the impressionism dataset with 780 images and the renaissance dataset consisting of 790 images. Figure 8 shows the kind of images that can be found in the datasets where the human figure is central.

Fig. 8: A selection from the different datasets. The human figure is central to all images.

2) Training: From the selected models only CycleGAN requires training. AdaIN can use any arbitrary style from a content image to do style transfer. This eliminates the need to train a new model for it and the pre-trained model can be used for the experiments. CycleGAN will be trained with the provided default parameters. No hyperparameter tuning will be done as the goal is to measure the performance between different approaches and not optimize a single model.

3) Qualitative Evaluation: As shown in Figure 9, AdaIN removes more of the details of the content than CycleGAN does, but as expected the style transfer is completely dependent on the style image used. CycleGAN does look like it is able to capture the general style of the learned art movements, e.g. baroque and renaissance are dark, and impressionism is colorful. All in all, the results are very disappointing as none

(a) Content (b) CycleGAN. Left to right: Baroque, Impressionism and Renaissance

(c) AdaIN. Paired with style image from left to right: Baroque, Impressionism and Renaissance.

Fig. 9: AdaIN abstracts the features more than CycleGAN, while StarGAN experiences modal collapse.

of the images look like they're a painting from a different time.

4) Quantitative Evaluation: To evaluate the trained models, Perceptual Distance (PD), Fréchet Inception Distance (FID), Inception Score (IS) and Learned Perceptual Image Patch Similarity (LPIPS) are used. Before applying the evaluation metrics there needs to be an adequate dataset to do meaningful measurements on first. Two datasets are considered for this purpose, AST-IQAD, but since it works with different kinds of content, and the content of the problem only focuses around persons, a custom dataset is created that focuses around persons. This is created the same way the style transfer datasets were created. In Table I, the results of the evaluation are available. AdaIn performs better than CycleGAN, impressionism does the best out of all of the styles and the custom dataset has a slightly better evaluation.

TABLE I: Performance comparison of Style Transfer measured by various metrics grouped by dataset; Perceptual Distance (PD), Inception score (IS), Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS).

Method	PD	IS	FID	LPIPS
AST-IQAD Dataset				
Baroque				
AdaIN	**10.734**	**8.975**	**265.036**	0.626
CycleGAN	14.670	10.850	272.652	**0.633**
Impressionism				
AdaIN	**10.671****	**8.453**	**246.736**	0.710
CycleGAN	14.160	10.046	247.468	**0.721**
Renaissance				
AdaIN	**10.671****	**8.453**	**246.736**	0.710
CycleGAN	14.160	10.046	247.468	**0.721**
Custom Dataset				
Baroque				
AdaIN	**10.507***	**6.639****	**195.487****	**0.654****
CycleGAN	13.435	7.137	200.299	0.635
Impressionism				
AdaIN	13.435	**4.974***	**177.581***	**0.737***
CycleGAN	**12.456**	6.047	190.658	0.711
Renaissance				
AdaIN	**11.472**	**5.156****	**197.560****	**0.693**
CycleGAN	12.962	7.825	200.920	0.678

* the best result overall.
** the best result for the style.

V. Experiments

To improve pose estimation, SWAHR and ViTPose will be trained on several different stylized datasets. A combination of the COCO dataset and the stylized dataset is used, and one with only the stylized dataset. The stylized datasets are created by applying both CycleGAN and AdaIN to the COCO dataset. One with a mixture of the baroque, impressionism and renaissance models, and one with only the impressionism model. This results in a combination of 16 models. The experiments will be conducted on the COCO-dataset as well as the Human-Art dataset. While the problem specifically tries to improve the performance on artworks, it's still interesting to also validate the results on the COCO-dataset.

1) Styled COCO Datasets: Since AdaIN requires a style image, the images used for training the CycleGAN models were used for this purpose. The style dataset was cycled through to transform the COCO dataset with AdaIN. The decision to not use one image as a representation for each style was made so that the dataset is more generalized. Afterwards, a dataset was created from a mixture of baroque, impressionism and renaissance stylized images, and one of only the impressionism style. For each, a version was made which is appended to the COCO dataset and one that stands on its own.

2) Training: All models are trained with the default parameters initiating the weights with the pre-trained model. They're trained for 200 epochs. As a control, 2 models are trained without initiating weights these were trained for the default 300 epochs.

3) Results: For the SWAHR network, shown in table **??**, the best results are found with the model trained on the COCO + AdaIN mixed style transfer dataset. The second best

TABLE II: Comparing the best models from the experiments on the COCO dataset with the baseline metrics.

Method	AP	AP^{50}	AR	AR^{50}
Pre-trained SWAHR	**0.687***	**0.881***	0.737	0.904
Pre-trained ViTPose	0.588	0.832	0.723	0.906
SWAHR	0.620	0.830	0.710	0.891
ViTPose	0.609	0.847	0.740	0.918
SWAHR COCO + AdaIN Mixed	**0.679****	**0.874****	0.732	0.902
ViTPose COCO + CycleGAN Mixed	0.635	0.861	**0.763***	**0.925***

* the best result overall.
** the best result without pre-trained models.

TABLE III: Comparing the best models from the experiments on the Human-Art dataset with the baseline metrics.

Method	AP	AP^{50}	AR	AR^{50}
Pre-trained SWAHR	0.528	0.759	0.593	0.635
Pre-trained ViTPose	0.380	0.656	0.571	0.803
SWAHR	0.492	0.742	0.563	0.784
ViTPose	0.406	0.682	0.591	0.818
SWAHR COCO + CycleGAN Mixed	**0.553***	**0.789***	**0.629***	0.839
ViTPose COCO + CycleGAN Mixed	0.439	0.726	0.617	**0.844***

* the best result overall.
** the best result without pre-trained models.

network was trained on the COCO + CycleGAN mixed style transfer dataset. For the ViTPose network, the best results are with COCO + CycleGAN mixed and COCO + CycleGAN impressionism being the second best. For AdaIN, there's a falloff of 7 to 10% AP between the datasets with COCO and the ones without. For CycleGAN, this falloff is less; between 0.2 and 4% AP. The best precision is found using the SWAHR model, while ViTPose has the honor of having the best recall. Table II compares the best models with the baseline. It shows that the pre-trained SWAHR model has the best precision of all of the models and trained on the COCO + AdaIN Mixed style transfer dataset, SWAHR also has the second best precision. ViTPose trained on COCO + CycleGAN mixed style transfer dataset has the best recall. The networks trained from the ground up don't have any significant difference between the other networks.

The results on the Human-Art dataset are shown in table **??**. Here, one dataset takes the crown. Both SWAHR as well as ViTPose have the best results for the models trained on the COCO + CycleGAN mixed style transfer dataset. The second best model for SWAHR is trained on the COCO + AdaIN mixed dataset and for ViTPose this is the one trained on COCO + CycleGAN impressionism. The falloff between the COCO and non-COCO datasets is between 5 to 9% AP for AdaIN, and 2 to 3% AP for CycleGAN. The best precision and recall belongs to the SWAHR models. Comparing the best models to the baseline (Table III), they still remain the best models overall with an increase of 3 to 5% AP. There's no significant difference between the non-initialized and bootstrapped networks.

REFERENCES

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

[2] H. Chen, R. Feng, S. Wu, H. Xu, F. Zhou, and Z. Liu, "2d human pose estimation: a survey," *Multimedia Systems*, pp. 1–24, 2022.

[3] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *CoRR*, vol. abs/2012.13392, 2020. [Online]. Available: https://arxiv.org/abs/2012.13392

[4] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," *CoRR*, vol. abs/1602.00134, 2016. [Online]. Available: http://arxiv.org/abs/1602.00134

[5] V. Ramakrishna, D. Munoz, M. Hebert, J. Andrew Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 33–47.

[6] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," *CoRR*, vol. abs/1603.06937, 2016. [Online]. Available: http://arxiv.org/abs/1603.06937

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[8] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," *CoRR*, vol. abs/1902.09212, 2019. [Online]. Available: http://arxiv.org/abs/1902.09212

[9] U. Iqbal and J. Gall, "Multi-person pose estimation with local joint-to-person associations," *CoRR*, vol. abs/1608.08526, 2016. [Online]. Available: http://arxiv.org/abs/1608.08526

[10] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[11] H. Fang, S. Xie, and C. Lu, "RMPE: regional multi-person pose estimation," *CoRR*, vol. abs/1612.00137, 2016. [Online]. Available: http://arxiv.org/abs/1612.00137

[12] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. P. Murphy, "Towards accurate multi-person pose estimation in the wild," *CoRR*, vol. abs/1701.01779, 2017. [Online]. Available: http://arxiv.org/abs/1701.01779

[13] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," *CoRR*, vol. abs/1711.07319, 2017. [Online]. Available: http://arxiv.org/abs/1711.07319

[14] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *CoRR*, vol. abs/1612.03144, 2016. [Online]. Available: http://arxiv.org/abs/1612.03144

[15] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," *CoRR*, vol. abs/1604.00600, 2016. [Online]. Available: http://arxiv.org/abs/1604.00600

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[18] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," 2022.

[19] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," *CoRR*, vol. abs/1511.06645, 2015. [Online]. Available: http://arxiv.org/abs/1511.06645

[20] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," *CoRR*, vol. abs/1605.03170, 2016. [Online]. Available: http://arxiv.org/abs/1605.03170

[21] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *CoRR*, vol. abs/1812.08008, 2018. [Online]. Available: http://arxiv.org/abs/1812.08008

[22] S. Kreiss, L. Bertoni, and A. Alahi, "Pifpaf: Composite fields for human pose estimation," *CoRR*, vol. abs/1903.06593, 2019. [Online]. Available: http://arxiv.org/abs/1903.06593

[23] A. Newell and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," *CoRR*, vol. abs/1611.05424, 2016. [Online]. Available: http://arxiv.org/abs/1611.05424

[24] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Bottom-up higher-resolution networks for multi-person pose estimation," *CoRR*, vol. abs/1908.10357, 2019. [Online]. Available: http://arxiv.org/abs/1908.10357

[25] Z. Luo, Z. Wang, Y. Huang, T. Tan, and E. Zhou, "Rethinking the heatmap regression for bottom-up human pose estimation," *CoRR*, vol. abs/2012.15175, 2020. [Online]. Available: https://arxiv.org/abs/2012.15175

[26] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:206593710

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[28] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," *CoRR*, vol. abs/1603.03417, 2016. [Online]. Available: http://arxiv.org/abs/1603.03417

[29] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, 2016. [Online]. Available: http://arxiv.org/abs/1603.08155

[30] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4105–4113.

[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[32] S.-C. Zhu, X. Liu, and Y. N. Wu, "Exploring texture ensembles by efficient markov chain monte carlo-toward a 'trichromacy' theory of texture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 554–569, 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:3194236

[33] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," *CoRR*, vol. abs/1703.06868, 2017. [Online]. Available: http://arxiv.org/abs/1703.06868

[34] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, 2016. [Online]. Available: http://arxiv.org/abs/1611.07004

[35] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *CoRR*, vol. abs/1611.02200, 2016. [Online]. Available: http://arxiv.org/abs/1611.02200

[36] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," *CoRR*, vol. abs/1704.02510, 2017. [Online]. Available: http://arxiv.org/abs/1704.02510

[37] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," *CoRR*, vol. abs/1703.05192, 2017. [Online]. Available: http://arxiv.org/abs/1703.05192

[38] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *CoRR*, vol. abs/1703.10593, 2017. [Online]. Available: http://arxiv.org/abs/1703.10593

[39] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. W. Fieguth, L. Liu, and M. S. Lew, "Deep image retrieval: A survey," *CoRR*, vol. abs/2101.11282, 2021. [Online]. Available: https://arxiv.org/abs/2101.11282

[40] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *CoRR*, vol. abs/1711.02512, 2017. [Online]. Available: http://arxiv.org/abs/1711.02512

[41] T. Jenícek and O. Chum, "Linking art through human poses," *CoRR*, vol. abs/1907.03537, 2019. [Online]. Available: http://arxiv.org/abs/1907.03537

[42] P. Madhu, A. Villar-Corrales, R. Kosti, T. Bendschus, C. Reinhardt, P. Bell, A. K. Maier, and V. Christlein, "Enhancing human pose estimation in ancient vase paintings via perceptually-grounded style transfer learning," *CoRR*, vol. abs/2012.05616, 2020. [Online]. Available: https://arxiv.org/abs/2012.05616

[43] D. Kadish, S. Risi, and A. S. Løvlie, "Improving object detection in art images using only style transfer," *CoRR*, vol. abs/2102.06529, 2021. [Online]. Available: https://arxiv.org/abs/2102.06529

[44] P. by Numbers. (2016) Kaggle. [Online]. Available: https://www.kaggle.com/c/painter-by-numbers/

[45] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[46] X. Ju, A. Zeng, J. Wang, Q. Xu, and L. Zhang, "Human-art: A versatile human-centric dataset bridging natural and artificial scenes," 2023.

[47] H. Chen, F. Shao, X. Chai, Y. Gu, Q. Jiang, X. Meng, and Y.-S. Ho, "Quality evaluation of arbitrary style transfer: Subjective study and objective metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 7, p. 3055–3070, Jul. 2023. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2022.3231041

[48] B. Saleh and A. M. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," *CoRR*, vol. abs/1505.00855, 2015. [Online]. Available: http://arxiv.org/abs/1505.00855