FACULTY OF ENGINEERING AND ARCHITECTURE

Improving Pose Estimation on Art Collections with Style Transfer

Tristan Verheecke
Student number: 20043518

Supervisors: Prof. dr. ir. Dieter De Witte, Prof. dr. Steven Verstockt
Counsellor: Dr. ir. Kenzo Milleville

Master's dissertation submitted in order to obtain the academic degree of
Master of Science in Information Engineering Technology

Academic year 2023-2024

GHENT UNIVERSITY

Preface

I've been interested in Art my entire life. In fact, I've a degree in the Fine Arts from LUCA School of Arts. There, I was known for my technological ability and one of my professors at the time asked me why I didn't do anything with that in my artworks. That remark has since stuck with me and was part of my motivation to apply for readmission for my Master of Science. With all the advancements in AI, I started thinking more and more about doing work with that. Like Matisse and Turner, I'm not satisfied with the tools available but want to create my own.

It was therefor to my delight that I was able to work on this thesis which has provided me the opportunity to acquire more insight in the subject. I would like to thank my supervisors Dieter De Witte and Steven Verstockt for this wonderful opportunity, and my counsellor Kenzo Milleville for his great guidance. As well as all the other people at IDLab for their feedback. I also want to thank Karine Lacaracina, Lies Van De Cappelle and the other people at RMFAB for providing help with the artistic sensibilities of the thesis.

Enjoy the read,

Tristan Verheecke
Ghent, June 2024

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 thesis 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 thesis 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

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.

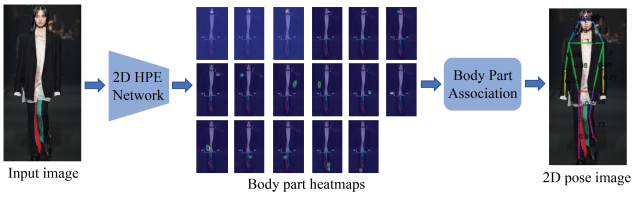
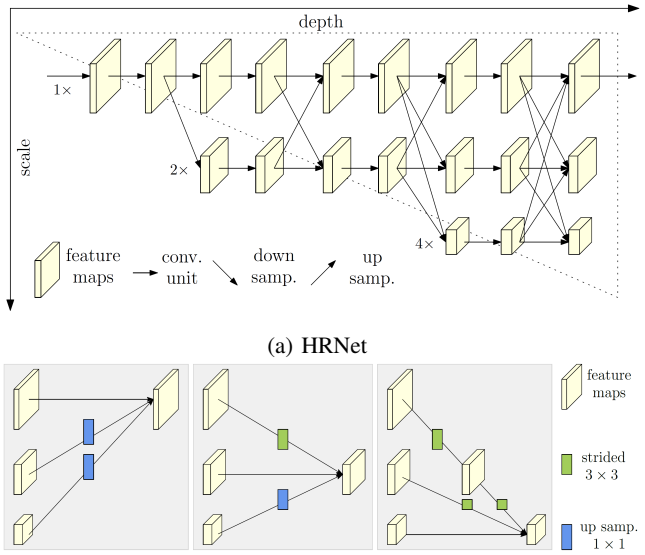


Fig. 1: Heatmap-based Methods as presented in [3]

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.

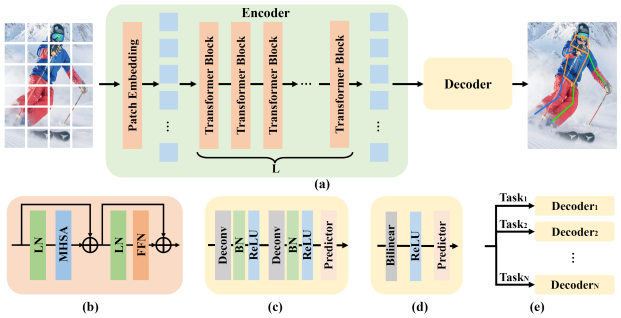


Fig. 3: (a) The framework of ViTPose. (b) The transformer block. (c) The classic decoder. (d) The simple decoder. (e) The decoders for multiple datasets. [18]

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 key-points 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.

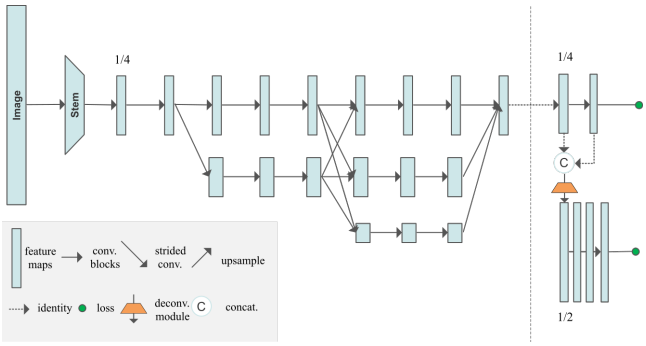


Fig. 4: The architecture of HigherHRNet. It uses HRNet as backbone. [24]

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.

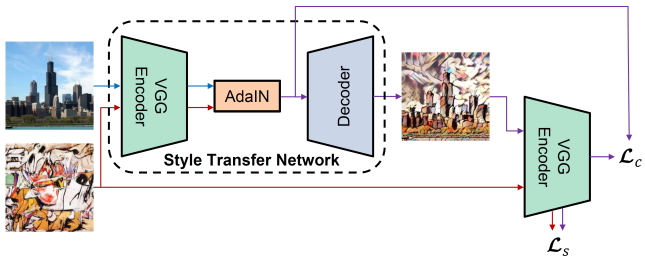


Fig. 5: Adaptive Instance Normalization network by Huang et al. [33].

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

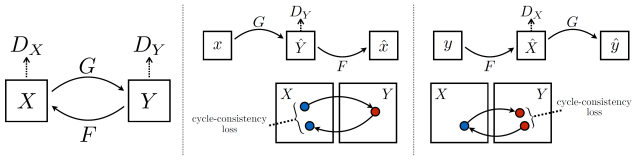


Fig. 6: The cycle-consistent network by Zhu et al. [38]. Unsupervised image-to-image translation between domains X and Y is established by training the generators G, F and discriminators D_X, D_Y. During training cycle-consistency loss is calculated under the assumption that $F(G(x)) \stackrel{!}{=}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

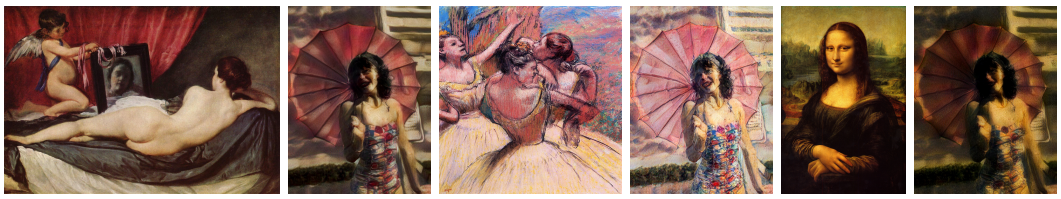
1) Datasets: None of the selected models has any pre-trained weights for certain art movements, so new models need to be made. The WikiArt dataset will be used to assemble different datasets and specifically on the Baroque, Renaissance and Impressionism styles. The impressionist style is chosen because 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) [29], Inception Score (IS) [49], Fréchet Inception Distance (FID) [50] and Learned Perceptual Image Patch Similarity (LPIPS) [51] 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. The first is 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 as well which focuses around people. This is created the same way the style transfer datasets were created.

For the evaluation, AdaIN cycles through the style images which it was trained on to use as input style images. The perceptual distance needs a content and style image to be able to make an evaluation. For AdaIN, it is clear what needs to be used here, but for CycleGAN this metric seems useless. However, the dataset it was trained on can be used as style images for this. The style features of the generated images should still be similar as the ones it was trained on. These same datasets are also used for the real image distribution needed for FID and LPIPS. 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. A pattern arises where AdaIN clearly does well with IS and FID, CycleGAN does not do well in any, and LPIPS has similar results for all.

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. Model Training

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, as this was the best scoring model from the quantitative evaluation. This results in a combination of 16 models.

A. Styled COCO Datasets

The next step in the process is to create the stylized COCO datasets on which the pose estimation models will be trained. A version was created from each style transfer model that CycleGAN was trained; baroque, impressionism and renaissance. AdaIN requires a style image to transform the images and for this purpose the images from the datasets used for training the CycleGAN models were used. From these

style datasets, images were sampled randomly to transform the COCO dataset with AdaIN. The decision to not use one image as a representation of 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.

B. Training

For each created styled COCO dataset, VitPose and SWAHR models are trained with the default parameters provided by their respective papers. They also use the default learning rate. No human detection model is trained, instead, the ground truth bounding boxes will be used to extract the poses for top-down algorithms. The weights are initiate with the pre-trained models. They're trained for 200 epochs. As a control, two models are trained without initiating weights. These were trained for the default 300 epochs.

VI. Experiments

The experiments will be conducted on the validation set of the COCO-dataset as well as the Human-Art dataset. While the problem specifically focuses on improving the performance on artworks, it's still interesting to also validate the results on the COCO-dataset. For the experiments, there will be two baselines that the results will be compared against. First, the pre-trained models that have been made available courtesy of the respective authors, and second, a control model trained on the same parameters that the augmented models are trained on. The control models are trained on the default COCO dataset like the pre-trained models are. Object Keypoint Similarity (OKS) is used as evaluation metric for the experiments to determine AP and AR.

A. Results

1) Performance on COCO dataset: For the SWAHR network, the best results are found for the model trained on the COCO + AdaIN mixed styled dataset. The second best network was trained on the COCO + CycleGAN mixed styled dataset. For the ViTPose network, the best results are for 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 different baselines. It shows that the pre-trained SWAHR model has the best precision of all of the models and trained on the COCO + AdaIN Mixed styled dataset, SWAHR also has the second best precision. ViTPose trained on COCO + CycleGAN mixed styled dataset has the best recall. The networks trained from the ground up don't have any significant difference between the other networks.

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
Control SWAHR	0.620	0.830	0.710	0.891
Control 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
Control SWAHR	0.492	0.742	0.563	0.784
Control 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.

2) Performance on Human-Art dataset: In the results on the Human-Art dataset one dataset takes the crown. Both SWAHR as well as ViTPose have the best results for the models trained on the COCO + CycleGAN mixed styled 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 two to 3% AP for CycleGAN. The best precision and recall belongs to the SWAHR models. Comparing the best models to the baselines (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. The full results of the experiments can be found in the Appendix A.

3) Performance compared to related papers: Similar experiments have already been run by others. A comparison with their results is appropriate. First, Madhu et al. [42] compare three different methods with their baseline: a styled model that's trained completely on their stylized COCO dataset, and two fine-tuned models on their artwork dataset. Since fine-tuning was not used in this thesis, the results will only be compared with their styled models. Second, Kadish et al. [43] only fine-tune a Faster R-CNN object detection network with their stylized COCO dataset. They perform the tests on the People-Art dataset [52]. This dataset is only labelled with bounding-boxes. Table IV shows that both their findings are similar as what was found in this thesis, except that their results are more pronounced. This shows that as a general trend, augmenting a dataset with synthetic artworks will give better results for art collections.

TABLE IV: Improvements made by Madhu et al. [42] and Kadish et al. [43] compared to the results in this thesis. For Madhu et al., the best results of their styled models on the COCO-dataset and their specialized dataset are used. Kadish et al. only compare their results with other papers. So, their baseline is from Gonthier et al. [53]. All values are in terms of AP, except for the values from Kadish et al. which are AP_{50}.

	Madhu et al.		**Kadish et al.**	**This paper**	
Model	**COCO**	**Theirs**	**People-Art**	**COCO**	**Human-Art**
Baseline	76.5	24.7	58	68.7	52.8
Styled	74.3	32.3	68	67.9	55.3
Difference	-2.2	+7.6	+10	-0.8	+2.5

VII. Conclusion

To achieve better pose estimation results on art collections, the COCO dataset is transformed with multiple style transfer methods to styled COCO datasets. With these styled datasets, new pose estimation models can be trained which work better than the already existing ones. This requires a style transfer method that is able to transform between artworks and photographs. Therefor, several datasets were created by using CBIR on the WikiArt dataset. The focus of these datasets is mainly around the human figure as this is the domain of pose estimation. Several style transfer models were trained and several new styled COCO datasets were created. After training the pose estimation networks successfully, it was found that they were able to increase the performance on art collections. To establish this, evaluation on the Human-Art dataset was very helpful. It was found that training the models on an augmented dataset can increase the performance by at least 2% AP. This is in line with related works who reported a similar increase. While ViTPose is considered state-of-the-art, in the experiments run, it does not outperform SWAHR for any of the evaluations. This shows again, that even though a model can be state-of-the-art in one task, this does not translate to other tasks. During training, the pose estimation networks converged very quickly and training them for 200 epochs wasn't needed, but instead only 100 epochs would have been enough.

To get better results, a first recommended improvement can be to use style transfer networks that have a higher fidelity. One promising technique is stable diffusion [54] which is able to synthesize high fidelity images. It would be useful if this could be used for style transfer. The created datasets for style transfer can also be more refined. Instead of having crowded images, the dataset could only focus on the human figure front and central, and crop them out as well. While this reduces generalization, the problem is only about pose estimation. This work only focuses on realistic artworks to run pose estimation on, but this leaves out more abstract works. Another area that deals with this is style transfer with geometric transformations. Further research into this can extend the now limited approach. During the experiments, the evaluations were compared with two baselines; the pre-trained model and a control model. The control was trained the same way the styled models were and had a worse performance than the pre-trained baseline. This means that the networks could possibly be fine-tuned to perform better. The difference between the pre-trained and the control is as high as 6% for SWAHR. This is a noteworthy difference and enough to warrant further research. All in all, there are still a lot of areas that can be improved upon.

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

[49] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *CoRR*, vol. abs/1606.03498, 2016. [Online]. Available: http://arxiv.org/abs/1606.03498

[50] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a nash equilibrium," *CoRR*, vol. abs/1706.08500, 2017. [Online]. Available: http://arxiv.org/abs/1706.08500

[51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *CoRR*, vol. abs/1801.03924, 2018. [Online]. Available: http://arxiv.org/abs/1801.03924

[52] H. Cai, Q. Wu, T. Corradi, and P. Hall, "The cross-depiction problem: Computer vision algorithms for recognising objects in artwork and in photographs," *CoRR*, vol. abs/1505.00110, 2015. [Online]. Available: http://arxiv.org/abs/1505.00110

[53] N. Gonthier, S. Ladjal, and Y. Gousseau, "Multiple instance learning on deep features for weakly supervised object detection with extreme domain shifts," *CoRR*, vol. abs/2008.01178, 2020. [Online]. Available: https://arxiv.org/abs/2008.01178

[54] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR*, vol. abs/2112.10752, 2021. [Online]. Available: https://arxiv.org/abs/2112.10752

APPENDIX

A. Results Experiments

TABLE V: Establishing a baseline for Pose Estimation on Artworks; Average Precision/Recall (AP/AR). The table shows the performance of the pre-trained and control models measured on The COCO dataset and the Human-Art dataset.

Method	**AP**	**AP**^{50}	**AP**^{75}	**AP**^{M}	**AP**^{L}	**AR**	**AR**^{50}	**AR**^{75}	**AR**^{M}	**AR**^{L}
COCO dataset										
Pre-trained SWAHR	**0.687**	**0.881**	**0.748**	**0.639**	**0.757**	0.737	0.904	0.788	0.670	**0.828**
Control SWAHR	0.620	0.830	0.684	0.604	0.653	0.710	0.891	0.765	0.640	0.803
Pre-trained ViTPose	0.588	0.832	0.641	0.573	0.629	0.723	0.906	0.782	0.682	0.786
Control ViTPose	0.609	0.847	0.680	0.597	0.644	**0.740**	**0.918**	**0.810**	**0.703**	0.795
Human-Art Dataset										
Pre-trained SWAHR	**0.528**	**0.759**	**0.565**	0.099	**0.573**	**0.593**	0.635	0.629	0.177	**0.635**
Control SWAHR	0.492	0.742	0.536	0.058	0.539	0.563	0.784	0.606	0.109	0.605
Pre-trained ViTPose	0.380	0.656	0.385	0.108	0.420	0.571	0.803	0.620	0.279	0.599
Control ViTPose	0.406	0.682	0.415	**0.130**	0.445	0.591	**0.818**	**0.632**	**0.306**	0.619

TABLE VI: Performance of different Pose Estimation models trained on Style Transformed datasets on COCO dataset.

Method	**AP**	**AP**^{50}	**AP**^{75}	**AP**^{M}	**AP**^{L}	**AR**	**AR**^{50}	**AR**^{75}	**AR**^{M}	**AR**^{L}
AdaIN										
Trained on COCO + Mixed Style Transfer										
SWAHR	**0.679***	**0.874***	0.735	**0.628***	0.751	**0.732**	**0.902**	0.782	0.651	0.824
ViTPose	0.618	0.859	0.685	0.599	0.661	0.748	0.924	0.816	0.709	0.805
Trained on COCO + Impressionism Style Transfer										
SWAHR	0.669	0.862	0.733	0.607	**0.755***	0.729	**0.902**	0.782	0.651	**0.834**
ViTPose	0.609	0.843	0.664	0.590	0.654	0.742	0.916	0.801	0.702	0.799
Trained on Mixed Style Transfer										
SWAHR	0.603	0.843	0.661	0.535	0.704	0.676	0.882	0.726	0.586	0.794
ViTPose	0.518	0.783	0.557	0.492	0.573	0.669	0.880	0.726	0.617	0.739
Trained on Impressionism Style Transfer										
SWAHR	0.591	0.830	0.654	0.527	0.688	0.663	0.873	0.716	0.574	0.780
ViTPose	0.497	0.784	0.531	0.463	0.564	0.650	0.874	0.710	0.594	0.728
CycleGAN										
Trained on COCO + Mixed Style Transfer										
SWAHR	0.672	0.863	**0.737***	0.618	0.747	**0.732**	**0.902**	**0.787**	**0.660**	0.827
ViTPose	**0.635**	**0.861**	0.697	0.616	**0.681**	**0.763***	**0.925***	**0.825***	0.723	**0.820**
Trained on COCO + Impressionism Style Transfer										
SWAHR	0.663	0.862	0.724	0.606	0.743	0.714	0.889	0.764	0.637	0.815
ViTPose	0.633	0.859	**0.701**	**0.618**	0.670	0.761	0.922	0.828	**0.725***	0.812
Trained on Mixed Style Transfer										
SWAHR	0.653	0.858	0.711	0.609	0.714	0.716	0.898	0.764	0.647	0.807
ViTPose	0.595	0.844	0.654	0.586	0.628	0.731	0.912	0.795	0.698	0.780
Trained on Impressionism Style Transfer										
SWAHR	0.661	0.864	0.717	0.621	0.719	0.718	0.896	0.765	0.656	0.802
ViTPose	0.591	0.841	0.643	0.582	0.619	0.727	0.910	0.790	0.695	0.773

* the best result overall.

TABLE VII: Performance of different Pose Estimation models trained on Style Transferred datasets on Human-Art dataset.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
AdaIN										
Trained on COCO + Mixed Style Transfer										
SWAHR	0.549	**0.791***	0.600	0.065	**0.602***	0.622	0.834	0.668	0.141	0.667
ViTPose	0.420	0.724	0.440	**0.151***	0.460	0.600	0.843	0.650	0.300	0.630
Trained on COCO + Impressionism Style Transfer										
SWAHR	0.540	0.779	0.576	0.071	0.591	0.612	0.822	0.646	0.156	0.655
ViTPose	0.421	0.706	0.430	0.149	0.458	0.600	0.831	0.641	0.303	0.629
Trained on Mixed Style Transfer										
SWAHR	0.492	0.750	0.525	0.048	0.547	0.581	0.811	0.625	0.142	0.622
ViTPose	0.332	0.627	0.316	0.079	0.372	0.522	0.784	0.559	0.223	0.551
Trained on Impressionism Style Transfer										
SWAHR	0.488	0.738	0.524	0.058	0.543	0.581	0.804	0.624	0.153	0.621
ViTPose	0.321	0.600	0.302	0.094	0.355	0.514	0.765	0.539	0.232	0.542
CycleGAN										
Trained on COCO + Mixed Style Transfer										
SWAHR	**0.553***	0.789	**0.604***	**0.122**	0.598	**0.629***	0.839	**0.677***	**0.208**	**0.669***
ViTPose	**0.439**	**0.726**	**0.458**	0.140	**0.481**	0.617	0.844	0.661	0.324	**0.646**
Trained on COCO + Impressionism Style Transfer										
SWAHR	0.522	0.778	0.556	0.113	0.565	0.590	0.819	0.628	0.173	0.630
ViTPose	0.438	0.724	0.448	0.147	0.479	**0.619**	**0.846***	**0.664**	**0.358***	0.645
Trained on Mixed Style Transfer										
SWAHR	0.524	0.779	0.559	0.102	0.569	0.613	**0.843**	0.645	0.200	0.652
ViTPose	0.405	0.696	0.419	0.148	0.442	0.590	0.829	0.639	0.338	0.615
Trained on Impressionism Style Transfer										
SWAHR	0.505	0.761	0.539	0.116	0.546	0.587	0.822	0.622	**0.208**	0.623
ViTPose	0.407	0.694	0.412	**0.151***	0.444	0.590	0.828	0.631	0.341	0.615

* the best result overall.

Contents

List of Figures

List of Tables

List of Acronyms

A

AdaIN	Adaptive Instance Normalization , 14, 21
AFHQ	Animal Face High Quality xviii, 22, 33
AIC-HKD	AI Challenger Human Keypoint Detection , 6
AP	Average Precision , 11
AR	Average Recall , 11
AST-IQAD	Arbitrary Style Transfer Image Quality Assessment Database , 31

B

BN	Batch Normalization , 13

C

CBIR	Content Based Image Retrieveal , 17, 18, 23, 32, 45
cGAN	conditional Generative Adversarial Network , 3, 8, 14
CIN	Conditional Instance Normalization , 13, 14
CIR	Category Image Retrieval , 17
CNN	Convolutional Neural Network , 3, 9
COCO	Common Object in Context , 6, 12
CPMs	Convolutional Pose Machines , 7, 9
CPN	Cascaded Pyramid Network

F

Faster R-CNN	Faster Region-based Convolutional Neural Network , 8
FID	Fréchet Inception Distance , 17, 31
FLIC	Frames Labeled In Cinema , 6

G

GAN	Generative Adversarial Network , 3, 14, 16, 21

H

HPE	Human Pose Estimation , 1
HRNet	High-Resolution Net

I

IIR	Instance Image Retrieval , 17
ILP	Integer Linear Programming , 9
IN	Instance Normalization , 13
IoU	Intersection over Union
IS	Inception Score , 16, 17, 31, 32

L

LPIPS	Learned Perceptual Image Patch Similarity , 17, 31
LSP	Leeds Sports Pose , 5

M

MPII	Max Planck Institute for Informatics xvii, 4–6
MSE	Mean Square Error , 14, 16

N

NMS	Non-Maximum-Suppression , 9
NST	Neural Style Transfer , 3, 14

O

OKS	Object Keypoint Similarity , 11

P

PAF	Part Affinity Field , 9
PAF	Part Association Fields , 10
PCk	Percentage of Correct Keypoints , 36
PCkh	Percentage of Correct Keypoints head
PCP	Percentage of Correct Parts , 36
PD	Perceptual Distance , 16, 31
PDJ	Percentage of Detected Joints , 10
PIF	Part Intensity Fields , 10

R

ResNet	Residual Network , 8, 9
RMFAB	Royal Museums of Fine Arts of Belgium , 1
RPME	Regional Multi-person Pose Estimation , 8

S

SAHR	Scale-adaptive Heatmap Regression , 10

SIFT Scale-Invariant Feature Transform , 18

T

TF-IDF Term Frequency-Inverse Document Frequency , 18

V

VAE Variational Autoencoder , 16

ViT Vision Transformer , 9

W

WAHR Weight-adaptive Heatmap Regression , 10

1

Introduction

1.1 Problem definition

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. While viewing the original artworks firsthand still provides a unique experience, the rarest and most beloved works are sealed behind glass. Through digitalization, a museum can offer people the experience of viewing them from up close. It also allows them to make more of their collection available as now they can only show a small percentage of their work at once. Some people might never have the chance to see these works, but through digitalization this becomes possible [20].

Digitization can also 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. These algorithms scan the images in search of recognizable objects and add their labels to the metadata. Even the latest state-of-the-art technology, struggles to recognize objects when pointed at a painting in a museum.

A solution may be to start over and have paintings be annotated by humans. This has been done in two recent projects: Saint-George-On-A-Bike [21] and INSIGHT [22]. However, paintings are very complex while manual annotation doesn't scale and is very expensive. For example, 10,000 paintings were annotated by the Royal Museums of Fine Arts of Belgium (RMFAB) with no clear return on investment [23]. They spent a year on this and this is not something they want to repeat. How can we automate this process and ensure that state-of-the-art computer vision models give good results on paintings and artworks?

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 thesis, 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.

1.2 Proposed solution

To make the vast quantity of research around HPE available to art collections, there are two proposed solutions that will be explored. The efficacy of these methods wil be analyzed in this thesis.

1. The input artwork is first converted to photographic realism on which pose estimation is then executed. With this method the pre-trained models from the state-of-the-art architectures can be reused without need to do any new adjustments. This method does require the style transfer network to have a high fidelity to realism.

2. 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.

1.3 Thesis Outline

Chapter 2 discusses the related literature. The research in style transfer is looked at in depth, as well as the extensive body of work in pose estimation. Content-Based Image Retrieval is shortly described and the chapter closes with a look at deep learning on art collections.

Chapter 3 discusses the use of different style transfer models. Novel datasets are created to train several new models. The models are evaluated with different evaluation metrics and a selection of the best is made.

Chapter 4 discusses the creation of a baseline for the experiments. Two different methods are used to achieve this.

Chapter 5 discusses two improvement proposals. The first transforms the input for existing pose estimation models to photorealistic images. The second augments the COCO-dataset with style transfer to train new pose estimation models. Afterwards, the results are discussed.

Chapter 6 wraps up the thesis with a summary and suggestions for future research.

2

Literature study

In order to correctly implement a solution, we need to first understand the fundamentals. These consist of two research fields: HPE and Neural Style Transfer (NST). The former will be used to detect poses in the art collections, but not before the latter has tried to make an improvement. Following will be an overview of the available research in these domains. Discussing what the goals of them are, how they achieve it, what their challenges are and their limitations.

2.1 Generative Adversarial Network

Convolutional Neural Networks (CNNs) have become the default go-to for many visual tasks and, in order to train a CNN correctly, it needs to minimize a loss function. This is something that still requires a great amount of effort in manually crafting loss-functions. For example, when using Euclidean distance to calculate loss during image generation, it will create a network that outputs blurry images, as minimizing Euclidean distance is achieved by averaging the output. Having a loss-function that does what it should is a difficult problem to solve. To sidestep these complications, Goodfellow et al. [24] propose a framework where the generative model competes against a discriminative model that learns to make a distinction between a sample from the real distribution and one from the generative model's distribution. Figure 2.1 shows the architecture of a Generative Adversarial Network (GAN). It consists of a generator G which takes in random noise and attempts to output a sample from a specific distribution. The discriminator D will then try to distinguish between the sample from the generator and a real sample from that distribution. A popular variation of GAN is the conditional Generative Adversarial Network (cGAN), 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.

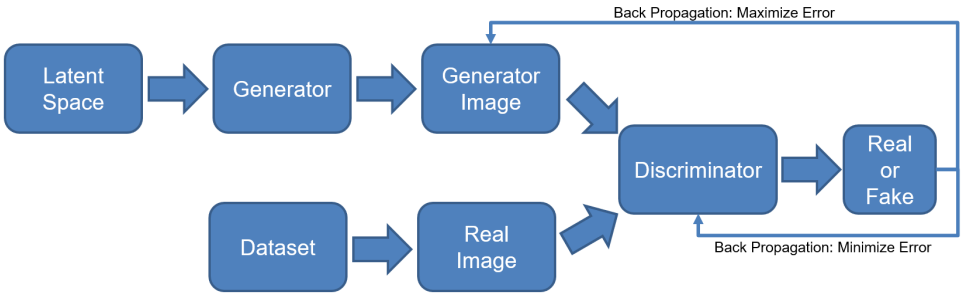


Figure 2.1: The architecture of the Generative Adversarial Network [1].

2.2 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. This chapter will give an overview of all the many challenges and proposed solutions. The focus will be on deep learning models, which have surpassed classical solutions significantly. Specifically, around 2D HPE [25, 4, 26, 27].

Figure 2.2: The various challenges HPE solutions face. Images from MPII dataset. [2, 3]

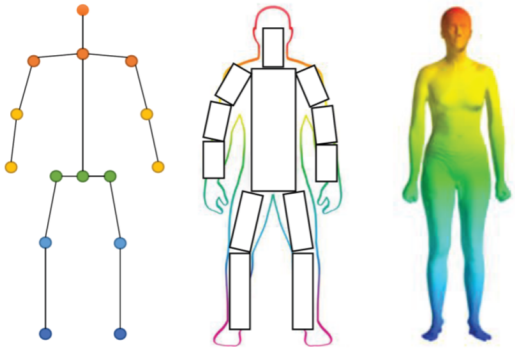
The human body has a high degree-of-freedom due to all the limbs, self-similar parts and body types, which may cause self-occlusion or rare/complex poses. The variations in configuration are made even larger due to clothing, lighting, foreground occlusion, as well as viewing angles and truncation, among others. Examples of this complexity are shown in Figure 2.2. This makes HPE one of the most difficult tasks in computer vision [28, 3].

2.2.1 Representation

An important factor in HPE is how the pose will be represented. Depending on the needs of the problem you can have a skeleton-based, contour-based, or volume-based solution [3] as seen in Fig. 2.3.

Skeleton-based model

The skeleton is made of a tree-structured set of keypoints that represent the joints of the human body. These can be explicitly described by their coordinates in 2D or 3D space [29]. More suitable for a CNN, however, is a heatmap which constructs a 2D Gaussian kernel around a keypoint [26, 30]. As they are easily implemented, they became the dominant representation. While the skeleton-based model is a compact and flexible representation, it suffers in this aspect by not being able to hold texture or shape information [4].



(a) Skeleton (b) Contour (c) Volume

Figure 2.3: Models for pose representation [4]

Contour representation

To capture the shape of the body parts, contour representation uses rectangles to estimate the body contours. These methods include cardboard models [31], which assumes people can be represented as a group of planar patches, and Active Shape Models [32], which tries to fit body part shapes to an image. They were used in earlier HPE methods [3].

Volume representation

Volumetric geometric shapes can also be used as a method of representation. Earlier methods used simple shapes like cylinders, conics, and other shapes [33]. Volume representation is a 3D mesh that represents the human body. The most used model is Skinned Multi-Person Linear, which includes natural pose-dependent deformations imitating soft-tissue dynamics [34].

For the purpose of this research, a simple model is more than adequate. Only the most essential joints are needed to label a pose. This makes the skeleton-based model the ideal representation to work with and will be the focus of further study.

2.2.2 Datasets

There are several publicly available datasets. Some are outdated and will be left out, focusing only on datasets used for deep learning.

1. **Leeds Sports Pose (LSP) Dataset [35]** contains 2,000 images found on Flickr using 8 different tags looking for sport activities (athletics, badminton, baseball, gymnastics, parkour, soccer, tennis, and volleyball). Each person has 14 keypoints. An extended version was later introduced [36], now consisting of 10,000 images. For this set, they only focused on the more challenging tags (parkour, gymnastics, and athletics).
2. **MPII Human Pose Dataset [2]** contains 24,290 images with 40,522 labeled people. They were extracted from YouTube videos found by querying for physical activities. Each person has 16 keypoints and it also includes occlusion labels.

3. **Common Object in Context (COCO) Dataset [37]** 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.

4. **Frames Labeled In Cinema (FLIC) Dataset [38]** contains 5,003 images extracted from Hollywood movies. They ran a person detector which collected 20,000 images from 30 movies. Occluded and difficult poses were then removed leaving only 5,000 images to be annotated. Only the upper body received 10 keypoints.

5. **AI Challenger Human Keypoint Detection (AIC-HKD) Dataset [38]** contains 300,000 images found using Internet search engines. In these, over 700,000 humans are annotated. Each person has 14 keypoints, a bounding-box, as well as visibility and left/right labels.

6. **CrowdPose Dataset [39]** puts an emphasis on crowded images. 30,000 images from MPII, glsCOCO and glsAIC-HKD were measured with a Crowd Index, which evaluates the crowdedness. Finally, 20,000 images are selected and 80,000 persons annotated. Each person has 14 keypoints and a full-body bounding box.

7. **Human-Art Dataset [40]** 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.

2.2.3 Discriminative Methods and Generative Methods

Before deep learning became prominent in HPE, there were already a number of different methods in use. Some of these methods are compatible with the deep learning methods and were promptly adopted. An early distinction is between generative and discriminative methods.

Generative Models work with prior beliefs about the pose. More information about this can be found in the section about representation 2.2.1. It will project the pose on the image and verify it with the image data. If it doesn't comply, the pose is adjusted using descent directions found by minimizing an error function to converge to a local optimization [41].

Discriminative Models on the other hand, try to map the pose on the image data with learned models. There are several methods in this category, among which are the deep learning-based methods. The deep-learning methods are further categorized by the following sections.

2.2.4 Single-Person Pose Estimation Methods

Single-person pose estimation tries to evaluate only one pose from an image. There are two major methods that are in use: regression methods and detection-based methods.

Regression-based Methods learn a network that maps all the body keypoints to the image directly, as shown in Figure 2.4. The first successful deep learning model came from Toshev and Szegedy [29] and is considered the switch in paradigm from classic approaches to deep learning HPE. Based on AlexNet for its simple but effective architecture [42], they use a seven-layered model with five convolution layers and two fully-connected layers for the pose regressor. They then cascade the resulting found keypoints to the next stage where it refines it using the area around the keypoints. While the network is the same, the different stages will have different learned parameters. With every stage, the found keypoints become more

accurate. Carreira et al. [43] introduce an Iterative Error Feedback which is a self-correcting model using top-down feedback. Using the image and a starting pose modeled as a heatmap, the model, based on GoogLeNet [44], will predict an error for each keypoint. The pose is then corrected based on the error and fed into the next module as a heatmap with the input image. With each iteration it converges towards the solution instead of making the prediction in one go. Regression-based methods map the keypoints directly on the image, making it a non-linear problem, which leads to a less robust generalization [26].

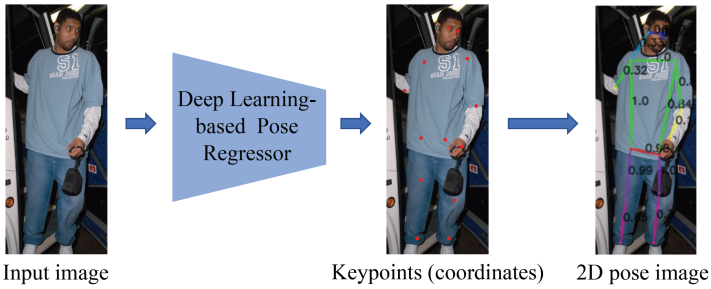


Figure 2.4: Single-Person HPE Regression Methods as presented in [4]

Heatmap and Detection-based Methods will first estimate the individual body parts using heatmaps. This method results in an easier optimization and a more robust generalization [27]. 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 2.5. Tompson et al. [45] proposed a hybrid architecture where the detection of body parts is handled by a CNN and a Spatial-Model to bring those together. The first step produces many false-positives which are removed in the second step by restricting joint inter-connectivity to enforce correct anatomy. They build on this in [46] where they use a cascade to refine predictions.

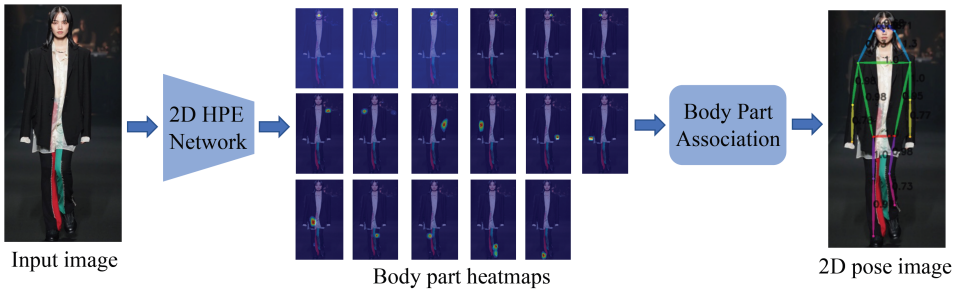
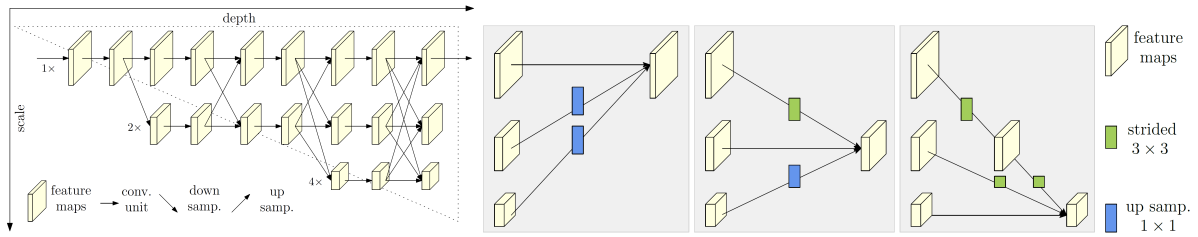


Figure 2.5: Single-Person HPE Heatmap-based Methods as presented in [4]

A fundamental work written by Wei et al. [47] combines convolution networks with Pose Machines [48]. 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. [49]. Similar to Convolutional Pose Machines (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. Several other works [50, 51] have since improved on the network design. 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 Residual Network (ResNet) [52] which allows better back-propagation at deeper levels through shortcuts.



(a) HRNet (b) Multi-Scale Fusion

Figure 2.6: The architecture of the High-Resolution network and how it applies multi-scale fusion [5].

A more recent work by Sun et al. [5] 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 2.6). This network has proven very effective and inspired several variations [7, 53, 54]. Chen et al. [55] propose using cGAN [56] to improve constraints of joint inter-connectivity and infer occluded body parts. A structure-aware convolution network using a stacked hourglass serves as generator which generates pose heatmaps as well as occlusion heatmaps for each joint. Two discriminators are used, one to discriminate between low- and high-confidence predictions, another for real and fake poses. A more classic GAN is used by Chou et al. [57], where they use a stacked hourglass network for both the generator as the discriminator. The generator predicts the heatmaps for each joint and the discriminator distinguished between the real and fake ones.

2.2.5 Multi-Person Methods

With multi-person methods comes an extra layer of difficulty; they need to detect each person separately. To solve this problem multi-person methods propose several solutions. The two most popular are top-down and bottom-up methods.

Top-Down Methods will first try to detect all persons in the image with a human detector. Each person is cropped by the bounding box and a single-person estimator predicts a pose. Occlusion and truncation are a regular occurrence in multi-person scenes and an inevitable problem. One of the early multi-person models, by Iqbal et al. [58], creates a robust model against occlusion. It uses Faster Region-based Convolutional Neural Network (Faster R-CNN) [59] 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. This technique is similar to [60], but, instead of working on all globally found joints, it only considers local joints. It can also handle any kind of occlusion or truncation. The use of a human detector comes with its own set of problems, which Fang et al. [61] try to remedy with Regional Multi-person Pose Estimation (RPME). Their solution consists of two components: They try to tackle inaccurate bounding boxes with a Symmetric Spatial Transformer Network and redundant detections with Parametric Pose Non-Maximum-Suppresion. They also propose a 3rd component, the Pose-Guided Proposals Generator, which can augment training samples. Papandreou et al. [62] use a two stage pipeline. In the first stage, they employ the Faster R-CNN detector. In the second stage, they estimate the pose in each found bounding box using their own network. It

predicts heatmaps using a fully convolutional ResNet and then uses their own novel aggregation procedure. Afterwards, they do post-processing using keypoint-based Non-Maximum-Suppression (NMS); a method of their own making. A continuous effort is taken by Chen et al. [63] to deal with occlusion and truncation. They suggest a two stage architecture, where first the "simple" keypoints are captured with GlobalNet, a feature pyramid network based on [64], and the "hard" keypoints are handled by their RefineNet. It integrates the information via upsampling and concatenating of HyperNet [65] and using an adapted stacked hourglass. They achieved great results and several others improved on their work [66, 67]. 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. [68], 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. [69] created a new model that can work with images; the Vision Transformer (ViT). Xu et al. [6] use the vision transformer to apply it to the HPE task. As seen in Figure 2.7, they try to keep the network simple and don't use certain optimizations that can increase complexity. 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. To show the strong representation ability of transformers the authors provide a simple decoder next to the classic decoder and show that even the simple decoder obtains competitive results.

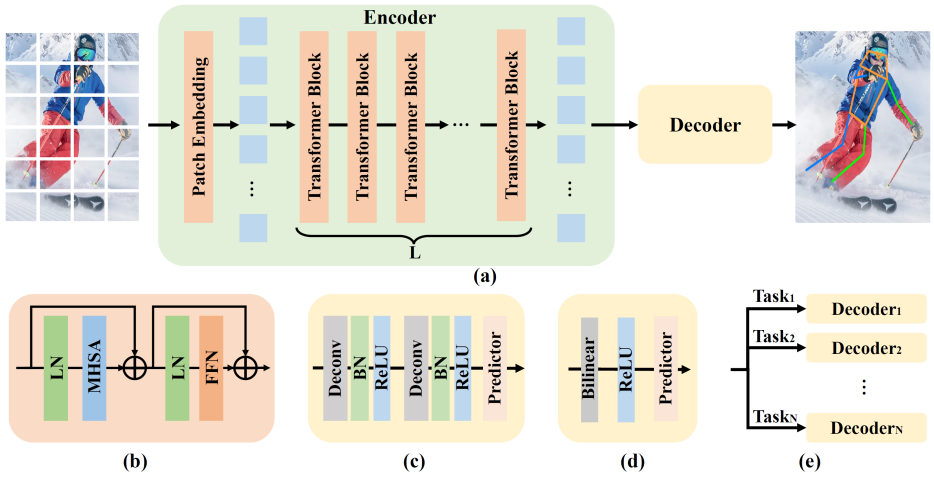


Figure 2.7: (a) The framework of ViTPose. (b) The transformer block. (c) The classic decoder. (d) The simple decoder. (e) The decoders for multiple datasets. [6]

Bottom-Up Methods use a different approach to predict the keypoints. They first locate all joints in the image and then afterwards assemble them in potential poses. DeepCut by Pishchulin et al. [60] is one of the first multi-person models using CNNs. Using Fast R-CNN [59], it detects the body parts and labels each. With the joints found, it then uses Integer Linear Programming (ILP) to assemble them. However, this method is very computationally expensive; NP-hard. Insafutdinov et al. [70] therefor introduce a stronger part detector and better optimization strategy with DeeperCut. CPMs make a return with OpenPose by Cao et al. [71]. They're used to predict the joints with heatmaps and Part Affinity Fields (PAFs). PAFs also encodes the position and orientation of the limb which makes the assembly of joints into different poses more reliable. They can achieve real-time results with this method, and several others have improved on their design [72, 73, 74]. The high performance is only applicable to high-resolution images and low-resolution images or images with occlusions perform poorly.

Kreiss et al. [75] continue on the idea of fields and introduce the Part Intensity Fields (PIF) and Part Association 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. [76] introduce associative embedding which is a new method to represent the output. This is a single-stage architecture as opposed to the two-staged architectures previously discussed. They make use of the stacked hourglass network from [49] with some small modifications, and produce joint heatmaps and associative embedding tags. Continuing on the idea of associative embedding, Cheng et al. [7] use HRNet [5] as backbone for their HigherHRNet (Figure 2.8). Their method focuses on the scale-variance problem; a problem which hasn't been studied much, so it can localize keypoints for small persons better. Lou et al. [77] introduce Scale-adaptive Heatmap Regression (SAHR) and Weight-adaptive Heatmap Regression (WAHR) 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.

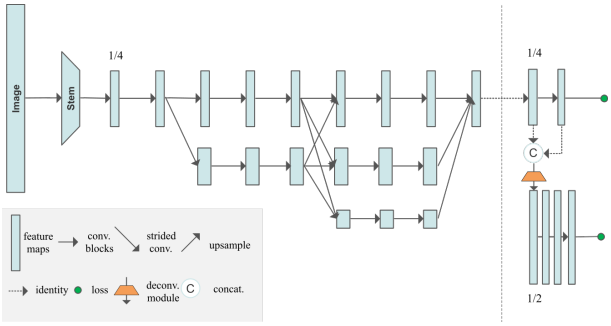


Figure 2.8: The architecture of HigherHRNet. It uses HRNet as backbone. [7]

Summary

An important challenge for HPE is making predictions in scenes with hight occlusions. Top-down models achieve state-of-the art performance on almost all benchmark datasets [3]. However, they have difficulty with overlapping bodies and human detectors add an extra layer of failure. To the same extent, bottom-up models have greater inaccuracy when grouping in occluded scenes. Computationally, the top-down model's speed is heavily bound to the number of people the human detector finds. The higher efficiency of bottom-up models make them more suitable for real-time applications.

2.2.6 Evaluation Metric

The evaluation of an HPE looks to measure the accuracy of the location of predicted joints. Because of the different number of features and tasks across datasets, there are also several different evaluation metrics in use. Explained next will be the most commonly used metrics.

1. **Percentage of Correct Parts (PCP)**, proposed by Ferrari at al. [78], measures the detection rate of limbs. A limb is considered the area between two joints and viewed as detected when the distance between the predicted joints and the real joints is less than halve the length of the limb. This method penalizes shorter limbs and to address this, Percentage of Detected Joints (PDJ) was introduced which instead measures it with a fraction of the torso diameter. The higher, the better.

2. **Percentage of Correct Keypoints (PCK)**, suggested by Yang et al. [79], measures the accuracy of the predicted keypoints. The keypoints should be within a certain threshold, which is a fraction of the person's bounding box size; denoted as PCK@0.2 when it should be less than 20%. It can also be 50% of the head's length; denoted as PCKh@0.5, which makes it "articulation independent". The higher, the better.

3. **Average Precision/Recall (AP/AR)**, by Yang et al. [79], is calculated by counting a keypoint that is within a certain threshold of the ground truth as a true positive. For Lin et al. [37], the AP is calculated by measuring the Object Keypoint Similarity (OKS) (Figure 2.9) which is similar to Intersection over Union in Object Detection. The OKS is defined as:

$$\text{OKS} = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{2.1}$$

Here, d_i is the distance between the predicted keypoint and the ground truth. The distance is run through a unnormalized Gaussian with a standard deviation of sk_i which yields a similarity that ranges between 0 and 1. s is the scale, calculated as the root of the segment area, and k_i is a constant for each keypoint that controls falloff. OKS is the mean of visible keypoints ($v_i > 0$). These can be used to calculate Average Precision (AP) and Average Recall (AR) at different thresholds. 10 different metrics are used to calculate the performance of a model: $\text{AP}^{0.5}$ (where the OKS threshold is 0.5), $\text{AP}^{0.75}$ and AP (the mean of 10 values from OKS $= 0.50$ to 0.95 with a 0.05 step), as well as, AP^M for medium scaled objects and AP^L for large scaled objects. The same metrics are calculated for AR. The higher, the better.

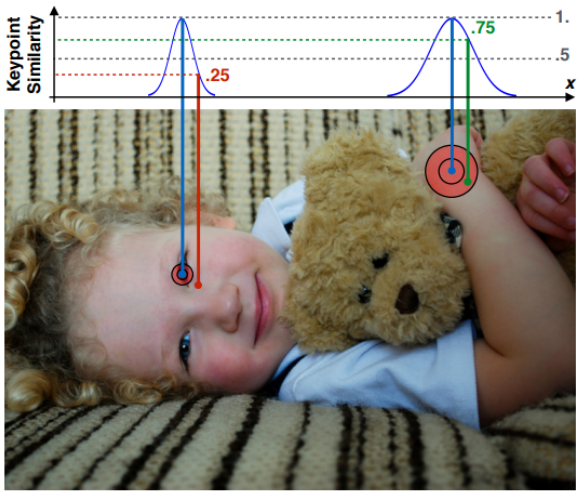


Figure 2.9: An illustration of keypoint similarity. The predicted values (red and green) are at the same distance from the ground-truth (blue). The wrist and eye have a different k_i causing a different falloff [8].

2.3 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. An overview of all the different challenges and solutions will be given in this chapter.

2.3.1 Datasets

Due to a lack of benchmark datasets, multiple papers will mix and match from different datasets, like COCO or ImageNet [80].

1. **Cityscape Dataset [81]** consists of 2975 images of cityscapes with semantic annotations.
2. **Facades Dataset [82]** consists of 400 images of building facades with architectural annotations.
3. **Maps Dataset [83]** consists of 1096 images of maps and areal photos gathered from Google Maps around New York City.
4. **Edges2shoes Dataset [84]** consists of 50,000 paired images between edges and photos of shoes.
5. **Edges2handbags Dataset [85]** consists of 137,000 paired images between edges and photos of handbags.
6. **Horse** $\leftrightarrow$ **Zebra [12]** consists of 2,500 images of 512x512 horses and zebras, that were sampled from ImageNet [80].
7. **Animal Face High Quality [86]** consists of 15,000 high quality images of 512x512 animal faces, including cat, dog and wildlife.
8. **Night2Day Dataset [87]** consists of 20,000 images taken from time-lapse datasets and annotated through crowdsourcing.
9. **WikiArt Dataset [16]** 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.

2.3.2 Optimization-based Networks

Gatys et al. [9] introduce deep neural networks to image style transfer. As seen in Figure 2.10, using a modified VGG-network [88], 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. They remark that the resolution affects the performance of the algorithm and is thus restricted to low resolutions. At the same time, the synthesized images contain some low-level noise, but this can be removed with a denoiser.

2.3.3 Feed-forward Generation Networks

To improve the performance, Ulyanov et al. [89] suggest using a feed-forward generation network instead of reconstruction. Reconstruction requires an iterative process to change the pixel values to match the desired statistics. A feed-forward network can do this in a single evaluation. 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 [9]. Johnson et al. [90] propose an almost

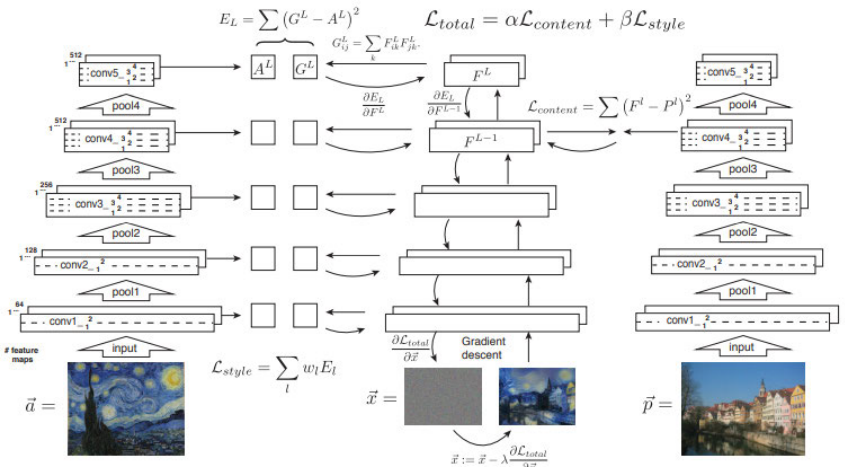


Figure 2.10: Style transfer algorithm by Gatys et al. [9].
The left side transfers the style from a given image, the right side the content.

identical framework independently. The work of Ulyanov et al. did increase the performance, but at the expense of quality, therefor they suggest further improvements to their network [10]. First, they replace Batch Normalization (BN) [91] with Instance Normalization (IN) which alone has a significant impact on quality as can be seen in 2.11. Second, they teach the generator to sample from the Julesz ensemble [92] which improves variation in the outputs.

(a) Content Image (b) Style Image (c) StyleNet with BN (d) StyleNet with IN

Figure 2.11: A comparison between (c) BN and (d) IN.[10]

Dumoulin et al. [93] observe that previous feed-forward networks are limited to one style. In order to facilitate many different styles, there would need to be a network trained separately for each which limits the applications for mobile devices. In order to make the network more memory efficient, they propose a conditional style transfer network; given a content image and a style name, it transforms the image to the corresponding style. They argue that after normalization each style can be distinguished by specializing scaling and shifting parameters. They call this Conditional Instance Normalization

(CIN). Since it only changes the scale and shift parameters for different styles, the network requires fewer parameters. Of the 1.6M parameters, only 3K are needed for the different styles. Another network that puts a focus on multiple styles comes from Chen et al. [94]. They propose a StyleBank which can store multiple convolution filter banks each representing a different style. They use an autoencoder with in between the encoder and decoder a StyleBank layer. During training, for each $T + 1$ iterations the entire network is first trained with a perception loss for the first T iterations. Then only the autoencoder network is trained with a Mean Square Error (MSE) loss. This way the autoencoder only retains the content and the StyleBank layer only the different styles. This also allows them to lock the encoder and decoder to learn a new style afterwards. While CIN allows for multiple styles, it's still limited to the ones that were seen during training. Huang et al. [11] 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 2.12 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. To make a comparison between different networks and how they deal with unseen styles, Figure 2.13 gives an overview.

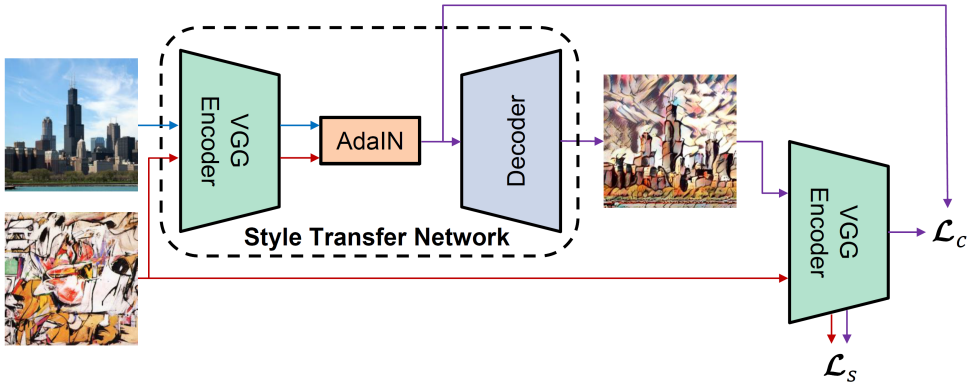
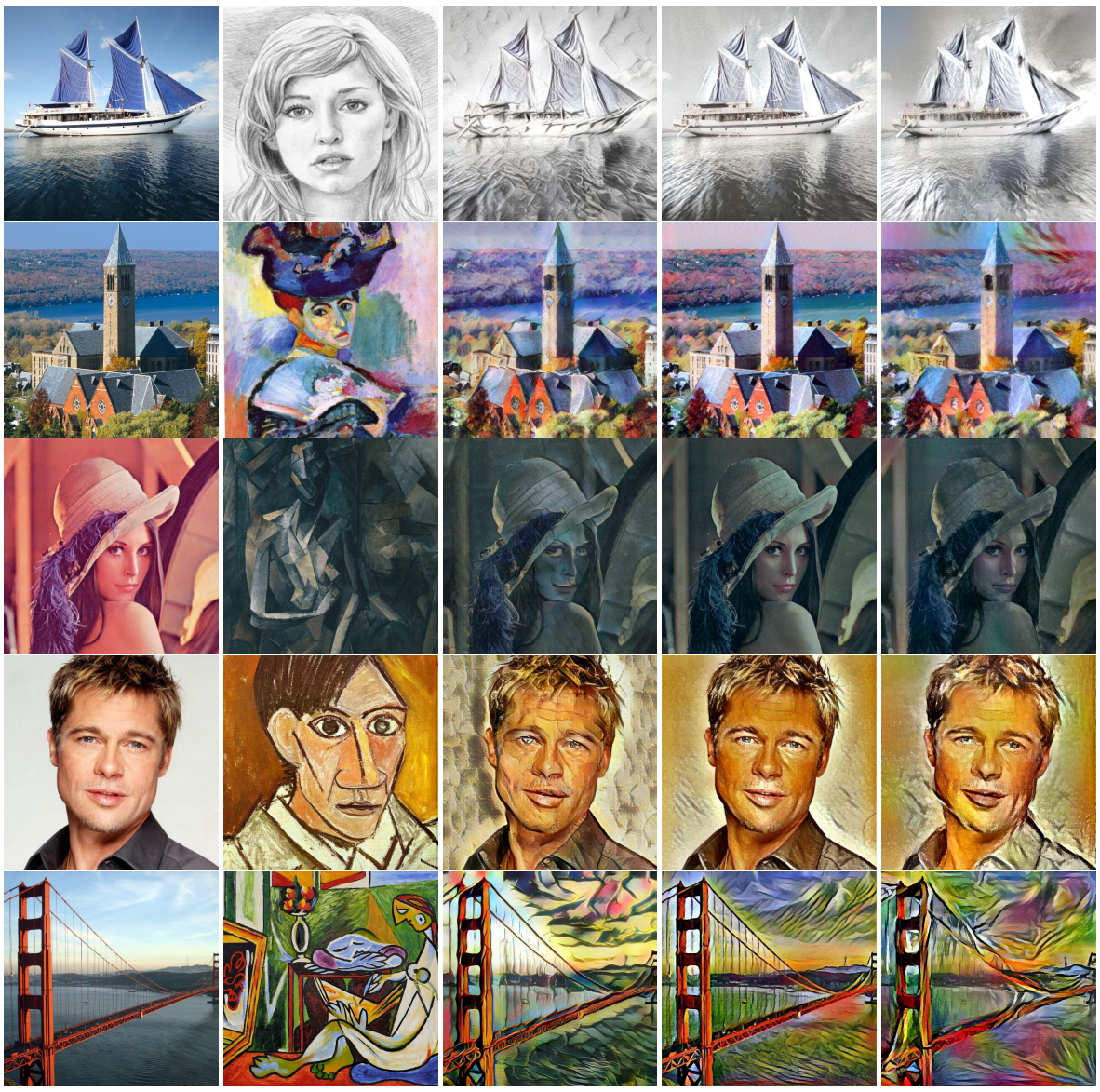


Figure 2.12: Adaptive Instance Normalization network by Huang et al. [11].

2.3.4 Generative Adversarial Networks

With the introduction of GANs, the quality of generative models has greatly increased. It is not surprising then that this got picked up in NST research. Among the first are Isola et al. [83], who use cGAN. They use the network from [95], that uses modules of the form convolution-BatchNorm-ReLu[91]. Additionally, in order to pass shared features in the generator they add skip connections like with "U-Net" [96]. For the discriminator, which they call PatchGAN, they validate $N \times N$ patches and take the average as output. They take this loss together with the $L1$ loss because $L2$ loss produces blurry results. However, this method still requires paired training samples. Meanwhile, Taigman et al. [97] are doing research in unsupervised domain transfer. Research in domain transfer can be easily adjusted for use in NST, but this is not possible the other way around. 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



(a) Content Image (b) Style Image (c) Huang et al. (d) Ulyanov et al. (e) Gatys et al.

Figure 2.13: A comparison between different style transfers where the style was not seen during training.

paired samples though. In order to make the network completely unsupervised, Yi et al.[98] propose DualGAN, Kim et al. [99] DiscoGAN and Zhu et al. [12] 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; the 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 2.14. The goal is to minimize the adversarial and cycle-consistency loss, while maximizing the discriminators' accuracy.

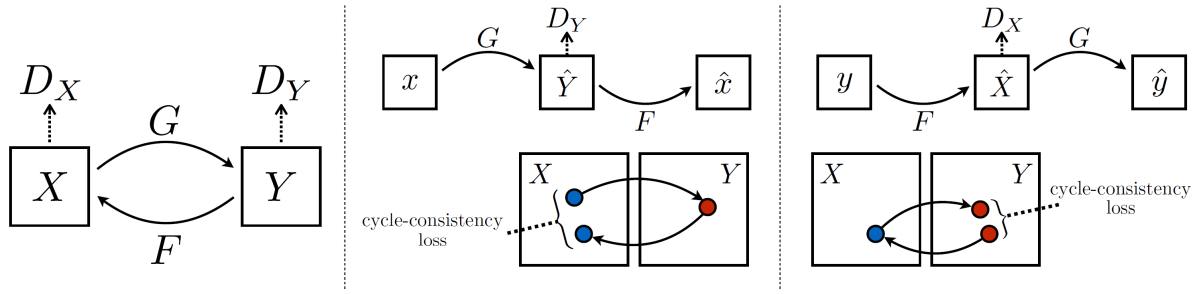


Figure 2.14: The cycle-consistent network by Zhu et al. [12]. Unsupervised image-to-image translation between domains X and Y is established by training the generators G, F and discriminators D_X, D_Y. During training cycle-consistency loss is calculated under the assumption that $F(G(x)) \stackrel{!}{=} x$ and $G(F(y)) \stackrel{!}{=} y$.

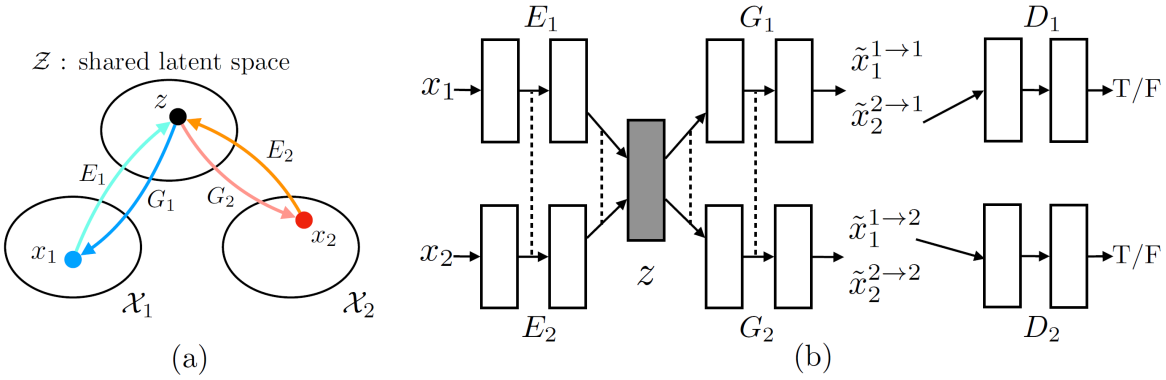
Zhu et al. [12] also introduce an identity loss. Liu et al. [13] introduce the latent space concept which assumes that paired images from different domains can be mapped to a shared latent space with the same latent representation. Their network consists of two domain image encoders E_1 and E_2, two domain image generators G_1 and G_2, and two domain discriminators D_1 and D_2, as can be seen in 2.15. The encoders and generators are paired and form a Variational Autoencoder (VAE) [100]. The encoder maps the input to latent space, and the generator reconstructs the image. This is the reconstruction loss. They use weight-sharing, which shares the weight of the last two layers of the encoders and of the first two layers of the generators. The generators and discriminators are paired to form a GAN. The generator can also construct an image from the latent code from the other encoder's input. This image is used to train the GAN. They also show that the shared-latent space assumption implies cycle-consistency, which is the final loss function of the network.

2.3.5 Evaluation Metric

There are several methods to evaluate the quality of a generated image. A first metric is through human evaluation, where a score is given based on generation quality. This proved to be inconsistent as a person's perception can change over time. Afterwards, new metrics were introduced which will be discussed here. [101]

1. **Perceptual Distance (PD)** is proposed by Johnson et al. [90]. It uses the VGG-16 network [88] trained on ImageNet [80] to define perceptual loss functions. These are extracted from the layers for the style and content images, and compared to the generated image. The lower the score, the better.

2. **Inception Score (IS)**, as described by Salimans et al. [102], uses a pre-trained Inception model [103] to describe the quality of the generated images. It prescribes that the entropy of the distribution of predicted labels for individual



(a) The shared latent space assumption.

(b) The unsupervised image-to-image translation network.

Figure 2.15: Liu et al. [13].

images needs to be minimized while the entropy of the distribution across all images need to be high. This equates to each image having generated a distinct label and the labels being equally distributed. The closer to 1, the better.

3. **Fréchet Inception Distance (FID)** is the most used measurement and suggested by Heusel et al. [104] to enhance IS, because it is only calculated on the distribution of the generated images. FID uses the Gaussian distribution of both real and generated images as it calculates the Fréchet distance [105] between them. The Gaussians are formed from the coding layer of the Inception network [103]. The lower, the better.

4. **Learned Perceptual Image Patch Similarity (LPIPS)** is a metric developed by Zhang et al. [106] and the second most popular. It calculates the distance between the activations of the hidden layers in an object detection model (several models are proposed). They show that this correlates closely to human perception. It can also be used to evaluate the diversity of a network by calculating the average LPIPS score of a pair of randomly generated outputs. The higher, the better.

Summary

There are plenty of other evaluation metrics available that also try to correlate closely to human evaluation, but they are mostly just attempts to improve previously discussed metrics. Until this day, image similarity metrics continue to be a challenging problem.

2.4 Content Based Image Retrieval

Content Based Image Retrieveal (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 [107]. Image retrieval can be categorized into two 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

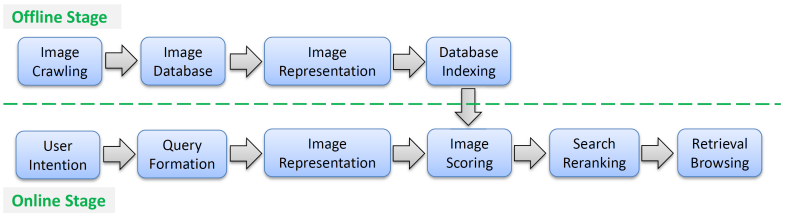


Figure 2.16: The general workflow of Content Based Image Retrieval. [14]

images with a particular instance given in the query image. The general workflow of CBIR is illustrated in **??**. This thesis will only discuss query formation, image representation, image scoring, and search re-ranking.

Query Formation can be done several ways. A user might want to find images based on keywords, which is your standard classification task. Instead of just giving a series of keywords, these can also be arranged in a layout. A query by concept layout will then search for an image with the same arrangement [108]. Similarly, a query by color layout will search for that arrangement of colors in the images [109]. It's also possible that a user wants to find images similar to a sketch (query by sketch) [110] or another image (query by example) [111]. An overview can be found in Fig. 2.17.

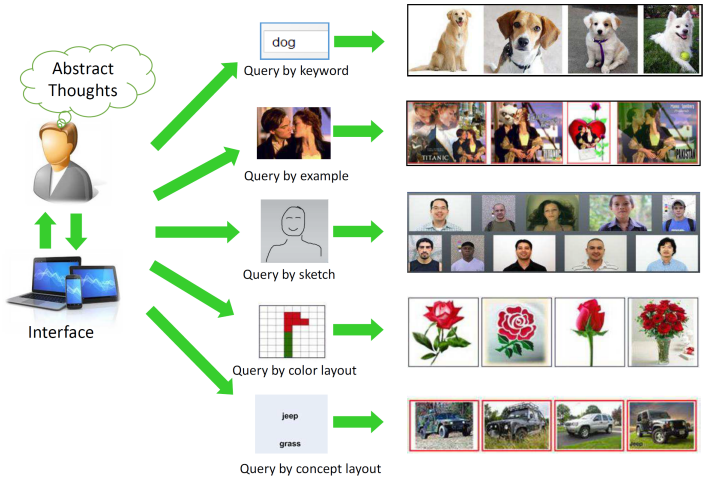


Figure 2.17: An overview of the different kinds of queries with corresponding retrieval results. [14]

Image Representation is a major challenge with image retrieval. It's goal is to proficiently measure similarity between images. Clearly, directly comparing pixels values is impracticable, so methods that extract visual features from images are used. They are transformed into a fixed-sized vector which form a representation of the image. Before deep learning, hand crafted feature algorithms were used. From these, Scale-Invariant Feature Transform (SIFT) [112] was the most popular. Thousands of features can be extracted this way, which is too much for an efficient query response. These features are further compressed for which various methods exist, called feature aggregation.

Image scoring gives the results during a search a relevance score for ranking. This score is determined by one of two ways: During feature aggregation, the image is represented with a fixed-sized vector. The relevance can be calculated with the L_p-normalized distance between the feature aggregation vectors. Scoring can also be quantified based on voting. This can be achieved by counting each similar feature to the vote [113], or through the calculation of Term Frequency-Inverse

Document Frequency (TF-IDF) [114].

Search Reranking are post-processing techniques that improve the accuracy of the query. Among them is geometric context verification, which will try to eliminate false similarities by looking at geometric context, like rotation, scale and the relations between local features. Another is to reuse the highest ranked results to create new queries which can expand on the original. Missed features in the original query can be found this way and improve recall performance. As a last improvement, different retrieval techniques can be merged together to give better results with retrieval fusion.

The model used in this work, by Radenovic et al. [111], trains 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.

2.5 Deep learning in the Art domain

Various papers have already discussed different techniques to improve object detection and pose estimation on artworks. In this section, three of those papers will be discussed. One paper discusses how the digitization of artworks can benefit the analysis of art collections. The other two discuss techniques of how existing models can be adapted to work on art collections.

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, like with religious paintings. In other cases, the inspiration is drawn from themes, which involve composition, lighting and poses. Jenicek et al. [115] 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.

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

Figure 2.18: Improvements to the state-of-the-art by Madhu et al. [15].

To improve these shortcomings on Greek vases, Madhu et al. [15] 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 2.18). Kadish et al. [19] 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 [116] 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, which is further discussed in Section 5.4. Part of this thesis will combine the previously discussed work.

While Madhu et al. only fine-tuned a network to work better for Greek vases, it would be more useful if the fine-tuned model could be used more generally. This is what Kadish et al. achieve, but they focused on the object detection task. Here, both methods will be combined to achieve improvements in pose estimation on non-specific art collections.

3

Style Transfer Model Selection and Building

The goal is to improve pose estimation on art collections. For this effort, two methods will be investigated: First, the input to existing models will be transformed from artworks to photographs. Second, the models will be fine-tuned on an COCO dataset which is augmented with synthetic COCO images. Both these methods need a style transfer model that is trained to do a transformation between an art movement and realistic images. Therefor, three algorithms for style transfer will be explored. The motivation for the choices of the algorithms will be explained in full detail. The different models will be trained on three datasets of different art movements and evaluated based on different metrics. There are several considerations to be made when choosing the right model. The focus will mainly be on methods that have code readily available. At the same time, there should be a wide variation in architectures. It makes little sense to analyze two similar architectures here, as the improvements between them has already been well documented in their corresponding papers and is not threading new ground. All these criteria are considered in the next sections as well as those uniquely for each section.

3.1 Training Style Transfer

3.1.1 Choice of Model

The most important criteria for style transfer is the quality. For the baseline, the photographs need to be inseparable from any artworks for the measurements to be useful. Pose estimation is trained on photographs, so style transfer needs to create accurate photographs. However, measuring the quality of an image is a difficult task. Numerous metrics each based on different criteria exist due to the absence of a universally agreed-upon metric [117], but there is a general consensus that it should closely resemble human evaluation. Of all the different models, the younger models aim more on finding a transformation mapping rather than merely doing texture transfer. To keep the complexity low, the focus will only be on the latter, while keeping to the main advancements. As previously mentioned, a diverse range of architectures should be selected. For these reasons, AdaIN [11] was selected from the feed-forward generation networks. It's also one of the networks which can transform from an arbitrary style unlike the other selected networks. CycleGAN [12] is a major breakthrough in the training scheme of GANs and cycle-consistency loss has since been incorporated in most new models. It also has several pre-trained networks in the styles of several important artists, which makes it an easy second choice. Another interesting concept is that of latent space, where the assumption is that there exists a common space that can encode information from several domains [13]. This is used in StarGANv2 [86] to implement a model that can transform images between several different domains using the same network. In the future, StarGANv2 will be described as StarGAN. Each of these models represent a significant contribution to the field of image-to-image translation and will be analyzed thusly.

3.1.2 Creation of datasets

None of the selected models has any pre-trained weights for certain art movements, so new models need to be made. The most popular and most used dataset for this seems to be the WikiArt dataset. It categorizes the artworks into several art movements, but also multiple genres as summarized in Table 3.1. To keep complexity low, the transformation between styles should be as small as possible, which eliminates the abstract styles as a potential choice, however they should not be hyper realistic either as then they would be so similar to photographs that the benefits of the improvements are meaningless. There are plenty of styles that are compatible with these criteria and also have plenty of images to create a well sized subset. The choice of style beyond that point is completely the result of the bias of the author. This results of the selection being: Baroque, Renaissance and Impressionism.

Table 3.1: List of the selected genres and names of the styles in the WikiArt dataset. [16]

Task Name	**List of Members**
Genre	abstract painting, cityscape, genre painting, illustration, landscape, nude painting, portrait, religious painting, sketch and study, still life
Style	Abstract Expressionism, Action Painting, Analytical Cubism, Art Nouveau-Modern Art, Baroque, Color Field Painting, Contemporary Realism, Cubism, Early Renaissance, Expressionism, Fauvism, High Renaissance, Impressionism, Mannerism-Late-renaissance, Minimalism, Primitivism- Naive Art, New Realism, Northern Renaissance, Pointillism, Pop Art, Post Impressionism, Realism, Rococo, Romanticism, Symbolism, Synthetic Cubism, Ukiyo-e

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. This was a deliberate choice to see if there's possibly a difference between these attributes. The Cezanne2photo dataset [12] was looked at to get an idea of what an adequate sized dataset should be. The conclusion is that it should at least be above 500 images. A bigger dataset is better, but there are only so many artworks available. In the end, the size for all except one are around 800 images. More details about this can be found in the Figures 3.6, 3.7, 3.9, and 3.8. When looking at the datasets mainly used by the unsupervised image-to-image models, there is a very specific focus on certain domains. AFHQ used for StarGANv2 or Horse↔Zebra show that the training images put the subject central in the image. This means that 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. At first glance, it seems that the genres "nude painting" and "portrait" would give a good set of images to use, however there are still multiple problems. The portraits are mostly zoomed in from the chest up (Figure 3.1). There should be a higher variation in poses than that. Like with the nude paintings, but those don't have as many images to create a dataset from (Figure 3.2). Another genre that might be promising, is "genre painting", but those don't always have the model central to the image (Figure 3.3). Overall, there is still a high variety of style even within the different art movements. There is also the presence of sketches or graphite drawings (Figure 3.4). As discussed previously, the art movements were deliberately chosen to see if certain attributes, e.g. color and abstraction, have an influence on the performance of style transfer. It is important then to have a consistent style in each dataset which is not possible to create by just splitting the genres provided by WikiArt. To achieve this, an algorithm was sought to find similar images.

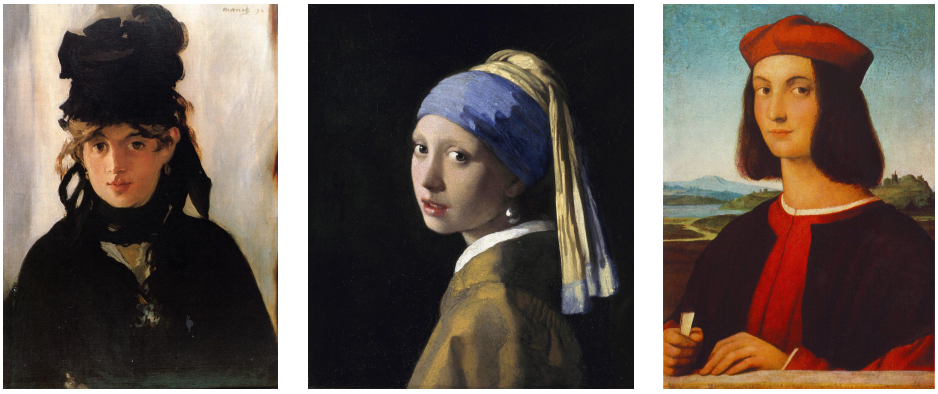


Figure 3.1: Portraits are mainly from the chest up.

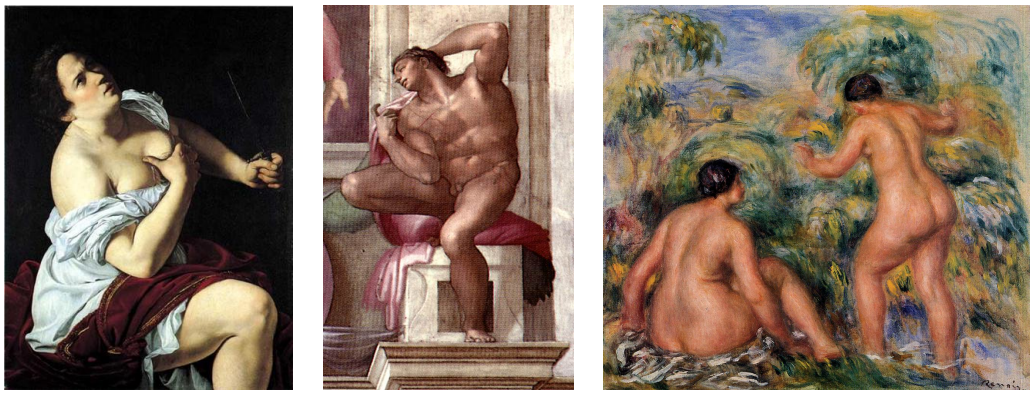


Figure 3.2: Nudes have a better variation of poses, but a small number of images.
23 for baroque, 247 for impressionism and 21 for renaissance.

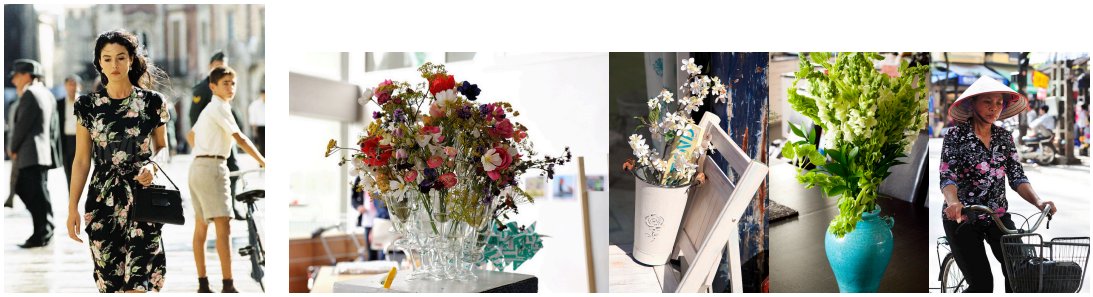
Feature extraction First, an algorithm that extracts features using VGG16-features from the images was used [118]. It calculates the cosine distance between the image features, and groups them using DBSCAN [119]. This did not yield any promising results. Instead of VGG16, YOLOv8 [120] was substituted for feature extraction, but this also didn't provide satisfactory results.

Content Based Image Retrieval Another way to find similar images is with CBIR. Using a query image it can find similar looking images. Because this algorithm is trained to recognize similar instances and not a specific style or genre, the query image needs to be carefully selected. When there's another recognizable instance besides a person in the image it will also score images with that instance highly. Figure 3.5 shows how a car or a flower pattern is enough to find different instances. On the other hand, some activities are so distinct that only instances of that activity are found, like tennis.

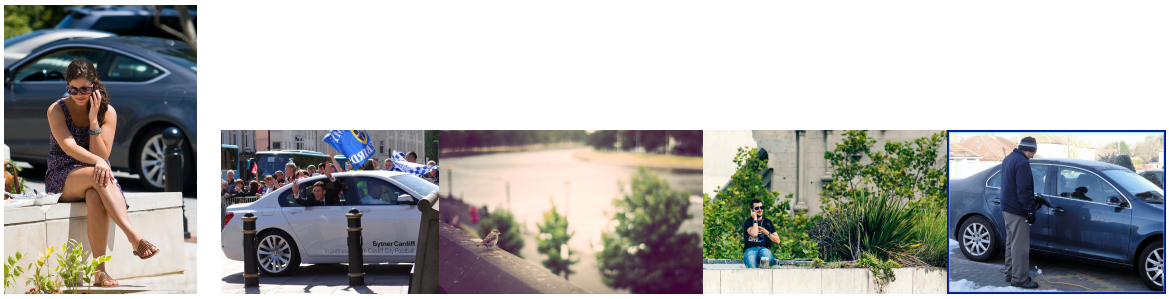
A photograph dataset is also needed to do proper training, therefore the same procedure is also applied to the COCO dataset to create a human central COCO subset. The figures 3.6, 3.7, 3.8 and 3.9 shows the query images used to construct the different datasets along with a selection of the dataset.

Figure 3.3: In genre paintings, humans are less central to the painting.

Figure 3.4: The variation in style within the different art movements.



(a) The flower pattern is isolated



(b) The car and concrete are isolated

Figure 3.5: Examples of failed queries for CBIR. The left image is the query image.

3.1.3 Training

From the selected models there are only two that require training, CycleGAN and StarGAN. 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. The other models 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.

CycleGAN was trained using a different number of epochs for each style to compare the performance . Baroque was trained for 200 and 2000 epochs, renaissance for 500 epochs and impressionism for 750 epochs.

StarGAN does not use epochs to determine the training progression, or, at least, the pytorch implementation doesn't. The model was trained to find a mapping between all different datasets for 100,000 iterations.

(a) Query images

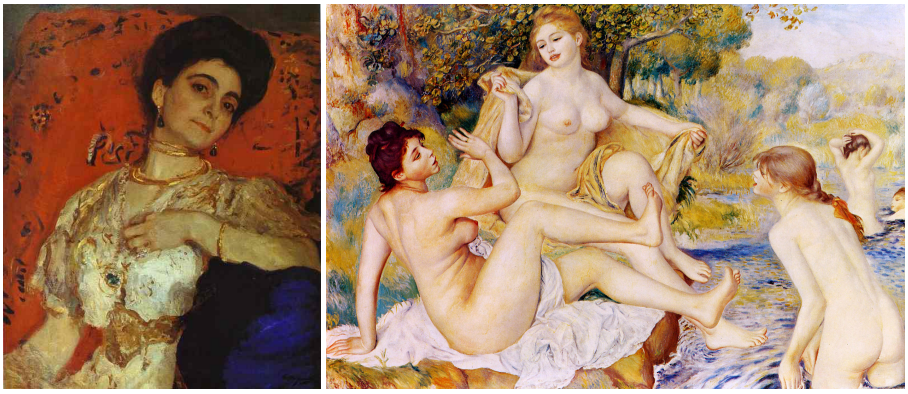
(b) Resulting dataset

Figure 3.6: The photograph dataset consists of 825 images.

(a) Query images

(b) Resulting dataset

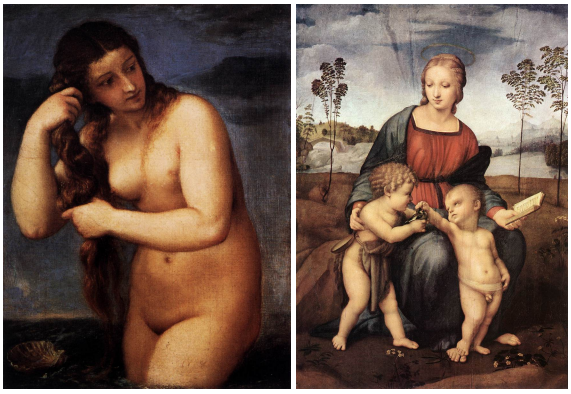
Figure 3.7: The baroque dataset consists of 518 images.



(a) Query images

(b) Resulting dataset

Figure 3.8: The impressionism dataset consists of 780 images.



(a) Query images

(b) Resulting dataset

Figure 3.9: The renaissance dataset consists of 790 images.

3.1.4 Results

Qualitative Evaluation

As shown in Figure 3.10, 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. StarGAN unfortunately experiences modal collapse. In the examples, either the images become complete random splatter, or it is not able to find a correct mapping between the content of different images, e.g. in one image the face is mapped to the back. Looking at the different epochs, it seems that after more epochs the stylization is stronger. All in all, the results are very disappointing as none of the images look like they're a painting from a different time.

Figure 3.10: Left is the content image. The middle-left is AdaIN using a renaissance style image. The middle-right is CycleGAN using the baroque style. The right is StarGAN impressionism. AdaIN abstracts the features more than CycleGAN, while StarGAN experiences modal collapse.

Figure 3.11: Left is the content image. The middle-left is AdaIN using a renaissance style image. The middle-right is CycleGAN using the baroque style. The right is StarGAN impressionism. AdaIN abstracts the features more than CycleGAN, while StarGAN experiences modal collapse.

Quantitative Evaluation

To evaluate the trained models, there exist several metrics, which are discussed in section 2.3.5. 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:

1. Arbitrary Style Transfer Image Quality Assessment Database (AST-IQAD) is a set specifically made to measure style transfer [121]. 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. While it also declares a set of style images, those will not be used.

2. Since the content of the problem of this thesis only focuses around persons and the AST-IQAD dataset works with different kinds of content, a custom dataset is created that focuses around people. This is created the same way the style transfer datasets were created.

For the evaluation, AdaIN cycles through the style images which it was trained on to use as input style images. The perceptual distance needs a content and style image to be able to make an evaluation. For AdaIN, it is clear what needs to be used here, but for the other models this metric seems useless. However, the dataset that CycleGAN and StarGAN were trained on can be used as style images for this. The style features of the generated images should still be similar as the ones it was trained on. These same datasets are also used for the real image distribution needed for FID and LPIPS.

In Table 3.2, the results of the evaluation are available. No model seems to distinct itself from the others. In fact, a pattern arises where AdaIN clearly does well with PD and FID, StarGAN does well with IS, CycleGAN does not do well in any, and LPIPS has similar results for all. Impressionism does the best out of all of the styles. Table 3.2 shows the same results, but grouped by model. This shows that the custom dataset has a slightly better evaluation.

3.1.5 Discussion

While the images are clearly stylized to look vaguely like the style of an art movement, it cannot be said that they belong in the same domain as actual artworks. The stylized images can still be useful to augment the COCO dataset as the question

Table 3.2: 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	Baroque				Impressionism				Renaissance			
	PD	IS	FID	LPIPS	PD	IS	FID	LPIPS	PD	IS	FID	LPIPS
AST-IQAD Dataset												
AdaIN	**10.734**	**8.975**	**265.036**	0.626	**10.671****	8.453	**246.736**	0.710	**10.746****	6.717	**255.062**	**0.696****
CycleGAN	14.670	10.850	272.652	0.633	14.160	10.046	247.468	**0.721**	13.453	9.878	263.348	0.689
StarGAN	13.453	9.878	263.348	**0.689****	17.920	**1.310***	399.215	0.712	18.467	**1.477**	412.430	0.687
Custom Dataset												
AdaIN	**10.507***	6.639	**195.487****	**0.654**	13.435	4.974	**177.581***	**0.737***	**11.472**	5.156	**197.560****	**0.693**
CycleGAN	13.435	7.137	200.299	0.635	**12.456**	6.047	190.658	0.711	12.962	7.825	200.920	0.678
StarGAN	19.302	**1.340****	434.779	0.646	18.028	**1.362**	376.450	0.715	19.608	**1.402****	380.034	0.683

* the best result overall.

** the best result for the style.

Table 3.3: Performance comparison of Style Transfer measured by various metrics grouped by model; Perceptual Distance (PD), Inception score (IS), Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS).

Method	Baroque				Impressionism				Renaissance			
	PD	IS	FID	LPIPS	PD	IS	FID	LPIPS	PD	IS	FID	LPIPS
AdaIN												
AST-IQAD Dataset	10.734	8.975	265.036	0.626	**10.671**	8.453	246.736	0.710	**10.746**	6.717	255.062	**0.696**
Custom Dataset	**10.507**	**6.639**	**195.487**	**0.654**	13.435	**4.974**	**177.581**	**0.737**	11.472	**5.156**	**197.560**	0.693
CycleGAN												
AST-IQAD Dataset	14.670	10.850	272.652	0.633	14.160	10.046	247.468	**0.721**	13.453	9.878	263.348	**0.689**
Custom Dataset	**13.435**	**7.137**	**200.299**	**0.635**	**12.456**	**6.047**	**190.658***	0.711	**12.962**	**7.825**	**200.920**	0.678
StarGAN												
AST-IQAD Dataset	**13.453**	9.878	**263.348**	**0.689**	**17.920**	**1.310**	399.215	0.712	**18.467**	1.477	412.430	**0.687**
Custom Dataset	19.302	**1.340**	434.779	0.646	18.028	1.362	**376.450**	**0.715**	19.608	**1.402**	**380.034**	0.683

whether stylized images can increase the evaluation results is still a useful one to ask. It is obvious that the used evaluation metrics for style transfer are not very helpful. Theoretically, they make complete sense, but they do not at all give a good reading on the quality of the images. The numbers vary greatly, but this variance cannot be seen in the qualitative evaluation. StarGAN, which experienced modal collapse, was still able to score high for IS. Ironically, StarGAN, while not retaining the content, does have the better oil painting characteristics. The identity image, as show in Figure. 3.12, looks like modern art. So, somewhere, the model does approach some kind of human like abstraction, or at least, as seen in abstract art. Perhaps, how artists make abstractions can be used as an inductive bias in future models. Another possible research area could be the use of CBIR models instead of the inception model for the evaluation metrics. CBIR models are more specialized in finding features for similarity measurements.

The question remains why the style transfer algorithms aren't able to make correct mappings between different styles. A first observation was discussed in section 3.1.2. It's a mistake to consider an art movement as a style, as even within the different art movements and realistic photographs there's a big variation in styles. There can be different lighting, different

(a) Input Image (b) Identity Image (c) Richter, Gerard, *Weiß*

Figure 3.12: An example of an image created by StarGAN that has oil painting qualities. A painting from Gerard Richter as comparison is shown.

brush stroke, different camera filter, different lines and different form. There are plenty of things that can vary to make a distinct style. It should be considered whether some things categorized as content now should instead be considered part of the style, like clothes. Whether clothes should be considered content or style can depend on which domains the mapping is searched for. Clothes change dramatically between the different time periods and this is clearly visible when comparing the artworks with photographs. In this context, they should be considered a style. While, when mapping within the same time period, they can be considered content. The same argument can be made for architecture.

When looking at the datasets that CycleGAN and StarGAN are trained on, it becomes clear that most success is made when the domain is extremely specific. As seen in Figure 3.13, all the images contain the subject in the center of the image without any other content. The custom datasets for the training contain a much higher disparity. Perhaps it would be useful to transform different patches where the content is very similar with high certainty at a time, and then combine those to create the transformed image. This can potentially be done by training on a dataset of 3d models where a shader is applied to simulate a different art style. Instead of having to manually label thousands of images, it is possible to have several 3d models act out different poses and render them with different shaders. A network can then be trained to recognize when patches have similar content and apply the style when they do. This will mean that when using an arbitrary style, it might not always find a high similarity and the style transfer will not benefit from this.

Figure 3.13: The AFHQ dataset consists of images that are close-ups of animals.

4

Pose Estimation Model Selection and Baseline

The goal is to improve pose estimation on art collections. For this effort, two methods will be investigated: First, the input to existing models will be transformed from artworks to photographs. Second, the models will be retrained on an COCO dataset which is augmented with synthetic COCO images. In the previous section, a style transfer model was trained for this, and now, a proper baseline needs to be established to compare to the results of the experiments. This chapter will determine that baseline. For this, two pose estimation algorithms will be explored. The motivation for the choices of the algorithms will be explained in full detail. The focus will be on quality instead of speed. The precision of pre-trained pose estimation models will be measured on the COCO dataset to establish a ground truth. Afterwards, the pre-trained models will be validated on the Human-Art dataset and the stylized COCO dataset. The results of that will give an indication of how well pose estimation works on art collections and where there's room for improvement.

There are several considerations to be made when choosing the right model. The focus will mainly be on methods that have code readily available. The model must also be compatible with the preferred dataset. Pose estimation has several datasets with different amounts of keypoints, bounding boxes or other metadata. The most popular dataset and supported by most models will be used, which is the COCO dataset. The main aspect of the problem is quality and speed. When setting up a database for querying, there needs to be qualitative results to search through. The search itself should be fast, but this is not the subject of this thesis. At the same time, there should be a wide variation in architectures as explained previously (Chapter 3). All these criteria are considered in the next sections as well as those uniquely for each section.

4.1 Baseline Pose Estimation

4.1.1 Choice of Model

Here again, quality is the most important criteria for performance. The current state-or-the-art is ViTPose [122]. The model is based on vision transformers. This makes it an obvious first choice. An overwhelming amount of models both in top-down as well as bottom-up architectures use HRNet [5] with the only difference being in pre-processing [123, 124] or post-processing [7, 125]. Since VitPose is a top-down architecture and to keep a variety of architectures, a bottom-up version of HRNet is selected. The best model in this family is SWAHR according to Chen et al. [27]. Other architectures were looked at, like KAPAO [126], which uses a single-stage architecture, but these were not performant enough to be considered.

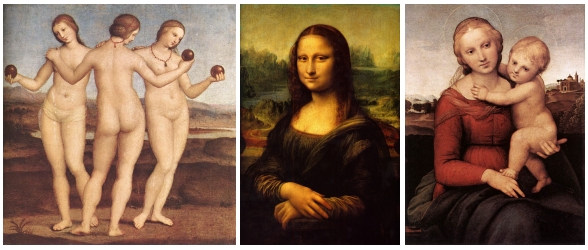
4.1.2 Training

For the sake of learning the different algorithms, the training methods were reverse engineered So, it was deemed appropriate to train the chosen models from scratch, so there's a plain network trained with the new setup for comparison. All training was done using the default parameters.

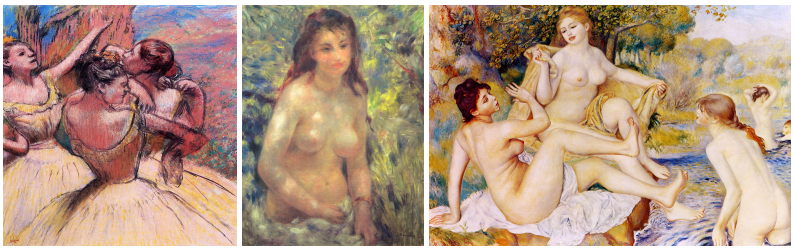
4.2 Pose Estimation after Applying Style Transfer to the COCO Dataset

Due to time constraints, the evaluation is only done on a subset of the COCO dataset. This set was created by randomly sampling 1000 images. The first baseline will establish how well the pre-trained models perform on a stylized COCO dataset. The evaluated pose estimators will be SWAHR and ViTPose, and each will use the trained weights mentioned in section 4.1.2. Since ViTPose is a top-down architecture, it will use the ground truth bounding box to extract the persons. They will both be tested on a styled version of the COCO dataset by CycleGAN and AdaIN. CycleGAN wil be applied for the 3 styles it was trained on; baroque, impressionism and renaissance. AdaIN uses the pre-trained model and uses 3 images of each of the previous styles to use as style image. The images were selected to best represent the style while also varying the content as shown in Figure 4.1. Each model uses the default parameters and at no time was the input image resized or otherwise distorted. This comes to a total of 24 combinations that will be assessed.

(a) Baroque style images

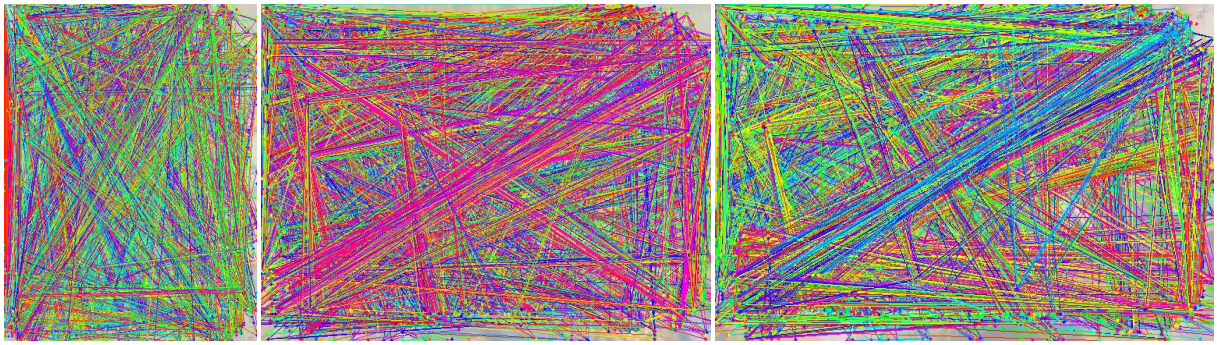


(b) Renaissance style images



(c) Impressionism style images

Figure 4.1: The style images used for AdaIN during evaluation.



(a) SWAHR

(b) ViTPose

Figure 4.2: Examples of the keypoints found by the pre-trained Pose Estimation networks.

4.2.1 Results

From the available metrics the only ones that were useful for these measurements were the Average Precision/Recall. They're implemented as part of the COCO dataset and work with any dataset that's compatible with the COCO format. Of the other metrics, Percentage of Correct Parts (PCP) is unusable because it only applies to networks that detect the limbs as boxes instead of keypoints. The chosen pose estimation networks only work with keypoints. Percentage of Correct Keypoints (PCk) looked like it could be useable. However, PCKh needs a head bounding box, which is only available for the MPII dataset. While all the implementations only work with top-down architectures. They each asserts that the length of predicted persons should be the same as that of the ground truth. In a bottom-up architecture, it is possible to find more or less persons. The results shown in Table 4.2 are the average of different evaluations, and it becomes immediately evident that this method is not going to work. As seen in Figure 4.2, SWAHR is completely lost and can't find any good keypoints while ViTPose, having a high recall, still found some of the poses.

4.3 Pose Estimation on the Human-Art Dataset

As a second baseline, the Human-Art dataset contains a subset of annotated oil paintings compatible with the COCO format. The evaluation dataset contains 250 images with 900 annotated persons. This will give a insight in the performance of the pose estimation models on artworks. SWAHR and ViTPose will be validated, and the trained weights mentioned in section 4.1.2 as well as the pre-trained weights from the original papers will be used. The input image will not be resized or otherwise

Table 4.1: Establishing a baseline for Pose Estimation on Artworks; measuring Average Precision/Recall (AP/AR). The COCO dataset is transformed with various Style Transfer models on which performance is measured from pre-trained pose-estimation models.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
AdaIN										
SWAHR	0	0	0	0	0	0	0	0	0	0
ViTPose	0.026	0.057	0.020	0.017	0.041	0.340	0.568	0.337	0.187	0.539
CycleGAN										
SWAHR	0	0	0	0	0	0	0	0	0	0
ViTPose	0.081	0.128	0.086	0.120	0.068	0.627	0.850	0.682	0.557	0.718

distorted, and the default parameters used. To confirm the premise that the pose estimation models perform less well on artworks, they are also validated on the COCO dataset. Because the other tests are only done on a subset of the COCO dataset, the models are also validated on this subset. This is a total of 8 combinations that will be validated.

4.3.1 Results

As mentioned in section 4.2.1, only the Average Precision/Recall will be measured. The table 4.2 shows the results of the measurements. It clearly shows that the models have inferior results on artworks than photographs by up to 20%. It also shows a significant difference between the pre-trained and self-trained models. However, for SWAHR the pre-trained model performed better by 6%, but for ViTPose, the self-trained model performs better by 2%. This difference well justifies the training of the models on the plain COCO dataset instead of using the pre-trained models to compare to. Thus going forward, the metrics will be compared to the self-trained models. This will give a more accurate picture of the improvements made. Notable as well is that despite ViTPose being the state-of-the-art, it performs worse than SWAHR on both datasets.

4.4 Discussion

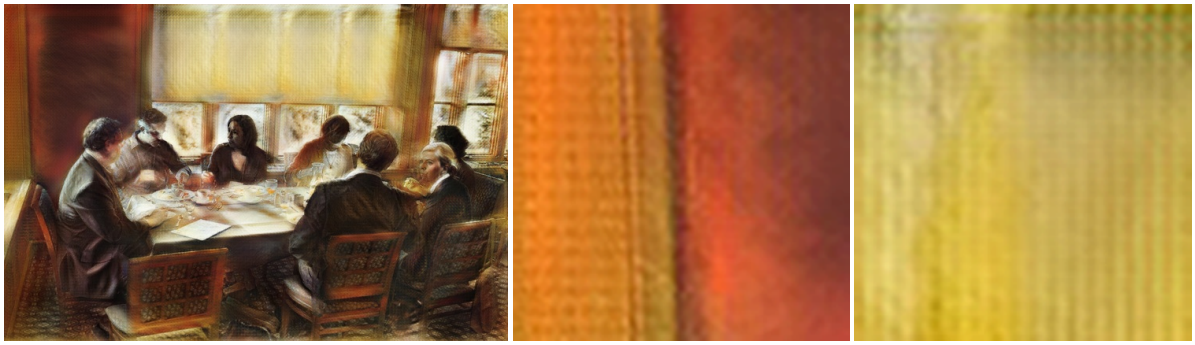
The results show that the use of style transfer on the input image will not yield any good results. The models don't perform well on the stylized images, likely because they don't produce high-fidelity transformations. Putting a second algorithm in the pipeline creates an extra chance for error. However, ViTPose was still able to discern most of the poses; having a high recall, but seems to have hallucinated others; giving a low precision. Which begs the question: What about this network makes it perform better than SWAHR here? Perhaps, it is merely able to deal with the artifacts left by the style transfer better, while SWAHR is completely confused by it? Figure 4.3 shows these artifacts. Or, perhaps, it is merely because as a top-down algorithm, it has an unfair advantage in that it used the ground-truth bounding boxes to crop the image.

The baseline on the plain COCO dataset confirms once more that the pose estimation models are inferior on artworks than photographs. It goes up as high as 50% for the medium areas, which makes sense as the smaller parts of an image will also be more abstract as brush strokes become more prominent.

Table 4.2: Establishing a baseline for Pose Estimation on Artworks; Average Precision/Recall (AP/AR). The table shows the performance of the pre-trained models measured on The COCO dataset and the Human-Art dataset.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
COCO dataset										
Pre-trained SWAHR	**0.687**	**0.881**	**0.748**	**0.639**	**0.757**	0.737	0.904	0.788	0.670	**0.828**
SWAHR	0.620	0.830	0.684	0.604	0.653	0.710	0.891	0.765	0.640	0.803
Pre-trained ViTPose	0.588	0.832	0.641	0.573	0.629	0.723	0.906	0.782	0.682	0.786
ViTPose	0.609	0.847	0.680	0.597	0.644	**0.740**	**0.918**	**0.810**	**0.703**	0.795
Human-Art Dataset										
Pre-trained SWAHR	**0.528**	**0.759**	**0.565**	0.099	**0.573**	**0.593**	0.635	0.629	0.177	**0.635**
SWAHR	0.492	0.742	0.536	0.058	0.539	0.563	0.784	0.606	0.109	0.605
Pre-trained ViTPose	0.380	0.656	0.385	0.108	0.420	0.571	0.803	0.620	0.279	0.599
ViTPose	0.406	0.682	0.415	**0.130**	0.445	0.591	**0.818**	**0.632**	**0.306**	0.619
Difference										
Pre-trained SWAHR	-0.159	-0.122	-0.183	-0.540	-0.184	**-0.144**	-0.269	**-0.159**	-0.493	-0.193
SWAHR	**-0.128**	**-0.088**	**-0.148**	-0.546	**-0.114**	-0.147	-0.107	-0.159	-0.531	-0.198
Pre-trained ViTPose	-0.208	-0.176	-0.256	**-0.465**	-0.209	-0.152	-0.103	-0.162	-0.403	-0.187
ViTPose	-0.203	-0.165	-0.265	-0.467	-0.199	-0.149	**-0.100**	-0.178	**-0.397**	**-0.176**

(a) AdaIN with impressionism as style.



(b) CycleGAN using impressionism as style

Figure 4.3: Examples of artifacts left by AdaIN and CycleGAN. The left images are stylized images, and the right images are close-ups of different patches.

5

Improving Pose Estimation with Style Transfer

Having established a baseline, it is now possible to search for improvements. In this chapter, two techniques will be explored to see if they can improve HPE. Using the same algorithms as seen in the previous chapter, they will now be used to: (1) transform an input artistic image to a photographic image to estimate poses on or (2) be trained with a dataset that is augmented with images that are transformed to different styles.

5.1 Pose Estimation after Style Transform

One option to predict poses on an artwork is to first transform it to photographic realism and let the plain model run on it. As previously seen in section 4.3.1, the results on photographs are dramatically better. If artworks are successfully transformed to that style, there is no need to train a new model and the extensive amount of datasets created for this task become available. To validate this, SWAHR and ViTPose are run on the Human-Art dataset after it was transformed using AdaIN and CycleGAN. For CycleGAN, three styles are used to perform this task, namely baroque, impressionism and renaissance. AdaIN uses three style images for each style as previously mentioned. Figure 4.2 shows the selection made for this. Before transforming the artwork, the size is checked and resized to 1024 if either of the sides is bigger than that, and only then. This is done because some artworks in the dataset are quite large and cause Out-Of-Memory errors. Otherwise, no other distortions are applied. This adds the total number of tests to be up to 24. As seen in section 4.2.1, here as well, style transfer is only more detrimental to the results. As seen in Table 5.1, the same observations can be made: Only ViTPose scores, but with very low precision and high recall.

5.2 Augmenting COCO Dataset for Pose Estimation Training

The second option that's been explored is the augmentation of the dataset with styled images. The chosen pose estimation algorithms, 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, as this was the best scoring model from the quantitative evaluation. This results in a combination of 16 models. The experiments will be conducted on the validation set of the COCO-dataset as well as the Human-Art dataset. While the problem specifically focuses on improving the performance on artworks, it's still interesting to also validate the results on the COCO-dataset.

Table 5.1: Performance of plain Pose Estimation models after Artwork is transformed with different Style Transfer models.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
AdaIN										
Trained on Baroque dataset										
SWAHR	0	0	0	0	0	0	0	0	0	
ViTPose	0.056	0.109	0.052	0.002	0.064	0.463	0.700	0.486	0.058	0.501
Trained on Impressionism dataset										
SWAHR	0	0	0	0	0	0	0	0	0	0
ViTPose	0.044	0.089	0.043	0.002	0.051	0.406	0.648	0.427	0.051	0.439
Trained on Renaissance dataset										
SWAHR	0	0	0	0	0	0	0	0	0	0
ViTPose	0.052	0.100	0.047	0.001	0.058	0.441	0.0679	0.457	0.045	0.477
CycleGAN										
Trained on Baroque dataset										
SWAHR	0	0	0	0	0	0	0	0	0	0
ViTPose	0.068	0.128	0.066	0.014	0.075	0.520	0.768	0.555	0.195	0.551
Trained on Impressionism dataset										
SWAHR	0	0	0	0	0	0	0	0	0	0
ViTPose	0.059	0.113	0.055	0.020	0.064	0.470	0.717	0.488	0.227	0.493
Trained on Renaissance dataset										
SWAHR	0	0	0	0	0	0	0	0	0	0
ViTPose	0.057	0.107	0.052	0.012	0.062	0.458	0.694	0.471	0.183	0.485

5.2.1 Creation of datasets

The next step in the process is to create the stylized datasets on which the pose estimation models will be trained. For each augmented dataset, the coco annotations file was used as a template. All metadata of the file was kept while only changing the id and file name. The file name points to the new location of the stylized image. A stylized version of COCO was created from each style transfer model that CycleGAN was trained for discussed in section 3.1.3, except impressionism for 2000 epochs. Other versions were created for AdaIN. Since AdaIN requires a style image, the images used for training the CycleGAN models were used for this purpose. The style dataset was sampled randomly 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 new annotation file 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. During training it was noticed that the stylized images were inverted, resulting in two models being trained on the inverted dataset. These were the COCO + mixed and mixed models.

5.2.2 Training

All models are trained with the default parameters provided by their respective papers. They also use the default learning rate. No human detection model is trained, instead, the ground truth bounding boxes will be used to extract the poses for top-down algorithms. The weights are initiate with the pre-trained models. They're trained for 200 epochs and the models are saved from the 100th epoch every 20 epochs. As a control, two models are trained without initiating weights, which are

the inverted mixed model and the COCO + impressionism model. These were trained for the default 300 epochs and were also saved from the 100th epoch every 20 epochs.

5.2.3 Results

For the SWAHR network, shown in table 1, the best results are found for the model trained on the COCO + AdaIN mixed style transfer dataset. The second best network was trained on the COCO + CycleGAN mixed style transfer dataset. For the ViTPose network, the best results are for 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 5.2 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.

Table 5.2: Comparing the best models from the experiments on the COCO dataset with the baseline metrics.

Method	**AP**	**AP**^{50}	**AP**^{75}	**AP**^{M}	**AP**^{L}	**AR**	**AR**^{50}	**AR**^{75}	**AR**^{M}	**AR**^{L}
Pre-trained SWAHR	**0.687***	**0.881***	**0.748***	**0.639***	**0.757***	0.737	0.904	0.788	0.670	**0.828***
Pre-trained ViTPose	0.588	0.832	0.641	0.573	0.629	0.723	0.906	0.782	0.682	0.7863
SWAHR	0.620	0.830	0.684	0.604	0.653	0.710	0.891	0.765	0.640	0.803
ViTPose	0.609	0.847	0.680	0.597	0.644	0.740	0.918	0.810	0.703	0.795
SWAHR COCO + AdaIN Mixed	**0.679****	**0.874****	**0.735****	**0.628****	**0.751****	0.732	0.902	0.782	0.651	**0.824****
ViTPose COCO + CycleGAN Mixed	0.635	0.861	0.697	0.616	0.681	**0.763***	**0.925***	**0.825***	**0.723***	0.820

* the best result overall.

** the best result without pre-trained models.

The results on the Human-Art dataset are shown in table 2. 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 two to 3% AP for CycleGAN. The best precision and recall belongs to the SWAHR models. Comparing the best models to the baseline (Table 5.3), 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.

5.3 Discussion

It is abundantly clear that trying to use style transfer to transform images in combination with pre-trained networks is a catastrophic failure. Style transfer, or at least the models used in this thesis, does not have the capabilities to convincingly transform a photograph to an artwork. It's difficult to believe that any of the styled images can be confused with an artwork by any reasonable person. There are several studies that confirm this: Chen et al. [17] and Wang et al. [18] calculate a

Table 5.3: Comparing the best models from the experiments on the Human-Art dataset with the baseline metrics.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
Pre-trained SWAHR	0.528	0.759	0.565	0.099	0.573	0.593	0.635	0.629	0.177	0.635
Pre-trained ViTPose	0.380	0.656	0.385	0.108	0.420	0.571	0.803	0.620	0.279	0.599
SWAHR	0.492	0.742	0.536	0.058	0.539	0.563	0.784	0.606	0.109	0.605
ViTPose	0.406	0.682	0.415	0.130	0.445	0.591	0.818	0.632	0.306	0.619
SWAHR COCO + CycleGAN Mixed	**0.553***	**0.789***	**0.604***	0.122	**0.598***	**0.629***	0.839	**0.677***	0.208	**0.669***
ViTPose COCO + CycleGAN Mixed	0.439	0.726	0.458	**0.140***	0.481	0.617	**0.844***	0.661	**0.324***	0.646

* the best result overall.

** the best result without pre-trained models.

deception score, which measure the believability of the fake images against the real images. Images from a set of stylized images and real artworks are shown to participants who need to determine whether it is real or fake. Table 5.4 shows that older networks have extremely bad performance on this with a meager 40% at best. While the newer models show a considerable improvement, they're still 20% below the real images. Other models, like Huang et al. [127] and Zhang et al. [128] ask participant to select the fake(s) from a group of images. They find that their own models were able to confuse participants; participants were not able to make a distinction between fake and real images, while older models did not achieve this. This confirms the observation that the used models are inadequate, but gives hopeful results for future research with state-of-the-art style transfer models.

Table 5.4: The deception score of different models calculated by Chen et al. [17] and Wang et al. [18].

Paper	WikiArt	Theirs	AdaIN	WCT [129]	LST [130]	SANet [131]
Chen et al.	0.875	0.624	0.363	0.099	0.125	0.161
Wang et al.	0.784	0.568	0.241	0.172	0.408	0.346

During training, a plain model was trained as a control for the fidelity of the reverse-engineered implementation. On the COCO-dataset, ViTPose improved on both the control and pre-trained models. The results for CycleGAN don't seem to be an improvement when comparing with the pre-trained model, but there are improvements compared to the control. This could mean that if the difference in training can be pin-pointed, the performance on the COCO-dataset for pre-trained models could potentially be increased. However, this is not a guarantee. On the other hand, the performances on the Human-Art dataset have increased compared to both baselines for both architectures. The models with the best results are those that combine the COCO dataset with a mixture of different styles and SWAHR shows the best performance of the pose estimation algorithms. An explanation for this could be that SWAHR generalizes better because during the creation of the styled COCO dataset, the style is more consistent than AdaIN, since AdaIN sampled random style images. The random images could make it more difficult for the network to converge as the styles are more dispersed. Nevertheless, the most successful models were the ones where the styles were mixed. This explanation might therefor not be correct. During the training, for every 20 epochs the networks were evaluated after 100 epochs. Table 5.5 show that after 100 iterations for both SWAHR and ViTPose the network only marginally increased; around 2%. While this is great for fine-tuning the network, for comparing architectures this does not seem necessary. The same conclusions would have been reached when keeping to only 100

epochs. Despite being state-of-the-art, ViTPose has a lower performance than SWAHR here. The evaluation is only on a subset of the COCO-dataset, which might explain it. According to Dosovitskiy et al., [69] Vision Transformers do not benefit from the inductive biases inherent to CNNs. To make up for that, they need to be trained on a bigger dataset. With the augmentation of the COCO-dataset, the training size was doubled. The increased performance might just only be because of a larger dataset. According to several surveys, bottom-up architectures are less precise then top-down architectures. Is this because top-down architectures are trained to only find one pose per ground truth in the found bounding box while bottom-up algorithms can find more poses per ground truth? This can skew the precision as there are now more false negatives. The cropping of the image also removes a lot of information that could potentially be relevant, like sitting on a horse or perspective. These could be clues that can help the algorithm more accurately do predictions, but how well can a network train for this? Perhaps 2D is limited in that sense.

Table 5.5: Marginal gains after 100 epochs. Trained on the COCO + Mixed and evaluated on the COCO dataset.

Paper	Epochs					
	100	**120**	**140**	**160**	**180**	**200**
SWAHR	0.669	0.672	0.666	0.669	0.675	0.676
VitPose	0.720	0.721	0.722	0.724	0.735	0.740

5.4 Related Papers

As discussed in section 2.5, similar experiments have already been run by others. To conclude, a comparison with the results of those papers is appropriate. First, Madhu et al. [15] compare three different methods with their baseline: a styled model that's trained completely on their stylized COCO dataset, and two fine-tuned models on their artwork dataset. Since fine-tuning was not used in this thesis, the results will only be compared with their styled models. Second, Kadish et al. [19] only fine-tune a Faster R-CNN object detection network with their stylized COCO dataset. Kadish et al. only compare their results with other papers. So, their baseline is from Gonthier et al. [132]. They perform the tests on the People-Art dataset [133]. This dataset is only labelled with bounding-boxes. Table 5.6 shows that both their findings are similar as what was found in this thesis, except that their results are more pronounced. This shows that as a general trend, augmenting a dataset with synthetic artworks will give better results for art collections.

Table 5.6: Improvements made by Madhu et al. [15] and Kadish et al. [19] compared to the results in this thesis. All values are in terms of AP, except for the values from Kadish et al. which are AP_{50}.

Model	Madhu et al.		Kadish et al.	This thesis	
	COCO	**Their Dataset**	**People-Art**	**COCO**	**Human-Art**
Baseline	0.765	0.247	0.580	0.687	0.528
Styled	0.743	0.323	0.680	0.679	0.553
Difference	-0.022	+0.076	+0.010	-0.008	+0.025

6

Conclusions

Because of the digitalization of art collections, museums are looking to improve their analytic tools to help them with their functions. Among them are the relationships between artworks depending on their themes of which poses are a big part. Unfortunately, the state-of-the-art pose estimation models have only been trained on photograph datasets and have a miserable performance on art collections. To achieve better results two methods of improvement were explored:

1. The input images for the pose estimation networks are first transformed from an artwork to a photograph. With this method, the already vast library of pose estimation methods is made available to the curators.

2. The COCO dataset is transformed with multiple style transfer methods to styled COCO datasets. With these styled datasets, new pose estimation models can be trained which work better than the already existing ones.

These methods require a style transfer method that is able to transform between artworks and photographs. Therefor, several datasets were created by using CBIR on the WikiArt dataset. The focus of these datasets is mainly around the human figure as this is the domain of pose estimation. Several style transfer models were trained, but it was found that they did not achieve high fidelity. Because of this, the first method was found to give unreliable results during both the baseline measurement as well as during the experiments. When transforming the images, artifacts of the method were left behind which confused the pose estimation models. However, ViTPose still had a high recall in these cases, but SWAHR did not have any good results. For the second method, several new styled COCO datasets were created with the newly trained style transfer models. After training the pose estimation networks successfully, it was found that they were able to increase the performance on art collections. To establish this, evaluation on the Human-Art dataset was very helpful, as the another method, which depended on transforming input images for evaluation, was broken. It was found that training the models on an augmented dataset can increase the performance by at least 2% AP. This is in line with related works who reported a similar increase.

6.1 Lessons Learned

During the experiments, it became clear that the chosen style transfer methods didn't have the right capabilities as was first thought. A potential culprit could be the different evaluation metrics which still do not adequately measure the similarity between images. Evidence of this is seen in the collapse of the StarGAN network, which still gives a good performance during quantitative evaluations while qualitatively it's subpar. While ViTPose is considered state-of-the-art, in the experiments run, it does not outperform SWAHR for any of the evaluations. This shows again, that even though a model can be state-of-the-art

in one task, this does not translate to other tasks. During training, the pose estimation networks converged very quickly and training them for 200 epochs wasn't needed, but instead only 100 epochs would have been enough.

6.2 Future Work

To get better results, a first recommended improvement can be to use style transfer networks that have a higher fidelity. During the discussion in section 5.3, a brief look was made at several papers that evaluated the believability of different models. They concluded that, although the older networks performed poorly, more recent models were able to perform better. This gives hope for future development within this area. One promising technique is stable diffusion [134] which is able to synthesize high fidelity images. It would be useful if this could be used for style transfer. Another improvement could be the evaluation metrics for style transfer. While there are a plethora of other metrics out there that haven't been mentioned, most of them are derivatives of the ones discussed in section 2.3.5. Their primary focus is also around generative methods and not specifically style transfer. A specialized metric might be something worth looking at, but maybe it's enough to update the current metrics with better feature extraction methods, like a CBIR method, which is more specialized in finding similarity. The created datasets for style transfer can also be more refined. Instead of having crowded images, the dataset could only focus on the human figure front and central, and crop them out as well. While this reduces generalization, the problem is only about pose estimation. It could even go so far that for each body part a dataset is made and style transfer is done in patches. However, this would increase the effort that needs to be undertaken to train a network to such an extend that it might just be easier to annotate paintings for pose estimation training. This work only focuses on realistic artworks to run pose estimation on, but this leaves out more abstract works. Another area that deals with this is style transfer with geometric transformations. Further research into this can extend the now limited approach. During the experiments, the evaluations were compared with two baselines; the pre-trained model and a control model. The control was trained the same way the styled models were and had a worse performance than the pre-trained baseline. This means that the networks could possibly be fine-tuned to perform better. The difference between the pre-trained and the control is as high as 6% for SWAHR. This is a noteworthy difference and enough to warrant further research. All in all, there are still a lot of areas that can be improved upon.

References

[1] J. Kalin, *Generative Adversarial Networks Cookbook*. Birmingham: Packt Publishing, 2018, p. 17.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[3] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *CoRR*, vol. abs/2006.01423, 2020. [Online]. Available: https://arxiv.org/abs/2006.01423

[4] 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

[5] 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

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

[7] 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

[8] M. R. Ronchi and P. Perona, "Benchmarking and error diagnosis in multi-instance pose estimation," *CoRR*, vol. abs/1707.05388, 2017. [Online]. Available: http://arxiv.org/abs/1707.05388

[9] 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

[10] 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.

[11] 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

[12] 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

[13] M. Liu, T. M. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *CoRR*, vol. abs/1703.00848, 2017. [Online]. Available: http://arxiv.org/abs/1703.00848

[14] W. Zhou, H. Li, and Q. Tian, "Recent advance in content-based image retrieval: A literature survey," *CoRR*, vol. abs/1706.06064, 2017. [Online]. Available: http://arxiv.org/abs/1706.06064

[15] 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

[16] 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

[17] H. Chen, l. zhao, Z. Wang, H. Zhang, Z. Zuo, A. Li, W. Xing, and D. Lu, "Artistic style transfer with internal-external learning and contrastive learning," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 26 561–26 573. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/df5354693177e83e8ba089e94b7b6b55-Paper.pdf

[18] Z. Wang, Z. Zhang, L. Zhao, Z. Zuo, A. Li, W. Xing, and D. Lu, "Aesust: Towards aesthetic-enhanced universal style transfer," 2022.

[19] 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

[20] T. F. of Art, G. O. Culture: How Digitization Is Shaking up Museums, and Artists. (2021) Digital values. [Online]. Available: https://digital-values.de/the-future-of-art-and-culture-how-digitization-is-shaking-up-museums-gallery-owners-and-artists-2/

[21] M.-C. Marinescu, A. Reshetnikov, and J. M. López, "Improving object detection in paintings based on time contexts," in *2020 International Conference on Data Mining Workshops (ICDMW)*, 2020, pp. 926–932.

[22] M. Sabatelli, N. Banar, M. Cocriamont, E. Coudyzer, K. Lasaracina, W. Daelemans, P. Geurts, and M. Kestemont, "Advances in digital music iconography: Benchmarking the detection of musical instruments in unrestricted, non-photorealistic images from the artistic domain," *Digital Humanities Quarterly*, vol. 15, no. 1, February 2021.

[23] R. M. of Fine Arts Belgium. (2024) Opac fabritius. [Online]. Available: https://www.opac-fabritius.be/

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

[25] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, "The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation," *IEEE Access*, vol. 8, pp. 133 330–133 348, 2020.

[26] W. Liu, Q. Bao, Y. Sun, and T. Mei, "Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective," *CoRR*, vol. abs/2104.11536, 2021. [Online]. Available: https://arxiv.org/abs/2104.11536

[27] 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.

[28] A. Jain, J. Tompson, M. Andriluka, G. W. Taylor, and C. Bregler, "Learning human pose estimation features with convolutional networks," 2014.

[29] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2014. [Online]. Available: https://doi.org/10.1109%2Fcvpr.2014.214

[30] 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

[31] S. Ju, M. Black, and Y. Yacoob, "Cardboard people: a parameterized model of articulated image motion," in *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 38–44.

[32] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314285710041

[33] H. Sidenbladh, F. De la Torre, and M. Black, "A framework for modeling the appearance of 3d articulated figures," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 368–375.

[34] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. Black, "Smpl: a skinned multi-person linear model," vol. 34, 11 2015.

[35] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *British Machine Vision Conference*, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:7318714

[36] ——, "Learning effective human pose estimation from inaccurate annotation," in *CVPR 2011*, 2011, pp. 1465–1472.

[37] 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

[38] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3674–3681.

[39] J. Li, C. Wang, H. Zhu, Y. Mao, H. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and A new benchmark," *CoRR*, vol. abs/1812.00324, 2018. [Online]. Available: http://arxiv.org/abs/1812.00324

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

[41] G. Pons-Moll and B. Rosenhahn, *Model-Based Pose Estimation*. London: Springer London, 2011, pp. 139–170. [Online]. Available: https://doi.org/10.1007/978-0-85729-997-0_9

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[43] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," *CoRR*, vol. abs/1507.06550, 2015. [Online]. Available: http://arxiv.org/abs/1507.06550

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014. [Online]. Available: http://arxiv.org/abs/1409.4842

[45] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *CoRR*, vol. abs/1406.2984, 2014. [Online]. Available: http://arxiv.org/abs/1406.2984

[46] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," *CoRR*, vol. abs/1411.4280, 2014. [Online]. Available: http://arxiv.org/abs/1411.4280

[47] 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

[48] 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.

[49] 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

[50] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," *CoRR*, vol. abs/1708.01101, 2017. [Online]. Available: http://arxiv.org/abs/1708.01101

[51] C. Chou, J. Chien, and H. Chen, "Self adversarial training for human pose estimation," *CoRR*, vol. abs/1707.02439, 2017. [Online]. Available: http://arxiv.org/abs/1707.02439

[52] 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

[53] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-hrnet: A lightweight high-resolution network," *CoRR*, vol. abs/2104.06403, 2021. [Online]. Available: https://arxiv.org/abs/2104.06403

[54] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution transformer for dense prediction," *CoRR*, vol. abs/2110.09408, 2021. [Online]. Available: https://arxiv.org/abs/2110.09408

[55] Y. Chen, C. Shen, X. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," *CoRR*, vol. abs/1705.00389, 2017. [Online]. Available: http://arxiv.org/abs/1705.00389

[56] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014. [Online]. Available: http://arxiv.org/abs/1411.1784

[57] C. Chou, J. Chien, and H. Chen, "Self adversarial training for human pose estimation," *CoRR*, vol. abs/1707.02439, 2017. [Online]. Available: http://arxiv.org/abs/1707.02439

[58] 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

[59] 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

[60] 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

[61] 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

[62] 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

[63] 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

[64] 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

[65] 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

[66] K. Su, D. Yu, Z. Xu, X. Geng, and C. Wang, "Multi-person pose estimation with enhanced channel-wise and spatial information," *CoRR*, vol. abs/1905.03466, 2019. [Online]. Available: http://arxiv.org/abs/1905.03466

[67] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," *CoRR*, vol. abs/1901.00148, 2019. [Online]. Available: http://arxiv.org/abs/1901.00148

[68] 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

[69] 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

[70] 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

[71] 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

[72] X. Zhu and Y. Jiang, "Multi-person pose estimation for posetrack with enhanced part affinity fields," 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:52563463

[73] G. Hidalgo, Y. Raaj, H. Idrees, D. Xiang, H. Joo, T. Simon, and Y. Sheikh, "Single-network whole-body pose estimation," *CoRR*, vol. abs/1909.13423, 2019. [Online]. Available: http://arxiv.org/abs/1909.13423

[74] J. Li, W. Su, and Z. Wang, "Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation," *CoRR*, vol. abs/1911.10529, 2019. [Online]. Available: http://arxiv.org/abs/1911.10529

[75] 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

[76] 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

[77] 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

[78] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[79] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[80] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[81] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *CoRR*, vol. abs/1604.01685, 2016. [Online]. Available: http://arxiv.org/abs/1604.01685

[82] R. Tylecek and R. Sára, "Spatial pattern templates for recognition of objects with regular structure," in *German Conference on Pattern Recognition*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:6060524

[83] 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

[84] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 192–199.

[85] J. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," *CoRR*, vol. abs/1609.03552, 2016. [Online]. Available: http://arxiv.org/abs/1609.03552

[86] Y. Choi, Y. Uh, J. Yoo, and J. Ha, "Stargan v2: Diverse image synthesis for multiple domains," *CoRR*, vol. abs/1912.01865, 2019. [Online]. Available: http://arxiv.org/abs/1912.01865

[87] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays, "Transient attributes for high-level understanding and editing of outdoor scenes," *ACM Transactions on Graphics (proceedings of SIGGRAPH)*, vol. 33, no. 4, 2014.

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

[89] 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

[90] 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

[91] 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

[92] 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

[93] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," *CoRR*, vol. abs/1610.07629, 2016. [Online]. Available: http://arxiv.org/abs/1610.07629

[94] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua, "Stylebank: An explicit representation for neural image style transfer," *CoRR*, vol. abs/1703.09210, 2017. [Online]. Available: http://arxiv.org/abs/1703.09210

[95] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016.

[96] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[97] 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

[98] 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

[99] 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

[100] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022.

[101] H. Hoyez, C. Schockaert, J. Rambach, B. Mirbach, and D. Stricker, "Unsupervised image-to-image translation: A review," *Sensors*, vol. 22, no. 21, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/21/8540

[102] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *CoRR*, vol. abs/1606.03498, 2016. [Online]. Available: http://arxiv.org/abs/1606.03498

[103] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, vol. abs/1512.00567, 2015. [Online]. Available: http://arxiv.org/abs/1512.00567

[104] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a nash equilibrium," *CoRR*, vol. abs/1706.08500, 2017. [Online]. Available: http://arxiv.org/abs/1706.08500

[105] M. Fréchet, "Sur la distance de deux lois de probabilité," *Annales de l'ISUP*, vol. VI, no. 3, pp. 183–198, 1957. [Online]. Available: https://hal.science/hal-04093677

[106] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *CoRR*, vol. abs/1801.03924, 2018. [Online]. Available: http://arxiv.org/abs/1801.03924

[107] 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

[108] H. Xu, J. Wang, X.-S. Hua, and S. Li, "Image search by concept map," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 275–282. [Online]. Available: https://doi.org/10.1145/1835449.1835497

[109] J. Wang and X. Hua, "Interactive image search by color map," *ACM Trans. Intell. Syst. Technol.*, vol. 3, pp. 12:1–12:23, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:6538567

[110] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang, "Mindfinder: interactive sketch-based image search on millions of images," 10 2010, pp. 1605–1608.

[111] 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

[112] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[113] W. Zhou, H. Li, Y. Lu, and Q. Tian, "Large scale image search with geometric coding," in *Proceedings of the 19th ACM International Conference on Multimedia*, ser. MM '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 1349–1352. [Online]. Available: https://doi.org/10.1145/2072298.2072012

[114] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *CVPR 2011*, 2011, pp. 809–816.

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

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

[117] E. Ioannou and S. Maddock, "Evaluation in neural style transfer: A review," 2024.

[118] Roman. (2023) Image similarity comparison using vgg16 deep learning model. [Online]. Available: https://medium.com/@developerRegmi/image-similarity-comparison-using-vgg16-deep-learning-model-a663a411cd24

[119] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Knowledge Discovery and Data Mining*, 1996. [Online]. Available: https://api.semanticscholar.org/CorpusID:355163

[120] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[121] 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

[122] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose++: Vision transformer for generic body pose estimation," 2023.

[123] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," *CoRR*, vol. abs/1910.06278, 2019. [Online]. Available: http://arxiv.org/abs/1910.06278

[124] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The devil is in the details: Delving into unbiased data processing for human pose estimation," *CoRR*, vol. abs/1911.07524, 2019. [Online]. Available: http://arxiv.org/abs/1911.07524

[125] Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang, "Bottom-up human pose estimation via disentangled keypoint regression," *CoRR*, vol. abs/2104.02300, 2021. [Online]. Available: https://arxiv.org/abs/2104.02300

[126] W. J. McNally, K. Vats, A. Wong, and J. McPhee, "Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation," *CoRR*, vol. abs/2111.08557, 2021. [Online]. Available: https://arxiv.org/abs/2111.08557

[127] S. Huang, J. An, D. Wei, J. Luo, and H. Pfister, "Quantart: Quantizing image style transfer towards high visual fidelity," 2023.

[128] Y. Zhang, F. Tang, W. Dong, H. Huang, C. Ma, T.-Y. Lee, and C. Xu, "A unified arbitrary style transfer framework via adaptive contrastive learning," 2023.

[129] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M. Yang, "Universal style transfer via feature transforms," *CoRR*, vol. abs/1705.08086, 2017. [Online]. Available: http://arxiv.org/abs/1705.08086

[130] X. Li, S. Liu, J. Kautz, and M. Yang, "Learning linear transformations for fast arbitrary style transfer," *CoRR*, vol. abs/1808.04537, 2018. [Online]. Available: http://arxiv.org/abs/1808.04537

[131] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," *CoRR*, vol. abs/1812.02342, 2018. [Online]. Available: http://arxiv.org/abs/1812.02342

[132] N. Gonthier, S. Ladjal, and Y. Gousseau, "Multiple instance learning on deep features for weakly supervised object detection with extreme domain shifts," *CoRR*, vol. abs/2008.01178, 2020. [Online]. Available: https://arxiv.org/abs/2008.01178

[133] H. Cai, Q. Wu, T. Corradi, and P. Hall, "The cross-depiction problem: Computer vision algorithms for recognising objects in artwork and in photographs," *CoRR*, vol. abs/1505.00110, 2015. [Online]. Available: http://arxiv.org/abs/1505.00110

[134] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR*, vol. abs/2112.10752, 2021. [Online]. Available: https://arxiv.org/abs/2112.10752

Bijlagen

Extended Experiments

Table 1: Performance of different Pose Estimation models trained on Style Transformed datasets on COCO dataset.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
AdaIN										
Trained on COCO + Mixed Style Transfer										
SWAHR	**0.679***	**0.874***	0.735	**0.628***	0.751	**0.732**	**0.902**	0.782	0.651	0.824
ViTPose	0.618	0.859	0.685	0.599	0.661	0.748	0.924	0.816	0.709	0.805
Trained on COCO + Impressionism Style Transfer										
SWAHR	0.669	0.862	0.733	0.607	**0.755***	0.729	**0.902**	0.782	0.651	**0.834**
ViTPose	0.609	0.843	0.664	0.590	0.654	0.742	0.916	0.801	0.702	0.799
Trained on Mixed Style Transfer										
SWAHR	0.603	0.843	0.661	0.535	0.704	0.676	0.882	0.726	0.586	0.794
ViTPose	0.518	0.783	0.557	0.492	0.573	0.669	0.880	0.726	0.617	0.739
Trained on Impressionism Style Transfer										
SWAHR	0.591	0.830	0.654	0.527	0.688	0.663	0.873	0.716	0.574	0.780
ViTPose	0.497	0.784	0.531	0.463	0.564	0.650	0.874	0.710	0.594	0.728
CycleGAN										
Trained on COCO + Mixed Style Transfer										
SWAHR	0.672	0.863	**0.737***	0.618	0.747	**0.732**	**0.902**	**0.787**	**0.660**	0.827
ViTPose	**0.635**	**0.861**	0.697	0.616	**0.681**	**0.763***	**0.925***	**0.825***	0.723	**0.820**
Trained on COCO + Impressionism Style Transfer										
SWAHR	0.663	0.862	0.724	0.606	0.743	0.714	0.889	0.764	0.637	0.815
ViTPose	0.633	0.859	**0.701**	**0.618**	0.670	0.761	0.922	0.828	**0.725***	0.812
Trained on Mixed Style Transfer										
SWAHR	0.653	0.858	0.711	0.609	0.714	0.716	0.898	0.764	0.647	0.807
ViTPose	0.595	0.844	0.654	0.586	0.628	0.731	0.912	0.795	0.698	0.780
Trained on Impressionism Style Transfer										
SWAHR	0.661	0.864	0.717	0.621	0.719	0.718	0.896	0.765	0.656	0.802
ViTPose	0.591	0.841	0.643	0.582	0.619	0.727	0.910	0.790	0.695	0.773

* the best result overall.

Table 2: Performance of different Pose Estimation models trained on Style Transferred datasets on Human-Art dataset.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
AdaIN										
Trained on COCO + Mixed Style Transfer										
SWAHR	0.549	**0.791***	0.600	0.065	**0.602***	0.622	0.834	0.668	0.141	0.667
ViTPose	0.420	0.724	0.440	**0.151***	0.460	0.600	0.843	0.650	0.300	0.630
Trained on COCO + Impressionism Style Transfer										
SWAHR	0.540	0.779	0.576	0.071	0.591	0.612	0.822	0.646	0.156	0.655
ViTPose	0.421	0.706	0.430	0.149	0.458	0.600	0.831	0.641	0.303	0.629
Trained on Mixed Style Transfer										
SWAHR	0.492	0.750	0.525	0.048	0.547	0.581	0.811	0.625	0.142	0.622
ViTPose	0.332	0.627	0.316	0.079	0.372	0.522	0.784	0.559	0.223	0.551
Trained on Impressionism Style Transfer										
SWAHR	0.488	0.738	0.524	0.058	0.543	0.581	0.804	0.624	0.153	0.621
ViTPose	0.321	0.600	0.302	0.094	0.355	0.514	0.765	0.539	0.232	0.542
CycleGAN										
Trained on COCO + Mixed Style Transfer										
SWAHR	**0.553***	0.789	**0.604***	**0.122**	0.598	**0.629***	0.839	**0.677***	**0.208**	**0.669***
ViTPose	**0.439**	**0.726**	**0.458**	0.140	**0.481**	0.617	0.844	0.661	0.324	**0.646**
Trained on COCO + Impressionism Style Transfer										
SWAHR	0.522	0.778	0.556	0.113	0.565	0.590	0.819	0.628	0.173	0.630
ViTPose	0.438	0.724	0.448	0.147	0.479	**0.619**	**0.846***	**0.664**	**0.358***	0.645
Trained on Mixed Style Transfer										
SWAHR	0.524	0.779	0.559	0.102	0.569	0.613	**0.843**	0.645	0.200	0.652
ViTPose	0.405	0.696	0.419	0.148	0.442	0.590	0.829	0.639	0.338	0.615
Trained on Impressionism Style Transfer										
SWAHR	0.505	0.761	0.539	0.116	0.546	0.587	0.822	0.622	**0.208**	0.623
ViTPose	0.407	0.694	0.412	**0.151***	0.444	0.590	0.828	0.631	0.341	0.615

* the best result overall.