Enhancing Human Pose Estimation in Ancient Vase Paintings via Perceptually-grounded Style Transfer Learning

PRATHMESH MADHU*, Pattern Recognition Lab, Friedrich Alexander University, Germany
ANGEL VILLAR-CORRALES*, Autonomous Intelligent Systems, University of Bonn, Germany
RONAK KOSTI, Pattern Recognition Lab, Friedrich Alexander University, Germany
TORSTEN BENDSCHUS, Institut für Klassische Archäologie, FAU, Germany
CORINNA REINHARDT, Institut für Klassische Archäologie, FAU, Germany
PETER BELL, German Studies and Art Studies, Philipps University of Marburg, Germany
ANDREAS MAIER, Pattern Recognition Lab, Friedrich Alexander University, Germany
VINCENT CHRISTLEIN, Pattern Recognition Lab, Friedrich Alexander University, Germany

Human pose estimation (HPE) is a central part of understanding the visual narration and body movements of characters depicted in artwork collections, such as Greek vase paintings. Unfortunately, existing HPE methods do not generalise well across domains resulting in poorly recognised poses. Therefore, we propose a two step approach: (1) adapting a dataset of natural images of known person and pose annotations to the style of Greek vase paintings by means of image style-transfer. We introduce a perceptually-grounded style transfer training to enforce perceptual consistency. Then, we fine-tune the base model with this newly created dataset. We show that using style-transfer learning significantly improves the SOTA performance on unlabelled data by more than 6% mean average precision (mAP) as well as mean average recall (mAR). (2) To improve the already strong results further, we created a small dataset (ClassArch) consisting of ancient Greek vase paintings from the 6–5th century BCE with person and pose annotations. We show that fine-tuning on this data with a style-transferred model improves the performance further. In a thorough ablation study, we give a targeted analysis of the influence of style intensities,

*Both authors contributed equally to this research.
revealing that the model learns generic domain styles. Additionally, we provide a pose-based image retrieval to demonstrate the effectiveness of our method. The code and pretrained models can be found at https://github.com/angelvillar96/STLPose.

CCS Concepts: • Applied computing → Digital libraries and archives; • Computing methodologies → Visual content-based indexing and retrieval; Supervised learning.

Additional Key Words and Phrases: Pose Estimation, Greek Vase Paintings, Style Transfer Learning, Digital Humanities

ACM Reference Format:
Prathmesh Madhu, Angel Villar-Corrales, Ronak Kosti, Torsten Bendschus, Corinna Reinhardt, Peter Bell, Andreas Maier, and Vincent Christlein. 2022. Enhancing Human Pose Estimation in Ancient Vase Paintings via Perceptually-grounded Style Transfer Learning. ACM J. Comput. Cult. Herit. 1, 1, Article 1 (January 2022), 18 pages. https://doi.org/10.1145/3569089

1 INTRODUCTION

Human pose estimation (HPE) is highly challenging as it is difficult to have one method that can generalise across all domains. Estimating human pose involves localising each visible body-keypoint (Fig. 1c), however the state-of-the-art (SOTA) methods underperform when tested on different domains, for example ancient Greek vase paintings (Fig. 1). HPE is central to understanding the visual narration and body movements of the characters depicted in these paintings and the recent rapid digitisation of art collections has created an opportunity to use HPE as a tool to digitally examine such artworks. These digital copies are usually either photographic reproductions [37] or scans of existing archives [48]. In addition to the preservation of cultural heritage, these digital collections allow remote access of the invaluable artistic data to the general public. However, due to content complexity and large size, navigating within such collections is often daunting.

To address the challenge of efficiently analysing large digital collections, several computer vision and image analysis techniques have been used for applications, such as artist identification [28], object recognition [10, 19, 55], character recognition [34], artistic image classification [6, 8, 47] and pose-matching [27]. However, when these methods are evaluated on a different domain, they show sub-optimal performance, c.f. Fig. 1. Hence, an important challenge is to learn effective representations using little data. Human pose representation is one such example.

The understanding of visual narration in Greek vase paintings is one of the main objectives in the field of Classical Archaeology. In order to display the actions and situations of a narrative, as well as to characterise the protagonists, ancient Greek artists made use of a broad variety of often similar image elements [17]. Some of the key aspects of the narrative are illustrated by meaningful interactions and compositional relationships (e.g., postures or gestures) between the characters displayed in the painting [36]. For example, the divine pursuit scene (Fig. 2, 1st & 2nd column) is a recurrent narrative in Greek vase paintings, often characterised by a character moving fast from left to right and reaching out with both his arms to catch a woman on her forearm or shoulder [49].

In this work, we propose to exploit these recurrent character interactions and postures in order to navigate semantically through collections of Greek vase paintings. We address image retrieval in such databases by measuring the similarity between character postures. Since ancient Greek artists made use of the postures to depict similar narratives, the retrieved images should display the same scene.

For human pose-based image retrieval in Greek vase paintings, we need a reliable human pose estimation algorithm. We propose a two-step approach: (1) First, we apply style-transfer [23] to the COCO [32] dataset to generate a synthetic annotated dataset with the style of Greek vase paintings and fine-tune the baseline person detection and pose estimation models on this dataset (Fig. 5 middle). (2) Second, we fine-tune these models on a
newly generated dataset for Classical Archaeology (Fig. 5 bottom). We show that both steps improve the person detection and pose estimation tasks and thus the retrieval performance considerably.

In particular, our main contributions are:

1. We introduce the Styled-COCO-Persons (SCP) and the ClassArch (CA) datasets. SCP is a synthetic dataset, generated by applying style transfer, with different style-intensities, to the images from COCO (only ‘person’ class) to mimic the style of the CA dataset (Greek vase paintings). The CA dataset consists of 1783 images (characters) from 1000+ Greek vase paintings along with pose keypoint annotations.

2. We show that by just using styles of the CA dataset on real images, one can improve the task of human pose estimation in Greek vase paintings without requiring any annotations. We also show that fine-tuning this model with the small CA dataset modestly enhances the performance compared to direct transfer learning. Moreover, styled-tuned models outperform state-of-the-art fine-tuned methods on the SCP and CA datasets.

3. We introduce a perceptual loss for style-transfer and show that this is beneficial for both person detection and pose estimation.

4. Additionally, we show that our styled transfer learning based pipeline is also beneficial for retrieving and discovering similar images based on poses of the character in narratives from ancient Greek vase paintings.

2 RELATED WORK

The task of representing human poses has been studied since the early days of computer vision [39]. However, its importance is much older. Pathosformeln [2], the iconic study of basic constructs (units) of body language is one of the firsts to view the body gesture (or posture) also as a way of voicing inner emotions. Body movement’s depiction is essential and central to the historian [3, 4], since it gives a way to recognise the inner emotions or expressions of the character. Impett et al. [24] studies it with a geometrical construct by operationalisation of...
body movements, which could be considered a way of Pathosformeln. McNiven et al. [36] also used human poses as the basis to study the interactions between characters in ancient vase paintings.

Since several years, Convolutional Neural Networks (CNNs) have been dominating computer vision tasks, and HPE is not an exception. After DeepPose [53], methods followed that improved HPE by using CNN cascades [54] and graphical models [52]. Two major types of approaches are popular with HPE: bottom-up and top-down.

**Bottom-up pose estimation** [5, 26, 42] directly estimate the location of all keypoints and assemble them into pose skeletons for all people in the image simultaneously. They use CNNs like ResNet [20] and DenseNet [22] as backbones to predict keypoints and optimisation-based matching techniques like DeepCut [42] and DeeperCut [26] for combining the keypoints into poses. Cao et al. [5] introduced Part Affinity Fields (PAFs) which is able to estimate poses in real-time by solving a bi-partite graph matching problem, as a way to solve the optimisation problem of aggregating poses. Bottom-up techniques’ lack of structural information leads to many false positives and often being outperformed by top-down pose estimation approaches.

**Top-down pose estimation** [9, 13, 41, 56] approach HPE in two steps (Fig. 5, first row). The first step addresses the problem of detecting all person instances in the image, whereas the second step aims at predicting the body-keypoints for each of the detected person. For person detection, a specific CNN, e. g., from the R-CNN family [15, 16, 43] is normally used. The second step involves using a single-person pose estimation model to process each of the person instances independently. Tompson et al. [52] first proposed the use of a CNN with multi-resolution receptive fields for the task of body joint localisation. More recent methods, however, focus on refinement techniques, such as Iterative Error Feedback [7], Stacked-hour-glass networks [40] and PoseFix [38]. Multi-scale approach by Sun et al. [51] uses a novel architecture to maintain high-resolution representations through the whole estimation process by repeated multi-scale feature fusions. These SOTA methods have improved the performance on their respective benchmark datasets, however, they fail to generalise to domains like Greek vase paintings.

**Domain adaptation** based approaches mainly aim at bringing the distributions of source domain closer to that of the target, so that a single model can be used for both domains. Methods using domain adaptation via style-transfer mainly transfer the style of target to the source while training online as a way of domain adaption, using style loss [44] or even adapting the domain progressively [25]. Some methods also use feature level
alignment for aligning the two domains [31], and others enforce it via self-similarity and domain-dissimilarity loss [12]. Recently, various methods have focused on reducing data bias in order to enhance the transfer-ability of features in task-specific layers [33, 35]. [45] and [50] proposed unsupervised domain adaptation techniques, using minimum entropy consensus and co-variance of source and target features respectively.

A few works using generative models have also been proposed. [46] proposed a method for optimising bi-directional image transformations and using class consistency loss, while [21] proposed a cycle GAN that adapts representations to combine feature-level and pixel level while enforcing structural (cycle loss) and semantic consistency (task-specific). A work closely related to ours [27], highlights the importance of using poses for artwork discovery and its subsequent analysis. Their study, however, is based on artworks with some case study images that are relatively easy for SOTA methods, such as OpenPose [5] to estimate the poses. One of the first work done on Greek Vase paintings was by Crowley et al. [11] to generate a correspondence between descriptions and unknown regions in the images of vase paintings. In their work, they describe the challenges of working with Greek vases and annotate the images automatically. On a similar note, estimating poses for characters in ancient Greek Vase paintings presents a completely different challenge, since OpenPose fails very often (Fig. 1).

Our work’s focus, instead, is on using the style transfer to generate a synthetic dataset from already existing labelled dataset like COCO to improve the pose estimation on unlabelled data like ancient Greek vases, by enforcing a pre-computed perceptual consistency loss.

3 DATASETS
In order to train deep networks in a supervised fashion, it is very important to have a high-quality annotated dataset. Hence we work with 3 main datasets viz. COCO-Persons (CP) with images of only the ‘person’ category, its corresponding styled counterpart called Styled-COCO-Persons (SCP), and we also introduce our own annotated dataset called ClassArch (CA). Each dataset is labelled with person bounding boxes and their corresponding body pose keypoints, details shown in Tab. 1. We focus on these datasets for training and evaluation of our models.

(a) COCO-Persons (CP) The Common Objects in CONtext dataset (COCO) [32] was specifically designed for the detection and segmentation of objects in their natural context. COCO has 328K images with over 2.5M labelled instances divided into 91 semantic object categories (e.g., car, person, dog, banana, etc.). We only consider images that include “person” instances, along with their corresponding bounding boxes and pose-keypoints. We call this split as COCO - Persons (CP). This split is taken across the training and validation sets only since the labels for the test set are not publicly available. Consequently, we use the validation set for testing our models. Tab. 1 shows the exact splits for the dataset in terms of images, persons and pose-keypoints. Figs. 4a & 4b illustrate some samples of CP dataset.

Table 1. Datasets used in our experiments. For COCO-Persons (CP), we use images from the person category of COCO [32]. Styled-COCO-Persons (SCP) is generated by using CP images as content and various splits of ClassArch (CA) images as styles (Images: images, Persons: person bounding boxes, Poses: pose annotations).

| Datasets → | CP & SCP | CA |
|-----------|----------|----|
| Split →   | Train   | Val | Total | Train | Val | Total |
| Images    | 64115   | 2693 | 66808 | 1210  | 303 | 1513  |
| Persons   | 257252  | 10777 | 268029 | 2098  | 531 | 2629  |
| Poses     | 149813  | 6352 | 156165 | 1425  | 303 | 1728  |
Fig. 3. Style transfer using AdaIN [23] with full style intensity ($\alpha = 1$). AdaIN adjusts the first and second order moments of the 'Content Image' to match those of the 'Style Image'. A 'Styled Image' (style-transferred) is generated with the semantic content of the 'content image' and style of the 'Style Image'.

(b) ClassArch (CA) We introduce a challenging dataset from the domain of Classical Archaeology, called ClassArch (CA) dataset. We chose five different recurrent narratives, viz. 'Pursuits', 'Leading of the Bride', 'Abductions', and 'Wrestling' in Agonal and Mythological contexts, taken from the period between the 6th and 5th century BCE. Pose-based analysis of such paintings is of critical importance for Classical Archaeology as discussed in Sec. 1. Each of the narratives in CA has its own set of characters, which appear recurrently and are depicted with similar features and in almost identical poses. Figs. 4e & 4f illustrate some images and their corresponding person bounding boxes and person keypoints of the CA dataset. Fig. 2 displays two examples from the 'Pursuit' (1st & 2nd column) and 'Leading the bride' (bottom row) narratives. In both scenes, the main characters ('persecutor'/'fleeing' & 'bride') are depicted with similar posture in every image. CA has different sets of labels associated with it. There are 1513 images, with 2629 person annotations and 1728 pose annotations. More detailed splits are shown in Tab. 1.

(c) Styled-COCO-Persons (SCP) The images in CP significantly differ in semantic content and style from the ancient Greek vase paintings, c.f. Fig. 4a vs. Fig. 4e). To bridge this domain gap between CP and CA, we use style transfer to adapt the style of the CP dataset to vase paintings.

Style transfer algorithms render a synthetic image that combines the semantic information from one input (denoted as content image) with the texture from the user-defined style image [14]. We apply an efficient and fast style transfer technique using adaptive instance normalisation (AdaIN) [23] to create SCP, a synthetic dataset that combines the semantic content of the CP with the style of CA. Fig. 3 illustrates the style transfer procedure. We can visually observe that images of Figs. 4c & 4d are more closer in styles with Fig. 4e than Fig. 4a, i.e., SCP is closer in style with CA, than CP is with CA. The SCP dataset will be released along with the code.

Alpha ($\alpha$) and Style-Sets Huang et al. [23] suggest a content-style trade-off technique to control the intensity of style transferred to the content image using $\alpha \in [0, 1]$. Based on this, we generate 2 groups of SCP. First with $\alpha = 0.5$, meaning that we only transfer half of the style intensity to the content images; and a second one in which $\alpha$ is chosen randomly from the uniform distribution $U[0, 1]$. The second group contains images across the whole spectrum, from no style ($\alpha = 0$) to full style ($\alpha = 1$).
Enhance Poses in Greek Vase Paintings

Fig. 4. Dataset Samples: (a) Images & (b) Labels of CP dataset; (c) & (d) are samples from the SCP dataset with $\alpha = 0.5$ and $\alpha = U$ respectively; (e) shows images with (f) the corresponding labels of our CA dataset. Each labelled example shows the corresponding person bounding boxes and their pose keypoints.

Additionally, we generate two more SCP dataset variations with the method described above using a different dataset of style images. We name this dataset as Red-Black figures (100 in total) or just RB, they are similar in style to the CA dataset but do not have any labels. Our hypothesis is that the model should be able to learn the styles and not the content of style images. In the end, we have four groups of SCP dataset with two different combinations of $\alpha$ with the two style-sets RB and CA.

Open Source Images (OSI) We will publicly release the CP and SCP datasets + annotations. 71 sample OSI links from the CA dataset are in the supplementary material, the rest can be downloaded from Beazley Archive Pottery Database\footnote{https://www.beazley.ox.ac.uk/pottery/default.htm} for research purposes. We’ll release the permanent links to those images, along with the pose and person annotations.

4 PROPOSED METHOD

In this section, we first present our proposed style-based transfer learning approach to enhance pose estimation. We then briefly explain our models that were trained and evaluated for all datasets mentioned in Sec. 3. Lastly, we propose a perceptual loss as a regularizer to improve the estimation of perceptually similar poses with different styles.

4.1 Pose Estimation Approach

We take a top-down approach to pose-estimation, which is divided into two stages. Fig. 5 details the two stages of the top-down approach. The first stage ($A/A^*$) detects all the persons in an image and then estimates the keypoints ($B/B^*$) for each person instance, then creating the poses for each instance by pose-parsing ($C/C^*$). The models without * are trained on styled datasets while the ones with * are fine-tuned on the CA dataset. We use Faster-RCNN [43] as our person detector that was trained on the COCO [32] dataset. Top-down pose estimation approaches dominate the COCO keypoint detection challenge in the past few years, and several use the HRNet [51] as their backbones. Hence, we chose HRNet-W32 (henceforth denoted as HRNet) as our pose estimation model.

2-Step Training Approach. We adopt a 2 step approach to enhancing pose estimation, as shown in Fig. 5. In the first step (Fig. 5, second row), we train our detector and pose estimation model on styled data (different groups of
Fig. 5. (first row, Top-down pose estimation) - (A*) styled person detector detects all instances, (B*) for which the body joint locations are predicted using a person keypoint detector, (C*) The pose skeletons are assembled by connecting the detected keypoints for each person. 2 Step Training Approach: Step 1 (second row, Styled Models) Person Detector trained on SCP persons data, and HRNet on SCP poses data; Step 2 (third row, Styled-Tuned Models) Styled Person Detector from second row is fine-tuned on CA persons data, and Styled HRNet is fine-tuned on CA pose data.

SCP) to generate styled models (A and B in Fig. 5). In this step, the styled models at the end of their training are expected to learn the styles of the target data, while trying to maintain the performance on the original task.

In the second step (Fig. 5, third row), we fine-tune these styled models on our CA data. During this step, the models that have learned the styles in the first step, now focus on improving their performance for the target dataset of CA. The final detector (A*) and pose estimator (B*) models are initialized with the A and B models from the first step and then fine-tuned on SCP persons and poses data.

We report all our experiments using four kinds of models for both tasks, person detection and pose estimation.

1. **Baseline models** are SOTA models. In case of the person detector, we drop the heads for all the classes except one, and fine-tune it on the ‘person’ class, further denoted as our baseline model.

2. **Tuned models** are SOTA models fine-tuned on the CA dataset. For the detector, we drop all the heads except one (similar as for baseline model) in Faster-RCNN and fine-tune it on persons data of the CA dataset. Likewise, for pose estimation, we take the SOTA HRNet and fine-tune it on pose data of CA.

3. **Styled models** are SOTA models trained on a particular group of the SCP dataset. As explained in Sec. 3, there are 4 different groups of SCP dataset. Depending on the values of $\alpha$ and style-set (RB or CA), the Styled models are trained on that particular group, for the detector as well as the pose estimator. Accordingly, there are 4 different Styled models for each of the detector and the pose estimator.

4. **Styled-Tuned (Sd→Td) models** are Styled models (3 above) fine-tuned on CA dataset. Accordingly, for the detector, Sd→Td model is a Styled Faster-RCNN model fine-tuned on CA persons data. Similarly, for the pose
estimator, $S_d \rightarrow T_d$ model is a styled HRNet model fine-tuned on CA poses data. Hence, depending on the group of the styled model, there is an equivalent $S_d \rightarrow T_d$ model.

4.2 Enforcing Perceptual Similarity
While training the styled models, the network is fed with styled data. The advantage of doing this is to allow the model to expand its capacity to recognise perceptually similar persons/poses with different styles. In order to achieve content consistency in the perceptual space, we enforce a pre-computed perceptual loss [29] while training, in addition to the regular loss. The model is penalised if it is not able to maintain perceptual consistency.

Let’s denote the task loss by $L_T$, where $L_T = L_{det}$ is for detector models, and $L_T = L_{pose}$ for pose models. We adopt two flavours of the combined loss, each for the detector as well as pose. In the first flavour ($L_{comb1}$), we adaptively weigh the perceptual loss ($L_{percept}$) with the corresponding detector or the pose loss, as shown in Eq. 1a. While, in the second flavour ($L_{comb2}$), we weigh each loss term with optimal values of $\lambda_1$ and $\lambda_2$ chosen using hyperparameter optimisation, as shown in Eq. 1b.

$$L_{comb1} = L_T + \lambda_1 * L_{percept}$$  \hspace{1cm} (1a)
$$L_{comb2} = \lambda_1 * L_T + \lambda_2 * L_{percept}$$  \hspace{1cm} (1b)

5 EXPERIMENTS AND ANALYSIS
The exact number of images, person bounding boxes and pose annotations, along with the corresponding train/val splits used for our experiments are mentioned in Tab. 1. In this section, we describe the evaluation protocol to train our detector and pose estimator. We also present the experimental results and discuss our findings.

5.1 Training Setup
In general, we use the standard parameters of the SOTA models and make adjustments to suit our experimental needs. For person detection (Faster-RCNN), we use an initial learning rate ($lr_{init}$) of 0.0001 with a scheduler $lr_{scheduler}$ on plateau (3 epochs) which reduces the lr by a factor of 0.33. We use Adam [30] with its default parameters and a batch-size ($bs$) of 8. Standard multi-task loss metric, a combination of log loss and regression loss ($L_{det} = L_{CLS} + L_{Reg}$), is used in our experiments for the detector. We train for 25 epochs on the CP dataset and 30 each on the SCP and CA datasets. However, we found that our models usually converge between 8–12 epochs.

Similar to the detector, we use Adam with its default values for pose models (HRNet) and $bs=64$. With an $lr_{init}$ of 0.01 and a $lr_{scheduler}$ on plateau (3 epochs) which reduces the lr by a factor of 0.1. We train all pose models for 100 epochs. Akin to the original HRNet [51], we also use the same configs (augmentations, image size) for fair comparison. Like HRNet, Object Keypoint Similarity (OKS) is used as an evaluation metric in our experiments as a simple Euclidean distance ($L_{pose} = L_{MSE}$) for pose estimation.

In both cases, person detection and pose estimation, we report the mean Average Precision ($mAP$) as well as the corresponding mean Average Recall ($mAR$).

5.2 Experiments
We compare different models in separate tables to give a clear understanding of our methods. As described in Sec. 3, Style-set (SS) represents two datasets CA and RB and alpha ($\alpha$) represents the amount style transferred in the SCP dataset from no style ($\alpha = 0$) to full style ($\alpha = 1$). Tab. 2 (Results A) compares the baseline models with styled models for detector as well as pose. Similarly, Tab. 3 (Results B) compares the tuned with styled-tuned. Tab. 4 (Results C) shows the influence of using different data quantities to fine-tune our models, where as Tab. 5 shows the advantage of using perceptual loss.
Madhu and Villar-Corrales, et al.  

Table 2. Results A: Comparing baseline model and styled models with different combinations of $\alpha$ (stylization factor) and SS (RB or CA), (a) for pose estimator and (b) person detector. $\alpha = 0.5$ or $\alpha = U$ (randomly sampled from a uniform distribution: $U = uniform(0, 1)$). All values in terms of mAP, except CA mAR (mAR). The (+/-) is in reference to the corresponding baselines.

| Model   | CP   | SCP | CA   | CA mAR | $\alpha$ | SS |
|---------|------|-----|------|--------|----------|----|
| Baseline | 76.5 | 46.2| 24.7 | 30.9   | -        | -  |
| Styled  |      |     |      |        |          |    |
|         | 73.4 (-3.1) | 54.4 (+8.2) | 29.7 (+5.0) | 36.0 | 0.5 | RB |
|         | 74.0 (-2.5) | 53.7 (+7.5) | 30.9 (+6.2) | 37.7 | $U$ | RB |
|         | 74.0 (-2.5) | 53.8 (+7.6) | 30.6 (+5.9) | 36.9 | 0.5 | CA |
|         | 74.3 (-2.2) | 53.5 (+7.3) | **32.3** (+7.6) | **39.0** | $U$ | CA |

(a) Pose Estimation

| Model   | CP   | SCP | CA   | CA mAR | $\alpha$ | SS |
|---------|------|-----|------|--------|----------|----|
| Baseline | 39.4 | 24.2| **10.4** | 9.8   | -        | -  |
| Styled  |      |     |      |        |          |    |
|         | 37.5 (-1.9) | 33.4 (+9.2) | 7.6 (-2.2) | 9.8 | 0.5 | RB |
|         | 36.9 (-2.5) | 32.1 (+7.9) | 6.5 (-3.3) | 8.5 | $U$ | RB |
|         | 37.7 (-1.7) | 35.2 (+9.0) | 8.2 (-1.6) | 10.0 | 0.5 | CA |
|         | 37.0 (-2.4) | 32.6 (+8.4) | 6.5 (-3.3) | 8.6 | $U$ | CA |

(b) Person Detection

Baseline vs Styled models. (Results A, Tab. 2) It is important to understand the impact of styles on the main task for detection and pose estimation. We study the impact of style-transfer by comparing baseline and styled models. As shown in Tab. 2 (SCP column), we observe that the styled models perform consistently much better than their baseline counterpart for detection and pose estimation. When tested on the CA dataset, counter-intuitively, these models underperform in detection. One potential reason is that the network has never seen the complex vase dataset during training. Conversely for pose estimation, styled models unambiguously are better for both SCP and CA datasets. Specifically, styled models, which were not trained on CA, give a considerable jump in performance: 7.62 (mAP) & 8.06 (mAR) when tested on CA.

Tuned vs. Styled-Tuned models. (Results B, Tab. 3) With the goal of enhancing pose estimation on our CA dataset, a naive approach is to fine-tune on this data, we call these models as Tuned models. Then, we take the styled models (Tab. 2), which have already learned the styles of CA data, and fine-tune them on our CA data (Styled-Tuned or $Sd \rightarrow Td$). As seen in Tab. 3, the $Sd \rightarrow Td$ models give a better performance as compared to their Tuned counterparts for pose estimation. Irrespective of the combination of $\alpha$ and SS, the pose models tend to perform better. We argue that this is partly because the models gradually learn the styles (SCP), while optimising for the main task. During training, the Styled models (Tab. 2) are able to see the different spectrum of style intensities. They adapt the styles while maintaining a consistent performance over the main task. However, for person detection the performance of $Sd \rightarrow Td$ model is detrimental in comparison to the Tuned counterpart. One reason for this can be attributed to the overlapping objects: animals and persons – which makes the person detection more difficult for the $Sd \rightarrow Td$ models as compared to their Tuned counter parts. Another reason for lower precision is the lack of ground truth annotations for side characters of the scene. For poses however, the overlap of keypoints when compared to the bounding boxes is very small and hence the model generalizes better from styled models in comparison to directly tuned models.
Table 3. **Results B**: Comparing tuned model with the styled-tuned (Sd→Td) model, with different combinations of $\alpha$ (stylization factor) and SS (RB or CA), for (a) pose estimator (b) and detector. $\alpha = 0.5$ or $\alpha = U$ (randomly sampled from a uniform distribution: $U = uniform(0, 1)$). All values in mAP, except CA$_{mAR}$ (mAR). The (+/-) is in reference to the corresponding baselines.

| Model | CP   | SCP  | CA    | CA$_{mAR}$ | $\alpha$ | SS |
|-------|------|------|-------|------------|----------|----|
| Tuned | 14.0 | 9.3  | 65.6  | 72.3       | -        | -  |
| Sd→Td |      |      |       |            |          |    |
|       | 11.8 (+2.2) | 10.5 (+1.2) | 66.8 (+1.2) | 73.3 | 0.5 | RB |
|       | 20.3 (+6.3) | 14.5 (+5.2) | **67.2 (+1.6)** | 73.3 | $U$ | RB |
|       | **34.9 (+20.9)** | **22.4 (+13.1)** | 66.6 (+1.0) | 73.1 | 0.5 | CA |
|       | 28.0 (+14.0) | 18.5 (+9.2) | 67.1 (+1.5) | **73.6** | $U$ | CA |

(a) Pose Estimation

| Model | CP | SCP | CA    | CA$_{mAR}$ | $\alpha$ | SS |
|-------|----|-----|-------|------------|----------|----|
| Tuned | -  | -   | 49.4  | 37.0       | -        | -  |
| Sd→Td | -  | -   | 44.3 (-5.1) | 32.9 (-4.1) | 0.5 | RB |
|       | -  | -   | **43.0 (-6.4)** | 32.6 (-4.4) | $U$ | RB |
|       | -  | -   | 43.7 (-5.7) | **33.4 (-3.6)** | 0.5 | CA |
|       | -  | -   | 43.9 (-5.5) | 32.9 (-4.1) | $U$ | CA |

(b) Person Detection

Table 4. **Results C**: Comparing tuned model with the styled-tuned (Sd→Td) model, by training on different quantities (25 %, 50 %, 75 %, 100 %) of the CA data for pose estimation. It clearly shows that the styled model learns quicker. All values in mAP, $\alpha = 0.5$ or $\alpha = U$ (randomly sampled from a uniform distribution: $U = uniform(0, 1)$), and SS = RB or CA.

| Model | 25 % | 50 % | 75 % | 100 % | $\alpha$ | SS |
|-------|------|------|------|-------|----------|----|
| Tuned | 61.3 | 65.0 | 65.1 | 65.6  | -        | -  |
| Sd→Td | 60.6 | 65.4 | 65.1 | 66.8  | 0.5      | RB |
|       | **62.1** | 64.7 | **66.5** | **67.2** | $U$ | RB |
|       | 60.6 | 65.2 | 65.4 | 66.6  | 0.5      | CA |
|       | 61.4 | 64.8 | 65.7 | 67.1  | $U$ | CA |

With Tab. 2 and Tab. 3, we were able to enhance the performance of pose models, with styled as well as styled-tuned models. Styled models can help to improve the performance with a 7.7 pp (mAP) jump in performance, without any labels. While the styled-tuned models show that fine-tuning with styled models is generally beneficial for the performance, the Sd→Td model for pose gives a significant 1.6 pp (mAP) performance improvement for CA dataset over its Tuned counterpart.

**Influence of data quantity.** (Results C, Tab. 4): Tab. 4 shows that Styled-Tuned models learn faster than their Tuned counterpart, for each of the corresponding splits of CA data (model with $\alpha = U$, SS=RB is more consistent). Specifically, the model with $\alpha = 0.5$, SS=RB and 50% of CA data gives equivalent performance to the Tuned model trained with whole CA data. We see that the deep learning based models converge faster when they have a suitable initialization of weights [18]. We argue that training on styled data helps the model to get a better initialization with respect to the dataset distribution. Consequently, the convergence is faster.
Table 5. **Perceptual Loss Comparison**: Comparing different loss combinations using the styled or styled-tuned (Sd→Td) model on the CA dataset for (a) pose estimation and (b) detection, i.e., just the detector or pose loss \( L_{\text{det}} \) or \( L_{\text{pose}} \) in combination with perceptual loss \( L_{\text{comb}} \) in two different variants (Eqs. (1a) and (1b). For (a): \( \alpha = 0.5 \) and for (b): \( \alpha = U \). Style-Set (SS) is CA for both.

| Model          | mAP  | mAR  | \( L_{\text{comb}} \) |
|----------------|------|------|-----------------------|
| Style          | 30.6 | 36.9 | \( L_{\text{pose}} \) |
| Posed          | 30.5 | 36.5 | \( L_{\text{comb1}} \) |
| Tuned          | 30.5 | 36.5 | \( L_{\text{comb2}} \) |
| Sd→Td          | 66.6 | 73.1 | \( L_{\text{pose}} \) |
| Tuned          | 67.2 | 73.6 | \( L_{\text{comb1}} \) |
| Tuned          | 67.2 | 73.6 | \( L_{\text{comb2}} \) |

(a) Pose Estimation

| Model          | mAP  | mAR  | \( L_{\text{comb}} \) |
|----------------|------|------|-----------------------|
| Posed          | 65.5 | 8.5  | \( L_{\text{det}} \) |
| Tuned          | 11.2 | 11.0 | \( L_{\text{comb1}} \) |
| Tuned          | 9.8  | 11.2 | \( L_{\text{comb2}} \) |
| Sd→Td          | 43.9 | 32.9 | \( L_{\text{det}} \) |
| Tuned          | 45.4 | 34.7 | \( L_{\text{comb1}} \) |
| Tuned          | 45.7 | 34.0 | \( L_{\text{comb2}} \) |

(b) Person Detection

**Perceptual Loss Comparison**: Tab. 5 shows the influence of perceptual loss for the model performance. \( L_{\text{comb1}} \) is the experiment with adaptively weighing the detector as well as pose losses (Eq. 1a). For \( L_{\text{comb2}} \) (Eq. 1b), we determined the values of \( \lambda \)s through a parameter search [1]: \( \lambda_1 = 0.43 \) and \( \lambda_2 = 0.92 \) for person detection and \( \lambda_1 = 0.47 \), \( \lambda_2 = 0.018 \) for pose estimation. We present results for the combination SS=CA & \( \alpha = U \) for the detector, and we chose SS=CA & \( \alpha = 0.5 \) for pose estimator.

Tab. 5 shows that the perceptual loss (adaptive: \( L_{\text{comb1}} \) or parameterised one: \( L_{\text{comb2}} \)) indeed helps the styled as well as Sd→Td models to improve their performance, in general. From Tab. 5a, we can see that the results for \( L_{\text{comb1}} \) and \( L_{\text{comb2}} \) are equal, however we have to note that this is an empirical observation for different values of lambda. Additionally, we note that perceptual loss does not harm the styled model for pose estimation, but actually helps since the styled model is fine-tuned on CA.

5.3 Qualitative Pose Estimation Results

Tabs. 3 & 5 show that styled-tuned models are consistently giving a better performance than any other method. We visualise the predictions of these models for comparison of their performances. Fig. 6 shows four characters (wrestler, fleeing, persecutor, bride) and their pose predictions from each of our 5 proposed models.

As shown in Fig. 6a, the baseline model is the poorest in pose predictions. It is not able to detect majority of keypoints, confuses between the limbs if multiple characters are present and incorrectly predicts the keypoint locations.

**Styled model** is generally much better (Fig. 6b) than the baseline model (also Tab. 2). It is able to predict more keypoints and does not get confused if multiple characters are present. However, it is not able to predict all the visible keypoints and sometimes (Fig. 6b, last row) gives worse performance than even a baseline model.

**Tuned, styled-tuned (Sd→Td)** and styled-tuned with perceptual loss (Sd→Td_p) models are overall quite superior to baseline and styled models. They are able to predict almost all of the visible keypoints, do not confuse between multiple characters and are quite precise with the keypoint locations. However, there are subtle differences that make Sd→Td models better. They are able to predict all the visible keypoints as shown in Fig. 6d and Fig. 6e, whereas tuned models miss some (e.g., Fig. 6c where the shoulder joints are missing). The keypoint location precision is also improved using Sd→Td models. Visually it is difficult to generalise if models with perceptual loss are better or not, however, they are more precise (Fig. 6e, third row ankle is corrected, but a shoulder is missed).
6 POSE-BASED RETRIEVAL

Our experiments (Sec. 5) showed that Styled models and Styled-Tuned models achieve better keypoint detection results, quantitatively and qualitatively, than their corresponding counterparts. In this section, we show that our two-step training pipeline is also beneficial for discovering similar images based on character poses. We call the process of retrieving images based on poses as pose-based retrieval.
Table 6. Retrieval Results: The (C)* models show the retrieval values based on characters, where as the (S)* models show for the scenes. P is Precision, and mAP is mean-Average Precision. α = U; SS: Style-Set. Sd→Td* are style-tuned models, where $p_1 = L_{comb1}$ (Eq. 1a) and $p_2 = L_{comb2}$ (Eq. 1b).

| Model       | P@1  | P@5  | mAP  | α  | SS    |
|-------------|------|------|------|----|-------|
| (C) Baseline| 31.7 | 25.5 | 21.5 | -  | -     |
| (C) Styled  | 37.3 | 30.2 | 23.1 | U  | CA    |
| (C) Tuned   | 43.0 | 39.8 | 27.5 | -  | -     |
| (C) Sd→Td   | 47.7 | 42.2 | 28.3 | U  | CA    |
| (C) Sd→Td_p1| 45.7 | 41.4 | 28.0 | 0.5 CA |
| (C) Sd→Td_p2| 48.3 | 41.1 | 28.4 | 0.5 CA |
| (S) Baseline| 43.6 | 43.2 | 35.9 | -  | -     |
| (S) Styled  | 46.9 | 43.6 | 37.3 | U  | CA    |
| (S) Tuned   | 56.4 | 52.7 | 41.5 | -  | -     |
| (S) Sd→Td   | 58.8 | 55.2 | 42.1 | U  | CA    |
| (S) Sd→Td_p1| 57.8 | 53.5 | 41.5 | 0.5 CA |
| (S) Sd→Td_p2| 58.8 | 53.4 | 41.8 | 0.5 CA |

6.1 Experimental Setup

The database for image retrieval and discovery is built from the CA validation dataset. The database consists of 303 images and their respective detected poses for best of baseline, styled, tuned, Sd→Td and Sd→Td* (with perceptual loss). We perform two retrieval experiments based on the class label for each image, which is either a character or scene. There are 15 unique characters (C) and 5 Scenes (S). Given a query image, we rank the retrieved images based on the OKS metric [32]. In order to evaluate the retrieval method, we use the precision as: $P^* = \frac{TP^*}{TP^* + FP^*}$, where $^* = @k$, consequently $P@k$ = Precision at k; TP := true positives, FP := false positives and FN := false negatives. We report P@k and mAP, for k=1 and k=5. In all our experiments, we exclude the self-retrieval (query itself) from the evaluation. For this task, we compare all the presented models to highlight the quality of our proposed models from an application perspective. The focus of this work is on enhancing poses for Greek vase paintings and not presenting a novel image retrieval method and hence we do not compare with SOTA image retrieval methods.

6.2 Retrieval Results and Discovery

Tab. 6 presents our pose-based retrieval results. We observe that styled-tuned models are consistently better for C and S, and the styled are better than baselines counterparts. Fig. 7 displays a query image (a) along with the top-5 ranked retrievals for the six different evaluated models. It can be observed that the tuned and the styled-tuned models outperform the baseline and the styled models. Fig. 7 row 1 shows poor retrieval results for the baseline model. For a pursuit scene, the first two retrieved samples belong to a wrestling scene while the last 2 belong to leading of the bride scene. The styled model (row 2) is already better wherein all the five retrievals belong to the pursuit scene, however at a character level, it retrieves a persecutor at the 5th retrieval. Tuned (row 3) and styled-tuned (row 4-6) models perform similarly well. On a closer look, we see that the 2nd and 4th retrieved samples of the styled-tuned model are closer to the query sample compared to the tuned model.

7 CONCLUSION

We presented a two-stage training approach for using style transfer and transfer learning in combination with perceptual consistency to improve pose estimation in ancient Greek vase paintings. We show that the use of styled transfer learning as a domain adaptation technique for such data significantly improves the performance.
of state-of-the-art pose estimation models on unlabelled data by 6% mean average precision (mAP) as well as mean average recall (mAR). We also analysed the impact of styles as progressive learning in a comprehensive manner showing that models learn generic domain styles. We experimentally showed that our proposed method outperforms their corresponding counterparts for human pose estimation. In general, our method can be applied to diverse unlabelled datasets without explicit supervised learning. Our method also provides a way for exploring diverse cross-domain datasets with low or no labels using human poses as a tool. Finally, we also show that our method can be used for pose-based image retrieval and discovery of similar, relevant poses and corresponding scenes in collections such as ancient Greek vase paintings. For future work, we plan to a) introduce the geometric structures of vases in COCO and Styled-COCO datasets as an augmentation technique during training, b) use shape information of the persons into our framework or using segmentation as a prior for vase paintings.
ACKNOWLEDGMENTS

This paper is partially funded by the FAU Emerging Fields Initiative (EFI) project “Iconographics. Computational Understanding of Iconography and Narration in Visual Cultural Heritage” as well as partially funded by the EU H2020 project “Odeuropa” under grant agreement No. 101004469. The authors would also like to thank NVIDIA for their hardware donation.

REFERENCES

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, United States, 2623–2631. 12

[2] Colleen M. Becker. 2013. Aby Warburg’s Pathosformel as Methodological Paradigm. The Journal of Art Historiography 9 (2013), 9–CB1. 3

[3] Peter Bell and Leonardo Impett. 2019. Iconographie und Interaktion. Computergestützte Analyse von Posen in Bildern der Heilsgeschichte. Das Mittelalter 24, 1 (2019), 31–53. 3

[4] Peter Bell, Joseph Schlecht, and Björn Ommer. 2013. Nonverbal communication in medieval illustrations revisited by computer vision and art history. Visual Resources 29, 1-2 (2013), 26–37. 3

[5] Zhe Cao, Tomas Simon, Shah-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE Computer Society, United States, 7291–7299. 3, 4, 5

[6] Gustavo Carneiro, Nuno Pinho da Silva, Alessio Del Bue, and João Paulo Costeira. 2012. Artistic Image Classification: An Analysis on the PRINTART Database. In Computer Vision – ECCV 2012, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 143–157. 2

[7] João Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. 2016. Human Pose Estimation with Iterative Error Feedback. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, United States, 4733–4742. https://doi.org/10.1109/CVPR.2016.512 4

[8] Eva Cetinic, Tomislav Lipic, and Sonja Grgic. 2018. Fine-tuning convolutional neural networks for fine art classification. Expert Systems with Applications 114 (2018), 107–118. 2

[9] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. 2018. Cascaded Pyramid Network for Multi-person Pose Estimation. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, United States, 7103–7112. https://doi.org/10.1109/CVPR.2018.00742 4

[10] Elliot Crowley and Andrew Zisserman. 2014. The State of the Art: Object Retrieval in Paintings using Discriminative Regions. In British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014. BMVA Press, UK. 2

[11] Elliot J Crowley and Andrew Zisserman. 2013. Of gods and goats: Weakly supervised learning of figurative art. learning 8 (2013), 14. 5

[12] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. 2018. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, US, 994–1003. https://doi.org/10.1109/CVPR.2018.00110 5

[13] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-person Pose Estimation. In 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, US, 2353–2362. https://doi.org/10.1109/ICCV.2017.256 4

[14] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, US, 2414–2423. https://doi.org/10.1109/CVPR.2016.265 6

[15] Ross Girshick. 2015. Fast R-CNN. In 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, US, 1440–1448. https://doi.org/10.1109/ICCV.2015.169 4

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, US, 580–587. https://doi.org/10.1109/CVPR.2014.81 4

[17] Luca Giuliani. 2003. Bild und Mythos: Geschichte der Bilderzählung in der griechischen Kunst. CH Beck, München. 2

[18] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 9). Yee Whye Teh and Mike Titterington (Eds.). PMLR, Chia Laguna Resort, Sardinia, Italy, 249–256. 11

[19] Peter Hall, Hongping Cai, Qi Wu, and Tadeo Corradi. 2015. Cross-depiction problem: Recognition and synthesis of photographs and artwork. Computational Visual Media 1, 2 (01 Jun 2015), 91–103. https://doi.org/10.1007/s41095-015-0017-1 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, US, 770–778. 4

[21] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In International conference on machine learning. PMLR, PMLR, Sweden, 1989–1998. 5
[47] Babak Saleh and Ahmed Elgammal. 2016. Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature. *International Journal for Digital Art History* 2, 2 (Oct. 2016), –. https://doi.org/10.11588/dah.2016.2.23376

[48] Benoit Seguin, Lisandra Costiner, Isabella di Lenardo, and Frédéric Kaplan. 2018. New techniques for the digitization of art historical photographic archives—the case of the cini foundation in venice. In *Archiving Conference (2018, 1)*. Society for Imaging Science and Technology, Society for Imaging Science and Technology, US, 1–5.

[49] Mark Stansbury-O’Donnell. 2009. Structural differentiation of pursuit scenes. *σ το Yatromanolakis -*, - (2009), 341–372.

[50] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*. Springer, Springer, Netherlands, 443–450.

[51] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, US, 5693–5703.

[52] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*. NIPS, Canada, 1799–1807.

[53] Alexander Toshev and Christian Szegedy. 2014. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, USA, 1653–1660.

[54] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. IEEE, US, 4724–4732.

[55] Nicholas Westlake, Hongping Cai, and Peter Hall. 2016. Detecting people in artwork with cnns. In *European Conference on Computer Vision*. Springer, Netherlands, 825–841.

[56] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*. ECCV, Germany, 466–481.