Domain-Adaptive Self-Supervised Face & Body Detection in Drawings

Barış Batuhan Topal[1] , **Deniz Yuret**[1] , **Tevfik Metin Sezgin**[1]
[1]Department of Computer Engineering, KUIS AI Center, Koç University
{baristopal20, dyuret, mtsezgin}@ku.edu.tr

Abstract

Drawings are powerful means of pictorial abstraction and communication. Understanding diverse forms of drawings, including digital arts, cartoons, and comics, has been a major problem of interest for the computer vision and computer graphics communities. Although there are large amounts of digitized drawings from comic books and cartoons, they contain vast stylistic variations, which necessitate expensive manual labeling for training domain-specific recognizers. In this work, we show how self-supervised learning, based on a teacher-student network with a modified student network update design, can be used to build face and body detectors. Our setup allows exploiting large amounts of unlabeled data from the target domain when labels are provided for only a small subset of it. We further demonstrate that style transfer can be incorporated into our learning pipeline to bootstrap detectors using a vast amount of out-of-domain labeled images from natural images (i.e., images from the real world). Our combined architecture yields detectors with state-of-the-art (SOTA) and near-SOTA performance using minimal annotation effort. Our code can be accessed from `https://github.com/barisbatuhan/DASS_Detector`.

1 Introduction

Drawings serve as a rich and expressive medium for communication. Here we focus on comic books and cartoons, which are relatively recent forms of media. They combine text and graphics in a unique format to convey narratives. Key problems such as extracting the visual structure of the scenes, understanding the accompanying text, and modeling how they connect to form the narrative pose significant challenges. Hence, understanding comics has been a problem of interest to computer vision, computer graphics, and NLP communities.

In drawings, the story is narrated primarily through the scene's main characters. Hence, we study on face and body detection, two primary problems for understanding drawings.

Figure 1: Examples on the adversity of this domain (left: non-human character, right: samples from different character designs and styles).

Training face and body detectors is complicated by two challenges. First, although a tremendous amount of unlabeled data is available (primarily as digitized comic book pages and animations), face and body annotations are largely lacking. Second, since character design and drawing style change substantially across artists, series, and cultures (see Figure 1), each domain inevitably requires domain-specific tuning to create detectors. In this work, we present a pre-training pipeline for creating domain-adapted detectors, which addresses both problems. Our pipeline has two major components. The first is a self-learning component that can exploit vast amounts of unlabeled data from the target domain to create detectors that can be tuned with minimal labeled data. More specifically, we introduce a modified version of teacher-student architecture to drawings, where we periodically update the student network's weights with teacher's after a specific number of iterations and utilize the OHEM [Shrivastava *et al.*, 2016] loss with an additional positive and negative confidence threshold limitation for a more stable training. We show that this self-learning model works best if it starts with a sufficiently good teacher. This component leads to the second key component of our pipeline, which uses style transfer to transform vast amounts of labeled natural images to create sufficiently good teacher models by utilizing 11 styles from 4 style transfer algorithms.

We employ a multi-tasking strategy by jointly training the model for faces and bodies to reduce inference time and to benefit from the contextual and spatial relationship. To uti-

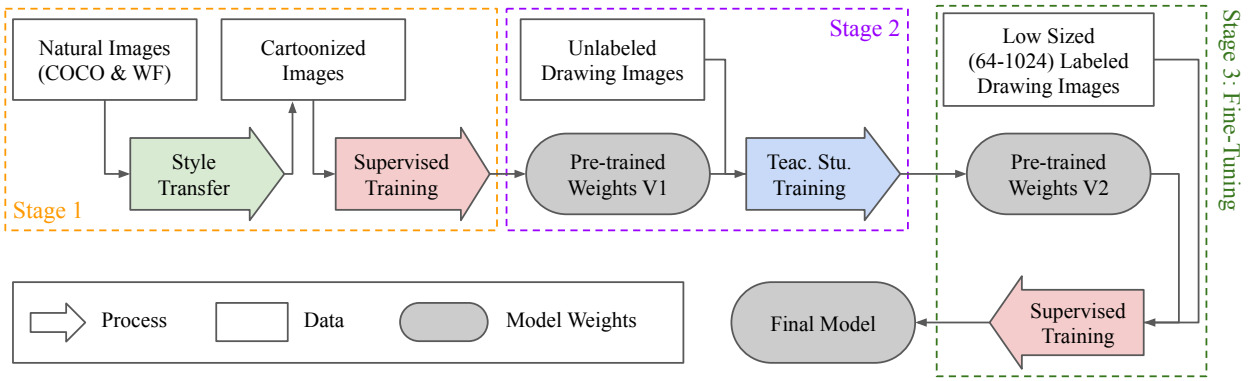

Figure 2: Summary of the proposed pipeline.

lize datasets with face-only and body-only annotations, we use two detection heads: one to predict the faces, and the other for bodies. Even without drawing domain supervision, our teacher-student model outperforms previous supervised SOTA of DCM 772 [Nguyen *et al.*, 2018] and weakly-supervised SOTA [Inoue *et al.*, 2018] in most datasets. When initialized with our pre-trained weights, our supervised model sets a new SOTA performance for most datasets, even if limited drawing data is used in training.

2 Related Works

2.1 Detection

With the increasing size of annotated data, models with high dependence on supervision were able to get good results (e.g., [Bochkovskiy *et al.*, 2020; Zhang *et al.*, 2020b; Ge *et al.*, 2021]). [Liu *et al.*, 2021] and [Xu *et al.*, 2021] introduced teacher-student training schemes and gained a significant performance boost with a low amount of labeled data. Unlike this work, these studies target natural images. Thus, cross-domain detection with these models is prone to false positives (FP) and negatives (FN). Several studies have improved the teacher-student scheme to work well in cross-domain detection. While MTOR [Cai *et al.*, 2019] exploits object relations in region-level consistency, inter-graph consistency, and intra-graph consistency, UMT [Deng *et al.*, 2021] tries to eliminate teacher and student network biases through distillation and style transferring, D-adapt [Jiang *et al.*, 2022] adopts an adversarial pipeline to the detector model, H^2FA R-CNN [Xu *et al.*, 2022] utilizes weak supervision and domain classifiers to create a more domain-invariant model. Although our solution is more similar to UMT compared to other cross-domain studies, we improve its style transferring part by mixing multiple styles, we modify the standard teacher-student training to compensate for the FP and FN cases, and we change the loss function to force the model to learn from more confident proposals.

Several studies have been done on face and object detection, specifically in drawings. [Zhang *et al.*, 2020a] proposed a fully-supervised face detector using only iCartoonFace; [Ogawa *et al.*, 2018] trained a detector from Manga 109; [Nguyen *et al.*, 2018] used DCM 772; [Inoue *et al.*, 2018] utilized Comic2k, Watercolor2k, and Clipart1k. However, these models are only trained on specific sub-domains of drawings (i.e., only utilized a single dataset with limited stylistic coverage). In this study, we leverage unlabeled drawing images from any sub-drawing domain and show that the performance on drawings can be significantly improved by using an effective pre-training pipeline and a better detector architecture.

2.2 Style Transfer

Conversion of natural images to drawings is an unpaired image-to-image translation task. SOTA models for this task have been designed with U-Net-like Generative Adversarial Networks (i.e., down-sampling first and then up-sampling). We use several cartoonization models to increase the stylistic variety of the pre-training data by selecting 11 styles from these works: Monet, Van Gogh, Cezanne from CycleGAN [Zhu *et al.*, 2017]; Shinkai, Hayao, Hosoda, Paprika from CartoonGAN [Chen *et al.*, 2018]; AS, KH, Miyazaki from GANILLA [Hicsonmez *et al.*, 2020]; and the default style in White-Box Cartoonization [Wang and Yu, 2020]. While previous detection studies on drawings have also utilized style transfer methods (e.g., [Inoue *et al.*, 2018; Deng *et al.*, 2021]), we improve on these results by combining multiple styles and analyzing which styles increase the performance more.

2.3 Datasets

Digitization has made millions of unlabeled drawings reachable on the internet. Thousands of old comic book series (e.g., Golden Age Comics between the 1930s - 1950s) have been published on several websites [1] and gathered as an unlabeled dataset named COMICS [Iyyer *et al.*, 2016]. Newer series can be obtained through web crawling. Unfortunately, annotated datasets only comprise a small subset of this domain in terms of stylistic variety and quantity. Regarding the stylistic distribution of labeled datasets, the majority of iCartoonFace [Zheng *et al.*, 2020] is retrieved from Asian products (∼74%), Manga 109 [Matsui *et al.*, 2017] only covers Japanese Manga

[1] comicbookplus.com and digitalcomicmuseum.com

styles, DCM 772 [Nguyen *et al.*, 2018] is limited to comics from Golden Age Era. Although [Inoue *et al.*, 2018] introduces Comic2k, Watercolor2k, and Clipart1k for body detection, they also remain stylistically bound in their sub-domain. Currently, none of the available datasets provide comprehensive stylistic coverage. In particular, contemporary US and Western comics have little if any annotated examples. In terms of dataset quantity, iCartoonFace contains a significant amount of face data with its 50,000 training and 10,000 validation images. The situation is a bit more challenging with body annotations: Manga 109 has $\sim$21,000 page images, but the style is limited to black and white mangas. DCM 772 consists of only 772 images. Comic2k, Watercolor2k, and Clipart1k increase the total labeled data by 2,500 instances. Building a body detector for drawings that is not fragile to different styles is challenging using only these datasets. Hence, self-supervised approaches are essential for creating suitable models for target instances with unseen styles.

3 Methodology

Our training consists of three stages. In the first stage, we use two large and annotated real-life image datasets, cartoonize them using style transfer methods, and perform pre-training for face and body detection. In the second stage, we utilize the extensive amount of unlabeled comic drawings available and perform self-supervised training on our pre-trained model with the modified form of the teacher-student architecture. In the final stage, we leverage the limited amount of annotated comic drawings to fine-tune our model. In Figure 2, you can see a demonstration of our complete pipeline. In the following subsections, we describe our base model and the three stages we propose in more detail.

3.1 Model Architecture

Since the challenge in our domain consists of stylistic variety in object representations (see Figure 1), we decide that adopting an object-detector-like model would provide greater performance, where the architecture is specifically designed to find multiple objects with various appearances. Secondly, we aim to use a more robust and simple model with low inference time to focus mainly on the effects of style transfer and self-supervised training. Therefore, we select one of the SOTA single-shot non-swin-transformer anchor-free object detectors, YOLOX [Ge *et al.*, 2021], as our baseline architecture. However, our pre-training pipeline does not depend on this specific baseline. Hence it can be applied to any detector.

As discussed in Section 2, COCO [Lin *et al.*, 2014], WIDER FACE [Yang *et al.*, 2016], and some of the available drawing datasets do not include both face and body annotations together. To train the model jointly for both face and body parts and benefit from all the available datasets, we separate the detection head of the original YOLOX model into two pieces. Each piece proposes bounding boxes with their confidence values only for a single class. Our overall architecture can be seen in Figure 3. During training, the heads are trained alternately at each forward pass.

3.2 Stage 1: Style Transferred Pre-Training

Preprocessing. We process COCO and WIDER FACE datasets with 11 different styles. We eliminate all the images in COCO that do not have people or animals. We also count animals as bodies during training because drawings may include animal-like characters. To the best of our knowledge, no dataset includes annotations for animal faces. Thus, facial training is solely done through human faces in WIDER FACE. We discard the images in which a person has a face with its maximum facial side length smaller than $\sim$2% of the image's minimum side length. These faces are not required in the dataset since characters in drawings mostly have a bigger appearance on the image.

Training Experiments. We create 5 different experiments to test our model's success at pre-training stage 1:

- **Single Styles:** We analyze the effect of each style on the detection performance by training individual models with only one style transferring method.
- **All Styles:** We train an additional model by combining all styles with random selection per each image to notice if using multiple styles increases the overall performance.
- **Best Styles:** We choose five styles that result in the greatest performance individually and train another model by combining only these to find if selecting the most effective styles is more logical instead of utilizing all styles.
- **No Style:** We train an extra model that uses the original images without any stylization to observe the benefit of style transferring.
- **No Animals Included:** We test the effect of including animal bodies to body annotations to the performance. We utilize all of the styles but exclude the animal boxes from the training data.

3.3 Stage 2: Self-Supervised Pre-Training

Model Architecture. The model consists of two different network parts: teacher and student. These networks are identical and initialized from the same pre-trained set of weights that we obtain from stage 1: style transferred pre-training with the mixture of all styles. The teacher network processes a non-augmented complete image and generates bounding box predictions along with their confidence values. The student network also generates predictions, but it processes a heavily augmented version of the same input image. High-confidence predictions of the teacher network are further processed with the non-maximum suppression (NMS) algorithm, and the outputs are considered as the pseudo-ground-truth labels of the image. The student network is trained with the loss computed by using the labels retrieved from the teacher network. The gradient flow of the teacher network is stopped, and it is updated at each iteration with respect to the Eq. 1, where TN is the teacher, SN is the student network weight, d is a hyper-parameter:

$$TN = d \cdot TN + (1 - d) \cdot SN \tag{1}$$

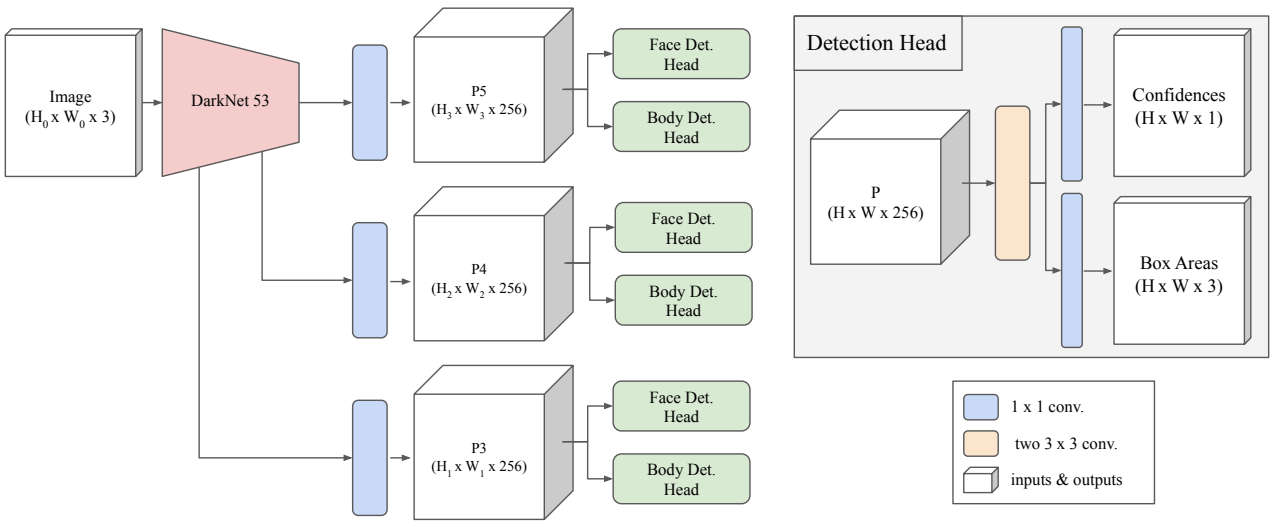

Figure 3: Our complete model architecture.

Although TN is updated with the student weights in earlier studies, student weights are only changed with backpropagation. In our experiments, we have seen that this design causes the development of both modules at the earlier stages but a significant performance drop in SN in the later iterations due to the noisy pseudo-ground-truth labels caused by the change in the input domain between pre-training stage 1 (i.e., cartoonized natural images) and self-supervised processes (i.e., drawings). This drop also affects the performance of TN. Hence, we load the weights of TN to SN per each Φ iteration to fix the deterioration of SN. Since this step manipulates the values without the gradient flow, an optimizer with the momentum information may mislead the overall model. Thus, we change our optimizer to Stochastic Gradient Descent (SGD). Our self-supervised architecture can be seen in Figure 4.

Loss. In Focal Loss, each prediction is included in the confidence loss calculation with a weight that balances the positive (i.e., predictions in which the actual ground truth object is present) and negative (i.e., predictions that point to a background area) boxes. This approach is advantageous in fully supervised training since the ground truth box areas of every object in the image are given to the model. On the other hand, in self-supervised detectors, the high-probability predictions of the teacher model are selected as pseudo-ground-truth values, which are prone to false positives (FP) and false negatives (FN). FP cases can be minimized by increasing the confidence threshold for ground truth selection. However, this choice also increases the FN rate. To further decrease the FN cases, we follow the OHEM loss [Shrivastava *et al.*, 2016], where only a subset of predictions are chosen to calculate the loss. We also modify this loss so that the predictions can be selected as positive predictions only above a specific confidence threshold and negative predictions below a particular threshold. Subset selection and this modification help the model to skip a subset of FN cases of the teacher model in loss calculations (e.g., if a face/body area is predicted but has a low confidence value). Loss calculation of a single selected box proposal can be seen in Eq. 2:

$$\begin{aligned}
\mathcal{L}_{conf} &= -p \cdot ct_{pos} \cdot log(\hat{p}) - (1-p) \cdot ct_{neg} \cdot log(1-\hat{p}) \\
\mathcal{L}_{reg} &= \sum_{i}^{\{w,h,x,y\}} smooth_{L_1}(i_{gt}, i_{pred}) \\
\mathcal{L}_{total} &= \mathcal{L}_{conf} + \beta \mathcal{L}_{reg}
\end{aligned} \tag{2}$$

$\mathcal{L}_{conf}$ is the confidence loss and $\mathcal{L}_{reg}$ is the regression loss. $p \in \{0, 1\}$ indicates if the box is selected as positive ($p = 1$) or negative ($p = 0$), $\hat{p} \in [0, 1]$ is the confidence value of the selected box, $ct_{pos} \in \{0, 1\}$ is 1 if the confidence of the proposed box is above the positive confidence threshold, $ct_{neg} \in \{0, 1\}$ is 1 if the confidence of the proposed box is below the negative confidence threshold, $\{w, h, x, y\}$ are the width, height, and the center points of the box, β is the balancing parameter between confidence and regression losses.

Unlabeled Datasets. We crawled 195,321 comic book pages from today's US and European series to train our model. We also utilized 198,657 pages from COMICS and leveraged iCartoonFace, Manga 109 pages, Comic2k, Watercolor2k, and Clipart1k images. At each forward pass, we select a random image from these image sets.

Experiments & Hyper-parameters. We run several experiments with different losses, Φ, β, d, positive and negative student confidence thresholds. In our final model, we set Φ to 500, β to 2, d to 0.9996, and positive and negative thresholds (ct_{pos}^{thres} and ct_{neg}^{thres}) to 0.5.

3.4 Stage 3: Fine-Tuning

We conduct experiments with three different pre-training methods: random initialization, style transferred pre-training in stage 1, and teacher-student network from stage 2. Since

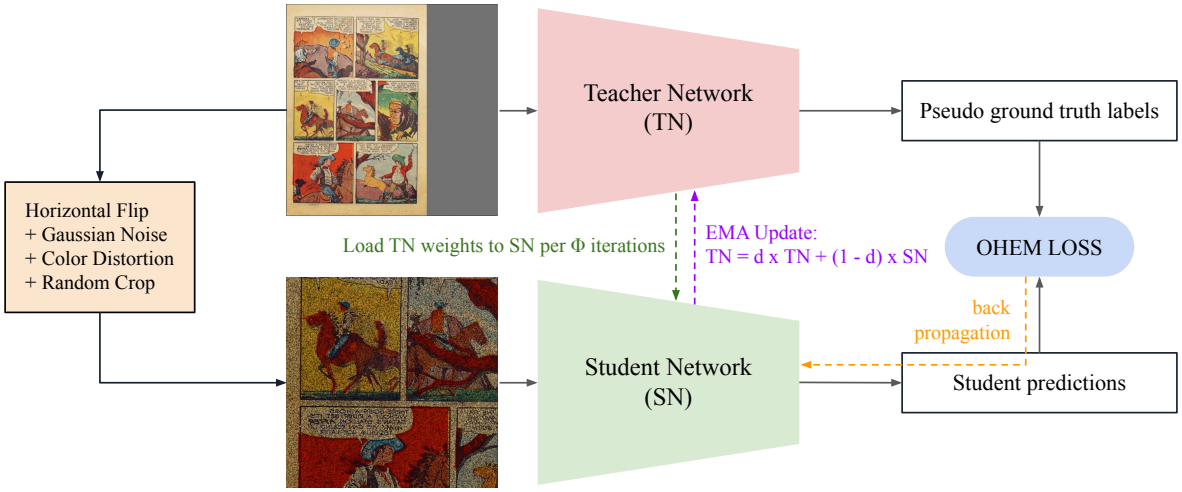

Figure 4: Our stage 2 teacher-student network training process.

each drawing dataset contains its own separate stylistic distribution, they should be fine-tuned separately to obtain the maximum performance on their test set. Thus, we fine-tune the model with single datasets for each pre-training variation by randomly selecting a limited number of image instances (i.e., 64, 128, 256, 512, 1024 images, or all data). As Manga 109 and DCM 772 consist of page images instead of individual panels, we separate panels during training to increase the number of input data and test the models with their page images.

4 Results & Discussion

In the following parts, we explain our training details, discuss the effect of style transferring in stage 1, analyze the experiments done by utilizing the teacher-student network, and present our results retrieved after fine-tuning with limited and unlimited drawing data. We will use abbreviations[2] of datasets in the given tables to save space since there are many datasets for evaluation. Average Precision (AP) is selected as the evaluation metric for detection, and the intersection of union value for evaluation is fixed at 0.5. At each table given in this section, the best result per column is marked in **bold** and the second is underlined.

4.1 Training Details

In all variations and experiments, the batch size is set to 16, and one Tesla T4 GPU is used. AP scores are calculated by running the same variation five times and computing the average of these runs. At stages 1 and 3, the learning rate is fixed at 0.001. The highest-scoring checkpoints in the evaluation set among 350 epochs are chosen as the final models. The first and the last 15 epochs include no augmentation. Otherwise, horizontal & vertical flips, the color distortion between $[-20°, 20°]$ degrees, shear, and mosaic augmentation (i.e., combining four random images and passing them as a single image) are applied randomly between the 15th and 335th epochs. For the teacher-student network, the learning rate is set as 0.0001, and the best checkpoints in 10000 iterations are taken as final models. While the input image of the teacher network is only horizontally flipped, Gaussian noise, color distortion, and random crop are applied additionally to the student network in all epochs.

[2]iCartoonFace as iCF, Manga 109 faces as M109-F, Manga 109 bodies as M109-B, DCM 772 faces as DCM-F, DCM772 bodies as DCM-B, Comic2k as C2k, Watercolor2k as W2k, and Clipart1k as C1k. If Manga 109 is used directly, then it means that the face and body AP scores are averaged.

4.2 Stage 1: Style Transferred Pre-Training

In this stage, we try to find the best combination to initialize the teacher-student network. For this purpose, we train the model variations with cartoonized natural images but evaluate them with drawing datasets. Scores retrieved after pre-training stage 1 are given in Table 2 for the individual top-5 styles (i.e., *Whitebox, Hosoda, KH, Hayao* and *Shinkai*) and other experiments.

In the drawing domain, characters can be drawn in various styles. Although texture and colors continuously change among products, key fragments of faces and bodies preserve their existence (e.g., faces include at least one eye, and bodies contain either a head, arms, or legs). In our case, we believe that using multiple styles instead of one forces the model to focus more on to shape of the object rather than texture. Consequently, the model learns more generalizable information rather than style-specific; the objects are detected more accurately when the model is tested with unseen examples. Therefore, while leveraging even a single style transferring method from top-5 ensures performance increase compared to using *No Styles*, *All Styles* outperforms both individual styles and *Top-5 Styles*. Furthermore, adding animal annotations to the ground truth during the style transferred pre-training stage

Index	Φ	Loss	ct_{pos}^{thres}	ct_{neg}^{thres}	ST	γ	iCF	Manga 109	DCM-B	AP Diff.
1	250	OHEM	0.15	0.85	Yes	0.0	49.10	69.21	77.52	0.12
2	500	OHEM	0.15	0.85	Yes	0.0	49.05	**69.32**	77.83	0.09
3	1000	OHEM	0.15	0.85	Yes	0.0	48.48	69.02	**77.93**	0.52
4	Never	OHEM	0.15	0.85	Yes	0.0	47.83	67.68	77.29	1.56
5	500	SimOTA	-	-	Yes	0.0	47.13	65.64	75.42	2.89
6	Never	SimOTA	-	-	Yes	0.0	47.10	65.71	75.48	2.87
7	500	OHEM	0.70	0.30	Yes	0.0	49.14	69.20	77.63	0.10
8	500	OHEM	0.50	0.50	Yes	0.0	<u>49.19</u>	**69.32**	<u>77.90</u>	**0.02**
9	500	OHEM	0.30	0.70	Yes	0.0	49.09	**69.32**	<u>77.90</u>	0.07
10	500	OHEM	0.00	1.00	Yes	0.0	**49.22**	69.20	77.75	<u>0.06</u>
11	500	OHEM	0.15	0.85	No	0.0	41.66	62.64	75.64	7.12
12	500	OHEM	0.15	0.85	Yes	0.9	48.72	65.72	76.59	2.05

Table 1: AP scores of different stage 2 configurations in the largest 3 drawing datasets. Φ is the number of iterations where teacher weights are loaded to student networks afterward, ct_{pos}^{thres} is the minimum confidence threshold for the student network prediction to be counted as positive in ohem loss, ct_{neg}^{thres} is the maximum confidence threshold for the student network prediction to be counted as negative in ohem loss. ST indicates if style transfer is applied in pre-training stage 1, γ is the momentum value that is used in the SGD optimizer (if used, nesterov SGD is utilized). The "AP Diff." column is calculated by averaging the maximum score in each dataset minus the experiment score.

Styles	iCF	Manga 109	DCM-B
Hayao	36.53 $\pm$ 0.77	35.30 $\pm$ 3.66	51.63 $\pm$ 4.48
Shinkai	34.88 $\pm$ 1.26	36.03 $\pm$ 2.33	56.40 $\pm$ 1.85
Hosoda	38.81 $\pm$ 0.40	43.05 $\pm$ 0.96	54.63 $\pm$ 3.13
KH	37.69 $\pm$ 0.62	36.39 $\pm$ 1.43	49.10 $\pm$ 1.41
Whitebox	42.22 $\pm$ 1.49	45.86 $\pm$ 1.93	52.46 $\pm$ 2.23
No Styles	33.00 $\pm$ 1.97	35.57 $\pm$ 2.82	58.94 $\pm$ 3.75
Top-5 Styles	42.04 $\pm$ 1.41	<u>47.90</u> $\pm$ 2.61	59.96 $\pm$ 1.82
No Animals	<u>42.31</u> $\pm$ 0.70	44.81 $\pm$ 1.30	<u>62.85</u> $\pm$ 1.16
All Styles	**42.50** $\pm$ 1.25	**48.73** $\pm$ 2.60	**65.46** $\pm$ 1.35

Table 2: AP scores after stage 1 in the largest 3 drawing datasets.

pushes the performance even further.

4.3 Stage 2: Self-supervised Pre-Training

In this Section, we discuss all of our experiments in the self-supervised stage. We will refer to Table 1 for the additional student network (SN) update interval (Φ), loss selection, positive (c_{pos}^{thold}) and negative (c_{neg}^{thold}) SN confidence thresholds, usage of momentum in the optimizer (γ), and for highlighting the importance of style transferring before the self-supervised stage (ST).

Loss. In experiments 2, 4, 5, and 6, our modified OHEM loss is compared with the SimOTA loss, which is the default loss method in YOLOX and an advanced variation of Focal loss. We believe that selecting a subset of predictions for backpropagation reduces the amount of misleading in FP and FN cases. Our results also validate that OHEM loss is more suitable for our self-supervised architecture. Models with OHEM loss outperform others with up to ~ 2.8 AP difference.

Updating SN per Φ Iterations. Between experiments 1 and 4, we try various iteration counts for Φ. We observe that the overall performance drops if $\Phi > 500$. The score is worst if there is no manual SN update (i.e., $\Phi = None$).

Student Confidence Thresholds (c_{pos}^{thold} and c_{neg}^{thold}). We test the influence of positive and negative SN confidence thresholds in experiments 2 and 7-10. With a threshold starting from too high for positive and too low for negative (exp. 7), the average performance is lower than the others. While the original OHEM loss corresponds to exp. 10, adding additional thresholds for SN results in greater or similar scores (e.g., experiments 8 and 9). The best performance is obtained by setting both c_{pos}^{thold} and c_{neg}^{thold} to 0.5.

Optimizer Selection. Our study states that manually changing SN's weights with TN's may mislead the overall model if an optimizer with momentum is utilized. To test our statement, we train two models with the same hyperparameter configurations but select standard SGD in one and Nesterov SGD in the other (exp. 2 and 12). In almost every dataset, standard SGD scores $\sim 1.5 - 2\%$ higher.

Style Transferring Before Self-supervised Stage 2. We investigate if style transferring is needed in stage 1 before applying self-supervised stage 2. We train two models with the same settings but initialize the pre-trained weights of these models in the teacher-student stage differently: one with the weights retrieved from pre-training stage 1, including style transferring, the other without style transferring (exp. 2 and 11). Overall, AP difference is $\sim 7\%$. Hence, applying style transferring in stage 1 has a significant positive effect on the self-supervised stage 2 model performance.

4.4 Stage 3: Fine-Tuning

We train our architecture for single datasets with limited instances to evaluate their behavior when only a low amount

Types	Models	iCF	M109-F	DCM-F	M109-B	DCM-B	C2k	W2k	C1k
NS	*All Styles*	42.50	54.74	69.93	42.72	65.46	56.80	67.36	55.65
SS	*Teacher-Student*	49.19	69.25	**82.45**	69.38	77.90	67.41	71.53	64.25
SS	UMT [Deng *et al.*, 2021]	-	-	-	-	-	-	69.90	70.50
SS	D-adapt [Jiang *et al.*, 2022]	-	-	-	-	-	53.50	68.90	69.30
WS	[Inoue *et al.*, 2018]	-	-	-	36.71*	41.89*	57.30	73.20	63.00
WS	H^2FA R-CNN [Xu *et al.*, 2022]	-	-	-	-	-	66.80	73.80	75.70
FS	*Train w/ 64 Images* **	65.47	80.41	69.80	77.72	77.28	68.36	71.24	58.74
FS	*Train w/ 256 Images* **	71.24	84.20	73.72	80.79	80.91	69.96	73.83	65.18
FS	*Train w/ 512 Images* **	74.39	85.15	74.85	82.32	82.40	71.05	77.63	-
FS	*Train w/ All Images* **	87.75	87.86	75.87	87.06	84.89	71.66	89.17	77.97
FS	*XL Model w/ All Images* **	90.01	**87.88**	77.40	**87.98**	**86.14**	**73.65**	**89.81**	**83.59**
FS	ACFD [Zhang *et al.*, 2020a]	**90.94**	-	-	-	-	-	-	-
FS	[Ogawa *et al.*, 2018]	-	76.20	-	79.60	-	-	-	-
FS	[Nguyen *et al.*, 2018]	-	-	74.94	-	76.76	-	-	-
FS	[Inoue *et al.*, 2018]	-	-	-	-	-	70.10	77.30	76.20

Table 3: Overall AP performances of our models and previous SOTA models. Our models are titled in *italic*. The teacher-student network is initialized with the style transferred pre-training, All of our supervised models are initialized with pre-training stage 2 weights. NS: no target domain supervision. SS: self-supervision, WS: weak-supervision, FS: full target domain supervision. Scores with "*" mean that they are evaluated by us using the model from the original project repository. "**" indicates that the results are retrieved from single-dataset trainings and each score is calculated by a separate model trained specifically with the particular dataset.

Pre-training	Image Instance Counts 64	512	All
None	47.79 $\pm$ 1.38	69.38 $\pm$ 0.82	80.60 $\pm$ 0.65
Stage 1	66.90 $\pm$ 1.40	75.34 $\pm$ 0.99	**82.87** $\pm$ 1.53
Stage 1 + 2	**71.13** $\pm$ 0.92	**77.44** $\pm$ 0.47	82.78 $\pm$ 0.93

Table 4: Average AP performance of our model when trained with a subset of individual datasets having annotations of a limited number of random images. Average is calculated by taking the mean of each score retrieved from each 6 datasets.

of data is available. The average scores for all datasets we evaluated are shared in Table 4. In the cases with extremely low instance counts (i.e., 64 and 512 images), utilization of natural images and self-supervised learning results in up to $\sim 24\%$ performance increase compared to starting from a random initial state. When trained with all available data, both style-transferring-based and teacher-student-based pre-training methods score similar values. We believe this is caused since there is sufficient data for these specific sub-domains to close the gap that emerged from the self-supervised stage. However, we still obtain a significant improvement ($\sim 2.2\%$) when models start from pre-trained weights instead of random initialization. This shows that leveraging style transferred pre-training enhances the performance independently from the amount of labeled fine-tuning data.

In Table 3, we compare previous SOTA models with our results from each stage checkpoint (i.e., style transfer, teacher-student, fine-tuning with individual datasets). Our model achieves close scores to ACFD [Zhang *et al.*, 2020a] and outperforms other SOTA models. Even with a low amount of training images, we obtain better or comparable results with [Nguyen *et al.*, 2018] and [Ogawa *et al.*, 2018]. Increasing the model size from the tiny version of YOLOX to the XL version also results in a further performance increase. *Our XL model* dominates our tiny version in each individual dataset.

5 Conclusion

In this study, we work on efficient pre-training for face and body detection models in drawings. First of all, we introduce a self-supervised teacher-student network to the domain of drawings. We propose a modified OHEM loss to overcome the false-negative cases caused by the teacher network and equalize the student network's weights to the teacher network's per 500 iterations to prevent distortions in the student network. By leveraging the existing style-transferring methods, we highlight the importance of using pre-trained weights for the domain adaptation task and the positive effects of using style-transfer on the pre-training data. Additionally, we show that using multiple style-transferring variations together provides higher performance. Lastly, we train fully supervised models with limited and available labeled data. Our model obtains the new SOTA score in most drawing datasets when pre-trained with our pipeline. This finding indicates that efficient pre-training is important where a low amount of data is available, and the teacher-student network is an effective way of pre-training.

6 Acknowledgements

This project is supported by Koç University & İş Bank AI Center (KUIS AI). We would like to thank KUIS AI for its support.

References

[Bochkovskiy *et al.*, 2020] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *ArXiv*, abs/2004.10934, 2020.

[Cai *et al.*, 2019] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11449–11458, 2019.

[Chen *et al.*, 2018] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. Cartoongan: Generative adversarial networks for photo cartoonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9465–9474, 2018.

[Deng *et al.*, 2021] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4089–4099, 2021.

[Ge *et al.*, 2021] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[Hicsonmez *et al.*, 2020] Samet Hicsonmez, Nermin Samet, Emre Akbas, and Pinar Duygulu. GANILLA: generative adversarial networks for image to illustration translation. *CoRR*, abs/2002.05638, 2020.

[Inoue *et al.*, 2018] Naoto Inoue, Ryosuke Furuta, T. Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5009, 2018.

[Iyyer *et al.*, 2016] Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan L. Boyd-Graber, Hal Daumé, and Larry S. Davis. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6478–6487, 2016.

[Jiang *et al.*, 2022] Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long. Decoupled adaptation for cross-domain object detection. In *International Conference on Learning Representations*, 2022.

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

[Liu *et al.*, 2021] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[Matsui *et al.*, 2017] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017.

[Nguyen *et al.*, 2018] Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. Digital comics image indexing based on deep learning. *Journal of Imaging*, 4(7), 2018.

[Ogawa *et al.*, 2018] Toru Ogawa, Atsushi Otsubo, Rei Narita, Yusuke Matsui, T. Yamasaki, and Kiyoharu Aizawa. Object detection for comics using manga109 annotations. *ArXiv*, abs/1803.08670, 2018.

[Shrivastava *et al.*, 2016] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[Wang and Yu, 2020] Xinrui Wang and Jinze Yu. Learning to cartoonize using white-box cartoon representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[Xu *et al.*, 2021] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[Xu *et al.*, 2022] Yunqiu Xu, Yifan Sun, Zongxin Yang, Jiaxu Miao, and Yi Yang. H^2FA R-CNN: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14329–14339, 2022.

[Yang *et al.*, 2016] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[Zhang *et al.*, 2020a] Bin Zhang, Jian Li, Yabiao Wang, Zhipeng Cui, Yili Xia, Chengjie Wang, Jilin Li, and Feiyue Huang. Acfd: Asymmetric cartoon face detector. *ArXiv*, abs/2007.00899, 2020.

[Zhang *et al.*, 2020b] Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. Dynamic R-CNN: Towards high quality object detection via dynamic training. In *ECCV*, 2020.

[Zheng *et al.*, 2020] Yi Zheng, Yifan Zhao, Mengyuan Ren, He Yan, Xiangju Lu, Junhui Liu, and Jia Li. Cartoon face recognition: A benchmark dataset. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2264–2272, 2020.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.