International Journal of Computer Vision
https://doi.org/10.1007/s11263-023-01927-2

Learning Portrait Drawing with Unsupervised Parts

Burak Tasdemir[1] · Mustafa Goktan Gudukbay[2] · Dogac Eldenk[1] · Adil Meric[3] · Aysegul Dundar[1]

Received: 3 November 2022 / Accepted: 9 October 2023

Abstract
Translating face photos into portrait drawings takes hours for a skilled artist which makes automatic generation of them desirable. Portrait drawing is a difficult image translation task with its own unique challenges. It requires emphasizing important key features of faces as well as ignoring many details of them. Therefore, an image translator should have the capacity to detect facial features and output images with the selected content of the photo preserved. In this work, we propose a method for portrait drawing that only learns from unpaired data with no additional labels. Our method via unsupervised feature learning shows good domain generalization behavior. Our first contribution is an image translation architecture that combines the high-level understanding of images with unsupervised parts and the identity preservation behavior of shallow networks. Our second contribution is a novel asymmetric pose-based cycle consistency loss. This loss relaxes the constraint on the cycle consistency loss which requires an input image to be reconstructed after transformations to a portrait and back to the input image. However, going from an RGB image to a portrait, information loss is expected (e.g. colors, background). This is what cycle consistency constraint tries to prevent and when applied to this scenario, results in learning a translation network that embeds the overall information of RGB images into portraits and causes artifacts in portrait images. Our proposed loss solves this issue. Lastly, we run extensive experiments both on in-domain and out-of-domain images and compare our method with state-of-the-art approaches. We show significant improvements both quantitatively and qualitatively on three datasets.

Keywords Portrait drawing · Unsupervised part segmentations · Unpaired image translation · Cycle consistency adversarial networks

1 Introduction

Image-to-image translation has been a popular topic and significant progress has been achieved (Isola et al., 2017; Park et al., 2019; Yi et al., 2019; Dundar et al., 2020; Liu & Tuzel, 2016; Yi et al., 2017; Liu et al., 2017; Mardani et al., 2020; Liu et al., 2022), especially in the area of translating face photos (Choi et al., 2020; Shen & Liu, 2017; Xiao et al., 2018; Zhang et al., 2018; Li et al., 2021; Huang et al., 2021; Yang et al., 2022; Dalva et al., 2022). Image translation of face photos into portrait drawings is one of them which requires detecting semantic parts of a face so that the drawing can exhibit abstraction of the face with an artistic touch.

Portrait drawing methods should capture the feelings of the person as well as the key facial features that define the identity of the person. On the other hand, irrelevant details in the input images (e.g. rich background, high-frequency details, color information) should not be transferred to the portrait. Such a drawing requires experience and skill where artists use a unique set of lines to represent the subject. Since portrait drawing requires hours of a skilled artist, automatic generation of them is desirable (Yi et al., 2019, 2020b, a).

The abstractness of portrait generation requires detecting important key features in the input image to translate and ignoring the rest of the features. These requirements make this task more challenging than the other image translation tasks, e.g. changing the painting style. Because of that neural style transfer networks (Gatys et al., 2016) and cycle consistent adversarial networks (cycleGAN) (Zhu et al., 2017) which are pretty successful at style transfer, fail on this task (Yi et al., 2020a). Previous methods for portrait generation require either paired data (Yi et al., 2019, 2020b) or part seg-

Communicated by Oliver Zendel.

✉ Aysegul Dundar
adundar@cs.bilkent.edu.tr

[1] Computer Science, Bilkent University, Ankara, Turkey

[2] Electrical Engineering and Computer Science, Penn State University, Pennsylvania, USA

[3] Computer Science, Technical University of Munich, Munich, Germany

Published online: 01 November 2023

mentations of faces and portraits to train local discriminators (Yi et al., 2019, 2020b, a). In this work, we are interested in learning a portrait generation method in a completely unsupervised way with unpaired data and no annotation of images.

We base our method on cycle-consistent adversarial network (Zhu et al., 2017). These methods use cycle consistency by minimizing the reconstruction loss between input images and images that are translated twice, translated to a different domain, and then back to their original domain. This regularizes the network to preserve the content of the input image. However, that is not enough to translate images with the content preserved since deep networks have the capacity to encode the details of the input image content in a non-recognizable way for humans in the translated images and achieve cycle consistency without translated images following the correct content. For example, in the portrait generation task, with cycle consistency training, the network can take an input image with eyeglasses and translate it to a portrait without glasses, and translate it back to the input image with eyeglasses. Even though visually it may seem that the information that carries the eyeglasses is erased when the translation happens, the network finds a way to encode such information and can translate the portrait back to the original image. This will result in satisfying the cycle-consistency constraint while changing the content. Therefore, researchers adopt another completing ingredient for content preservation which is to use shallow network architectures (Zhu et al., 2017; Liu et al., 2017; Huang et al., 2018; Yi et al., 2020a; Choi et al., 2018). Shallow networks cannot make drastic changes to the input images and this helps to preserve the identity of the photos. However, for high-quality portrait generation, the network should have the capacity to differentiate key facial features from irrelevant features. To achieve that we combine the advances in unsupervised part detection (Jakab et al., 2018; Lorenz et al., 2019; Dundar et al., 2021) with the cycle consistent adversarial networks.

Another challenge of learning portrait drawing is that while cycle consistency loss encourages content preservation, it also forces the portrait image to encode all the information of the face photo. However, portrait drawings should carry less information than their face photo counterparts since colors and many details are supposed to be removed. To relax this constraint, Yi et. al. (Yi et al., 2020b) propose a relaxed cycle consistency loss where cycle consistency is calculated between the edge maps of input photo images and reconstructed ones. Even though this removes color information, it still requires the portrait image to include many high-frequency details (all the edges). To solve this problem, we propose a novel asymmetric pose-based cycle consistency loss and with that, we encourage alignment between the high-level features of faces in photos and portraits.

Our contributions are as follows:

1. We propose a method for portrait drawing that only learns from unpaired data with no additional labels. Previous methods either require paired data or part labels.
2. We propose an image translation architecture that combines the high-level understanding of images with unsupervised parts and identity preservation behavior of shallow networks.
3. We propose a pose-based cycle consistency loss. This loss does not force the network to preserve the color and background information while converting an image to a portrait.
4. We extensively run experiments both on in-domain and out-of-domain images and compare our method with the state-of-the-art. We show significant improvements both quantitatively and qualitatively.

2 Related Work

Image to Image Translation Methods Image translation methods powered with adversarial losses have shown successful applications in image resolution (Yuan et al., 2018), dehazing (Engin et al., 2018), texture expansion (Mardani et al., 2020), and 3D inference (Bhattad et al., 2021; Dundar et al., 2022) to name a few. Many of those image translation methods require paired data (Isola et al., 2017; Wang et al., 2018; Park et al., 2019; Dundar et al., 2020; Mardani et al., 2020). However, for many tasks, paired training data may not be available. For example, one may want to learn a mapping function that can convert a photograph into a painting. Whereas it is not easy to capture a photograph that has the same content and layout as a painting, it is easy to collect unpaired images of photographs and paintings. For this reason, researchers propose to train image translation methods with unpaired data (Liu et al., 2017; Yi et al., 2017; Zhu et al., 2017) which are referred to as unsupervised image-to-image translation methods. This problem is inherently ill-posed since there is no guidance for one-to-one mapping. Therefore, researchers propose various regularizers and structures to the image translator training to achieve plausible results. Regularizers include shared-latent space assumption (Liu et al., 2017), weight-sharing (Liu & Tuzel, 2016), contrastive learning (Park et al., 2020), and most popularly cycle consistency constraint (Zhu et al., 2017; Yi et al., 2017; Liu et al., 2017).

For cycle consistency constraint, two image translators are trained simultaneously between two domains. If the image translators preserve the underlying representation, it should be possible to map an image from one domain to another and map it back to the original domain and to the original image. In terms of loss function, two reconstruction losses that force cycle consistency are added to the training. These reconstruction losses are implemented as either pixel-level L2 loss

or higher level VGG loss. The translation methods trained with cycle consistency achieve successful image translation on many domains (Zhu et al., 2017; Yi et al., 2017; Liu et al., 2017; Kim et al., 2020; Zheng et al., 2020). However, portrait drawing is more challenging than many other image translation tasks since it requires high abstraction and expressiveness. Because of that, researchers propose methods that are specialized for portrait drawing. We review those next.

Portrait Drawing Methods Portrait drawing of face photos is an image translation topic. APDrawingGAN (Yi et al., 2019, 2020b) is one of the first image translation methods that is specialized in artistic portrait generation. APDrawingGAN sets an encoder-decoder architecture and trains a model with adversarial and image reconstruction losses. Image reconstruction losses are possible since the model is trained on paired data. However, finding paired data for portrait generation is difficult and because of that, there is extensive interest in methods that can learn from unpaired data. Later on, Yi et al. (2020a, 2022) proposes an architecture that is able to learn from unpaired training data. Their method is based on cycle constraint but an asymmetric one; when a portrait is translated to a photograph and translated back to a portrait, a cycle consistency is applied such that the input and output portraits are the same. Whereas when a photograph is translated to a portrait and then back to a photograph, since the portrait translation lost details, a one-to-one reconstruction between the first and last photograph is not expected. Instead, a cycle consistency between the edge maps is used to guide the networks. This relaxes the cycle consistency constraint but still requires portrait translation to keep all the edge information. However, realistic portraits do not have crowded backgrounds and only key features of faces are translated while many details are ignored for abstraction. Therefore, the relaxed cycle consistency on the edges does not solve the issue completely. All these models that are proposed for portrait translation use local discriminators to preserve the facial structure of the drawings and these local discriminators require part labels of faces. Without local discriminators (e.g. a discriminator for nose, hair), the generated images contain defects around the eyes, nose, and mouth which make the portraits look unrealistic. Previous works relying on different annotations are given in Table 1. In addition to these annotations, UnpairPortrait++ (Yi et al., 2022) also requires human annotation to rank the quality of portraits and obtain a quality metric for training.

Different from these works, our work, to the best of our knowledge, is the first one that achieves portrait translation without requiring any annotated part labels. Our image translation architecture is different from previous methods as we combine the high-level understanding of images with unsupervised part maps and identity preservation behavior of shallow networks. Also different from previous works, we use a novel asymmetric pose-based cycle consistency loss. With these contributions, we achieve significantly better results than the previous works both on in-domain and out-of-domain images.

Table 1 Summary of methods that require supervision in the form of paired data or part labels. Our proposed method does neither require paired data nor part labels

Methods	Paired data	Part labels
APDrawingGAN (Yi et al., 2019)	✓	✓
APDrawGAN++ (Yi et al., 2020b)	✓	✓
UnpairPortrait (Yi et al., 2020a)	✗	✓
UnpairPortrait++ (Yi et al., 2022)	✗	✓
Ours	✗	✗

3 Method

Our method sets a carefully designed image translation architecture for this task. The image translator should have the capacity for high-level understanding of facial features and additionally should not have so much freedom to change the content completely. By freedom here, we mean it should not be a deep architecture as we show in our ablation study, they change the content drastically. For this design, our method consists of learning an unsupervised part detector and an image translator which are explained next.

Unsupervised Part Detector. Recently, there has been significant success in unsupervised part and landmark learning research via image reconstructions (Jakab et al., 2018; Lorenz et al., 2019; Dundar et al., 2021; Sardari et al., 2021; Hung et al., 2019). To learn an unsupervised part detector, also referred to as pose encoder, we set an encoder-decoder pipeline as shown in Fig. 1. Encoders' job is to encode poses (parts) and appearances and the job of the decoder is to reconstruct the input image from those information. Before feeding images to the pose and appearance encoders, we perturb the input images. For the pose encoder, by color jittering, we obtain appearance-transformed image $T_{cj}(x)$. For shape transformation, we use thin-plate-spline warping to create synthetic pose-changes $T_{tps}(x)$. The pipeline is forced to learn pose and appearance because while trying to reconstruct the input, neither the pose encoder nor the appearance encoder is ever given direct access to the original image x. Not enabling direct access to the original image is the first information bottleneck that encourages learning parts and learning disentangled pose and appearance features.

Following Lorenz et al. (2019), the pose encoder is executed on both the pose-perturbed and color-jittered input images to map the localized appearance information. Note that both pose and appearance encoders have the encoder-decoder architecture. Via the convolution blocks that down-

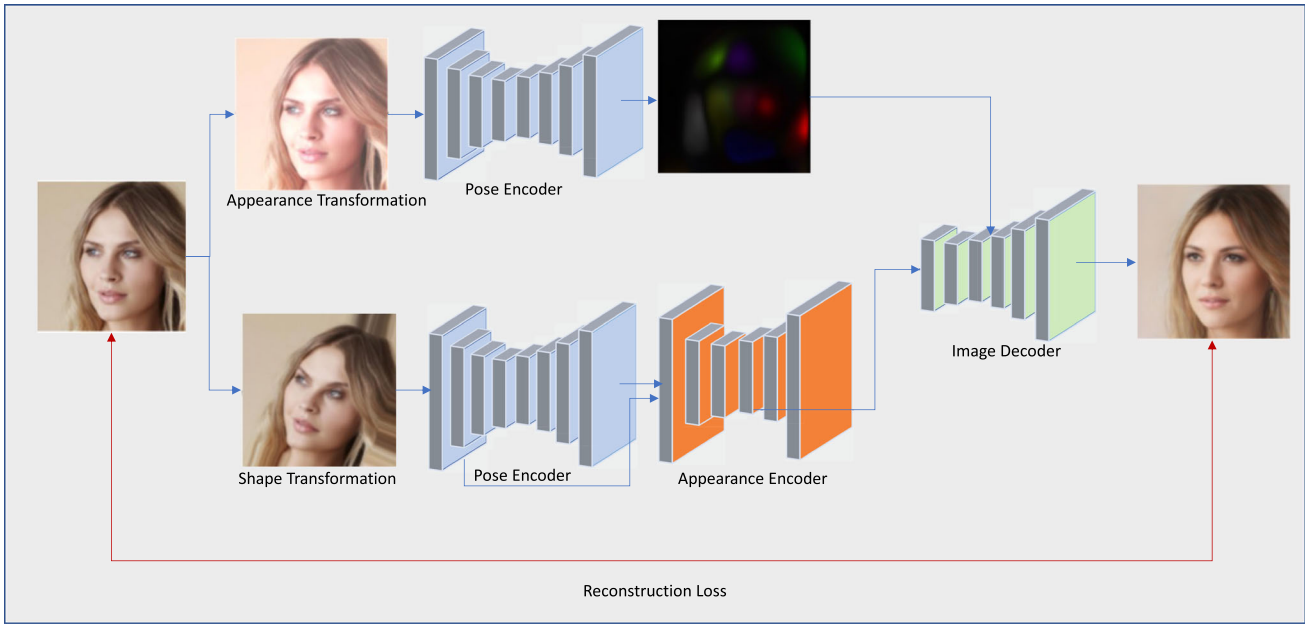

Fig. 1 Unsupervised part segmentation learning pipeline (Pose encoder). To learn a pose encoder, we learn an appearance encoder and an image decoder jointly to enable training. To achieve pose and appearance disentanglement, we perturb the input image before feeding it to the encoders. For the pose encoder, by color jittering, we obtain an appearance-transformed image while preserving the pose. For shape transformation, we use thin-plate-spline warping to create synthetic pose-changes while preserving the appearance. The pipeline is forced to learn pose and appearance because while trying to reconstruct the input, neither the pose encoder nor the appearance encoder is ever given direct access to the original image. The overall framework is trained with image reconstruction loss. Network architectures are taken from Dundar et al. (2021)

sample the features, our encoders increase their receptive field to have a more global view of the image. However, in the end, we want poses to be encoded as heatmaps that are in the same resolution as the input image to provide us with segmentation parts of objects. Therefore, we add convolution blocks with upsampling layers to reach to the same resolution as the input. We also want appearance information to be encoded to the same dimension so that we can pool the appearance features based on the predicted parts which will be explained next, more in depth. The details of the architectures are given in Sect. 4.

The other bottleneck that encourages learning parts is the Gaussian bottleneck. We fit the encoded pose into a Gaussian distribution in order to capture the key facial features. Following Lorenz et al. (2019), we fit a 2D Gaussian to each activation of the K activation maps by computing the mean over activation locations and using either an estimated or predefined covariance matrix. Each part is then written as: $\widetilde{\Phi}_k^{pose} = (\mu_k, \Sigma_k)$, where $\mu_k \in \mathbb{R}^2$ and $\Sigma_k \in \mathbb{R}^{2\times 2}$. The 2D Gaussian approximation forces each part activation map into a unimodal representation. This results in limiting the information flow from the encoder to the decoder. This bottleneck results in the keypoint-like or part-segmentation-like interpretation that each part appears in at most one location per image. We learn K number of distinct heatmaps after this process which corresponds to K number of parts.

The appearance encoder $\Phi^{app} = Enc^{app}(x; Enc^{pose}(x))$ extracts local appearance information. The encoded appearances should be projected to the correct locations to reconstruct the final image. This happens as follows; given an input image x, the pose encoder first provides $K \times H \times W$ part activation maps Φ^{pose}. The appearance encoder then projects the image to a $C \times H \times W$ appearance feature map M^{app}. Using the pose activation map to compute a weighted sum over the appearance feature map, we extract the reduced appearance vector for the kth part as:

$$\Phi_{k,c}^{app} = \sum_{i}^{H}\sum_{j}^{W} \Phi_{k,i,j}^{pose} M_{c,i,j}^{app} \text{ for } c = 1...C, \tag{1}$$

giving us K C-dimensional appearance vectors. Here, each activation map in Φ^{pose} is softmax-normalized.

The appearance vectors and encoded poses are fed to the image decoder. Our training procedure as depicted in Fig. 1 can be expressed as follows:

$$\Phi_{cj}^{pose} = Enc^{pose}(T_{cj}(x)) \tag{2}$$

$$\Phi^{app} = Enc^{app}(T_{tps}(x); Enc^{pose}(T_{tps}(x))) \tag{3}$$

$$\tilde{x} = Dec(\widetilde{\Phi}_{cj}^{pose}, \Phi^{app}). \tag{4}$$

Fig. 2 Examples of unsupervised part activation heatmaps. Here, we use a different color for each channel of the pose encoder. As can be seen, parts are specialized to detect certain parts (e.g. forehead, eye, nose) in face photos to achieve reconstruction and they are consistent among examples

where Enc^{pose}, Enc^{app} correspond to pose and appearance encoders, respectively. The image decoder, Dec takes the encoded pose and appearance codes and tries to reconstruct the original image. Our reconstruction loss between x and $\tilde{x}$ is a VGG perceptual-loss (Zhang et al., 2018) same as the previous works (Jakab et al., 2018; Dundar et al., 2021) with pre-trained ImageNet weights. We take the network architectures from Dundar et al. (2021) and the parameters are given in Sect. 4. Some examples of the part detections are shown in Fig. 2. The parts learn to track visually-consistent facial features across warps since parts are suited to represent them for reconstruction. As shown in Fig. 2, heatmap activations are consistent across examples.

We use the pose encoder in the image translator to provide the network with a high-level understanding of our input photos. Additionally, the same pose encoder is used for pose-based cycle consistency loss. Both are described in the next subsection. Not that the appearance encoder and image decoder from Fig. 1 are only trained to enable learning of the pose encoder via image reconstruction loss and discarded in the rest of the framework.

Image Translator for Portrait Drawing. For our image translator, we use the previously described pose encoder that has a deep architecture that can provide high-level understanding and a translation architecture that has a shallow architecture that can prevent drastic changes to the content as shown in Fig. 3. We use the pose encoder with frozen parameters in this pipeline and learn the parameters of the translation architecture. In our ablation study, we show why the image translator architecture should not be deep as it will not preserve the content.

In our setting, the image translation architecture has 2 downsampling and 2 upsampling layers. Before each downsampling and upsampling layer, there is a convolution block that consists of convolution, relu, and instance normalization layers. After 2 consecutive downsampling blocks, there are 9 convolutional blocks and then 2 consecutive upsampling blocks. Since there are only 2 downsampling layers, the receptive field of the architecture is small. On the other hand, the pose encoder has 4 downsampling and upsampling blocks which provides the network with a larger receptive field and a higher understanding of the content. Additional architectural details are provided in Sect. 4. We provide this high level of understanding extracted from images to our translator by concatenating the encoded poses with the input image and feeding them to the image translation network as shown in Fig. 3. We do not tune the pose encoder during image translation training.

For training the translation architecture, we use cycle consistent adversarial network training pipeline. Specifically, we follow the asymmetric CycleGAN training (Yi et al., 2020a) as our baseline. We set two image translators; G_p to translate face photos into portrait photos and G_r to translate portrait photos into face photos. These two translators are trained jointly with each having a discriminator, D_p and D_r, to guide the networks for translating images to correct domains. Models are trained with adversarial and cycle consistency losses. The adversarial losses are given as follows where a discriminator and a generator play a min-max game:

$$\begin{aligned} L_{adv} = &\min_{G_p}\max_{D_p} \lambda_{adv}\, \ell_{adv}(G_p, D_p) \\ &+ \min_{G_r}\max_{D_r} \lambda_{adv}\, \ell_{adv}(G_r, D_r) \end{aligned} \tag{5}$$

Additionally, we use a cycle consistency loss between portrait photos as follows:

$$L_{cyc} = ||x_p - G_p(G_r(x_p))|| \tag{6}$$

where x_p is a portrait drawing. This loss regularizes the network so that a portrait image when translated to a face photo with G_r can be translated back again with G_p. The final translated image should be the same as x_p, the very input image.

On the other hand, the same cycle consistency from face image to portrait to face image is too demanding for the task since when a face image is translated to portrait, there is a considerable loss of information. Because of this reason, Yi et al. (2020a) propose to use a relaxed cycle consistency as given in Eq. 7.

$$L_{rel-cyc} = ||E(x_r) - E(G_r(G_p(x_r)))|| \tag{7}$$

E is a deep learning based edge detector (Xie & Tu, 2015). The relaxed cycle consistency is calculated by only applying reconstruction loss between the edge information of the input and output RGB images since translating an image to portrait results in removing color information from the input image. In our first experiment, we use this loss as well. However, this relaxed cycle consistency still requires translated portrait

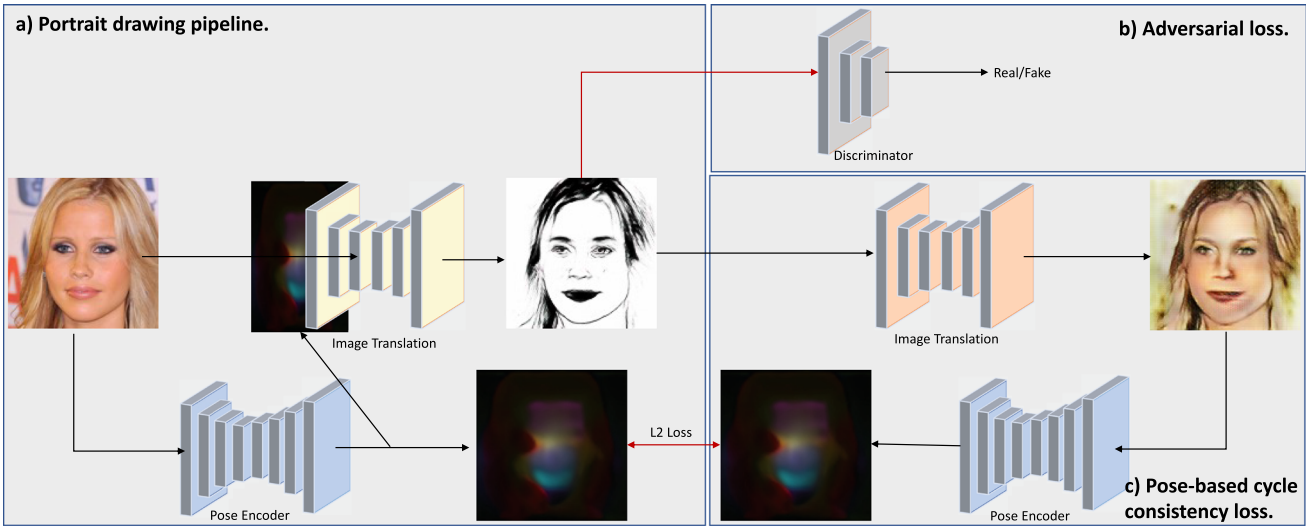

Fig. 3 Portrait drawing pipeline takes RGB image and outputs a portrait image (**a**). It includes a pose encoder and a shallow translation architecture. Pose encoder has a deep architecture that can provide a high-level understanding. It is trained with the pipeline given in Fig. 1 and its parameters are kept frozen in this pipeline. The image translation network is shallow which prevents drastic changes of the content. Encoded poses are concatenated with the input image and are fed into the image translation network. The portrait drawing pipeline is guided with adversarial (**b**) and pose-based cycle consistency losses (**c**). For adversarial guidance, we train a discriminator which learns to discriminate generated portraits from real portraits. For pose-based cycle consistency loss, the L2 loss is calculated between the encoded poses of input and the reconstructed image. This provides a relaxed cycle consistency constraint such that from a portrait, the original input image is not expected to be reconstructed perfectly since going from an RGB image to a portrait, information loss is expected (e.g. colors, background). Losses are highlighted with red arrows. We still expect the content of the input image to be preserved and achieve that with pose-based cycle consistency loss

images to preserve all the edge details from the input photo which can cause portrait images to be noisy with a lot of details.

Pose-based Cycle Consistency Loss. We propose to replace the relaxed cycle consistency introduced by Yi et al. (2020a) with pose cycle consistency loss as given in Eq. 8.

$$L_{pose-cyc} = ||Enc^{pose}(x_r) - Enc^{pose}(G_r(G_p(x_r)))|| \quad (8)$$

We use the same pose encoder that we use in the image translation architecture. With this loss which is calculated across the encoded features from the pose encoder, the translation from photos to portraits does not need to preserve edge details. Instead, it is encouraged to preserve the facial pose details. This can be seen in Fig. 3 where portraits are not forced to encode the writings on the background which would make the portrait very noisy. Since we do not force the output image to match the input image at a pixel level, the reconstructed image $G_r(G_p(x_r))$ does not look realistic but it is not important because our final goal is to obtain an accurate G_p and not G_r. We only learn G_r to facilitate cycle-consistent adversarial network training.

Combining all of the loss components described, we reach the overall objective for optimization as given in Eq. 9.

$$\min_{G_p,G_r} \max_{D_p,D_r} \lambda_a \mathcal{L}_{adv} + \lambda_c \mathcal{L}_{cyc} + \lambda_r \mathcal{L}_{pose-cyc} \quad (9)$$

with $\lambda_a = 0.5$, $\lambda_c = \lambda_r = 5$. We use these hyperparameters following Yi et al. (2020a). We do not tune the hyperparameters and obtain improvements with our proposed modules without any tuning. In our experiment provided in Ablation Study, first, we only update the translator and keep using the loss defined as $\mathcal{L}_{rel-cyc}$. In our final setting, we replace that relaxed loss with $\mathcal{L}_{pose-cyc}$.

4 Experiments

Set-up. We use the portrait drawing images of APDrawing set (Yi et al., 2019). There are 140 portrait images and we use them for training. For RGB images, we use CelebA training images (Liu et al., 2015) which includes 27000 images. Example images of these datasets are shown in Fig. 4. This setting is referred to as training with unpaired data and is considered to be very challenging (Liu et al., 2017; Yi et al., 2020a).

We report Frechet Inception Distance (FID) (Heusel et al., 2017) metric which looks at the realism and diversity by comparing the real and generated data distributions. To calculate FID scores, we use an Inception-v3 network (Szegedy et al., 2016) that is pre-trained on the ImageNet dataset (Deng et al., 2009). Inception features are extracted from the real and generated datasets via the Inception-v3 network

Fig. 4 Example samples from our training dataset. For portrait drawing images, we use APDrawing set (Yi et al., 2019) and for RGB images, we use CelebA training images (Liu et al., 2015). This setting is referred to as training with unpaired data

and multi-variate Gaussian is fitted to compute the mean and the covariance matrix of these feature sets. Frechet distance is calculated by the mean and covariance matrix of features extracted from real and generated sets. The smaller distance corresponds to a better match of real and generated data distributions.

We calculate FID between the translated CelebA validation images and portrait images. We refer to this evaluation as in-domain results since the training and validation images even though are different come from the same dataset. We also run evaluations on two challenging datasets, Metface dataset (Karras et al., 2020) and Fantasy image dataset (Yang et al., 2022). They are out-of-the-domain images since they exhibit domain gaps with the CelebA training set. These sets are not used during training and they test the generalization of our method to other domains.

Network Architectures and Training Parameters. We train our model with an input size of 128 × 128. Our image translation architecture has 2 downsampling convolutional blocks. Each block contains convolution, relu, and instance normalization layers. The channels of convolutional layers of these two blocks are 64 and 128, respectively, kernel size is 3, padding is 1, and downsampling is achieved by a stride of 2. After 2 consecutive downsampling blocks, there are 9 residual convolutional blocks. In these residual blocks, there are skip connections from input to output at each convolutional layer in the form of summation. Their channel sizes are 256, kernel sizes are 3, and padding sizes are set to 1. After that, there are two transposed convolution blocks that bring the features in the same spatial dimension as the input image. The channels size are 128 and 64 and their strides are set to 2. We concatenate the input image and encoded pose and feed them to this image translator. Therefore, the input channel of the first layer is 13. The first three channels are RGB images and the rest 10 channels are unsupervised heatmaps. The output channel of the last layer of the image translator is set to 3 to output an image. We set the number of landmarks to 10 following previous works (Lorenz et al., 2019; Dundar et al., 2021).

For the pose encoding, we train the pose encoder together with the appearance encoder and image decoder to enable the training with reconstruction loss as shown in Fig. 1. We use U-net architectures (Ronneberger et al., 2015) for the pose encoder and appearance encoders complete with skip connections. The pose encoder has 4 blocks of convolutional downsampling (strided conv) modules. Each convolutional downsampling module has a convolution layer-Instance Normalization-ReLU and a downsampling layer. At each block, the number of filters doubles, 64, 128, 256, 512. The upsampling portion of the pose encoder has 3 blocks of convolutional upsampling, and the number of channels is halved at every block, 512, 256, 128. The appearance encoder network has one convolutional downsampling and one convolutional upsampling module. The image decoder has 4 convolution-ReLU-upsample modules. We first downsample the appearance feature map by a factor of 16 in each spatial dimension. The number of output channels for each convolution-ReLU-upsampling module in the image decoder is 256, 256, 128, 64, and 3 respectively. Each layer also has an adaptive instance normalization layer. We also provide the output of encoded pose to the decoder via those adaptive normalization layers similar to SPADE architecture (Park et al., 2019). The encoded poses are scaled to match the size of the features at each layer of the decoder.

To train the image translation and pose encoding networks, we use an Adam optimizer with a learning rate of 0.0002. We train our model for 20 epochs on a single GPU with a batch size of 16.

Baselines. We set state-of-the-art baselines to compare our method with. We compare with CycleGAN (Zhu et al., 2017), Asymmetric CycleGAN (Yi et al., 2020a), Image2StyleGAN (Abdal et al., 2019), and DualStyleGAN (Yang et al., 2022). We train CycleGAN and Asymmetric CycleGAN on our data. CycleGAN is proposed for general image translation tasks and Asymmetric CycleGAN is proposed specifically for portrait generation. Image2StyleGAN is proposed to embed a face image to StyleGAN latent space and further shows a style transfer application by latent swapping; transferring a style between an embedded stylized image and other face images. We follow their proposed setup and replace the latent code of the last 9 layers of the base image with the embedded portrait image (style). The latent codes are estimated with IdInvert (Zhu et al., 2020). We also train DualStyleGAN on our portrait data which performs style transfer via a pre-trained StyleGAN model. DualStyleGAN additionally includes trainable style paths that feed features to the pre-trained StyleGAN architecture. Face photos and portraits are encoded to the latent space by a pSp encoder (Richardson et al., 2021). An encoded face photo is used as content and an encoded portrait photo is used as style. We follow the training

Table 2 Quantitative results of our and competing methods on CelebA (Liu et al., 2015), MetFace (Karras et al., 2020), and Fantasy datasets (Yang et al., 2022)

Methods	CelebA	MetFace	Fantasy
CycleGAN Zhu et al. (2017)	70.57	99.28	105.84
Asymmetric CycleGAN Yi et al. (2020a)	70.40	141.11	131.97
Image2StyleGAN Abdal et al. (2019)	93.35	250.53	136.59
DualStyleGAN Yang et al. (2022)	<u>67.06</u>	<u>92.07</u>	<u>94.79</u>
Ours	**62.33**	**88.57**	**89.16**

We highlight the best results in bold and the second-best results with underline. Our method achieves better scores than competing methods on all datasets

Fig. 5 Qualitative comparison on CelebA validation dataset (Liu et al., 2015). We compare our method with CycleGAN (Zhu et al., 2017), Asymmetric CycleGAN (Yi et al., 2020a), Image2StyleGAN (Abdal et al., 2019), and DualStyleGAN (Yang et al., 2022). All methods are trained using the same training dataset. DualStyleGAN suffers from semantic alignment, whereas CycleGAN, Asymmetric CycleGAN, and Img2StyleGAN suffer from artifacts on the faces. Our method achieves significantly better portraits

steps given in the author's released code. During translation, for each content image, we randomly sample from the portrait training images to generate styles.

Results. We show quantitative and qualitative results of our model and competing models in Table 2 and Fig. 5, respectively, for the CelebA validation dataset. Since models are trained on CelebA training images, these experiments represent the in-domain results. CycleGAN and Asymmetric CycleGAN both achieve reasonable output quality where the results preserve the content. However, the portrait results also contain a lot of irrelevant edges from the background making the portraits look noisy. Additionally, there are many artifacts around the facial features for both models while CycleGAN has more than Asymmetric CycleGAN. Image2StyleGAN even though is not proposed for portrait generation showcases applications that include style transfer with portraits. It is impressive that portraits can be translated with StyleGAN without any training, even though the results are not on par with the image translation methods that are trained end to end for this task. DualStyleGAN takes the Img2StyleGAN approach for portrait generation one step ahead and trains an intrinsic style path of StyleGAN and benefits from StyleGAN's ability to synthesize high-quality face images. DualStyleGAN shows very successful results for anime, cartoon, and caricature translations. On the other hand, for portrait drawing, the results again suffer from semantic alignment. Such weak alignment presents no issue when the style transfer is from a face photo to a cartoon or anime image since they are not expected to be aligned with facial features in detail. On the other hand, portrait generation requires a high resemblance to the input image and this makes portrait generation a difficult task where DualStyleGAN is not as quite successful.

Our results achieve the best FID score among the models that preserve the content as given in Table 2. Our model improves FID from 70.40 (Asymmetric CycleGAN) to 62.33. Visually, the results are significantly better than the others. The portraits are semantically aligned with their input images. The focus is on the face portraits and not on the irrelevant features such as the background.

Domain Generalization Results. Deep neural networks show reduced performance when trained on one domain (e.g. synthetic images) and tested on a different domain (e.g. real images) (Dundar et al., 2020; Zou et al., 2020; Shyam et al., 2021; Altindis et al., 2021; Xu et al., 2022). However, in the real world, test images may come from different data distributions and the robustness of the network is important for its usability. To measure, the robustness of the proposed methods to different domains (domain generalization), we run inferences on the Metface dataset and Fantasy image datasets. Metface dataset (Karras et al., 2020) includes face images extracted from the collection of the Metropolitan Museum of Art. The fantasy image dataset is a collection of images

Fig. 6 Qualitative comparison on MetFace dataset (Karras et al., 2020). We compare our method with CycleGAN (Zhu et al., 2017), Asymmetric CycleGAN (Yi et al., 2020a), Image2StyleGAN (Abdal et al., 2019), and DualStyleGAN (Yang et al., 2022). CycleGAN and Asymmetric CycleGAN suffer from artifacts on the faces. Img2StyleGAN does not output images with the portrait style. DualStyleGAN suffers from semantic alignment, Our method achieves significantly better portraits

generated by Stable diffusion and provided by Yang et al. (2022).

We show the quantitative and qualitative results of our model and competing models in Table 2 and Fig. 6, respectively, for the MetFace dataset. As can be seen from input images, this dataset shows a domain gap with the CelebA images. CycleGAN and Asymmetric CycleGAN again achieves reasonable output quality but with more artifacts on the faces. MetFace images do not have a rich background and therefore, during translation, we do not observe the noisy background generations of CycleGAN and Asymmetric CycleGAN on these images. However, these images have different hairstyles, and these methods poorly translate those parts. Img2StyleGAN also performs poorly and DualStyleGAN again suffers from semantic alignment even more dramatically on this dataset. Our method

Fig. 7 Qualitative comparison on Fantasy dataset (Yang et al., 2022). We compare our method with CycleGAN (Zhu et al., 2017), Asymmetric CycleGAN (Yi et al., 2020a), Image2StyleGAN (Abdal et al., 2019), and DualStyleGAN (Yang et al., 2022). DualStyleGAN suffer from semantic alignment, whereas CycleGAN, Asymmetric CycleGAN, and Img2StyleGAN suffer from artifacts on the faces. Our method achieves significantly better portraits

achieves significantly better than others both qualitatively and quantitatively improving FID from 141.11 (Asymmetric CycleGAN) to 88.57.

Next, we show quantitative and qualitative results of our model and competing models in Table 2 and Fig. 7, respectively, for the Fantasy image dataset. This dataset also shows a domain gap with the CelebA dataset. The faces have unusual coloring and illumination. The hairstyles are also different and unusual. We observe that CycleGAN and Asymmetric CycleGAN show even more artifacts on these images as can be seen from the second-row example of Fig. 7. Especially, the results for the second-row examples have black spots on the portraits which make the resulting portraits look not realistic. Img2StyleGAN also performs poorly and DualStyleGAN again suffers from semantic alignment on this dataset. Our results do not generate such artifacts as CycleGAN and Asymmetric CycleGAN do. Our results improve FID from 131.97 (Asymmetric CycleGAN) to 89.16 and achieve significantly better portraits.

Ablation Study. In this section, we provide an ablation study. We base our framework on Asymmetric CycleGAN which is also proposed for portrait generation. We refer to the Asymmetric CycleGAN as our baseline. Next, we replace the baseline image translation architecture with our translator. As shown in Table 3, FIDs improve consistently on all datasets with our translator. Our translator combines deep feature extraction of pose encoders and shallow translation networks. We observe that portrait generation requires a deep understanding of an image which can also be achieved with a deep network architecture. Therefore, we also experiment with deep network architectures that will be presented later in this section.

Next, we experiment with different cycle consistency losses. We add our pose-based cycle consistency as given in Eq. 8 to our training. As shown in Table 3, this improves the metrics on all datasets consistently. Instead of pose-based cycle consistency, we also experiment with VGG-based cycle consistency. As shown in Table 3 and Fig. 8, VGG-based (Zhang et al., 2018) cycle consistency loss even harms the performance. That is because it causes a strict consistency whereas ours is a relaxed one that only expects the face parts/pose to be preserved from the input image and not the background or color information. In the training phase, VGG-based cycle consistency loss has retained the encoder from learning different poses compared to the pose-based cycle consistency. In Fig. 8, we provide comparisons of our translator trained with and without pose-based cycle consistency. The pose cycle consistency further improves the removal of the background and provides high-quality portrait

Table 3 Quantitative results of Ablation Study on CelebA (Liu et al., 2015), MetFace (Karras et al., 2020), and Fantasy datasets (Yang et al., 2022)

Methods	CelebA	MetFace	Fantasy
Baseline (Yi et al., 2020a)	70.40	141.11	131.97
w/ Baseline + pose-based consistency	75.28	104.80	103.56
w/ Our Translator	<u>66.31</u>	95.34	<u>100.23</u>
w/ Our Translator + VGG-based consistency	72.77	<u>94.98</u>	112.50
w/ Our Translator + pose-based consistency	**62.33**	**88.57**	**89.16**

We highlight the best results in bold and the second-best results with underline

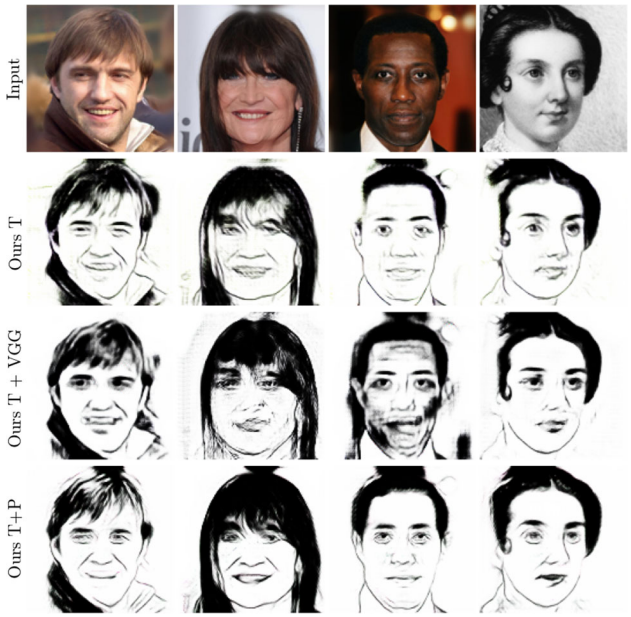

Fig. 8 Qualitative results of Ablation Study. The first three column input images are from the CelebA dataset and the last one is from the MetFace dataset. Deep Arch. refers to the experiments of baseline with deep architecture translator. Ours (T) is the setup of baseline with our proposed translator. Ours (T + VGG) set-up is trained with our architecture but with VGG-based cycle consistency loss. Ours (T + P) refers to our final model which is also trained with pose-based cycle consistency. Adding pose cycle consistency further improves the removal of the background and outputs more complete parts, e.g. the hair of the second and fourth columns. Overall, it achieves better results

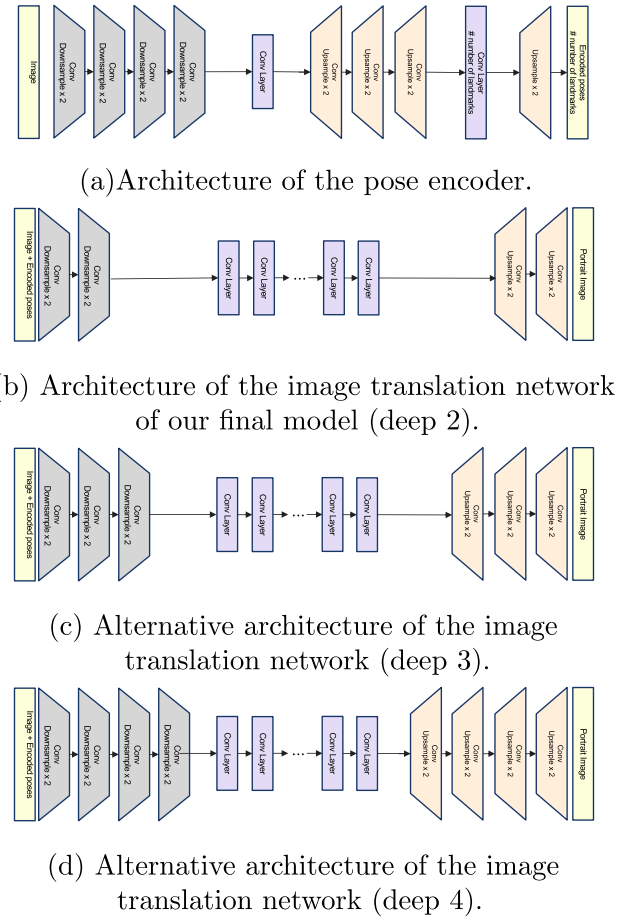

Fig. 9 We provide the architectures of the pose encoder and different variants of image translation networks with different depths that we use in our Ablation Study. We define depth as the number of downsampling layers. Each convolution layer is followed by instance normalization and ReLU layers which we remove from the figure for brevity. Details of each block are given in Sect. 4 for the pose encoder and our image translation network

results. Especially it improves hair completion in examples of columns two and four.

We also run experiments with the baseline architecture trained with pose-based cycle consistency. We find that it does not improve the scores on CelebA images but improves on MetFace and Fantasy images. Our translator combined with pose-based consistency constrain achieves significantly better scores. This is because pose-based consistency can better guide the network when the expected poses are given as input to the network. Because the original network is shallow, it does not have the capacity to extract the parts and so the pose-based consistency loss does not guide the network as well.

We further experiment with different architecture choices in the image translation network as shown in Fig. 9. The main motivation of our image translator is to combine deep feature extraction of pose encoders and shallow translation networks to preserve the content. Our pose encoder has 4 downsampling and upsampling blocks and has the capacity to extract high-level features. Because it is trained on a large number of CelebA images in an unsupervised way, it does not overfit and is able to encode poses. On the other hand, both our translation network and Asymmetric CycleGAN only use 2 downsampling and upsampling layers and convolution layers in the middle. We also run experiments with 3 and 4 downsampling - upsampling layers as shown in Fig. 9. The quantitative and qualitative results of these experiments are given in Table 4 and Fig. 10, respectively. In these experiments, we also use our improvements by both guiding the training with pose-based cycle consistency loss and combining input images with pose encodings. Downsampling of 4 layers becomes similar to our pose encoder in terms of complexity. We observe that the FID score becomes worse as we use a deeper architecture in the image translation network.

Looking at the qualitative results in Fig. 10, the portraits do not resemble the input images especially when there are 4 downsampling layers. They try to generate samples close to the training portrait images but the outputs are not semantically aligned with the input images. This also shows that the cycle consistency we set in poses alone is not enough to preserve the identity when the translation network has a

Table 4 Quantitative results of Ablation Study on the depth of image translation network. Methods use our translator and pose-base consistency with different depths of the network in the image translation architecture

Deep	CelebA	MetFace	Fantasy
2 (Ours Final)	**62.33**	**88.57**	**89.16**
3	93.01	111.80	109.78
4	106.78	114.91	116.47

The number in the first column corresponds to how many downsampling/upsampling layers there are in the image translation network

Fig. 10 Qualitative results of Ablation Study as we increase the complexity of image translation network by increasing the depth of downsampling and upsampling layers of the encoder-decoder architecture. The first three column input images are from the CelebA dataset and the last one is from the MetFace dataset. The deep architecture with 4 downsampling layers does not achieve semantic alignment with the input images. This misalignment becomes even more severe on out-of-domain images like the MetFace image in the last column. Our final model can achieve content preservation by keeping the image translator shallow and combining it with a deep understanding of the pose encoder

large capacity. Even though we provide the encoded pose information to the network, when the network has a deep architecture, it tends to make drastic changes to the content to better match the target distribution on the training images. This translates to poor results on the validation images. That is the motivation of our translator to employ a shallow translation architecture. We also observe that the model with deep architecture does not generalize well to the out-of-domain images. It can be seen that the outputs look similar for the third and last columns in Fig. 10 for the deepest architecture even though the input to the last column looks completely different and is from the MetFace dataset.

5 Conclusion

In this paper, we propose a method for translating face photos to portrait drawings that learns to perform this task only from unpaired data and with no additional labels. Our architecture while having the capacity to process images for high-level understanding does not change the identity of the photos while translating. This is achieved by the proposed image translation network which combines a pose encoder and a shallow translation network. Our pose-based cycle consistency regularizes the network to preserve facial details in the portraits and does not require the image translator to encode unnecessary information from face photos.

We provide an extensive ablation study and set various strong baselines to compare our method with. These baselines include general image translation method like CycleGAN, style transfer methods like Img2StyleGAN and DualStyleGAN, as well as image translation method that is proposed for portrait drawing like Asymmetric CycleGAN. Compared to a number of strong baselines, our method shows significant improvements both quantitatively and qualitatively and on both in-domain and challenging out-of-domain images.

References

Abdal, R., Qin, Y., & Wonka, P. (2019). Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, (pp. 4432–4441).

Altindis, S.F., Dalva, Y., Pehlivan, H., & Dundar A. (2021). Benchmarking the robustness of instance segmentation models. arXiv preprint arXiv:2109.01123

Bhattad, A., Dundar, A., Liu, G., Tao, A., & Catanzaro, B. (2021). View generalization for single image textured 3d models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 6081–6090).

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & J. Choo (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, (pp. 8789–8797).

Choi, Y., Uh, Y., Yoo, J., & Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dalva, Y., Altindis, S. F., & Dundar, A. (2022). Vecgan: Image-to-image translation with interpretable latent directions. In *European conference on computer vision (ECCV)*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 248–255. Ieee.

Dundar, A., Gao, J., Tao, A., & Catanzaro, B. (2022). Fine detailed texture learning for 3d meshes with generative models. arXiv preprint arXiv:2203.09362.

Dundar, A., Liu, M.-Y., Yu, Z., Wang, T.-C., Zedlewski, J., & Kautz, J. (2020). Domain stylization: A fast covariance matching framework towards domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(7), 2360–2372.

Dundar, A., Sapra, K., Liu, G., Tao, A., & Catanzaro, B. (2020). Panoptic-based image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 8070–8079).

Dundar, A., Shih, K., Garg, A., Pottorff, R., Tao, A., & Catanzaro, B. (2021). Unsupervised disentanglement of pose, appearance and background from images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*.

Engin,D., Genç,A., & Kemal Ekenel, H. (2018). Cycle-dehaze: Enhanced cyclegan for single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, (pp. 825–833).

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, (pp. 2414–2423).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*.

Huang, J., Liao, J., & Kwong, S. (2021). Unsupervised image-to-image translation via pre-trained stylegan2 network. *IEEE Transactions on Multimedia, 24*, 1435–1448.

Huang, X., Liu, M.-Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. *European Conference on Computer Vision (ECCV)*.

Hung, W.-C., Jampani, V., Liu, S., Molchanov, P., Yang, M.-H., & Kautz, J. (2019). Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 869–878).

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1125–1134.

Jakab, T., Gupta, A., Bilen, H., & Vedaldi, A. (2018). Unsupervised learning of object landmarks through conditional image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, .

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems (NeuRIPS), 33*, 12104–12114.

Kim, G., Park, J., Lee, K., Lee, J., Min, J., Lee, B., Han, D. K., & Ko, H. (2020). Unsupervised real-world super resolution with cycle generative adversarial network and domain discriminator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 456–457.

Li, X., Zhang, S., Hu, J., Cao, L., Hong, X., Mao, X., Huang, F., Wu, Y., & Ji, R. (2021). Image-to-image translation via hierarchical style disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8639–8648.

Liu, G., Dundar, A., Shih, K. J., Wang, T.-C., Reda, F. A., Sapra, K., Yu, Z., Yang, X., Tao, A., & Catanzaro, B. (2022). Partial convolution for padding, inpainting, and image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(5), 6096–6110.

Liu, M.-Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. *Advances in neural information processing systems (NeurIPS)*, 30.

Liu, M.-Y., & Tuzel, O. (2016). Coupled generative adversarial networks. *Advances in Neural Information Processing Systems (NeurIPS), 29*, 469–477.

Liu, Z., Luo, P., Wang, X., & Tang, X. (December 2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lorenz, D., Bereska, L., Milbich, T., & Ommer, B. (2019). Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mardani, M., Liu, G., Dundar, A., Liu, S., Tao, A., & Catanzaro, B. (2020). Neural ffts for universal texture image synthesis. *Advances in Neural Information Processing Systems (NeurIPS), 33*, 14081–14092.

Park, T., Efros, A. A., Zhang, R., & Zhu, J.-Y. (2020). Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision (ECCV)*, pages 319–345. Springer.

Park, T., Liu, M.-Y., Wang, T.-C., & Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346.

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., & Cohen-Or, D. (2021). Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2287–2296.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*.

Sardari, F., Ommer, B., & Mirmehdi, M. (2021). Unsupervised view-invariant human posture representation. arXiv preprint arXiv:2109.08730.

Shen, W., & Liu, R. (2017). Learning residual images for face attribute manipulation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4030–4038.

Shyam, P., Yoon, K.-J., & Kim, K.-S. (2021). Towards domain invariant single image dehazing. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, 9657–9665.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2818–2826.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 8798–8807.

Xiao, T., Hong, J., & Ma, J. (2018). Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–184.

Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 1395–1403.

Xu, M., Wang, H., & Ni, B. (2022). Graphical modeling for multi-source domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*.

Yang, S., Jiang, L., Liu, Z., & Loy, C. C. (2022). Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7693–7702.

Yi, R., Liu, Y.-J., Lai, Y.-K., & Rosin, P. (2022). Quality metric guided portrait line drawing generation from unpaired training data. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*.

Yi, R., Liu, Y.-J., Lai, Y.-K., & Rosin, P. L. (2019). Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10743–10752.

Yi, R., Liu, Y.-J., Lai, Y.-K., & Rosin, P. L. (2020). Unpaired portrait drawing generation via asymmetric cycle mapping. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8217–8225.

Yi, R., Xia, M., Liu, Y. J., Lai, Y. K., & Rosin, P. L. (2020). Line drawings for face portraits from photos using global and local structure based gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, pages 1–1.

Yi, Z., Zhang, H., Tan, P., & Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., & Lin, L. (2018). Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, pages 701–710.

Zhang, G., Kan, M., Shan, S., & Chen, X. (2018). Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–432.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Zheng, Z., Wu, Y., Han, X., & Shi, J. (August 2020). Forkgan: Seeing into the rainy night. In *The IEEE European Conference on Computer Vision (ECCV)*.

Zhu, J., Shen, Y., Zhao, D., & Zhou, B. (2020). In-domain gan inversion for real image editing. In *European conference on computer vision (ECCV)*, pages 592–608. Springer.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Zou, Y., Yang, X., Yu, Z., Kumar, B., & Kautz, J. (2020). Joint disentangling and adaptation for cross-domain person re-identification. In *European Conference on Computer Vision (ECCV)*, pages 87–104. Springer.