LEARNING PORTRAIT DRAWING OF FACE PHOTOS FROM UNPAIRED DATA WITH UNSUPERVISED LANDMARKS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Burak Taşdemir
December 2023

Learning Portrait Drawing of Face Photos from Unpaired Data with Unsupervised Landmarks
By Burak Taşdemir
December 2023

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Ayşegül Dündar Boral(Advisor)

Uğur Güdükbay

Mehmet Erkut Erdem

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

LEARNING PORTRAIT DRAWING OF FACE PHOTOS FROM UNPAIRED DATA WITH UNSUPERVISED LANDMARKS

Burak Taşdemir
M.S. in Computer Engineering
Advisor: Ayşegül Dündar Boral
December 2023

Translating face photos to artistic drawings by hand is a complex task that typically needs the expertise of professional artists. The demand for automating this artistic task is clearly on the rise. Turning a photo into a hand-drawn portrait goes beyond simple transformation. This task contemplates a sophisticated process that focuses on highlighting key facial features and often omits small details. Thus, designing an effective tool for image conversion involves selectively preserving certain elements of the subject's face. In our study, we introduce a new technique for creating portrait drawings that learn exclusively from unpaired data without the use of extra labels. By utilizing unsupervised learning to extract features, our technique shows a promising ability to generalize across different domains. Our proposed approach integrates an in-depth understanding of images using unsupervised components and the ability to maintain individual identity, which is typically seen in simpler networks. We also present an innovative concept: an asymmetric pose-based cycle consistency loss. This concept introduces flexibility to the traditional cycle consistency loss, which typically expects an original image to be perfectly reconstructed after being converted to a portrait and then reverted. In our comprehensive testing, we evaluate our method with both in-domain and out-domain images and benchmark it against the leading methods. Our findings reveal that our approach yields superior results, both numerically and in terms of visual quality, across three different datasets.

Keywords: Portrait drawing, Unsupervised part segmentations, Unpaired image translation, Cycle consistency adversarial networks, Generative adversarial networks.

ÖZET

DENETLENMEYEN YER İŞARETLERİYLE EŞLEŞTİRİLMEMİŞ VERİLERDEN YÜZ FOTOĞRAFLARININ PORTRE ÇİZİMİNİ ÖĞRENME

Burak Taşdemir
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Ayşegül Dündar Boral
Aralık 2023

Yüz fotoğraflarını sanatsal portre çizimlerine dönüştürmek, uzman sanatçılar için zaman alıcı bir iştir, dolayısıyla bu sürecin otomatikleştirilmesine ihtiyaç vardır. Bir fotoğraftan portre oluşturmak yalnızca basit bir görüntü dönüştürme işlemi değildir; karmaşıklıklarla doludur. Süreç, birçok karmaşık ayrıntıyı gözden kaçırırken temel yüz özelliklerini vurgulamalıdır. Bu nedenle, yetenekli bir görüntü dönüştürme aracının yüzün seçilmiş yönlerini algılaması ve koruması gerekir. Araştırmamızda, yalnızca eşleştirilmemiş verilerden öğrenen ve hiçbir ekstra etiketleme içermeyen bir portre çizimi yöntemi öneriyoruz. Denetimsiz özellik öğreniminden yararlanan yöntemimiz, başarılı genelleştirme davranışı gösterir. İlk katkımız, görüntülerin üst düzey detaylarını denetlenmeyen parçalarla ve sığ ağların kimlik koruma davranışını birleştiren bir görüntü çeviri mimarisidir. İkinci katkımız yeni bir asimetrik poz temelli döngü tutarlılığı kayıp fonksiyonudur. Bu kayıp, bir giriş görüntüsünün bir portreye dönüştürüldükten ve tekrar giriş görüntüsüne dönüştürüldükten sonra yeniden oluşturulmasını gerektiren döngü tutarlılığı kaybı üzerindeki kısıtlamayı hafifletir. Son olarak hem alan içi hem de alan dışı görüntüler üzerinde kapsamlı deneyler yaptık ve yöntemimizi en son teknoloji yaklaşımlarla karşılaştırdık. Üç veri kümesinde hem niceliksel hem de niteliksel olarak önemli gelişmeler gösteren modelimizi sunuyoruz.

Anahtar sözcükler: Portre çizimi, Denetimsiz parça bölümlemeleri, Eşleştirilmemiş görüntü çevirisi, Döngü tutarlılığı rakip ağlar, Çekişmeli üretken ağlar.

Acknowledgement

I wish to extend my deepest gratitude to my advisor, Ayşegül Dündar Boral, whose invaluable guidance and unwavering support have been pivotal throughout my academic journey. Her profound insights and extensive expertise have fundamentally shaped my scholarly development, for which I am immensely grateful.

I would also like to express my sincere appreciation to Prof. Uğur Gudukbay and Prof. Mehmet Erkut Erdem for their readiness to participate in my thesis jury. Their expert opinions and constructive critiques have played a significant role in the enhancement and refinement of my research.

I would like to extend a heartfelt note of thanks to my parents, Filiz Taşdemir and İbrahim Taşdemir. Their steadfast guidance and unwavering support, especially during the most challenging times, have been the cornerstone of my resilience and determination. I am equally grateful to my brother, Tahsin Taşdemir, for his consistent encouragement throughout my academic endeavors. Their unwavering belief in me has been an inexhaustible source of motivation and strength.

The journey I have undertaken would not have been conceivable without the invaluable contributions of these remarkable individuals and countless more friends of mine. Their role in my academic pursuits and personal development is deeply appreciated, and for this, I am profoundly thankful.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The research area of image-to-image translation has captivated the interest of researchers, manifesting substantial progress across varied sectors [1, 2, 3, 4, 5, 6, 7, 8, 9]. In this realm, the transformation of facial photographs to artistic portraits has garnered particular attention [10, 11, 12, 13, 14, 15, 16, 17]. The ability to infuse an artistic flair into the generated portraits by semantically analyzing facial components is a testament to this field's complexity.

Portrait drawing transcends simple replication, requiring a deep understanding of the subject's persona and distinctive facial characteristics. These processes purposefully neglect irrelevant features present in the original photographs, such as complex backdrops and minute textural details. The fine line between capturing a likeness and creating art has historically necessitated the expertise of accomplished artists. Thus, the automation of portrait drawing presents a fascinating challenge [3, 18, 19].

The abstract requirements of generating artistic portraits add layers of complexity to the translation process. This task requires a selective discernment of pivotal facial features to be translated while discarding ancillary details. It represents a significant leap in difficulty compared to simpler image translation tasks, such as modifying painting styles. Techniques renowned for their style

transfer capabilities, like neural style transfer [20] and cycle-consistent adversarial networks (CycleGAN) [21], are found lacking in this specialized context [19].

Previous methods for automating portrait drawings have largely depended on datasets comprising directly correlated pairs [3, 18] or required detailed segmentation of facial parts for training discriminative models [3, 18, 19]. The present work endeavors to develop an unsupervised approach to portrait generation that minimizes reliance on paired data and dispenses with the need for detailed image annotations.

This work is an extension of cycle-consistent adversarial network methodologies [21]. Such methods employ a cycle consistency logic to minimize the loss when an image is translated to an alternate domain and then back again. They allow domain translation without paired samples by positing an inherent link between the domains, which often depict different interpretations of the same underlying scene. The central goal is to uncover and learn this link without direct comparative examples.

Employing cycle-consistent frameworks to translate unpaired data can be fraught with complications, such as mode collapse, where diverse inputs might culminate in indistinguishable outputs. These frameworks promote the preservation of original content, yet they may not ensure the content's meaningful translation, especially since deep networks have the capacity to internalize information in modes beyond human perception.

Despite their potential, cycle-consistent approaches may not fully resolve issues inherent to domain-specific translations. For example, the method might incorrectly maintain elements such as eyeglasses throughout the translation cycles, thus achieving cycle consistency but distorting the artistic content in the portrait representation.

To confront this issue, integration of shallow network architectures has been explored [21, 7, 22, 19, 23]. Shallow networks have a reduced propensity for drastic alterations, which facilitates the preservation of original image content.

However, the quality of the portrait generation hinges on the network's capacity to distinguish critical facial attributes from extraneous information. In addressing this need, the present study merges the latest developments in unsupervised feature detection [24, 25, 26] with the principles of cycle-consistent adversarial networks to propose a system that harmonizes the essential elements of abstraction with the retention of content for portrait drawings.

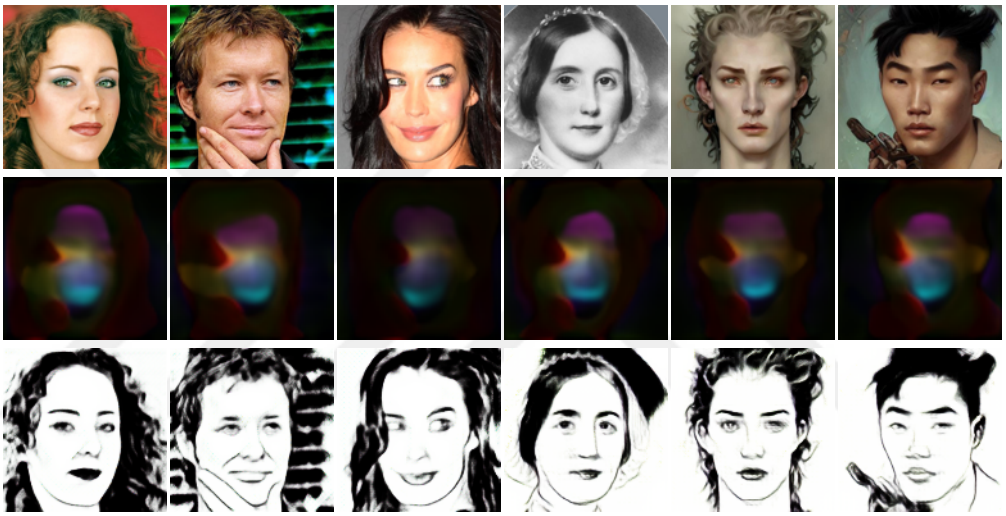


Figure 1.1: Examples of Image translation results of our method. The first row shows the input images, the second row shows the result of our pose encoder, and the last row shows the result of our network on different types of input images (real celebrity portraits, old photography portraits, and digital art-drawn portraits, respectively). Our desired style is hand-drawn portrait images which require consistency in the face attributes while keeping the style elements without overflowing with black shadows. Our method enables cycle consistency loss to be effective in promoting content preservation, it inadvertently compels portrait images to retain excessive facial information from the source photo.

Learning portrait drawing presents a unique challenge: while cycle consistency loss is effective in promoting content preservation, it inadvertently compels portrait images to retain excessive facial information from the source photo. In the field of learning portrait drawing through image translation, one encounters a paradoxical challenge: the traditional employment of cycle consistency loss for content preservation in the translated images may lead to an unintended retention of excessive facial details from the original photographs. Within the spectrum of image-to-image translation, the essence of cycle consistency loss lies in its

capacity to ensure resemblance between the twice-translated image and its original counterpart, a principle that is fundamental for the integrity of content during domain transformations. Nevertheless, while this loss is efficacious in maintaining content, it can be paradoxically detrimental when the artistic abstraction of portraits is the desired outcome.

Our exploration in this domain reveals that portraits can be varied in nature, not strictly confined to realism. The efficacy of our proposed model and method is evidenced by its ability to generate coherent results across a wide array of inputs, encompassing celebrity images, historical photography, and even digital art. Despite this versatility, concerns arise regarding the possibility of subpar outcomes contingent upon the dataset and type of image used. Illustrative results, as shown in Figure 1.1, attest to the robustness and generalizability of our method. Remarkably, some non-facial elements, such as the brickwork in the background of the second image, are rendered in an artistically consistent style without an overabundance of detail, thanks to our style-capturing generative model which complements the cycle-consistent framework.

The key to our model's success lies in the dynamic extraction of pose structures from the source images facilitated by the pose encoder. This encoder provides pivotal key points that inform the characteristic areas of the face, which our model then uses to predict portrait outcomes that consider these salient facial regions. The cycle consistency loss, when coupled with heatmaps from the recreated asymmetric portraits derived from the translation network employing the same pose encoder, essentially dictates the extent to which the portraits will resemble the facial details of the original images. Consequently, the heatmaps generated from pose-encoded data are integral to the efficacy of our model's architecture.

In the pursuit of creating portrait drawings, the goal is often to communicate a distilled essence, deliberately eschewing color and numerous detailed elements inherent in the original face photographs. To address this artistic dichotomy, Yi et al. [18] have introduced an adapted form of cycle consistency loss that considers the consistency between the edge maps of the original photos and their reconstructed versions. Although this approach proficiently eliminates color, it

inadvertently necessitates the incorporation of high-frequency details, such as all visible edges, in the final portrait rendering.

To mitigate this limitation, we present an innovative asymmetric pose-based cycle consistency loss. This new loss function is designed to foster an alignment between the overarching features of faces in photographs and their artistic portrait renditions. It does so without the burden of preserving unnecessary details like color and background elements. Our methodology delicately navigates between the realms of abstraction and content retention in portrait artistry, ensuring that crucial facial features are depicted accurately while permitting artistic freedom and the elimination of redundant detail. This novel take on cycle consistency loss resonates with our larger objective of producing expressive and evocative portraits. We aim to fabricate hand-drawn portraits that exemplify consistency in facial attributes and maintain stylistic authenticity without the overemphasis on dark shadows. While the bottom row of Figure 1.1 illustrates the traditional role of cycle consistency loss in preserving content, our research ventures beyond this to present a nuanced approach that facilitates the creation of high-quality portraits. This thesis explores the sophisticated methods and novel contributions that empower us to overcome these challenges, facilitating the creation of superior-quality portrait artwork.

The key contributions of this thesis are manifold and can be summarized as follows:

1. Our innovative approach to portrait drawing eschews the conventional need for paired data or explicit part labeling, a notable departure from previous methodologies. Traditional techniques in this arena have been heavily reliant on paired data or specific part labels to inform the drawing process.

2. We introduce an advanced image translation architecture. This architecture adeptly merges a deep comprehension of image content with the ability to recognize unsupervised parts while maintaining the essence of the subject's identity, akin to the capabilities of shallow networks.

3. We have developed a novel pose-based cycle consistency loss. This groundbreaking loss function is designed to guide the network in focusing on the crucial aspects of image translation without holding on to irrelevant color and background details when converting an image into a portrait drawing.

4. Our empirical studies are thorough, encompassing both in-domain and out-of-domain images. In our comparative analysis with leading-edge techniques, our method exhibits superior performance, which is evident through both qualitative and quantitative enhancements.

Chapter 2

Related Work

2.1 Image to Image Translation Methods

Image translation methodologies fueled by adversarial losses have carved a niche across a range of tasks, including enhancing image resolution [27], clearing haze from images [28], broadening texture details [8], and deducing 3D structures [29, 30]. Numerous such techniques depend on matched datasets for efficient training [1, 31, 2, 4, 8]. Nevertheless, the procurement of matched datasets for training remains a daunting task in various practical contexts. For example, formulating a mapping function to transform a photo into a painting is impeded by the complexities of obtaining images that mirror content and composition across the two domains. Contrarily, it is more feasible to collect unmatched sets of photos and paintings. To tackle this, unsupervised image-to-image translation models, trained on unpaired data, have been put forward by researchers [7, 6, 21], addressing the inherently challenging task of training without one-to-one correspondences by integrating multiple regularizers and structural advancements into the image translation models. These include shared-latent space constructs [7], weight-sharing strategies [5], and notably, cycle consistency conditions [21, 6, 7].

The principle of cycle consistency is pivotal in image translation, entrusting

Table 2.1: Image translation tasks have conventionally relied on supervised methods necessitating either paired examples or specific part labels for adequate training. Contrary to such approaches, the method proposed in our study eliminates the need for both paired data and part labels.

Methods	Paired Data	Part Labels
APDrawingGAN [3]	✓	✓
APDrawGAN++ [18]	✓	✓
UnpairPortrait [19]	✗	✓
UnpairPortrait++ [34]	✗	✓
Ours	✗	✗

the mutual training of image translators for interconnecting disparate domains. It posits that if an image can be translated from one domain to another and reciprocally converted back, the integrity of the original image should be restorable, assuming the preservation of essential content. To reinforce cycle consistency, image translation training incorporates two types of reconstruction losses: pixel-wise L2 loss and the more sophisticated VGG loss. These training methods have shown exceptional effectiveness in various domains [21, 6, 7, 32, 33]. However, portrait drawing as an image translation task stands apart, necessitating a unique blend of high-level abstraction and artistic expressiveness. Acknowledging these specialized needs, researchers have proposed targeted techniques specifically for the art of portrait drawing, which we will explore in subsequent sections.

2.2 Style Transfer in Image-to-Image Translation

The exploration of style transfer in the field of image-to-image translation represents a compelling intersection of traditional artistry and modern machine-learning techniques. This area of research is fascinating, as it focuses on bestowing ordinary images with artistic flair, transforming everyday visuals into captivating art pieces. Indeed, the study of style transfer transcends aesthetic changes, venturing into the core of artistic expression and interpretation [35, 36]. Style transfer techniques

have evolved remarkably since the pioneering work of Gatys et al. [35], which revolutionized the field with the concept of neural style transfer. This innovative method employed convolutional neural networks (CNNs) to distinctively separate and recombine the content and style of images. Successive breakthroughs have included the advent of fast style transfer [36] and the concept of arbitrary style transfer [37]. These developments brought about the capability for instantaneous style application and a broader range of style choices, enhancing the flexibility and speed of style transfer.

However, when it comes to applying these techniques to portrait drawing and artistic style transfers, there are distinct challenges. The intricate task lies in accurately capturing artistic styles and effectively applying them to facial features while maintaining the subject's identity. Present methodologies often find themselves in a quandary, struggling to strike a delicate balance between the faithful application of style and the preservation of content, especially in instances involving complex facial expressions and details [38, 39].

Fusing style transfer with image-to-image translation offers an innovative pathway to address these hurdles. By adopting unsupervised learning approaches, such as those in cycle-consistent adversarial networks [21], and integrating them with principles of style transfer, the ambition is to craft translations of images that are both artistically nuanced and technologically advanced. Recent progress in this area includes the introduction of StyleGAN [40] and Image2StyleGAN [41], which have significantly broadened the horizons for generating high-fidelity, stylistically diverse imagery. Image2StyleGAN, introduced by Abdal et al. [41], represents a significant advancement in the field of style transfer. This research extends the capabilities of StyleGAN, enabling the embedding of external images into the latent space of StyleGAN. This method allows for detailed manipulation of the embedded images, offering a way to perform style transfer, image blending, and even image corruption repairs.

Future research in this domain could explore deeper integration of style transfer with unsupervised learning methods to further enhance the artistic quality of image translations. Particularly in the research area of portrait generation, there's

immense potential for models that can more adeptly interpret and replicate specific artistic styles, offering a fertile ground for future exploration and innovation [3, 18].

2.3 Portrait Drawing Methods

The field of translating facial photos into portrait drawings is an intriguing aspect of image translation research. APDrawingGAN, introduced by Yi et al. [3, 18], emerges as a pioneering approach in this domain, specifically engineered for crafting artistic portraits. It utilizes an encoder-decoder structure and trains via a synergy of adversarial forces and image reconstruction losses. Traditional reconstruction losses in the training schema require paired data, a significant hurdle in portrait drawing that has led to an increased focus on techniques that learn from unpaired data.

Yi et al. [19, 34] have proposed a system proficient in learning from unpaired datasets. This system incorporates a cycle constraint with an innovative asymmetrical approach to enforce cycle consistency. It is expected that the starting and resulting portraits should be congruent after a cycle of translations, portrait to photo and back to portrait. However, translating a photo to a portrait and back to a photo is challenging due to detail loss in the portrait step, which hinders exact reconstruction. To address this, a cycle consistency is used focusing on edge maps, offering a degree of flexibility in the cycle constraint. Yet, the translation process must still ensure the retention of critical edge details in portraits.

In realistic portrait rendering, the focus is on translating essential facial features against simplistic backgrounds, abstracting finer details. The relaxed cycle consistency regarding edges is helpful but not wholly sufficient for this complex task. All models for portrait translation depend on local discriminators that maintain the integrity of facial structure within the drawings, which requires part labels for effective operation. In the absence of such discriminators, portraits suffer from inaccuracies around key facial features, affecting the authenticity of the produced portraits. Table 2.1 succinctly summarizes existing works, indicating

their dependency on assorted annotations for facilitating portrait translation. Additionally, UnpairPortrait++ [34] introduces human judgment to gauge and order the quality of portraits, creating a beneficial metric for training evaluation.

Contrasting with the prior studies, our research stands out as the foremost, to our knowledge, to attain portrait translation devoid of the necessity for annotated part labels. Our image translation framework diverges from conventional techniques, as it integrates a sophisticated perception of images with unsupervised part maps and the identity-preserving properties inherent in shallow network structures. Furthermore, our approach diverges from established practices by incorporating a new asymmetric pose-based cycle consistency loss. These advancements have enabled us to markedly surpass the outcomes of previous studies, delivering substantially improved results on both in-domain and out-of-domain images.

Chapter 3

Method

Our approach is anchored by a specially designed image translation framework, which is fine-tuned for the precise task of portrait translation. At the core of our strategy is the imperative for the image translator to possess an advanced understanding of facial attributes, coupled with a disciplined approach that avoids unwarranted alterations to the content. This disciplined approach is particularly highlighted in our ablation study, which reveals the inherent risks associated with deep architectural models that tend to induce significant content modifications. To achieve our design goals, our methodology incorporates two critical elements: an unsupervised part detector and an image translator. The following sections will provide a detailed exposition of these components.

3.1 Preliminaries: Generative Adversarial Networks and Cycle-Consistency

3.1.1 Overview of Generative Adversarial Networks (GANs)

The inception of Generative Adversarial Networks (GANs), as conceptualized by Goodfellow et al. [42], involves a duality of neural networks—the Generator (G) and the Discriminator (D)—that evolve through adversarial training.

Generator (G): This network, represented by G, synthesizes data that mimics a genuine dataset. From a latent noise variable z, it generates an output $G(z)$ that aims to be indiscernible from real data in the eyes of the discriminator.

Discriminator (D): The discriminator network, denoted by D, is tasked with differentiating real data from the synthetic outputs of G. It processes an input x and yields a probability score reflecting its authenticity. The discriminator is trained to perfect its ability to recognize real data while detecting data generated by G.

This adversarial interaction is akin to a strategic game where the generator fabricates false artifacts, and the discriminator works to identify them. This dynamic reaches a Nash Equilibrium when the generator's outputs become indistinguishable from real data, leaving the discriminator guessing at chance levels [43].

3.1.2 Cycle-Consistent Adversarial Networks (CycleGAN)

On the heels of GANs, Zhu et al. [21] introduced CycleGANs to enable the translation between unpaired image domains. CycleGAN builds on the GAN structure with two sets of generators and discriminators: Generators: $G : X \rightarrow Y$

and $F : Y \rightarrow X$. Discriminators: D_X (assesses images from domain X) and D_Y (evaluates images from domain Y). Here, X and Y represent the source and target image domains, respectively. The training process involves the discriminators, D_X and D_Y, learning from authentic and synthesized images, and the generators, G and F, are trained to both fool the discriminators and preserve cycle consistency.

CycleGAN is a breakthrough for Unpaired Image Translation as it does not require corresponding images in both domains, which is ideal for applications like translating photos to artistic renderings [21]. It has shown efficacy in a variety of image translation tasks, and with further modifications to the basic structure, it serves as a formidable framework for tackling a broad spectrum of image processing tasks.

3.2 Unsupervised Part Detector

In recent explorations, the field of unsupervised learning has seen substantial advancements, particularly concerning the autonomous identification of parts and landmarks within images [24, 25, 26, 44, 45]. A cornerstone of our study involves the deployment of an encoder-decoder model, which we term the pose encoder, as illustrated in Figure 3.1. This model is instrumental in our research, providing a robust platform for further analysis.

At the heart of this model lies the encoders, tasked with capturing both the stance and the visual traits of the subject, and the decoder, whose responsibility is to reconstruct the initial image using the encoded details. We instill a perturbation phase prior to encoding, which is critical for the encoders to discern and prioritize the salient features from the altered input. Such a step is vital to our methodology, accentuating its importance.

Specifically, for pose detection, we introduce an alteration via color jittering, creating a visually modified image, which we denote as $T_{cj}(x)$. This alteration is a quintessential element of our technique, enriching the pose encoding process.

Similarly, we employ thin-plate-spline warping for shape alteration, denoted as $T_{tps}(x)$, which is essential for disentangling the shape attributes from the appearance, further enhancing our method's efficacy.

Crucially, we designed our system so that neither the pose nor the appearance encoder has direct exposure to the original image x. This strategic limitation acts as an information bottleneck, a critical component of our architecture. Such a bottleneck is designed to stimulate the separate learning of parts and the clear distinction between pose and appearance, propelling our system toward the intended research objectives. This constraint fosters indirect learning paths, yielding a deeper and more refined comprehension of the images.

Through this innovative pipeline, which incorporates strategic perturbations and information bottlenecks, this pipeline enables significant headway in the realm of unsupervised parts and landmark detection via image reconstruction. The insights garnered from this work illuminate the path forward for continued innovation in the wider domain of image processing.

In line with the techniques proposed by Lorenz et al. [25], our pose encoder processes images that have been perturbed in pose and altered in color to accurately map localized appearance features. This method is at the forefront of contemporary research in the field. The pose and appearance of encoders embedded within our system conform to the encoder-decoder architecture, a blueprint that is widely recognized in the landscape of computer vision research. These encoders utilize convolutional blocks to methodically downsample features, thereby broadening their scope of perception to include a more global representation of the input image. This global understanding is imperative for the precise encoding of parts and their appearances.

Yet, the aspiration is to construct pose representations in the form of heatmaps that preserve the original resolution of the input image to allow for the exact segmentation of individual object components. In the pursuit of this objective, the architecture includes convolutional blocks designed with upsampling layers to refine the feature dimensions back to the resolution of the initial image. This

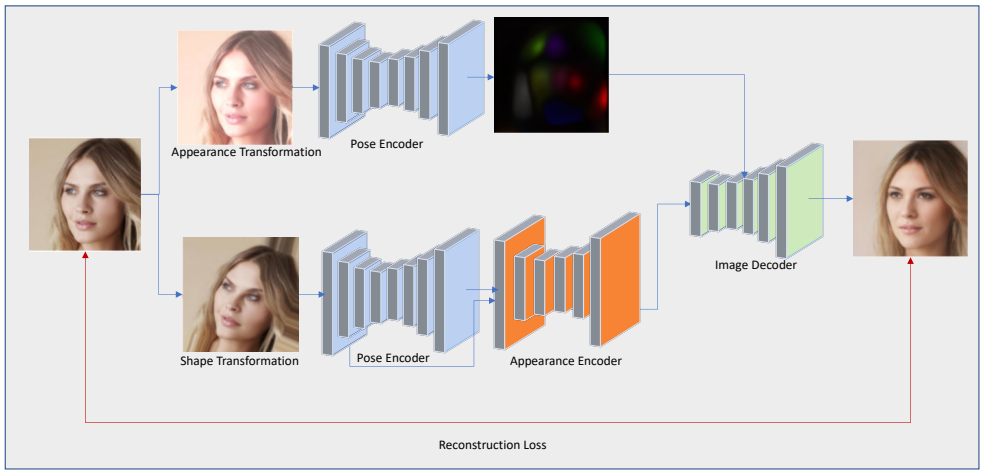


Figure 3.1: Unsupervised Part Segmentation Learning Pipeline (Pose Encoder): The objective is to train a pose encoder alongside an appearance encoder and an image decoder, facilitating the learning process. To effectively disentangle pose and appearance attributes, the input image undergoes perturbation before encoder processing. In the context of the pose encoder, color jittering is employed to generate an appearance-altered image, while maintaining the original pose. Conversely, for shape manipulation, the thin-plate-spline warping technique is utilized to generate synthetic pose variations, keeping the appearance intact. This methodology compels the learning of distinct pose and appearance features, as the process of reconstructing the input does not allow the pose or appearance encoder to directly access the original image. The training of this comprehensive framework hinges on the image reconstruction loss metric. This approach ensures a more robust and accurate learning of separate pose and appearance characteristics. Network architectures are taken from Dundar et. al. [26].

detail is particularly noteworthy as it underscores our commitment to precision in part segmentation. Concurrently, the design aims to align the dimensionality of appearance information, thereby enabling a more effective aggregation of appearance traits based on the delineated parts. The depth of this methodology will be further discussed in the following chapters of this documentation.

For a thorough grasp of our structural design and its particular arrangements, readers are directed to consult the expansive discourse in Chapter 4. Here, the discussion extends to the nuances of our experimental framework and the specificities of the network architectures. Our system, drawing inspiration from the groundwork laid by Lorenz et al., integrates encoder-decoder schemes, convolutional

Figure 3.2: Examples of Unsupervised Part Activation Heatmaps: In our approach, each channel of the pose encoder is represented by a unique color. This visualization demonstrates that the channels are specialized in detecting specific facial features, such as the forehead, eyes, and nose, in photographs. This specialization is crucial for achieving accurate face reconstruction. Consistency in part detection across various examples is a key observation, indicating the effectiveness of the encoder in identifying and differentiating distinct facial features. This consistent detection plays a vital role in enhancing the overall performance of the pose encoder in unsupervised learning scenarios.

layers, and upsampling strategies to capture and articulate pose and appearance data, which are pivotal to our investigative pursuits.

A critical element in our framework's ability to differentiate between distinct features is the Gaussian bottleneck. This bottleneck is central to our methodology and has profound implications for the identification of key facial attributes. As per the processes outlined by Lorenz et al. [25], our model utilizes a Gaussian distribution to encapsulate the encoded pose information. This statistical approach is adept at isolating prominent facial features, which are vital for achieving the goals of our study. More precisely, for each of the K activation maps, we fit a 2D Gaussian distribution by calculating the mean across activation locations and applying an assumed or pre-defined covariance matrix.

This formulation is succinctly expressed as $\widetilde{\Phi}_k^{pose} = (\mu_k, \Sigma_k)$, with $\mu_k \in \mathbb{R}^2$ and $\Sigma_k \in \mathbb{R}^{2\times 2}$, offering a compact yet descriptive analysis of individual facial components. The implementation of the 2D Gaussian approximation imposes a significant restriction by defining each part activation map as unimodal, which

is essential to the structure of our information bottleneck. This restriction deliberately limits the flow of data between the encoder and the decoder, reinforcing the method's efficacy.

As a direct consequence of this bottleneck, we obtain a representation akin to keypoint or part-segmentation, with each distinct feature being represented in a singular location within each image. This limitation is not merely a byproduct but a deliberate design choice, reflecting our commitment to the separation and accurate portrayal of facial features. Following the application of the Gaussian bottleneck, our method moves forward with the derivation of a set of K discrete heatmaps, each signifying a particular facial element. This step is fundamental to our objective of capturing an intricate portrayal of facial features.

The introduction of the Gaussian bottleneck is a deliberate strategy within our framework, essential for disambiguating the learning of discrete parts. This strategic decision is a cornerstone of our methodology, culminating in the delineation of K unique heatmaps that each correspond to a specific facial feature.

Ensuring the encoded appearances are appropriately leveraged for the final image reconstruction necessitates a systematic approach. Within this framework, when an input image x is introduced, the pose encoder $\Phi^{app} = Enc^{app}(x; Enc^{pose}(x))$ embarks on generating $K \times H \times W$ part activation maps Φ^{pose}. These maps are adept at capturing the nuanced spatial arrangement of facial features. Subsequently, the focus shifts to the appearance encoder, which adeptly transposes the input image onto a $C \times H \times W$ appearance feature map, succinctly represented as M^{app}.

This procedure paves the way for the seamless integration of localized appearance attributes, as discerned by the appearance encoder, with the spatial bearings inferred by the pose encoder. The importance of this integration cannot be overstated, as it is critical for the efficacy of the subsequent image reconstruction endeavor.

To elaborate, the pose activation maps serve as a guide to perform a weighted

aggregation across the appearance feature map, culminating in the isolation of a condensed appearance vector corresponding to the kth feature, which is meticulously calculated as:

$$\Phi^{app}_{k,c} = \sum_{i}^{H} \sum_{j}^{W} \Phi^{pose}_{k,i,j} M^{app}_{c,i,j} \text{ for } c = 1...C, \tag{3.1}$$

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

The intricacies of image decoding are underpinned by the meticulous integration of appearance vectors alongside the encoded pose data. The training regimen adopted in our study, which is illustrated in Fig. 3.1, can be succinctly articulated as a sequential process:

$$\Phi^{pose}_{cj} = Enc^{pose}(T_{cj}(x)) \tag{3.2}$$

$$\Phi^{app} = Enc^{app}(T_{tps}(x); Enc^{pose}(T_{tps}(x))) \tag{3.3}$$

$$\tilde{x} = Dec(\Phi^{pose}_{cj}, \Phi^{app}), \tag{3.4}$$

where Enc^{pose}, Enc^{app} correspond to pose and appearance encoders, respectively. The narrative of this process encapsulates the essence of our methodological advancements and emphasizes the concerted effort to marry the dual aspects of appearance and pose, resulting in the nuanced reconstruction of images.

Our image processing methodology depicts a pivotal role to the image decoder, symbolized as Dec. This component is essential, as it brings together the pose and appearance codes—meticulously encoded—to recreate the original image. It is a testament to the rigorous nature of our study that the image decoder is integral to achieving our scientific ambitions. In our pursuit of excellence, we employ a VGG perceptual loss function [46] for the objective evaluation of our image reconstruction efforts. This choice mirrors the benchmarks set by prior groundbreaking research [24, 26] and harnesses the power of pre-trained ImageNet weights to measure the fidelity of our reconstructions with precision. This metric

stands as a bulwark of quality assurance within our research paradigm.

The architectural underpinnings of our network are adopted from the renowned work of Dundar et al. [26]. Those seeking the granular details of our network parameters will find them exhaustively chronicled in Chapter 4. Such meticulous documentation ensures that our methodologies are both replicable and robust, thereby setting a benchmark for future explorations.

Exemplifying the efficacy of our approach, Fig. 3.2 presents a series of part detections that illuminate our ability to consistently track facial features across a variety of image manipulations. This visual anthology speaks volumes, offering indisputable proof of our methodology's adeptness in identifying and reconstituting critical facial components. The heatmap activations depicted in Fig. 3.2 maintain a remarkable consistency across differing samples, which buttresses the credibility of our part detection technique.

The integration of the pose encoder within our image translation framework serves not one, but two pivotal roles. This dual utility not only heightens the network's grasp of the complex information presented by input images but also paves the way for the application of pose-based cycle consistency loss. These strategic enhancements will be dissected and discussed in further depth in ensuing subsections. It is crucial to underscore that the appearance encoder and image decoder, while fundamental in the initial learning phase of the pose encoder via image reconstruction loss, are gracefully retired from active duty in later stages of our framework, as illustrated in Fig. 3.1. This refined focus empowers us to concentrate on the seminal components of our research, optimizing computational resources in the process.

3.3 Image Translator for Portrait Drawing

For image translator in our proposed framework, we use the described pose encoder that has a deep architecture that is which endows it with the capacity to

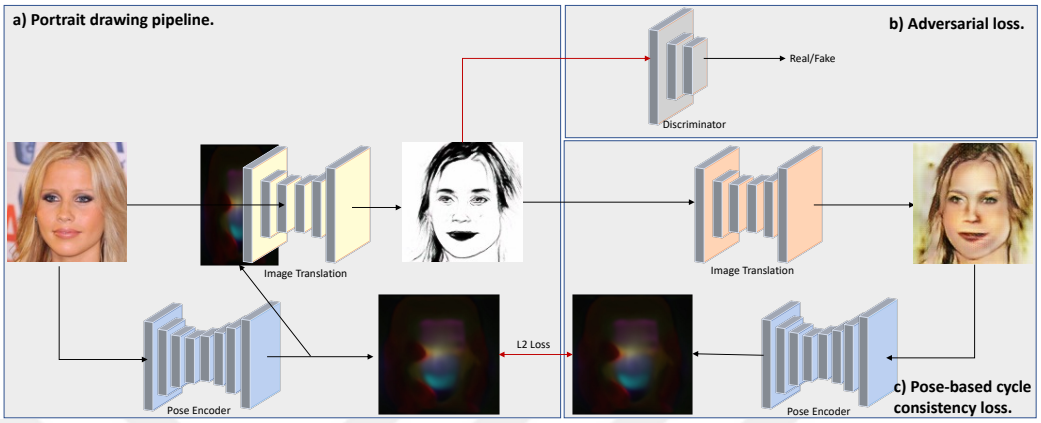


Figure 3.3: Portrait drawing translation network receives an RGB image and outputs a portrait drawing (Figure a). The pipeline contains a pose encoder part and a shallow neural translation architecture. Pose encoder has a deep architecture that can provide a high-level understanding. It is trained with the pipeline given in Fig. 3.1 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 (Figure b) and pose-based cycle consistency losses (Figure c). To harness adversarial guidance, we initiate the training process by introducing a discriminator that is tasked with distinguishing between the generated portraits and authentic ones. In the context of pose-based cycle consistency loss, we employ an L2 loss metric to quantify the disparity between the encoded poses of the input image and the subsequently reconstructed image. These loss components are visually indicated by red arrows in our framework. Nevertheless, our primary expectation is for the content of the input image to be adequately preserved, and we employ the pose-based cycle consistency loss as a pivotal means to achieve this objective.

comprehend high-level abstractions, coupled with a translation architecture that remains shallow to safeguard the original content integrity, as visually interpreted in Fig. 3.3. Our methodology hinges on utilizing the pose encoder with immutable parameters while dynamically learning the parameters of the translation architecture through extensive training. The underpinnings and rationale for this architectural depth are elaborated upon in our comprehensive ablation study.

Within our framework, the translation architecture is composed of a symmetrical structure: two downsampling and two upsampling layers, augmented by a convolutional block preceding each sampling layer. These blocks integrate

convolution, ReLU, and instance normalization layers, providing a balanced and streamlined processing pathway. After two sequential downsampling phases, a series of nine convolutional blocks lead into the final upsampling stages. This architectural design, featuring a mere two downsampling layers, inherently restricts the receptive field, ensuring delicate handling of the input data. In contrast, the pose encoder's architecture, with its four downsampling and upsampling layers, affords a larger receptive field, fostering a deeper comprehension of the image content. We disclose further architectural nuances in Chapter 4, offering a clear window into our design philosophy.

Our approach imparts a sophisticated level of understanding, distilled from the images, to the translation network by concatenating the encoded poses with the input image, as depicted in Fig. 3.3. Notably, throughout the training of the image translation, the parameters of the pose encoder remain untuned, a testament to the encoder's foundational strength.

The training of the translation architecture is orchestrated through a cycle-consistent adversarial network framework. This strategy recognizes the inherent loss of information when translating from a portrait back to the original image—a concession to the pragmatic limits of transformation. We base our training on the asymmetric CycleGAN approach outlined by Yi et al. [19], establishing a two-fold image translation mechanism: the G_p translates face photos to portrait drawings, and the G_r performs the reverse. Each translator operates in tandem with a respective discriminator, D_p and D_r, ensuring images are translated within the realms of their intended domains.

3.4 Learning Objectives

3.4.1 Adversarial and Cycle Consistency Losses

The adversarial losses are articulated through a min-max game between discriminators and generators, as detailed in the following equation:

$$\begin{aligned} L_{adv} = &\min_{G_p} \max_{D_p} \ \lambda_{adv} \ \ell_{adv}(G_p, D_p) \\ &+ \min_{G_r} \max_{D_r} \ \lambda_{adv} \ \ell_{adv}(G_r, D_r) \end{aligned} \tag{3.5}$$

Supplementary to this, we deploy a cycle consistency loss to maintain fidelity in portrait photo translations, as per the equation below:

$$L_{cyc} = ||x_p - G_p(G_r(x_p))|| \tag{3.6}$$

However, the reciprocal cycle consistency when translating from face photo to portrait and back is considered excessively stringent due to the significant loss of information. Hence, a relaxed cycle consistency, as proposed by Yi et al. [19] and denoted in Eq. 3.7, becomes instrumental:

$$L_{rel-cyc} = ||E(x_r) - E(G_r(G_p(x_r)))|| \tag{3.7}$$

Here, E signifies an advanced edge detection algorithm [47], allowing the cycle consistency to focus solely on the edge details, a nod to the selective information retention typical in portrait translations.

3.4.2 Pose-based Cycle Consistency Loss

To circumvent the limitations of the relaxed cycle consistency, we introduce a pose-based cycle consistency loss, as illustrated in Eq. 3.8.

$$L_{pose-cyc} = ||Enc^{pose}(x_r) - Enc^{pose}(G_r(G_p(x_r)))|| \tag{3.8}$$

Leveraging the same pose encoder used in the image translation architecture, this loss shifts the focus from preserving minute edge details to maintaining essential facial pose information. Figure 3.3 exemplifies this approach, where the portrait translation is not burdened with reproducing background details that could potentially clutter the image with noise. Our goal is not to reproduce a photorealistic output but to perfect G_p through the cycle-consistent adversarial training process.

3.4.3 Overall Objective Loss

Integrating the individual loss components, we define the holistic objective for our optimization process as given in Eq. 3.9:

$$\min_{G_p, G_r D_p, D_r} \max \lambda_a \mathcal{L}_{adv} + \lambda_c \mathcal{L}_{cyc} + \lambda_r \mathcal{L}_{pose-cyc} \tag{3.9}$$

For our experimentation, we adopt the hyperparameter configuration suggested by Yi et al. [19], with $\lambda_a = 0.5$ and $\lambda_c = \lambda_r = 5$. This decision aligns with our commitment to validate the efficacy of our proposed modules without engaging in hyperparameter optimization. Our ablation study first observes the translator's performance using $\mathcal{L}_{rel-cyc}$, subsequently transitioning to our proposed $\mathcal{L}_{pose-cyc}$ for the conclusive experimental setup.

Chapter 4

Experiments

4.1 Set-up

In the conducted experiment, we harness the capabilities of the APDrawing dataset [3] for our portrait drawing image requirements. This dataset is an exclusive compendium that encompasses 140 exquisitely curated portrait images for this task, all designated for the purpose of training. The selection of the APDrawing dataset represents a strategic decision, recognizing its divergence from conventional image collections. As a point of comparison, we also incorporate the CelebA training images [48] into our training protocol. This dataset constitutes a vast array of 27,000 RGB images, offering a comprehensive range for the training process. To facilitate an understanding of the diversity and nature of these datasets, we have included representative images in Fig. 4.1.

The paradigm we embrace for this experiment is known in the literature as training with unpaired data. This technique is particularly challenging, as underscored by previous investigations [7, 19]. The complexity arises from the endeavor to identify and construct a logical mapping between two disparate image sets that lack a natural correspondence. Given the complexities inherent in this method, it is essential to meticulously apply our established methodologies. This

ensures that the model training is executed with precision, leading to a successful outcome.

Figure 4.1: Specifically, we utilized the APDrawing set [3] for portrait drawing images, while the CelebA training images [48] served as our source for RGB images, as can be seen in the example samples provided.

4.2 Metrics

4.2.1 Frechet Inception Distance (FID)

The Frechet Inception Distance (FID) metric, as introduced by Heusel et al. [49], serves as a gauge for assessing the realism and variety of images by juxtaposing the data distributions of authentic and synthesized datasets. For the computation of FID scores, the Inception-v3 architecture [50], pre-trained using the ImageNet dataset [51], is employed. The Inception-v3 network facilitates the extraction of features from both authentic and generated datasets. Subsequently, these feature sets are utilized to fit multivariate Gaussians, allowing for the computation of their respective means and covariance matrices.

The concept of Frechet distance, harnessed in this context, relies on the means and covariance matrices derived from the extracted features. A diminutive Frechet distance signifies a closer resemblance between the data distributions of genuine

and generated images. The FID effectively quantifies the discrepancy between the data distributions of actual and fabricated images within a feature space that is defined by an intermediary layer of a sophisticated neural network model, such as the Inception network.

If we denote two multivariate Gaussian distributions corresponding to real and generated images as P_r and P_g, with their means represented by μ_r, μ_g and covariance matrices by Σ_r, Σ_g, the squared form of the FID is articulated mathematically as:

$$FID^2(P_r, P_g) = ||\mu_r - \mu_g||^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{0.5}) \tag{4.1}$$

The operation $Tr(.)$ indicates the trace operation applied to the matrices. FID's strength lies in its ability to encapsulate both the fidelity and diversity of the synthesized images. Moreover, empirical evidence suggests that FID exhibits a higher correlation with human evaluations of image quality when compared to alternative metrics such as the Inception Score.

We undertake the computation of FID by comparing translated images from the CelebA validation set with the original portrait images. This evaluation is considered as in-domain since the images for both training and validation, albeit distinct, originate from the identical dataset. Additionally, we conduct evaluations utilizing two more demanding datasets, the Metface dataset [52] and the Fantasy image dataset [16]. As these datasets display distinct domain characteristics compared to the CelebA training set, they serve as a litmus test for our method's generalizability to disparate domains, thus categorizing them as out-of-domain images.

4.3 Network Architectures and Training Parameters

4.3.1 Translation Network

In the quest to refine our model, the decision was made to adopt an input resolution of 128×128 pixels. The constructed image translation framework features two initial downsampling convolutional blocks. Each block comprises a convolutional layer, activated via a ReLU function, and is normalized by instance normalization layers. The convolutional strata within the first and second blocks possess channel dimensions of 64 and 128, respectively. Moreover, each convolutional layer is characterized by a kernel size of 3, a padding parameter of 1, and a stride of 2, which effectively reduces the spatial dimensions of the input image.

Advancing beyond the downsampling stages, our model architecture is further embellished with nine residual convolutional blocks. It is within these blocks that skip connections are implemented, bridging each convolutional layer and promoting the direct flow of information by adding the input to the output. The channels within these residual segments are uniform in size, measuring 256, and retain a kernel span of 3 alongside a padding magnitude of 1. Progressing from the residual blocks, the structure incorporates a pair of transposed convolutional blocks, tasked with upsampling the feature maps to the original input size. These subsequent blocks maintain channel sizes of 128 and 64, with stride parameters steadfastly set to 2.

To bring our architecture to fruition, a fusion of the original input image with the encoded pose information is executed, feeding this amalgamated input into our image translation network. Such a scheme dictates the necessity for an input channel dimension of 13 in the introductory convolutional layer. Here, the triad of initial channels is dedicated to RGB image data, while the remaining ten channels are allocated for the unsupervised pose heatmaps. Correspondingly, the output facet of the image translator's terminal layer is configured to output three

channels, facilitating the synthesis of a color image. In determining the number of landmarks for our model, we have elected to establish 10 key points. This choice aligns with the benchmarks established by seminal studies in the field [25, 26].

4.3.2 Pose Encoder Network

In the endeavor of pose representation, our methodology embraces a synergetic training strategy. This strategy incorporates the integration of the pose encoder, appearance encoder, and image decoder within our training pipeline, employing reconstruction loss as demonstrated in Fig. 3.1. The acclaimed U-net structure [53] has been adopted for both the pose and appearance encoders, enhanced by skip connections that fortify the network's capabilities.

Precisely, the pose encoder embodies four segments of convolutional downsampling units. Each segment is composed of a convolution layer, instance normalization, a ReLU activation function, and a downsampling component. The filter count of these units exhibits a progressive doubling with each segment, assuming values in the sequence of 64, 128, 256, 512. In juxtaposition, the pose encoder's upsampling facet comprises three segments of convolutional upsampling, where the channels diminish by half in each ensuing segment, resulting in channel dimensions of 512, 256, 128.

Regarding the architecture of the appearance encoder, it includes one convolutional downsampling and one convolutional upsampling module. The image decoder, conversely, is architected with four modules. Each module embodies a convolution layer, a ReLU activation function, and an upsampling process. The initial step involves the downsampling of the appearance feature maps by a magnitude of 16 across each spatial dimension. The number of output channels designated for each module in the image decoder aligns sequentially at 256, 256, 128, 64, and lastly 3. Each layer within the decoder is endowed with an adaptive instance normalization layer. These adaptive layers serve as conduits for the encoded pose outputs into the decoder, an approach that mirrors the SPADE architecture [2]. Additionally, the encoded poses undergo scaling adjustments to

harmonize with the feature dimensions at each layer of the decoder.

4.3.3 Training Parameters

With regard to the training parameters, the Adam optimizer was selected [54], with the learning rate established at 0.0002. The training duration extended over 20 epochs, executed on a solitary GPU. For our model, a batch size of 16 was determined to be optimal.

4.4 Baselines

To ensure a comprehensive assessment of our approach, we juxtaposed it with an array of leading-edge baselines. These included CycleGAN [21], Asymmetric CycleGAN [19], Image2StyleGAN [41], and DualStyleGAN [16].

Within the scope of our experimental design, we undertook the training of both CycleGAN and Asymmetric CycleGAN on our designated dataset. CycleGAN emerges as a multifaceted instrument tailored for a variety of image translation scenarios, whereas Asymmetric CycleGAN is engineered with a focus on achieving proficiency in generating portrait images. Additionally, our analysis incorporates Image2StyleGAN, an innovator in the integration of facial images into the StyleGAN latent space, thereby enabling style modification via latent exchange. This process permits the seamless transfer of stylistic elements between a stylized image embedded in the space and other facial images. In conformity with the established methodology of Image2StyleGAN, we substituted the latent vectors of the foundational image's last 9 strata with those of the embedded portrait image. This action is aimed at the stylistic alteration of the foundational image to mirror the portrait's style. To facilitate the estimation of latent vectors, we employed the IdInvert algorithm [55].

Moreover, we allocated resources to the training of DualStyleGAN utilizing our

portrait dataset. This innovative method relies on the adeptness of style transference executed via a pre-trained StyleGAN framework, enhanced by the integration of modifiable style pathways that infuse features into the pre-established StyleGAN structure. In this process, facial photographs and portraits are transmuted into the latent domain through the application of a pSp encoder [56]. Herein, the encoded facial photograph is delineated as the content, whilst the encoded portrait photograph embodies the style. We adhered to the training guidelines meticulously as delineated by the code made public by the authors. During the translation phase, for every content image, we engaged in the random selection of portrait images from the training set to engender a spectrum of stylistic variations.

Chapter 5

Analysis and Discussions

5.1 Results

This section delineates the evaluation of our proposed model in both quantitative and qualitative dimensions, contrasting it with the outcomes of the rival models. The encapsulation of this assessment is presented in Table 5.2 and Figure 5.1, respectively, for the validation dataset originating from CelebA. The models under consideration have been subjected to training on CelebA datasets, hence the results are demonstrative of the proficiency within this specific domain.

In the aspect of output integrity, CycleGAN alongside Asymmetric CycleGAN has exhibited laudable efficacy, successfully maintaining content fidelity. Notwithstanding this achievement, a limitation emerges from the unintended incorporation of superfluous background edges. These contribute to the visual disturbance, manifesting as noise within the resultant portraits. In addition, the emergence of artifacts around facial structures is discernible in the outputs of both models. These are more pronounced in CycleGAN compared to the Asymmetric CycleGAN.

While Image2StyleGAN is not intrinsically fashioned for portrait creation, it

performs admirably in illustrating the potential of style transfer in the domain of portraits. The competency of Image2StyleGAN in translating portraits via StyleGAN, exempt from any additional training, is quite remarkable. Nonetheless, it is pertinent to recognize that the outputs procured from Image2StyleGAN do not parallel the benchmark established by dedicated image translation methods trained specifically for this task.

Further, DualStyleGAN embodies a progressive adaptation of the Image2StyleGAN paradigm, tailored for portrait generation. By refining an inherent style path within StyleGAN, DualStyleGAN exploits the capability of StyleGAN to fabricate portraits of exquisite quality. This confers appreciable results, especially in the translation of anime, cartoon, and caricature images. Despite these advancements, DualStyleGAN encounters shortcomings in generating portrait drawings, marked by a deviation in semantic congruence. This semantic discrepancy, while negligible in the transition from a photograph to a cartoon or anime representation, becomes critical in portrait generation, where a stringent resemblance to the source image is imperative. Herein, DualStyleGAN's performance does not meet the anticipated standard.

In stark relief, our proposed model emerges as the preeminent contender, achieving the most favorable FID score among the models geared towards content conservation, corroborated by the data in Table 5.2. This denotes an enhancement in the FID score, progressing from 70.40, the score achieved by Asymmetric CycleGAN, to 62.33. On the visual front, our model's outputs are indisputably more refined than those of its counterparts, as the generated portraits exhibit semantic coherence with the input images. There is an evident focus on facial attributes while effectively discarding extraneous background details.

5.2 Domain Generalization Results

Deep learning architectures often experience a degradation in performance when there is a discrepancy between the training and testing domains, such as when

Table 5.1: Quantitative Outcomes of the Ablation Study on CelebA [48], MetFace [52], and Fantasy Datasets [16]: This table presents the statistical findings from our ablation study conducted on the aforementioned datasets. The resulting scores are calculated according to the FID metric of the generated images over the portrait dataset. We emphasize the top-performing results by using boldface, and the runner-up outcomes are underlined for a clear distinction.

Methods	CelebA	MetFace	Fantasy
Baseline [19]	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**

models are trained on synthetic imagery and evaluated on authentic images [57, 58, 59, 60, 61]. In real-world applications, it is not uncommon for test images to emerge from varied data distributions, rendering the resilience of the network crucial for its practical deployment. To assess the ability of the proposed methods to generalize across different domains (domain generalization), inferences were drawn using the Metface dataset and the Fantasy image dataset. The Metface dataset [52] is constituted of facial images sourced from the Metropolitan Museum of Art's collection. In contrast, the Fantasy image dataset, assembled by Yang et al. [16], consists of visuals generated by Stable Diffusion.

The qualitative and quantitative analyses of our approach in comparison to other models are chronicled in Table 5.2 and Figure 5.2, in relation to the MetFace dataset. The images from this dataset exhibit a significant domain shift when compared to the CelebA images. Both CycleGAN and Asymmetric CycleGAN are able to generate outputs of acceptable quality; however, they do so with an increased presence of artifacts within the facial regions. Owing to the MetFace images' absence of intricate backgrounds, the typical noisy background generations characteristic of CycleGAN and Asymmetric CycleGAN are mitigated in this scenario. Nonetheless, these images display a wide array of hairstyles, which is an aspect that both methods struggle to replicate accurately. Meanwhile, Image2StyleGAN does not meet the expected standards of performance, and DualStyleGAN's semantic misalignment becomes increasingly apparent with this particular dataset. In a notable divergence, our proposed methodology surpasses

Table 5.2: Quantitative results of our and competing methods on CelebA [48], MetFace [52], and Fantasy datasets [16]. The resulting scores are calculated according to the FID metric of the generated images over the portrait dataset. We highlight the best results in bold and the second-best results in underlining. Our method achieves better scores than competing methods on all datasets.

Methods	CelebA	MetFace	Fantasy
CycleGAN [21]	70.57	99.28	105.84
Asymmetric CycleGAN [19]	70.40	141.11	131.97
Image2StyleGAN [41]	93.35	250.53	136.59
DualStyleGAN [16]	<u>67.06</u>	<u>92.07</u>	<u>94.79</u>
Ours	**62.33**	**88.57**	**89.16**

its competitors, in both qualitative and quantitative measures, improving the FID score from 141.11 (noted with Asymmetric CycleGAN) to 88.57.

Subsequently, we delineate the quantitative and qualitative analyses of our proposed model in comparison to the competing models, delineated in Table 5.2 and Figure 5.3, corresponding to the Fantasy image dataset. This dataset accentuates the domain divergence from the CelebA dataset, particularly due to the unconventional coloration and lighting of the facial features, in addition to the distinctive hairstyles that are not typically encountered.

A detailed examination reveals that CycleGAN and Asymmetric CycleGAN suffer from an intensification of distortions within images from this dataset, as the second-row examples in Figure 5.3 illustrate. In particular, the portraits generated for these examples are compromised by the presence of black anomalies, which detract from their realism. In a similar vein, the performance of Image2StyleGAN is less than ideal, and DualStyleGAN continues to encounter semantic alignment issues when applied to this dataset.

In contrast, our approach yields results that are untainted by the imperfections that afflict the outputs produced by CycleGAN and Asymmetric CycleGAN. It is of particular note that our methodology advances the Frechet Inception Distance (FID) from 131.97 (as noted with Asymmetric CycleGAN) to 89.16, thus enabling the production of portraits that are markedly more authentic and refined.

Table 5.3: Quantitative Results of Ablation Study on the Depth of Image Translation Network: This segment presents the outcomes from an ablation study focusing on the depth variation within the image translation network. The resulting scores are calculated according to the FID metric of the generated images over the portrait dataset. The methodologies employed in this study utilize our translator and pose-based consistency, differing primarily in the depth of the network within the image translation architecture. The values in the first column of the results table denote the quantity of downsampling and upsampling layers present in the image translation network. This quantitative analysis aims to explore how varying the network depth impacts the overall performance, providing insights into the optimal layer configuration for effective image translation.

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

5.3 Ablation Study

In the ensuing segment, we engage in an ablation study to quantitatively ascertain the contributions of distinct components within our proposed model. We adopt Asymmetric CycleGAN, crafted with a specialization in portrait generation, as a reference point for our exploration. Throughout this discourse, 'baseline' shall denote the Asymmetric CycleGAN. This study involves the integration of our bespoke image translator in lieu of the image translation mechanism inherent in the baseline.

The empirical data, as chronicled in Table 5.1, elucidates that there is a uniform enhancement of the Fidelity Indices (FIDs) across varied datasets upon the application of our image translator. It merits emphasis that our translator augments the intrinsic feature discernment of the pose encoders while harnessing the agility of succinct translation networks. Such a synergy is crucial, considering that the essence of portrait generation lies in the nuanced comprehension and rendition of the image, a task to which a deep network architecture is well-suited. In pursuit of substantiating this premise, we further extend our experimentation to incorporate deep network architectures, the insights from which are expounded in subsequent sections of this study.

5.3.1 Comparison Regarding the Consistency Loss

We continue our ablation study by evaluating the impact of different cycle consistency losses on our method. Our initial experiment incorporates the pose-based cycle consistency outlined in Eq. 3.8, leading to notable advancements in performance metrics across diverse datasets, as reflected in Table 5.1. In comparison, a cycle consistency loss founded on VGG-based principles [46] yields less than optimal outcomes. This is due to the VGG-based loss's rigorous consistency requirements, which encompass not only facial features and pose but also background and color retention, as evidenced by the data in Table 5.1 and visualizations in Fig. 5.4. The implemented pose cycle consistency aids in removing background distractions, thus facilitating the generation of portraits with enhanced hair details and overall quality.

Moving forward, the investigation incorporates pose-based cycle consistency within the baseline architecture. This enhancement does not significantly alter the scores for images derived from the CelebA dataset; however, it does improve results on images from the MetFace and Fantasy datasets. Remarkably, the fusion of our translation approach with the pose-based consistency loss surpasses alternative configurations, reinforcing the premise that pose-based guidance is instrumental when the model is provided with expected poses as inputs.

The advantages of our pose-based consistency approach are showcased in Fig. 1.1. The variety of input images in the figure demonstrates the adaptability and strength of our framework. The second image's background, composed of bricks, is interpreted with a finesse that avoids excessive edge detailing. This stylistic consistency stems from our model's aptitude for discerning subtle stylistic elements, which is realized through the synergy of a generative model and cycle consistency approach.

Our pose encoder's dynamic capabilities shine as it adeptly extracts pose features from the source images, providing us with critical facial keypoints. This information proves essential in forecasting prominent facial regions in the generated portraits.

The cycle consistency loss, working in concert with the heatmaps generated by our translator—originating from the same pose encoder—acts as a barometer for the fidelity of the facial details in the created portraits. It is clear that the heatmaps informed by pose encoding are not simply advantageous; they are fundamental to the success and structural soundness of our framework.

5.3.2 Deliberation on Deep Network Architecture

In the quest for the most effective architecture, we explore various configurations of the image translation network, as illustrated in Fig. 5.5. Our image translator, aiming to integrate the profound feature extraction capacity of pose encoders with the content conservation of shallow networks, contrasts with the deeper structure of the Asymmetric CycleGAN baseline. Our network, alongside its baseline equivalent, incorporates 2 layers each for downsampling and upsampling, interspersed with convolutional layers. We extend our study to configurations with 3 and 4 layers of downsampling and upsampling, with findings detailed in Table 5.3 and Fig. 5.6.

Through this analysis, we observe a trend: increasing the image translation network's complexity correlates with a decline in FID scores. This decline is accentuated in Fig. 5.6, where the generated portraits display a lack of semantic congruence with the input imagery. This issue is especially pronounced with images from domains not represented in the training set, like the MetFace dataset, highlighting the shortcomings of excessively deep architectures.

In contrast, our optimal model, characterized by a shallow translator paired with a perceptive pose encoder, excels in maintaining content integrity. This balance between depth for feature extraction and shallowness for content fidelity results in outputs that are visually and semantically coherent. The introduction of elements such as the pose-based cycle consistency loss and the use of pose encodings suggests that while a 4-layer downsampling process might mirror the intricacies of our pose encoder, it does not enhance performance, as the FID scores worsen and the disparities between the generated and input images become more

pronounced in Fig. 5.6. The qualitative results indicate a disconnection from the original images, particularly when the architecture employs 4 downsampling layers. Despite aiming to produce images that resemble the training portraits, the network falls short in maintaining semantic alignment with the inputs. This illustrates a key limitation in our method; the pose-focused cycle consistency is inadequate for preserving identity when the translation network possesses extensive capacity.

Furthermore, a deeply structured network, when provided with encoded pose data, tends to over-adjust content in an effort to conform to the target distribution outlined by the training data, often resulting in subpar validation image performance. This shortfall informed our decision to employ a shallow translation structure within our translator.

Additionally, a deep architectural design's deficiency becomes evident when considering its generalization capability, or lack thereof, concerning out-of-domain images. This is strikingly apparent in Fig. 5.6, where the output in the third and final columns bear a resemblance despite the distinctly different input images in the latter, sourced from the MetFace dataset. The model's struggle to adapt to images outside its training scope is undeniable.

Hence, our comprehensive experimentation and analysis culminate in a model that skillfully marries the necessity for deep feature extraction with the imperative of preserving content. The resulting framework, with its shallow translation network and insightful pose encoder, epitomizes the efficacy of our proposed methodology in producing semantically aligned and visually coherent portraits.

5.4 Limitations

In this section, we present the failure cases of our algorithm. For example, in Figure 5.7, instances where our method does not perform optimally are worth noting. The first row in the accompanying figures illustrates the input photographs, while the second row depicts the outcomes produced by our network when processing diverse

categories of input images—these include authentic celebrity portraits, vintage photographic portraits, and portraits created using digital artistic techniques. Our methodology exhibits limitations under certain conditions, particularly when processing images where the subject's face is subjected to uneven lighting. This issue is notably evident in the results of such scenarios. Additionally, the two images positioned on the right, which fall outside the scope of our training set and feature distinct alignments, present further challenges. Alignment change is a notable limitation for generative networks, hence the compatibility of our network with out-domain images is a reputable test for our methodology. These images, deviating from the typical configurations and conditions captured within our training dataset, highlight the limitations of our model's adaptability to novel and varied image compositions. On the other hand, one can see that subjects in these images also have integrated facial hair with the face, this situation could indeed fuse poor results around the mouth of the resultant image. In summary, while our approach demonstrates competence in a broad range of scenarios, it encounters difficulties in accurately rendering portraits under atypical lighting conditions and when confronted with images that significantly diverge from the training set characteristics.

5.5 Additional Results

In this section, we demonstrate additional results with our pipeline both in-domain and out-domain test images.

5.5.1 In-domain Results

Figure 5.8 demonstrates examples of the inference images from the test set of our training CelebA dataset. The first two input images in the figure are examples of challenging small elements for our network with hat and beret, yet our results captured correct characteristics from the input images. The third and fourth images have glasses and sunglasses respectively, and our network again performed

acceptably even though sunglasses are out-domain elements. In the last image, there is a scar and blood on the face of the actor and in the resulting image there is a subtle shading over the blood-marked area, hence our method can also capture challenging small details of the input image.

5.5.2 Out-domain Results

Figure 5.9 presents results on the inference images from the test set of our training CelebA dataset. Even though both of the datasets are out-domain and very different in means of art style and detail level compared to our dataset, the resulting images in the second row contain the correct level of detail to maintain the input image style and results in the fourth row presents acceptable face details to capture human emotions even though the input images are oil painting style images. Therefore, our approach has proven successful in both in-domain and out-domain test images.

Figure 5.1: Qualitative comparison on CelebA validation dataset [48]. We undertake a comparative analysis between our approach and existing methodologies such as CycleGAN [21], Asymmetric CycleGAN [19], Image2StyleGAN [41], and DualStyleGAN [16], each having been subjected to training utilizing an identical dataset. The DualStyleGAN approach reveals shortcomings in maintaining semantic congruity. In parallel, CycleGAN, Asymmetric CycleGAN, and Img2StyleGAN are prone to the emergence of distortions upon facial regions within their outputs. In this competitive context, our proposed technique secures a notable advantage, rendering portraits of considerably enhanced quality.

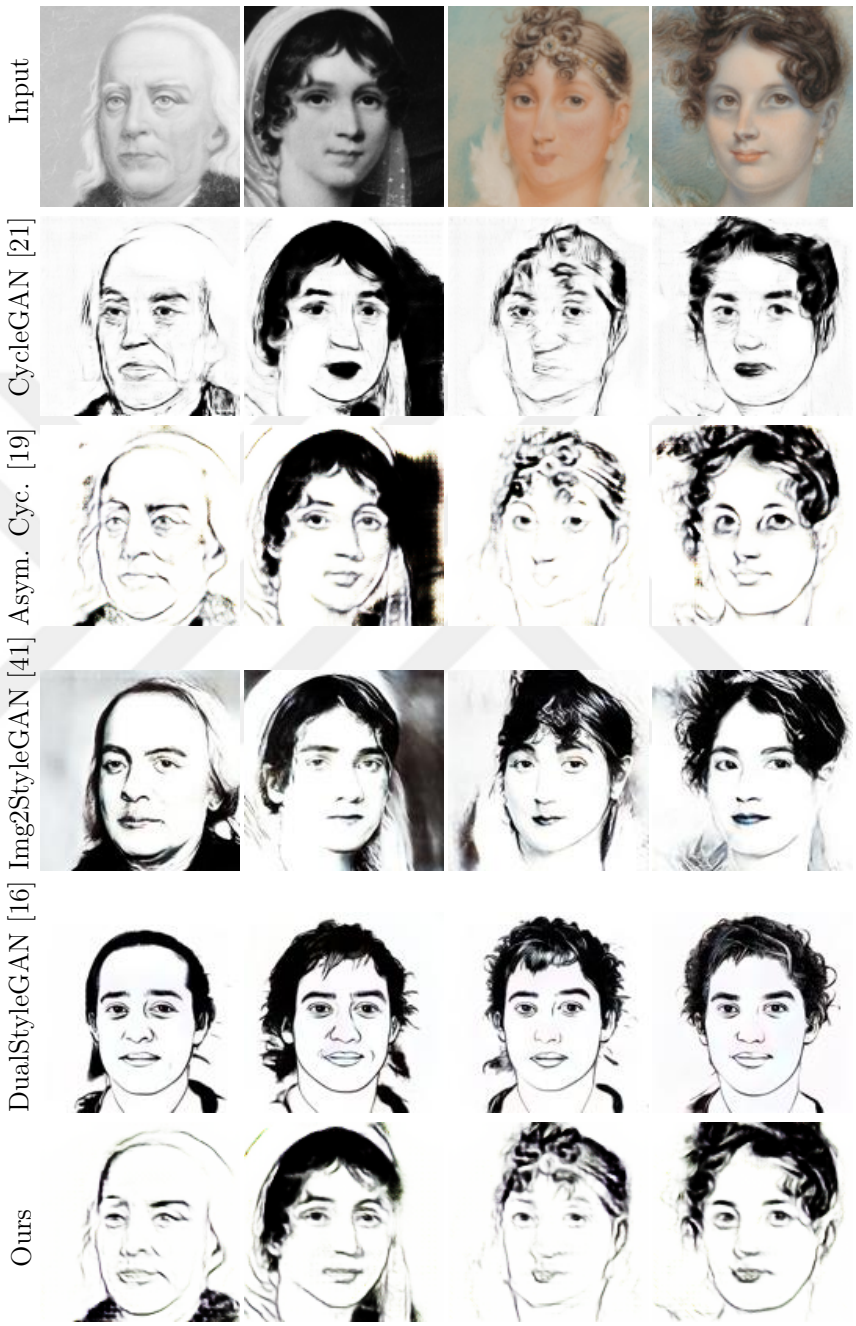


Figure 5.2: Qualitative Comparison on the MetFace Dataset [52]: In this analysis, our method is juxtaposed with several notable techniques, including CycleGAN [21], Asymmetric CycleGAN [19], Image2StyleGAN [41], and DualStyleGAN [16]. It is observed that both CycleGAN and Asymmetric CycleGAN encounter issues with artifacts in facial regions. Image2StyleGAN, while effective in certain aspects, fails to accurately reproduce the desired portrait style. DualStyleGAN, on the other hand, shows limitations in maintaining semantic alignment. In contrast, our approach demonstrates a marked improvement, yielding significantly enhanced portrait results.

Figure 5.3: Qualitative comparison on Fantasy dataset [16]. We compare our method with CycleGAN [21], Asymmetric CycleGAN [19], Image2StyleGAN [41], and DualStyleGAN [16]. DualStyleGAN suffers from semantic alignment, whereas CycleGAN, Asymmetric CycleGAN, and Img2StyleGAN suffer from artifacts on the faces. Our method achieves significantly better portraits.

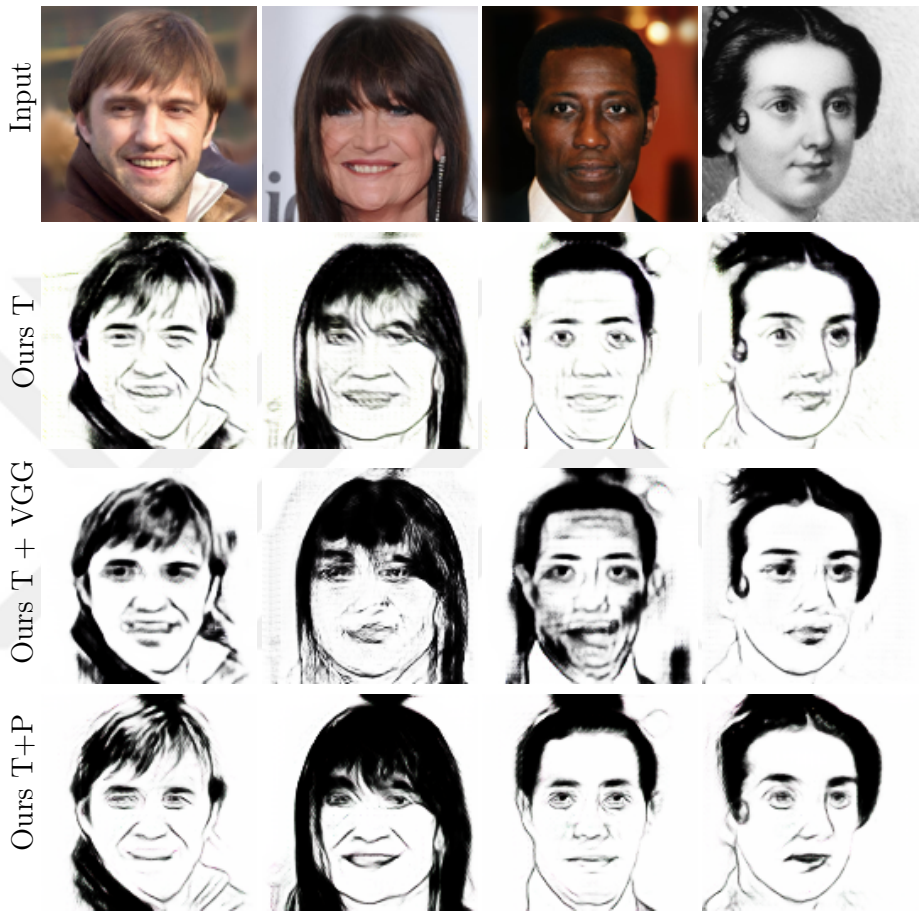
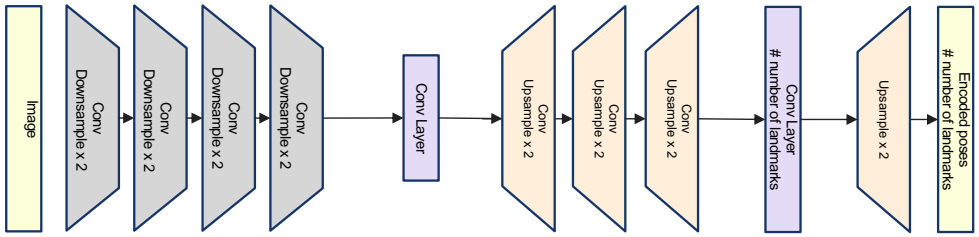
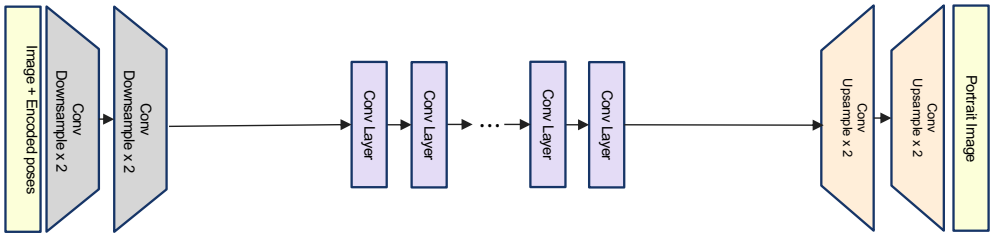


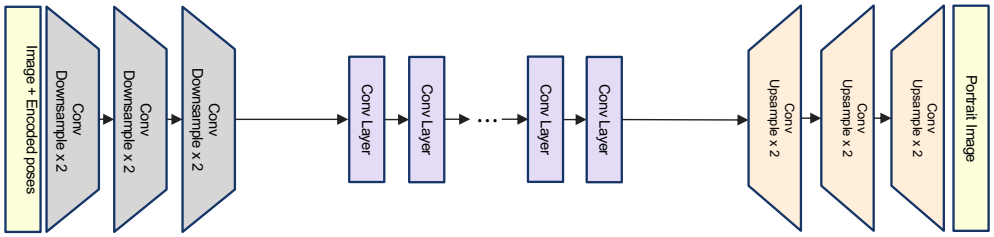
Figure 5.4: Qualitative Outcomes of the Ablation Study: This study showcases results where the initial three columns of input images originate from the CelebA dataset, with the final column representing images from the MetFace dataset. 'Deep Arch.' denotes the baseline experiments utilizing a deep architecture translator. 'Ours (T)' represents the baseline configuration integrated with our proposed translator. The 'Ours (T + VGG)' setup is developed using our architecture, complemented by a VGG-based cycle consistency loss. Meanwhile, 'Ours (T + P)' signifies our ultimate model configuration, which incorporates pose-based cycle consistency. The addition of pose cycle consistency markedly enhances the model's ability to eliminate background elements, resulting in more comprehensive depictions of certain features, such as the hair in the second and fourth columns. Collectively, this leads to a noticeable improvement in the overall results, demonstrating the effectiveness of our refinements in the model's performance, particularly in the context of detailed feature representation.



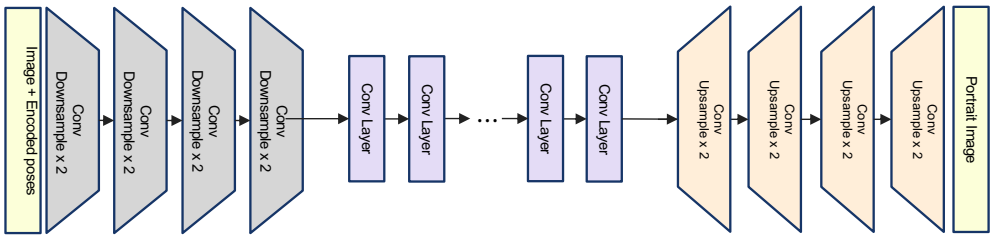
(a)Architecture of the pose encoder.



(b) Architecture of the image translation network of our final model (deep 2).



(c) Alternative architecture of the image translation network (deep 3).



(d) Alternative architecture of the image translation network (deep 4).

Figure 5.5: Architectures of the Pose Encoder and Variants of Image Translation Networks: Architectural frameworks of the pose encoder and various versions of the image translation networks, which differ in depth as utilized in our Ablation Study. Here, 'depth' is characterized by the number of downsampling layers included. Each convolutional layer in these architectures is augmented with instance normalization and ReLU layers. These layers are omitted from the illustrative figures. Comprehensive details about the individual blocks of the pose encoder and our image translation network can be found in Chapter 4. This information provides a clear understanding of the structural nuances that underpin the functionality of each network, enabling a deeper appreciation of their roles within the overall system.

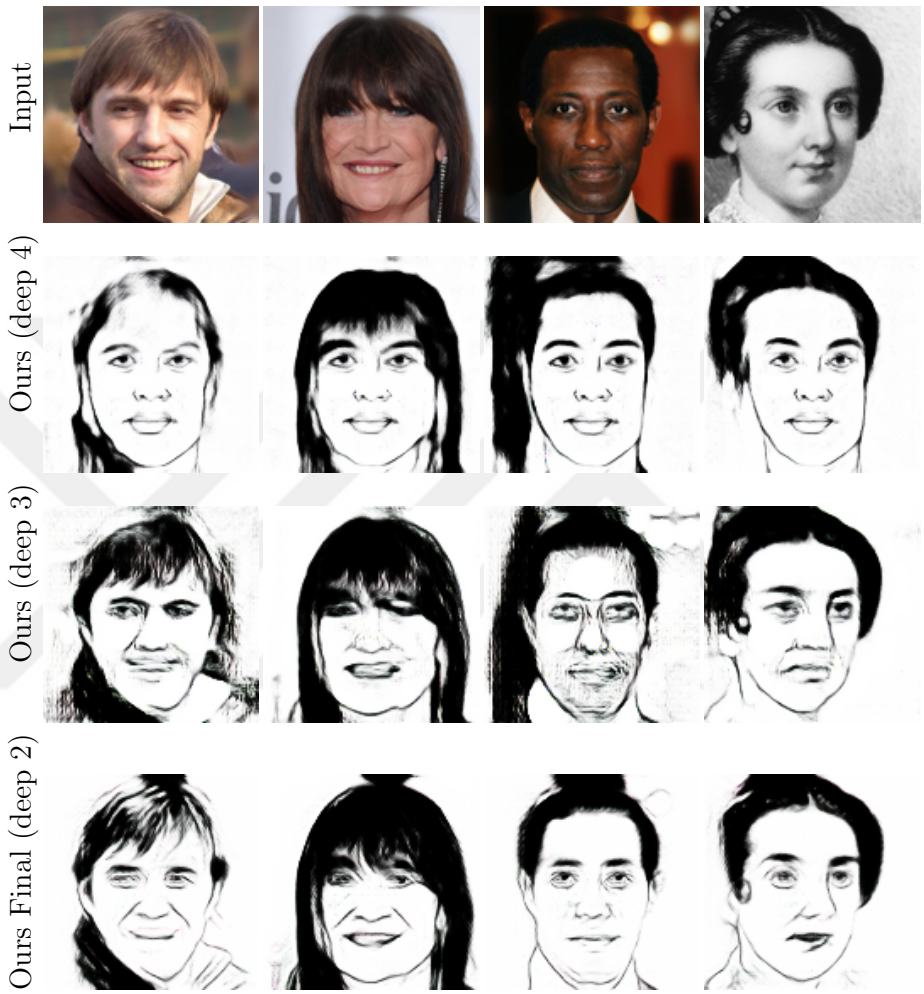


Figure 5.6: Qualitative Results of Ablation Study with Increasing Complexity in Image Translation Network: This study explores the effects of augmenting the complexity within the image translation network by enhancing the depth of the downsampling and upsampling layers in the encoder-decoder architecture. The initial three columns of input images are sourced from the CelebA dataset, with the concluding column featuring images from the MetFace dataset. It is observed that a deeper architecture, specifically with four downsampling layers, fails to maintain semantic alignment with the input images, a discrepancy that becomes increasingly pronounced with out-of-domain images, exemplified by the MetFace image in the last column. Conversely, our refined model successfully preserves content integrity by employing a shallower image translator, combined with a more profound comprehension inherent in the pose encoder. This balance between translator depth and encoder understanding is pivotal in achieving superior results, particularly in the context of content preservation and semantic alignment.

Figure 5.7: Examples of failure cases of our proposed method. The first row shows the input images, and the second row shows the result of our network on different types of input images (real celebrity portraits, old photography portraits, and digital art-drawn portraits, respectively). Our results seem to struggle when the face of the subject has uneven lighting and for the two images at the right which are not from our training set and have different alignments.

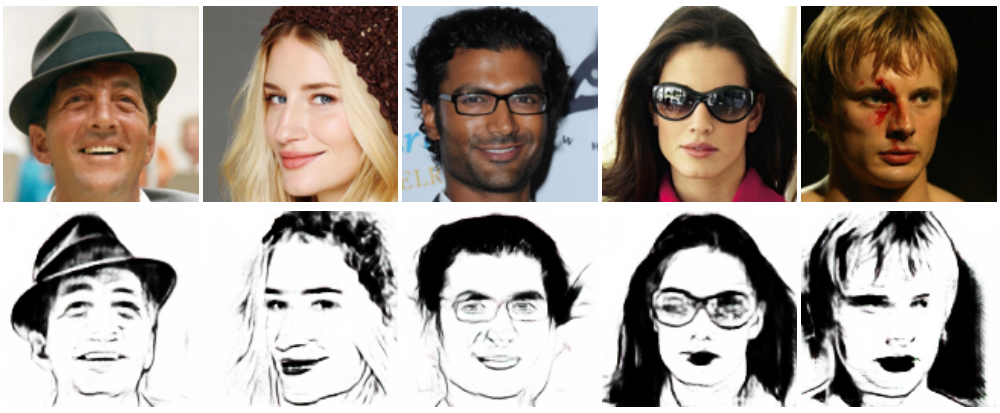


Figure 5.8: Additional in-domain test dataset (CelebA) results with our method. The first row shows the input images, and the second row shows the result of our network on input images. Our method performed acceptably in capturing small details of the various input images such as hats, berets, glasses, sunglasses, and scars.

Figure 5.9: Additional out-domain test datasets (Fantasy and MetFace) results with our method. The first and the third row shows the input images, and the second and the fourth row shows the result of our network on input images for Fantasy and MetFace datasets respectively. Even though both of the datasets are out-domain and very different in means of art style and detail level compared to our dataset, the results are satisfactory.

Chapter 6

Conclusion

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. This methodology is instrumental in facilitating a deeper comprehension of images. Our architecture, endowed with a profound capacity for high-level image interpretation, also ensures that the intrinsic identity of photos remains inviolate during the translation process. 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 methods like CycleGAN, style transfer methods like Img2StyleGAN and DualStyleGAN, as well as image translation methods that are proposed for portrait drawing like Asymmetric CycleGAN. Upon meticulous evaluation against this diverse array of strong baselines, our method emerges as a clear frontrunner, registering substantial advancements both from quantitative and qualitative standpoints. Impressively, our model's excellence transcends boundaries, demonstrating unparalleled prowess on both conventional in-domain and more intricate out-of-domain images.

Bibliography

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, pp. 1125–1134, 2017.

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

[3] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "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)*, pp. 10743–10752, 2019.

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

[5] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, pp. 469–477, 2016.

[6] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 2849–2857, 2017.

[7] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

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

[9] G. Liu, A. Dundar, K. J. Shih, T.-C. Wang, F. A. Reda, K. Sapra, Z. Yu, X. Yang, A. Tao, and B. Catanzaro, "Partial convolution for padding, inpainting, and image synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6096–6110, 2022.

[10] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8188–8197, 2020.

[11] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proceedings of the IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, pp. 4030–4038, 2017.

[12] T. Xiao, J. Hong, and J. Ma, "Elegant: Exchanging latent encodings with GAN for transferring multiple face attributes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 168–184, 2018.

[13] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative adversarial network with spatial attention for face attribute editing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 417–432, 2018.

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

[15] J. Huang, J. Liao, and S. Kwong, "Unsupervised image-to-image translation via pre-trained stylegan2 network," *IEEE Transactions on Multimedia*, vol. 24, pp. 1435–1448, 2021.

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

[17] Y. Dalva, S. F. Altindis, and A. Dundar, "VecGAN: Image-to-image translation with interpretable latent directions," in *European Conference on Computer Vision (ECCV)*, pp. 153–169, 2022.

[18] R. Yi, M. Xia, Y. J. Liu, Y. K. Lai, and P. L. Rosin, "Line drawings for face portraits from photos using global and local structure based GANs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[19] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "Unpaired portrait drawing generation via asymmetric cycle mapping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8217–8225, 2020.

[20] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, pp. 2414–2423, 2016.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 2223–2232, 2017.

[22] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," *European Conference on Computer Vision (ECCV)*, pp. 172–189, 2018.

[23] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, pp. 8789–8797, 2018.

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

[25] D. Lorenz, L. Bereska, T. Milbich, and B. Ommer, "Unsupervised part-based disentangling of object shape and appearance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10955–10964, 2019.

[26] A. Dundar, K. J. Shih, A. Garg, R. Pottorf, A. Tao, and B. Catanzaro, "Unsupervised disentanglement of pose, appearance and background from images and videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3883–3894, 2021.

[27] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "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)*, pp. 701–710, 2018.

[28] D. Engin, A. Genç, and H. Kemal Ekenel, "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, 2018.

[29] A. Bhattad, A. Dundar, G. Liu, A. Tao, and B. Catanzaro, "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, 2021.

[30] A. Dundar, J. Gao, A. Tao, and B. Catanzaro, "Fine detailed texture learning for 3D meshes with generative models," *arXiv preprint arXiv:2203.09362*, 2022.

[31] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proceedings of the IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, pp. 8798–8807, 2018.

[32] G. Kim, J. Park, K. Lee, J. Lee, J. Min, B. Lee, D. K. Han, and H. Ko, "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*, pp. 456–457, 2020.

[33] Z. Zheng, Y. Wu, X. Han, and J. Shi, "ForkGAN: Seeing into the rainy night," in *The IEEE European Conference on Computer Vision (ECCV)*, August 2020.

[34] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. Rosin, "Quality metric guided portrait line drawing generation from unpaired training data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[35] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[36] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *European Conference on Computer Vision*, pp. 694–711, 2016.

[37] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.

[38] A. Selim, M. Elgharib, and L. Doyle, "Painting style transfer for head portraits using convolutional neural networks," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–18, 2016.

[39] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4990–4998, 2017.

[40] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

[41] R. Abdal, Y. Qin, and P. Wonka, "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, 2019.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[43] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

[44] F. Sardari, B. Ommer, and M. Mirmehdi, "Unsupervised view-invariant human posture representation," *arXiv preprint arXiv:2109.08730*, 2021.

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

[46] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Patter Recognition*, pp. 586–595, 2018.

[47] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1395–1403, 2015.

[48] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

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

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, pp. 2818–2826, 2016.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, pp. 248–255, IEEE, 2009.

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

[53] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[55] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain GAN inversion for real image editing," in *European Conference on Computer Vision (ECCV)*, pp. 592–608, Springer, 2020.

[56] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Patter Recognition (CVPR)*, pp. 2287–2296, 2021.

[57] A. Dundar, M.-Y. Liu, Z. Yu, T.-C. Wang, J. Zedlewski, and J. Kautz, "Domain stylization: A fast covariance matching framework towards domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2360–2372, 2020.

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

[59] P. Shyam, K.-J. Yoon, and K.-S. Kim, "Towards domain invariant single image dehazing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 9657–9665, 2021.

[60] S. F. Altindis, Y. Dalva, H. Pehlivan, and A. Dundar, "Benchmarking the robustness of instance segmentation models," *arXiv preprint arXiv:2109.01123*, 2021.

[61] M. Xu, H. Wang, and B. Ni, "Graphical modeling for multi-source domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.