

Modern Augmented Reality: Applications, Trends, and Future Directions

Shervin Minaee*, Xiaodan Liang**, Shuicheng Yan†

*Snap Inc

**Sun Yat-sen University

†Sea AI Lab

Abstract—Augmented reality (AR) is one of the relatively old, yet trending areas in the intersection of computer vision and computer graphics with numerous applications in several areas, from gaming and entertainment, to education and healthcare. Although it has been around for nearly fifty years, it has seen a lot of interest by the research community in the recent years, mainly because of the huge success of deep learning models for various computer vision and AR applications, which made creating new generations of AR technologies possible. This work tries to provide an overview of modern augmented reality, from both application-level and technical perspective. We first give an overview of main AR applications, grouped into more than ten categories. We then give an overview of around 100 recent promising machine learning based works developed for AR systems, such as deep learning works for AR shopping (clothing, makeup), AR based image filters (such as Snapchat's lenses), AR animations, and more. In the end we discuss about some of the current challenges in AR domain, and the future directions in this area.

I. INTRODUCTION

Augmented reality (AR) is an interactive experience of real-world environments, where the objects of the real world are enhanced by computer-generated perceptual information, sometimes across multiple modalities, including visual, auditory, haptic, and somatosensory [1]. It provides an enhanced version of the real physical world. Augmented reality (AR) and virtual reality (VR) are closely related to each other, but in VR the users' perception of reality is completely based on virtual information.

Despite the huge popularity of AR in recent years, its history goes back to more than 50 years ago. Of course early AR applications were very basic, and AR technology has come a long way with a growing list of use cases in recent years. Here we provide a brief history augmented reality systems, from concepts to the new applications. In 1968, Ivan Sutherland, a Harvard professor and computer scientist, created the first head-mounted display called ‘The Sword of Damocles’. In 1974, a lab dedicated to artificial reality was created at the University of Connecticut, called “Videoplace”. The term “augmented reality” was later coined by Tom Caudell, a Boeing researcher. Later in 1992, a researcher (Louis Rosenberg) in the USAF Armstrong’s Research Lab, created ‘Virtual Fixtures’, which was one of the first fully functional augmented reality systems. This system allowed military personnel to virtually control and guide machinery to perform tasks like training their US Air Force pilots on safer flying practices. And in 1994, Julie Martin, a writer and producer, brought augmented reality to the entertainment industry for the first time with

the theater production titled Dancing in Cyberspace. In 1999, NASA created a hybrid synthetic vision system of their X-38 spacecraft. The system leveraged AR technology to assist in providing better navigation during their test flights.

AR systems started to get broader interests and more real-world applications around 2000. In 2000, Hirokazu Kato developed an open-source software library called the ARToolKit. This package helps other developers build augmented reality software programs. In 2003, Sportvision enhanced the 1st & Ten graphic to include the feature on the new Skycam system, providing viewers with an aerial shot of the field with graphics overlaid on top of it. In 2009, Esquire Magazine used augmented reality in print media for the first time in an attempt to make the pages come alive. In 2013, Volkswagen debuted the MARTA app which primarily gave technicians step-by-step repair instructions within the service manual. In 2014, Google unveiled its Google Glass devices, a pair of augmented reality glasses that users could wear for immersive experiences. In 2016, Microsoft started shipping its version of wearable AR technology called the HoloLens. In 2017, IKEA released its augmented reality app called IKEA Place that was a new experience in the retail industry. Also during past few years, Snapchat has introduced several AR lenses in their apps, which have made image and video communications much more fun.

Hardware components for AR includes a processor, display, sensors and input devices. Modern mobile computing devices like smartphones and tablet computers contain these elements, making them suitable AR platforms. In terms of display, various technologies are used in AR rendering, including optical projection systems, monitors, handheld devices, and display systems, which are worn on the human body. A head-mounted display (HMD) is a display device worn on the forehead, such as a harness or helmet-mounted. AR displays can be rendered on devices resembling eyeglasses (such as Google Glass, and Snapchat's new Spectacles).

Some of the popular tools for developing augmented reality related solutions includes: ARKit developed by Apple and used by iOS developers to build mobile AR apps and games for iPhones, iPads, and other Apple devices, ARCore developed by Google and has many features that help integrate AR elements into the real environment, including motion tracking, surface detection, and lighting estimation (supports development in Android, iOS, Unreal, and Unity), SnapML and Lens Studio developed by Snap and used by the lens developers for Snapchat app, echoAR (a cloud platform for augmented reality and virtual reality), Unity, SparkAR, Vuforia, Wikitude, and ARToolKit.

In this work we provide a high level review of modern augmented reality from both application and technical perspectives. We first provide an overview of the main current applications of augmented reality, grouped into more than 10 categories. We then provide an overview of the recent machine learning based algorithms developed for various AR applications (such as clothing, make-up try on, face effects). Most of these works are based on deep learning models. We also mention the popular public benchmarks for each of those tasks, for cases where a public dataset is available. After that, we provide a detailed section on the main challenges of AR systems, and some of the potential future directions in AR domain, for the young researchers in this area. The main AR applications discussed in this paper includes:

- 1) Games
- 2) Social Networks and Communications
- 3) Education
- 4) Healthcare
- 5) Shopping
- 6) Automotive Industry
- 7) Television and Music Industry
- 8) Art and Museum Galleries
- 9) Constructions
- 10) Advertisement and Financial Companies
- 11) Other Areas (Archaeology, Industrial Manufacturing, Commerce, Literature, Fitness and Sport Activities, Military, and Human Computer Interaction)

The structure of the rest of this paper is as follows: In Section II, we review some of the prominent AR applications, grouped into several categories. In Section III, we provide an overview of the prominent Machine/Deep learning based models developed for AR applications. In Section IV, some of the challenges of the current AR systems, and some of the potential future directions in AR areas are discussed. In the end, we conclude this paper in Section V.

II. CURRENT APPLICATIONS

With the rising popularity of augmented reality in recent years, it has been used in more and more new applications everyday, which makes it hard to list all possible AR applications here. Instead, we try to cover the main applications of AR in today's world, grouped into several categories. We review their high-level applications in this section, and leave the technical/modeling part of those works for the next section.

A. Games

Gaming is bigger than it has ever been, driven by the growth of mobile gaming, and now makes up 20-26 percent of all media consumption hours. AR gaming is the integration of visual and audio content of the game with the user's environment in real time. Unlike virtual reality gaming, which usually requires a separate room or confined area to create an immersive environment, augmented reality gaming uses the existing environment and creates a playing field within it, which makes it simpler for both users and developers. An augmented reality game often superimposes a pre-created environment on top of a user's actual environment.

Some of the prominent AR gaming apps includes **Pokémon GO** (which uses a smartphone's camera, gyroscope, clock

and GPS and to enable a location-based augmented reality environment) shown in Fig 1, **Jurassic World Alive** (which brings dinosaurs into the real world and players can head out in search of the prehistoric monsters and capture them), **Harry Potter: Wizards Unite** (that sets players to walk around in the real world and collect various wizarding items, battle with foes and deal with a calamity that has hit wizards and witches across the world), **The Walking Dead: Our World** (that the undead zombies from the popular television series out of the TV screen and into our surrounding environment).



Fig. 1. A sample snapshot of Pokemon app, for finding the Pokemon from within the tall grass. Courtesy of Pokemon App.

B. Social Networks and Communications

Augmented Reality is one of the trending additions to the social networks and communication applications features. AR can make communications with friends and celebrities more entertaining. As an example, Snapchat provides various AR lenses for people, from simply adding hats/horns/eyeglasses to making popular landmarks move (some examples shown in Fig 2)

The AR effects are also used in image and video communication tools, such as Zoom, Microsoft Teams, and Google Meet (with the help of SnapCamera), in which people can augment their videos during a meeting by applying various AR effect.

C. Education

Augmented reality is great material for education and learning/training platforms. It can be used to make the education and training platforms more engaging and fun. Children often enjoy learning new experiences and technology, so AR can motivate students to learn and make the classes more entertaining and engaging. AR based platforms for education have been in huge demand after the COVID-19 pandemic, which shifted most of the education systems to the remote phase. Since AR has become more accessible and affordable, instead of buying physical supplies, AR may be more cost-effective for schools in the future.

As an example, in 4D Anatomy [3], students can explore more than 2,000 anatomical structures and discover 20 different dissection specimens of real anatomy. They can improve understanding anatomy by manipulating and observing virtual 3D objects from different angles.



Fig. 2. The AR effect made by Snapchat on Eiffel Tower, and White House. Courtesy of TheVerge [2].

D. Healthcare

Medical and healthcare industry are another place which augmented reality can be very effective and useful. AR is already used in simulations for surgeries and diseases to enhance patient treatments. It is also used in education for patient and doctor. But their potential scope could go well beyond these.

One prominent AR based solution in healthcare is AccuVein [4], which uses projection-based AR that scans a patient's body and shows doctors the exact location of veins. This leads to improvement in the injection of the vein on the first stick by 3.5 times and reduces the need to call for assistance by 45%.

E. Shopping

Shopping is perhaps one of the main areas in which AR can have a huge impact. With the advent of e-commerce, some retail stores have already adopted the newest of AR technologies to enhance the customer's shopping experience to get an edge over other stores. They have transformed the whole experience of a shopper from entering a store to opening the final product at home in unimaginable ways.

AR application in shopping is very broad, from virtual clothing try-on (either on the app, or using the in-store magic mirror), virtual makeup try-on, to virtual in-store navigation. Some of the AI and machine learning applications in this space, includes techniques for clothing/make-up trying, object understanding, human parsing, object segmentation, size estimation, scene understanding, and many more.

Some of the popular apps which are using AR for shopping includes, Home Depot (which expanded the functionality in



Fig. 3. A woman tries on virtual garments using virtual mirror in-store. Courtesy of [5].



Fig. 4. User experience with a sample virtual makeup augmented reality system. Courtesy of [6].

its main mobile app to allow users to overlay Home Depot merchandise on any room in their home), IKEA place (this app take a picture of your living room, and automatically measures the space, then it provides recommendations on furniture that fits in the space), Wayfair, Target, Sephora (this app contains a Virtual Artist, that uses facial recognition technology to allow customers to digitally try on products), Nike (the app has Nike Fit feature that allows customers to find their true shoe size), Warby Parker (its app allows customers to digitally try on glasses from the comfort of the customer's home), Amazon, and many more.

F. Automotive Industry

Augmented reality in the automotive industry is in high demand, and car manufacturers are planning to incorporate AR into cars in near future. AR in automobiles is expected to have a value of more than \$600 billion by 2025.

As a prominent example, Nissan has developed Invisible-to-Visible (I2V) solution using AR and AI [7], which makes drivers aware of potential hazards like nearby objects, and redirect drivers' focus to the road if they are not concentrated. AR can be very helpful for safety because it may decrease the number of accidents and drivers can drive comfortably.

G. Television and Music Industry

Another big AR applications are in TV and Music industries. It can help the producer to enrich their contents by providing more information about the show, program, music, and creators.

AR has already been used in various TV programs for while. As an example, when you're watching a show on the television, you may receive additional information about it, e.g. for a baseball match you receive match scores, player information, and related information. There could also be some pointers showing the position of some objects in sport games, such as balls, or players.

Also music has been transforming a lot recently, and music is more than just listening to some favorite tracks put together in playlists. AR can help us grab information like the artist bio, cover up videos, dance videos on the track and so much more. It can help us enhance live performance streaming events by telling us a story which couldn't have been possible without AR.

H. Art and Museum Galleries

AR can make seeing an artwork or a museum much easier for people, and help people overcome location/distance barrier.

We're seeing more art galleries incorporating AR experiences. In December 2017, the first ever AR-powered art exhibition by Perez Art Museum Miami (PAMM), was released. Another popular app along this use case is the Civilizations app by BBC. The app creators gathered more than 280 artifacts from famous museums and galleries and turned them into 3D models. This app allows exploring artifacts in exhibitions and learning their history and specific details. One example is shown in Fig 5.



Fig. 5. An example of AR experience with Civilizations app. Courtesy of [8].

In addition to the museums, many artists have come ahead with AR mobile apps that let users around the world view their artwork the way it is meant to be seen. This can help artists to better promote their artworks, and make it accessible to more people around the world.

I. Constructions

AR has several applications in construction areas, and has been already used by many of the biggest construction firms around the world. Its applications ranges from simple use cases such as safety training of the works, to more advanced use-cases such as team collaboration, real-time project information, and project planning and building modeling. With the help of AR technology an empty shell of a building floor can come

to life with the location, style and size of windows and doors, pipes, and HVAC systems. Using an AR headset, the worker sees these details as if they were right in front of them; they can compare what they see to the building plan to ensure everything is in order.

AR can also be used to showcase 3D models and even provide tours, giving clients a solid idea of what a building would look like before it's built. If an owner wants to show the client what a new installation would look like on-site, AR can also bring that vision to life.

J. Advertisement and Financial Companies

Since advertisement remains one of the biggest source of revenue for tech companies, many of them are using AR to produce a more engaging and informative ad, to lure customer to buy different products. AR allows brands to interact with your customers by giving them a 'free taste' of the product before making a purchase.

Augmented reality trends in banking aim to help consumers keep track of their finances better. AR in banking offers a rich visualization of their data and other services. As an example, Wells Fargo designed and built an AR system for consumers to interact with bank tellers within a virtual space placed over reality. Moreover, it comes with gamification like AR games and puzzles.

In addition to the financial companies, insurance companies are also adopting AR. Through the use of AR, insurance companies can better communicate and explain their service to their customers, and help them. As an example, Allianz uses AR to make its customers aware of possible dangers within their homes. Using their smartphone, they can see such hazards. These range from an overheating toaster to crashed upstairs bathroom floors due to sink flooding, and much more.

K. Other Areas

It is obvious that the AR applications are not limited to the above items, and AR can be useful in many more areas. In addition to the areas listed above, AR has applications in Archaeology (to augment archaeological features onto the modern landscape), industrial manufacturing, Commerce, Literature (as an example AR was blended with poetry by ni ka from Sekai Camera in Tokyo, Japan), fitness and sport activities, and human computer interaction.

III. POPULAR DEEP LEARNING BASED MODELS

In this section, we are going to review some of the recent prominent machine/deep learning algorithms developed for various AR applications. Many of the deep learning based models for AR applications are focused on:

- AR for Shopping (clothing try-on, makeup try-on)
- AR for Face/Body Transformations
- Tracking and Pose Estimation for AR Applications
- AR for 3D human Reconstruction
- Geometry Applications
- Scene Understanding and Reconstruction

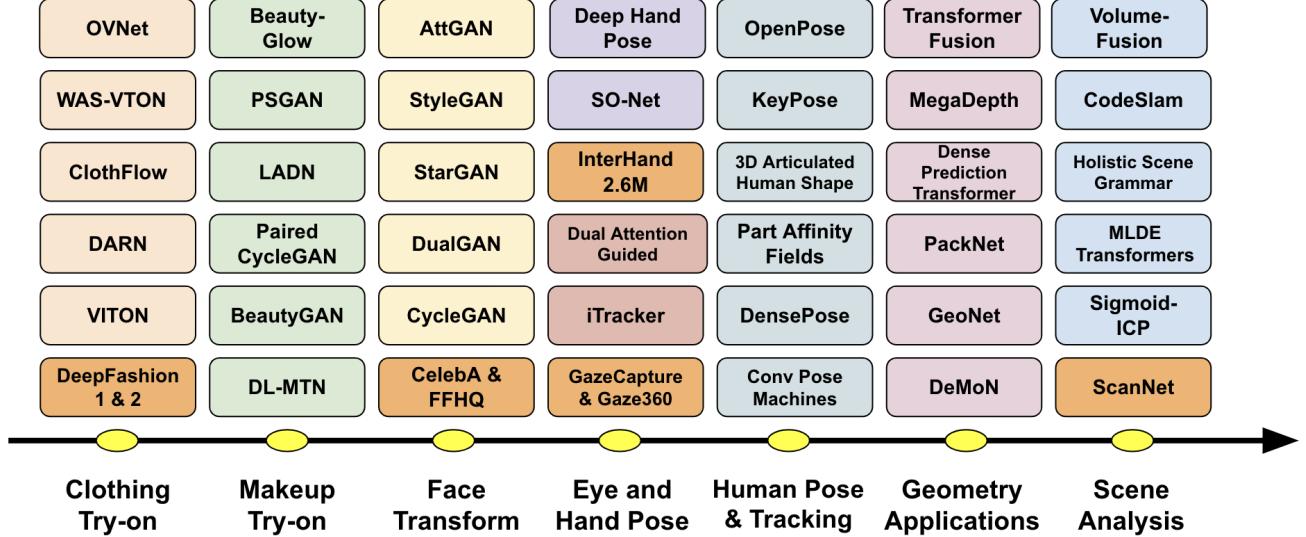


Fig. 6. An illustration of some of the most representative deep learning works related to various AR applications. Those blocks shown in dark orange refer to potential dataset to use for the corresponding task.

Fig 6 provides an illustration of a subset of the representative works, which are going to be discussed in the following sections.

A. Models for Clothing Shopping and Try On

In this section we will provide an overview of some of the recent works for clothing retrieval and try-on. We first cover some of the prominent works for clothing retrieval/matching, and then discuss about the models developed for clothing try-on.

Matching a real-world clothing/garment to the same item in an online shopping website could be the first step in finding a desired garment (one example shown in Fig 7. This is an extremely challenging task due to visual differences between street photos (pictures of people wearing clothing in everyday uncontrolled settings) and online shop photos (pictures of clothing items on models, mannequins, or in isolation, captured by professionals in more controlled settings).



Fig. 7. An example of clothing matching. Courtesy of [9].

In [9], Kiapour et al. collected a dataset for this application containing 404,683 shop photos collected from 25 different online retailers and 20,357 street photos, providing a total of 39,479 clothing item matches between street and shop photos, and developed three different methods for Exact Street to Shop retrieval, including two deep learning baseline methods, and a method to learn a similarity measure between the street and shop domains. The overview of their proposed model is shown in Fig 8.

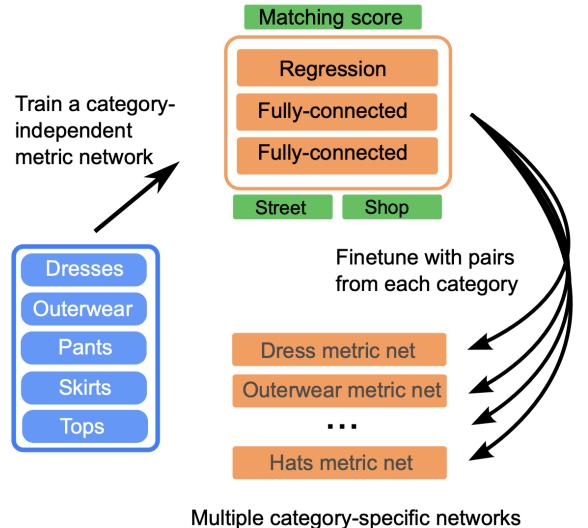


Fig. 8. The overview of the training, followed by fine-tuning procedure for training category-specific similarity for each category in [9]. Courtesy of [9].

In [10], Liu et al. introduced DeepFashion1, a large-scale clothes dataset with comprehensive annotations. It contains over 800,000 images, which are richly annotated with massive attributes, clothing landmarks, and correspondence of images taken under different scenarios including store, street snapshot, and consumer. Fig 9, shows some of the sample images from this dataset. To demonstrate the advantages of DeepFashion, they proposed a new deep model, namely FashionNet, which learns clothing features by jointly predicting clothing attributes and landmarks. FashionNet pipeline is shown in Fig 10.

In [11], Dong et al. developed a deep learning framework capable of model transfer learning from well-controlled shop clothing images collected from web retailers to in-the-wild images from the street. They formulated a novel MultiTask Curriculum Transfer (MTCT) deep learning method to explore



Fig. 9. Example images of different categories and attributes in DeepFashion. The attributes form five groups: texture, fabric, shape, part, and style. Courtesy of [10].

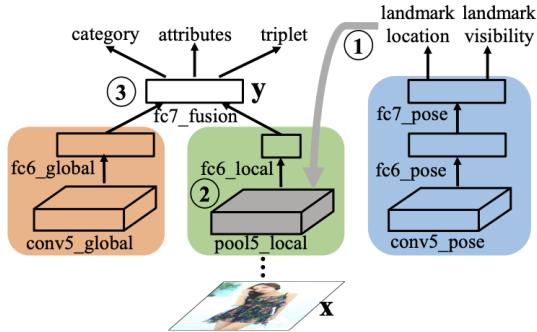


Fig. 10. Pipeline of FashionNet, which consists of global appearance branch (in orange), local appearance branch (in green) and pose branch (in blue). Shared convolution layers are omitted for clarity. Courtesy of [10].

multiple sources of different types of web annotations with multi-labelled fine-grained attributes. The architecture of the proposed framework is shown in Fig 11.

In [12], Han et al. proposed an image-based Virtual Try-On Network (VITON) without using 3D information in any form, which seamlessly transfers a desired clothing item onto the corresponding region of a person using a coarse-to-fine strategy. Conditioned upon a new clothing-agnostic yet descriptive

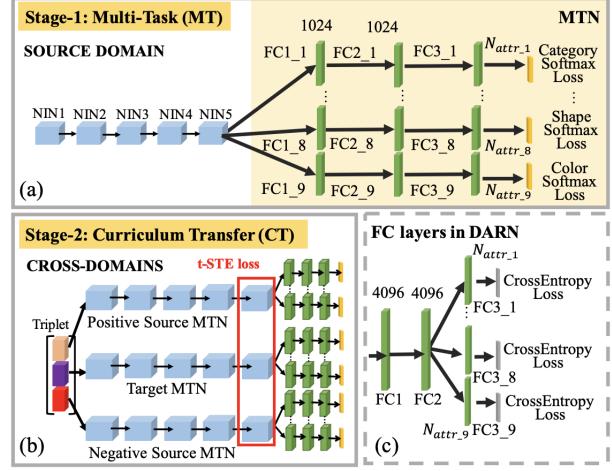


Fig. 11. (a),(b) show the MTCT network design. (c) illustrates the FC layers of DARN. Courtesy of [11]

person representation, this framework first generates a coarse synthesized image with the target clothing item overlaid on that same person in the same pose. They further enhance the initial blurry clothing area with a refinement network. This network is trained to learn how much detail to utilize from the target clothing item, and where to apply to the person in order to synthesize a photo-realistic image in which the target item deforms naturally with clear visual patterns. The architecture of this framework is shown in Fig 12.

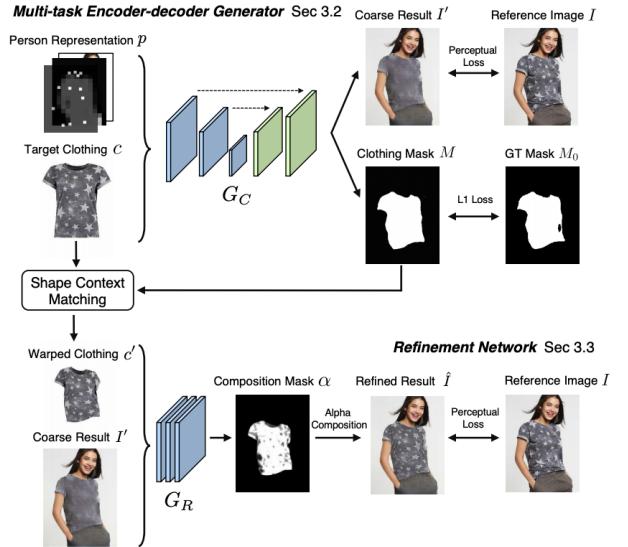


Fig. 12. An overview of VITON. VITON consists of two stages: (a) an encoder-decoder generator stage, and (b) a refinement stage. Courtesy of [12]

In [13], Ge et al. presented DeepFashion2, which is new dataset for clothing. It is a versatile benchmark of four tasks including clothes detection, pose estimation, segmentation, and retrieval. It has 801K clothing items where each item has rich annotations such as style, scale, viewpoint, occlusion, bounding box, dense landmarks (e.g. 39 for ‘long sleeve outwear’ and 15 for ‘vest’), and masks. There are also 873K Commercial-Consumer clothes pairs. The comparison between

DeepFashion and DeepFashion2 for some images is shown in Fig 13.



Fig. 13. Comparisons between (a) DeepFashion and (b) DeepFashion2. (a) only has single item per image, which is annotated with 4-8 sparse landmarks. The bounding boxes are estimated from the labeled landmarks, making them noisy. In (b), each image has minimum single item while maximum 7 items. Each item is manually labeled with bounding box, mask, dense landmarks (20 per item on average), and commercial-customer image pairs. Courtesy of [13]

In [14], Han et al. presented ClothFlow, an appearance-flow-based generative model to synthesize clothed persons for pose-guided person image generation and virtual try-on. By estimating a dense flow between source and target clothing regions, ClothFlow effectively models the geometric changes and naturally transfers the appearance to synthesize novel images. They achieve this with a three stage framework: 1) Conditioned on a target pose, they first estimate a person semantic layout to provide richer guidance to the generation process. 2) Built on two feature pyramid networks, a cascaded flow estimation network then accurately estimates the appearance matching between corresponding clothing regions. The resulting dense flow warps the source image to flexibly account for deformations. 3) Finally, a generative network takes the warped clothing regions as inputs and renders the target view. The architecture of ClothFlow framework is shown in Fig 14.

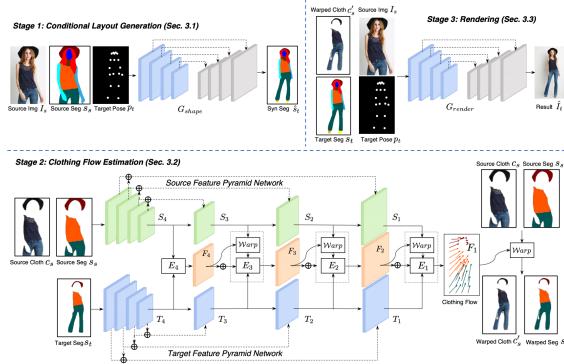


Fig. 14. Overview of the proposed ClothFlow architecture. Courtesy of [14]

In [15], Xie et al. proposed WAS-VTON that employs the Neural Architecture Search (NAS) to explore the garment-category-specific warping network and the optimal garment-

person fusion network for the virtual try-on task. To meet this end, WAS-VTON introduces NAS-Warping Module and NAS-Fusion Module, each of which is composed of a network-level (i.e., with different network architecture) and an operation-level (i.e., with different convolution operations) search space. Specifically, the search space of NAS-Warping Module covers various sub-networks with different warping ability which is defined by the number of warping blocks within each warping cell, while the search space of NAS-Fusion Module consists of various sub-networks with skip connections between different scale features. Furthermore, to support two searchable modules, WAS-VTON introduces Partial Parsing Prediction to estimate the semantic labels of the replaced region in the try-on result. Finally, WAS-VTON applies the one-shot framework in [16] to separately search the category-specific network for garment warping, and search the optimal network with particular skip connection for garment-person fusion. The search space of each module and the overall framework of WAS-VTON are shown in Fig 15 and Fig 16, respectively.

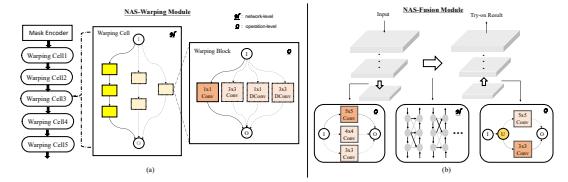


Fig. 15. The search space for the NAS-Warping Module and NAS-Fusion Module of WAS-VTON. Courtesy of [15].

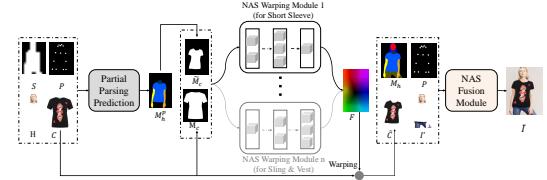


Fig. 16. The overall framework of WAS-VTON. Courtesy of [15].

In [17], Neuberger et al. presented a new image-based virtual try-on approach (Outfit-VITON) that helps visualize how a composition of clothing items selected from various reference images form a cohesive outfit on a person in a query image. Fig 17 illustrates the high-level ideal of this framework. The O-VITON framework has three main steps. The first shape generation step generates a new segmentation map representing the combined shape of the human body in the query image and the shape feature map of the selected garments, using a shape auto-encoder. The second appearance generation step feed-forwards an appearance feature map together with the segmentation result to generate an a photo-realistic outfit. An online optimization step then refines the appearance of this output to create the final outfit. This is shown in Fig 18.

In [19], Li et al. proposed Outfit Visualization Net (OVNet) to capture these important details (e.g. buttons, shading, textures, realistic hemlines, and interactions between garments) and produce high quality multiple-garment virtual try-on images. OVNet consists of 1) a semantic layout generator and 2) an image generation pipeline using multiple coordinated warps. We train the warper to output multiple warps using a cascade loss, which refines each successive warp to focus



Fig. 17. The O-VITON algorithm is designed to synthesize images that show how a person in a query image is expected to look with garments selected from multiple reference images. Courtesy of [18]

on poorly generated regions of a previous warp and yields consistent improvements in detail. In addition, they introduce a method for matching outfits with the most suitable model and produce significant improvements for both our and other previous try-on methods. The high-level architecture of OVNet is shown in Fig 19.

In [20], Zhao et al. presented a Monocular-to-3D Virtual Try-On Network (M3D-VTON), which also takes as input an in-shop clothing image and a person image, but output the 3D try-on mesh instead of image in 2D space (Fig 20). By efficiently exploiting the merits of 2D non-rigid deformation and 3D non-parametric body estimation, this work successfully recovers high-fidelity 3D try-on result, being as well much faster than pure 3D methods. Fig 21 depicts its overall architecture that contains three modules, namely the preparatory Monocular Prediction Module (MPM), the Texture Fusion Module (TFM) to generate try-on texture and the Depth Refinement Module (DRM) to estimate the fine-detail body depths. M3D-VTON highlights the first attempt to bridge the 2D try-on and 3D human reconstruction, leading to an effective solution of 3D try-on problem in a novel view.

Some of the other promising works for AR shopping for clothing includes, SwapNet [21], Learning-based animation of clothing for virtual try-on [22], GarNet: A two-stream network for fast and accurate 3d cloth draping [23], 360-degree textures of people in clothing from a single image

[24], M2e-try on net: Fashion from model to everyone [25], Fw-gan: Flow-navigated warping gan for video virtual try-on [26], LA-VITON: a network for looking-attractive virtual try-on [27], Fashion++: Minimal edits for outfit improvement [28], TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style [29], ViBE: Dressing for diverse body shapes [30], Cloth Interactive Transformer for Virtual Try-On [31], VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization [32], Parser-Free Virtual Try-on via Distilling Appearance Flows [33], and Complementary Transferring Network (CT-Net) [34].

B. Models for Makeup Try On

There have been several deep learning based frameworks proposed for make-up try-on. Here we provide an overview of some of the most popular ones.

In [35], Liu et al. proposed a novel Deep Localized Makeup Transfer Network to automatically recommend the most suitable makeup for a female and synthesize the makeup on her face. Given a before-makeup face, her most suitable makeup is determined automatically. Then, both the before makeup and the reference faces are fed into the proposed Deep Transfer Network to generate the after-makeup face. The makeup recommendation for one sample image is shown in Fig 22.

The proposed Deep Localized Makeup Transfer Network contains two sequential steps. (i) the correspondences between the facial part (in the before-makeup face) and the cosmetic (in the reference face) are built based on the face parsing network. (ii) Eye shadow, foundation and lip gloss are locally transferred with a global smoothness regularization. The high-level architecture of this work is shown in Fig 23.

In [36], Alashkar et al. developed a fully automatic makeup recommendation system and proposed a novel examples-rules guided deep neural network approach. The framework consists of three stages. First, makeup-related facial traits are classified into structured coding. Second, these facial traits are fed into examples-rules guided deep neural recommendation model which makes use of the pairwise of Before-After images and the makeup artist knowledge jointly. Finally, to visualize the recommended makeup style, an automatic makeup synthesis system is developed as well. Fig 24 illustrates the block-diagram of the proposed model by this work.

In [37], Li proposed an instance-level facial makeup transfer with generative adversarial network, called BeautyGAN. Some of the sample result generated by this framework are shown in Fig 25. They first transfer the non-makeup face to the makeup domain with a couple of discriminators that distinguish generated images from domains' real samples. On the basis of domain-level transfer, they achieve instance-level transfer by adopting a pixel-level histogram loss calculated on different facial regions. To preserve face identity and eliminate artifacts, they also incorporate a perceptual loss and a cycle consistency loss in the overall objective function. The overall architecture of this framework is shown in Fig 25.

In [38], Chang et al. introduced an automatic method for editing a portrait photo so that the subject appears to be wearing makeup in the style of another person in a reference

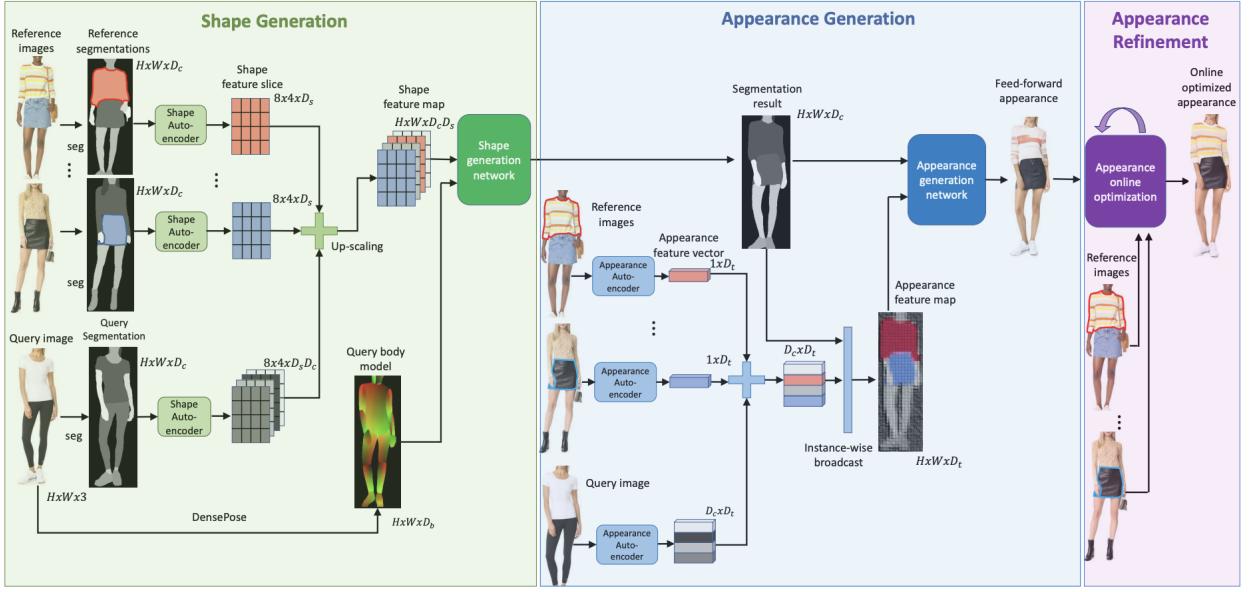


Fig. 18. The model architecture of O-VITON framework. Courtesy of [18]

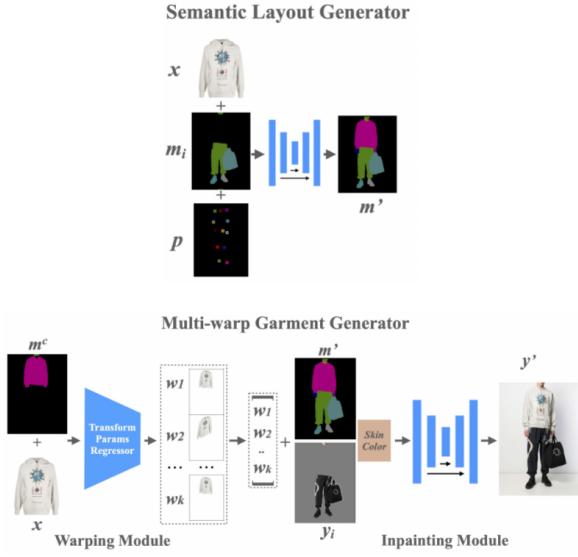


Fig. 19. The model architecture of Outfit Visualization Net. Courtesy of [19]

photo. The proposed unsupervised learning approach relies on cycle-GAN. Different from the image domain transfer problem, this style transfer problem involves two asymmetric functions: a forward function encodes example-based style transfer, whereas a backward function removes the style. They constructed two coupled networks to implement these functions – one that transfers makeup style and a second that can remove makeup – such that the output of their successive application to an input photo will match the input. Fig 26 shows some of the sample makeup transferred images by this model. For each image, they applied face parsing algorithm to segment out each facial component. And they trained three generators and discriminators separately for eyes, lip and skin considering the unique characteristics of each regions. This is



Fig. 20. 3D try-on results from M3D-VTON. Courtesy of [20]

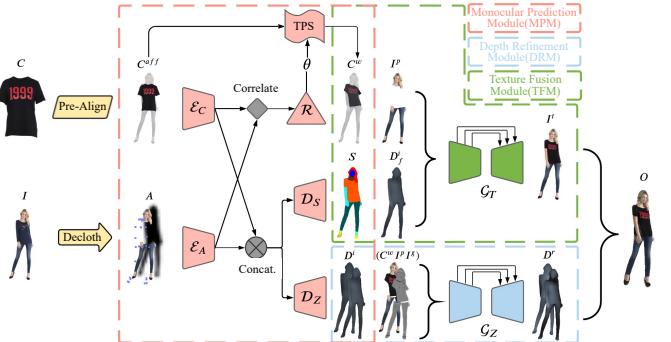


Fig. 21. Overview of M3D-VTON architecture. Courtesy of [20]

shown in Fig 27.

In [39], Gu et al. proposed a local adversarial disentangling network (LADN) for facial makeup and de-makeup. Central to their method are multiple and overlapping local adversarial discriminators in a content-style disentangling network for achieving local detail transfer between facial images, with the use of asymmetric loss functions for dramatic makeup styles with high-frequency details. Fig 28 shows the result of this framework on two sample images. The high-level structure of the generator part of LADN is shown in Fig 29.

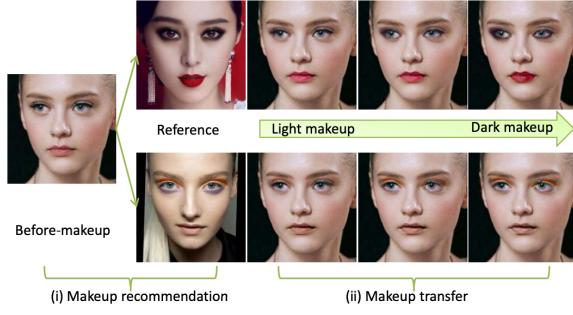


Fig. 22. The makeup recommendation and synthesis for an example image by DL-MTN. Courtesy of [35]

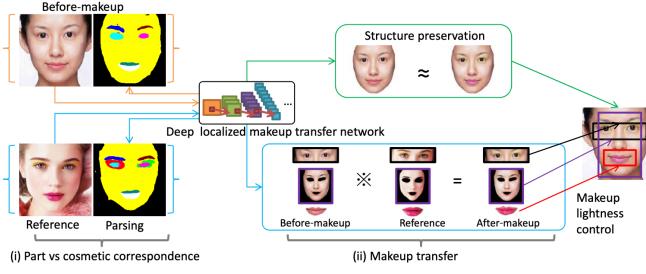


Fig. 23. The model architecture of Deep Localized Makeup Transfer Network. Courtesy of [35]

In [40], Jiang et al. tried to address the issues with previous existing methods for facial makeup transfer, which transferring between images with large pose and expression differences, and also not being able to realize customizable transfer that allows a controllable shade of makeup or specifies the part to transfer, which limits their applications. They proposed Pose and expression robust Spatial-aware GAN (PSGAN). It first utilizes Makeup Distill Network to disentangle the makeup of the reference image as two spatial-aware makeup matrices. Then, Attentive Makeup Morphing module is introduced to specify how the makeup of a pixel in the source image is morphed from the reference image. The model architecture of PSGAN framework is shown in Fig 30.

In [41], Nguyen et al. proposed a holistic makeup transfer framework that can handle all the mentioned makeup components. It consists of an improved color transfer branch and a novel pattern transfer branch to learn all makeup properties, including color, shape, texture, and location. To train and evaluate such a system, we also introduce new makeup datasets for real and synthetic extreme makeup. Fig 31 shows the high level architecture of the proposed framework.

Some of the other promising works for virtual makeup try-on includes: makeup removal via bidirectional tunable de-makeup network [42], face beautification: Beyond makeup transfer [43], BeautyGlow [44], face beautification via dynamic skin smoothing, guided feathering, and texture restoration [45], and weakly supervised color aware GAN for controllable makeup transfer [46].

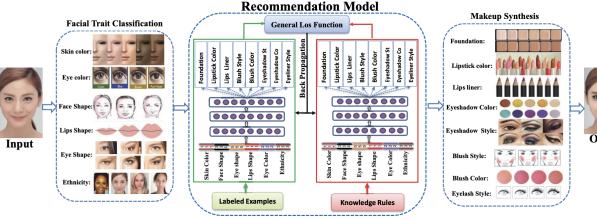


Fig. 24. The model architecture of the Examples-Rules Guided DNN for makeup recommendation. Courtesy of [36]

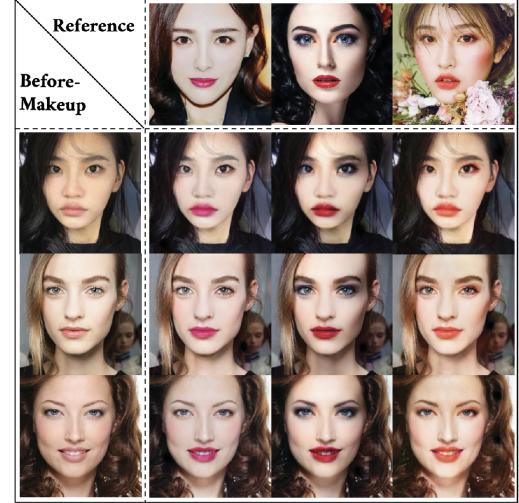


Fig. 25. Example results of our BeautyGAN model for makeup transfer. Courtesy of [37]

C. Models for Face/Body Transformations

Face style transfer or (face transformation) is another active research area, with huge applications in social media such Snapchat Lenses, Instagram Filters, TikTok lenses/effects. Although the algorithm used by those companies is not known, there are several research works which have developed algorithms for applying various effects on faces. Since acquiring paired training data for face transformation is not very easy in most cases, we are going to mostly focused on algorithms which would work in an unpaired fashion here.

In [47], Zhu et al. presented an approach for learning to translate an image from a source domain X to a target domain Y in the absence of paired examples. Their goal is to learn a mapping $G : X \rightarrow Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution Y using an adversarial loss. Because this mapping is highly under-constrained, they coupled it with an inverse mapping $F : Y \rightarrow X$ and introduce a cycle consistency loss to push $F(G(X)) \approx X$ (and vice versa). The high-level idea of CycleGAN framework is shown in Fig 32. Some of the sample images generated via CycleGAN model are shown in Fig 33.

In [48], Yi et al. developed dual-GAN mechanism, which enables image translators to be trained from two sets of unlabeled images from two domains. In their architecture, the primal GAN learns to translate images from domain U to those in domain V , while the dual GAN learns to invert the task. The



Fig. 26. Three source photos are each modified to match makeup styles of three reference photos to produce nine different outputs. Courtesy of [38]

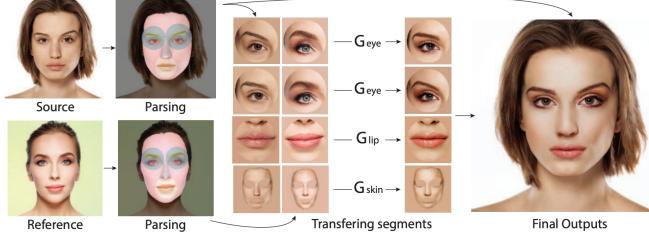


Fig. 27. Illustration of generator per segment in PairedCycleGAN. Courtesy of [38]

closed loop made by the primal and dual tasks allows images from either domain to be translated and then reconstructed. Hence a loss function that accounts for the reconstruction error of images can be used to train the translators. On high-level, dual-GAN and CycleGAN share a lot of similarities.

In [49], Choi et al. proposed StarGAN, a novel and scalable approach that can perform image-to-image translations for multiple domains using only a single model. Such a unified model architecture of StarGAN allows simultaneous training of multiple datasets with different domains within a single network. This leads to StarGAN's superior quality of translated images compared to existing models as well as the novel capability of flexibly translating an input image to any desired target domain. This is specially useful when the number of domains are large, as shown in Fig 34. Some of the generated models via StarGAN model for different human emotions are shown in Fig 35.

In [50], Huang et al. proposed a Multimodal Unsupervised Image-to-image Translation (MUNIT) framework. They assumed that the image representation can be decomposed into a content code that is domain-invariant, and a style code that captures domain-specific properties. To translate an image to another domain, they recombine its content code with a random style code sampled from the style space of the target domain. The high-level overview of this framework is shown in Fig 36.

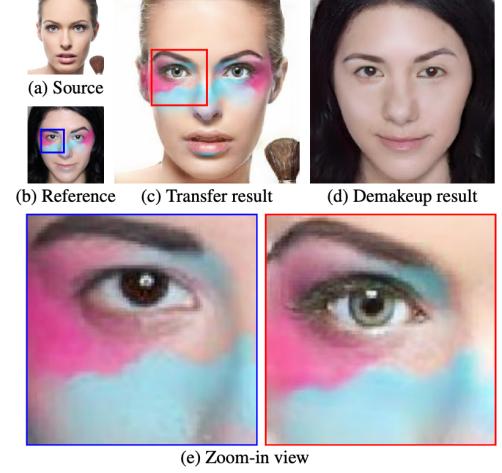


Fig. 28. Facial makeup and de-makeup with dramatic makeup style using LADN framework. Courtesy of [39]

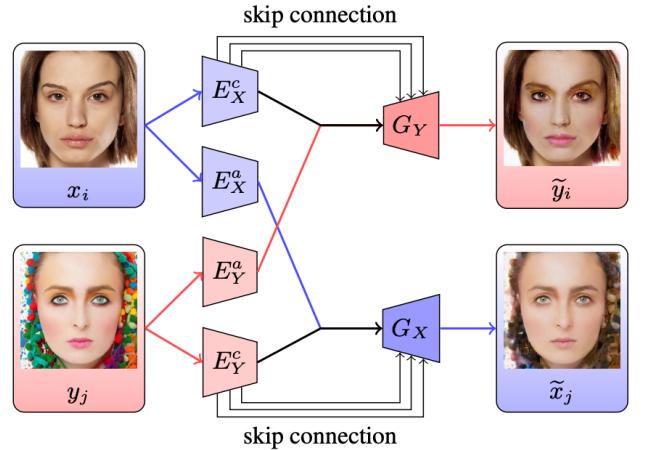


Fig. 29. The generator part of LADN framework. The outputs of E^c and E^a are C and A, which are concatenated at the bottleneck and fed into generators. Skip connections are added between E^c and G to capture more details in generated results. Courtesy of [39]

In [51], Karras et al. proposed an alternative generator architecture for generative adversarial networks (which is also called StyleGAN), borrowing from style transfer literature. The new architecture leads to an automatically learned, unsupervised separation of high-level attributes (e.g., pose and identity when trained on human faces) and stochastic variation in the generated images (e.g., freckles, hair), and it enables intuitive, scale-specific control of the synthesis. When it was proposed, the new generator improved the state-of-the-art in terms of traditional distribution quality metrics, led to demonstrably better interpolation properties, and also better disentangles the latent factors of variation. StyleGAN model opened up the door for many of deep learning based (realistic) AR effects on human face and body images. Some of the sample images generated by StyleGAN model are shown in Fig 37.

In [52], He et al. developed AttGAN, which applies an attribute classification constraint to the generated image to

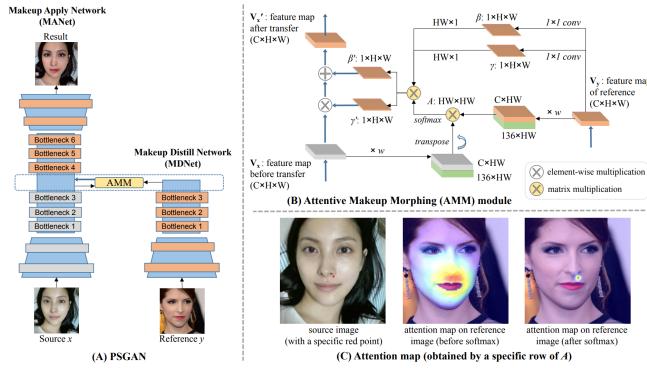


Fig. 30. Illustration of PSGAN framework. Courtesy of [40]

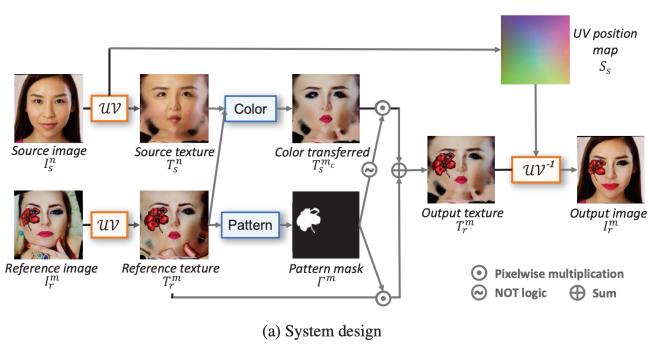


Fig. 31. The high-level architecture of . Courtesy of [41]

just guarantee the correct change of desired attributes, i.e., to “change what you want”. Meanwhile, the reconstruction learning is introduced to preserve attribute-excluding details, in other words, to “only change what you want”. Besides, the adversarial learning is employed for visually realistic editing. These three components cooperate with each other forming an effective framework for high quality facial attribute editing. Fig 38 shows the high-level architecture of AttGAN framework. Some of the sample results of this model are shown in Fig 39.

In [53], Choi et al. proposed StarGAN v2, a single framework that tackles the following properties and shows significantly improved results over the baselines. On one hand it tries to have a good diversity among the generated images and on the other hand it tries to achieve scalability over multiple domains.

In [54], Karras et al. proposed StyleGAN-v2, which introduces changes in StyleGAN’s both model architecture and training methods to address some of the previous issues. In particular, they redesigned the generator normalization, revisit progressive growing, and regularize the generator to encourage good conditioning in the mapping from latent codes to images. In addition to improving image quality, this path length regularizer yields the additional benefit that the generator becomes significantly easier to invert. This makes it possible to reliably attribute a generated image to a particular network. Some of the example images and their projected and re-synthesized counterparts with StyleGAN and StyleGAN2 are shown in Fig 40.

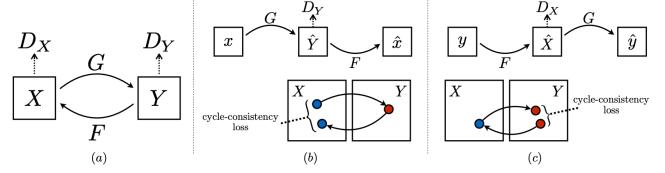


Fig. 32. CycleGAN model contains two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators D_Y and D_X . D_Y encourages G to translate X into outputs indistinguishable from domain Y , and vice versa for D_X , F , and X . To further regularize the mappings, they introduced two “cycle consistency losses” that capture the intuition that if they translate from one domain to the other and back again we should arrive where they started. Courtesy of [47]

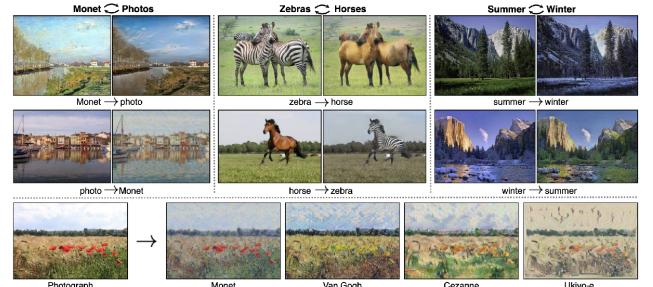


Fig. 33. The sample images transferred via CycleGAN model. Courtesy of [47]

In [55], Wu et al. explored and analyzed the latent style space of StyleGAN2, a state-of-the-art architecture for image generation, using models pre-trained on several different datasets. They first showed that StyleSpace, the space of channel-wise style parameters, is significantly more disentangled than the other intermediate latent spaces explored by previous works. They also described a method for discovering a large collection of style channels, each of which is shown to control a distinct visual attribute in a highly localized and disentangled manner. Furthermore, they proposed a simple method for identifying style channels that control a specific attribute, using a pre-trained classifier or a small number of example images. The comparison of StyleSpace with some of the other frameworks are shown in Fig 41.

In [56], Karras et al. discussed that despite their hierarchical convolutional nature, the synthesis process of typical generative adversarial networks depends on absolute pixel coordinates in an unhealthy manner. This manifests itself as, e.g., detail appearing to be glued to image coordinates instead of the surfaces of depicted objects. They traced the root cause to careless signal processing that causes aliasing in the generator network. Interpreting all signals in the network as continuous, they derive generally applicable, small architectural changes that guarantee that unwanted information cannot leak into the hierarchical synthesis process. The resulting networks match the FID of StyleGAN2 but differ dramatically in their internal representations, and they are fully equivariant to translation and rotation even at subpixel scales. Some of the examples of “texture sticking” using this model and also StyleGAN2 are shown in Fig 42.

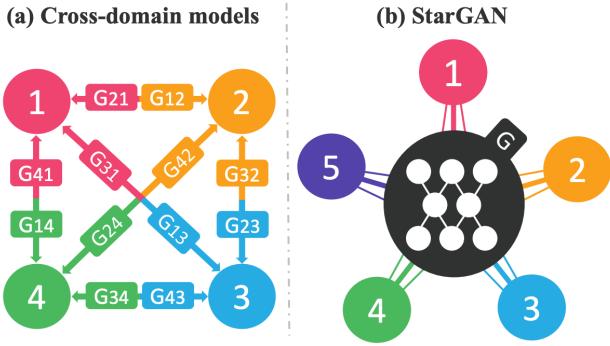


Fig. 34. Comparison between cross-domain models and the proposed StarGAN model. Courtesy of [49]

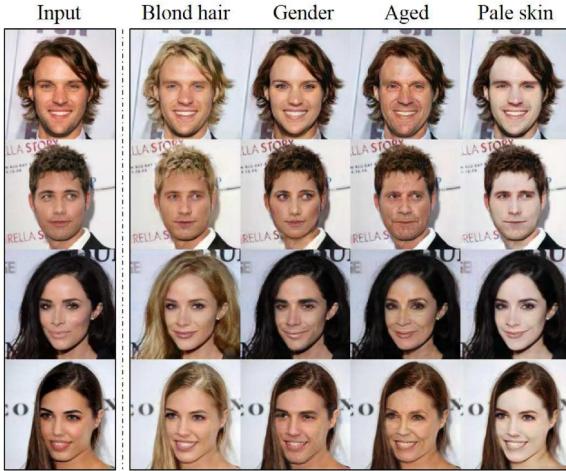


Fig. 35. The first column shows input images, while the remaining columns are images generated by StarGAN. Courtesy of [49]

D. Tracking and Pose Estimation for AR

Augmented reality has revolutionized the gaming industry, and there have been several AR based games which have been developed in the past decade, such as Pokemon Go, Jurassic World Aliev, The Walking Dead: Our World, and many more. There are various algorithms which are the core of AR based games, such as tracking, scene understanding, and reconstruction. In this part, we focus on the tracking frameworks, which involve algorithms for tracking a target object/environment via cameras and sensors, and estimating viewpoint poses. Although vision is not the only modality used for tracking in AR applications, given the scope of this paper, we mainly focus on vision based tracking frameworks.

1) *Eye Tracking and Gaze Estimation*: In [57], Kafka et al. introduced GazeCapture, the first large-scale dataset for eye tracking, containing data from over 1450 people consisting of almost 2.5M frames. Using GazeCapture, they trained iTracker, a convolutional neural network for eye tracking, which achieved a significant reduction in error over previous approaches while running in real time (10–15fps) on a modern mobile device. Their model achieved a prediction error of 1.71cm and 2.53cm without calibration on mobile phones and tablets respectively. An overview of iTracker is shown in Fig

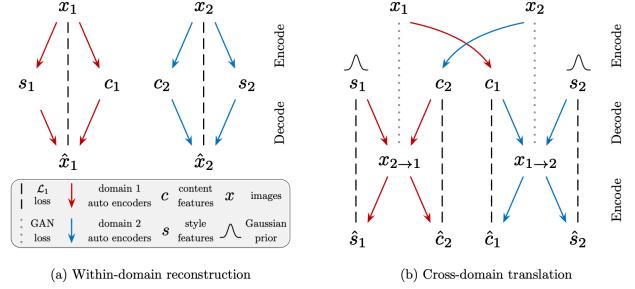


Fig. 36. The overview of MUNIT framework. Their image-to-image translation model consists of two autoencoders (denoted by red and blue arrows respectively), one for each domain. The latent code of each autoencoder is composed of a content code c and a style code s . They train the model with adversarial objectives (dotted lines) that ensure the translated images to be indistinguishable from real images in the target domain, as well as bidirectional reconstruction objectives (dashed lines) that reconstruct both images. Courtesy of [50]

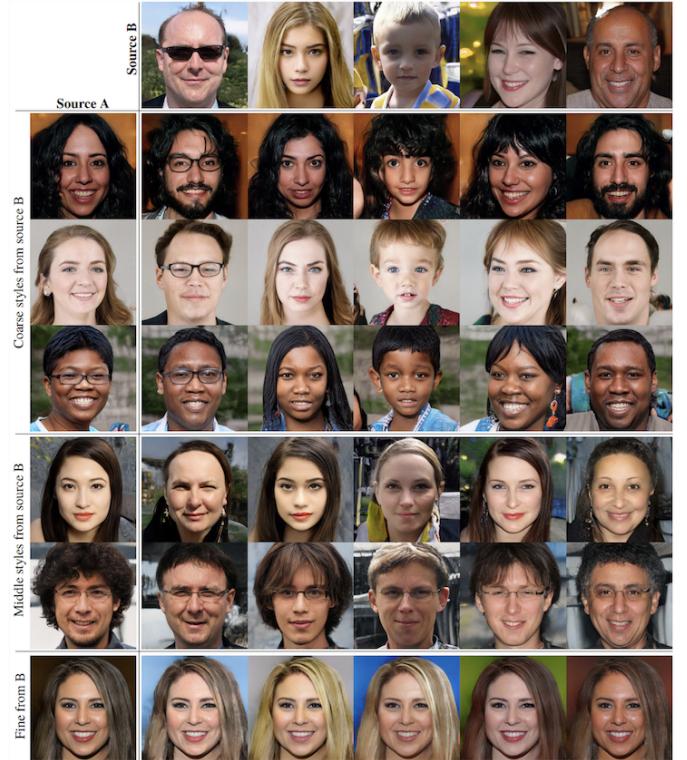


Fig. 37. Two sets of images were generated from their respective latent codes (sources A and B); the rest of the images were generated by copying a specified subset of styles from source B and taking the rest from source A. Courtesy of [51]

43.

In [58], Zhang et al. proposed an appearance-based method for eye gaze estimation that, in contrast to a long-standing line of work in computer vision, only takes the full face image as input. This method encodes the face image using a convolutional neural network with spatial weights applied on the feature maps to flexibly suppress or enhance information in different facial regions.

In [59], Fischer et al. tried to address two limitations of

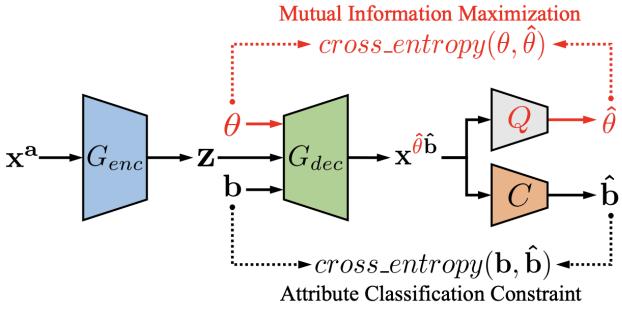


Fig. 38. Illustration of AttGAN extension for attribute style manipulation. Courtesy of [52]



Fig. 39. Facial attribute editing results from our AttGAN. Courtesy of [52]

the previous gaze estimation frameworks, which are: hindered ground truth gaze annotation and diminished gaze estimation accuracy as image resolution decreases with distance. They introduced a novel dataset of varied gaze and head pose images in a natural environment, and also presented a new real-time algorithm involving appearance-based deep convolutional neural networks with increased capacity to cope with the diverse images in the new dataset. The architecture of this model is shown in Fig 44.

In [60], Kellnhofer et al. presented Gaze360, a large-scale gaze-tracking dataset and method for robust 3D gaze estimation in unconstrained images. Their dataset consists of 238 subjects in indoor and outdoor environments with labeled 3D gaze across a wide range of head poses and distances. It was the largest publicly available dataset of its kind by both subject and variety, at the time. Some of the sample images from this dataset are shown in Fig 45. They also proposed a 3D gaze model that extended existing models to include temporal information and to directly output an estimate of gaze uncertainty.

In [61], Yu and Odobez proposed an effective approach to learn a low dimensional gaze representation without gaze annotations. The main idea is to rely on a gaze redirection



StyleGAN — generated images



StyleGAN2 — generated images

Fig. 40. Example images and their projected and re-synthesized counterparts. For each configuration, top row shows the target images and bottom row shows the synthesis of the corresponding projected latent vector and noise inputs. With the baseline StyleGAN, projection often finds a reasonably close match for generated images, but especially the backgrounds differ from the originals. The images generated using StyleGAN2 can be projected almost perfectly back into generator inputs, while projected real images (from the training set) show clear differences to the originals, as expected. Courtesy of [54]

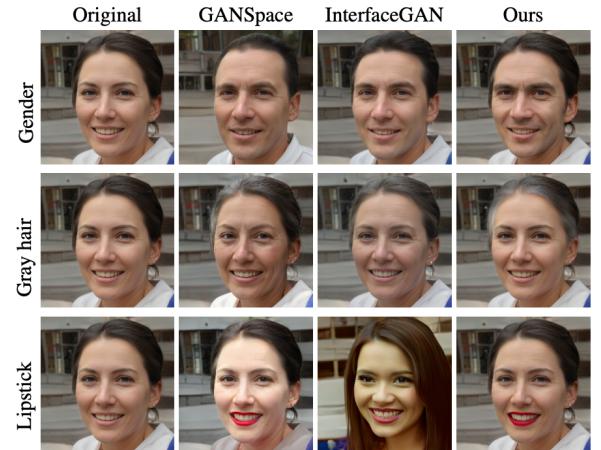


Fig. 41. Comparison with state-of-the-art methods using the same amount of manipulation. Courtesy of [55]

network and use the gaze representation difference of the input and target images (of the redirection network) as the redirection variable. A redirection loss in image domain allows the joint training of both the redirection network and the gaze representation network. In addition, they propose a warping field regularization which not only provides an explicit physical meaning to the gaze representations but also

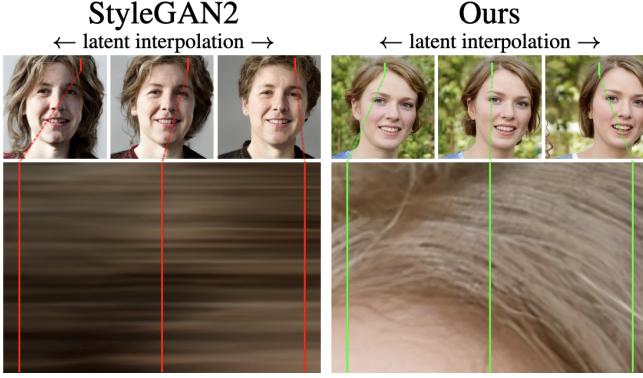


Fig. 42. Examples of “texture sticking”. From a latent space interpolation (top row), they extract a short vertical segment of pixels from each generated image and stack them horizontally (bottom). The desired result is hairs moving in animation, creating a time-varying field. With StyleGAN2 the hairs mostly stick to the same coordinates, creating horizontal streaks instead. Courtesy of [56]

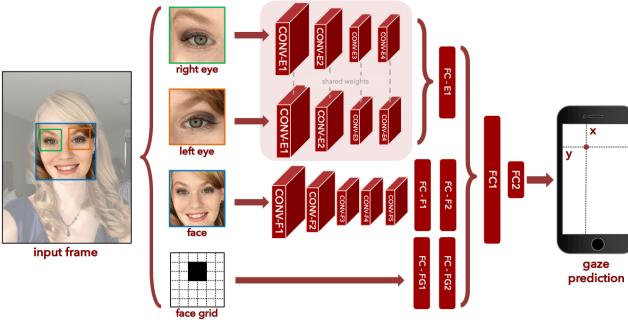


Fig. 43. An overview of iTracker framework. Courtesy of [57]

avoids redirection distortions. The high level architecture of this framework is shown in Fig 46.

In [62], Fang et al. proposed a three-stage method to simulate the human gaze inference behavior in 3D space. In the first stage, they introduced a coarse-to-fine strategy to robustly estimate a 3D gaze orientation from the head. The predicted gaze is decomposed into a planar gaze on the image plane and a depth channel gaze. In the second stage, they develop a Dual Attention Module (DAM), which takes the planar gaze to produce the field of view and masks interfering objects regulated by depth information according to the depth-channel gaze. In the third stage, they use the generated dual attention as guidance to perform two sub-tasks: (1) identifying whether the gaze target is inside or out of the image; (2) locating the target if inside. The architecture of this model is shown in Fig 47.

Some of the other works for eye tracking and gaze estimation includes: Few-shot adaptive gaze estimation [63], TH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation [64], towards end-to-end video-based eye-tracking [65], weakly-supervised physically unconstrained gaze estimation [66].

2) Hand Tracking and Pose Estimation: In [67], Oberweger et al. introduced and evaluated several architectures

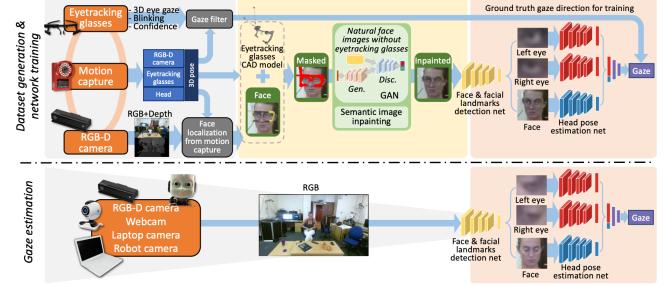


Fig. 44. An overview of RT-GENE architecture. Courtesy of [59]

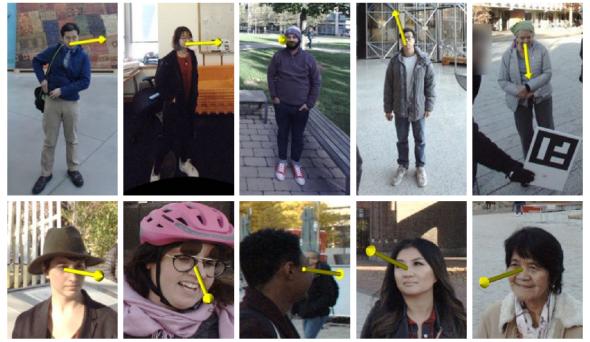


Fig. 45. Some of the sample images from Gaze360 dataset. Courtesy of [60]

for Convolutional Neural Networks to predict the 3D joint locations of a hand given a depth map. They introduced a prior on the 3D pose and significantly improved the accuracy and reliability of the predictions. They also showed how to use context efficiently to deal with ambiguities between fingers.

In [68], Zhou et al. proposed a model based deep learning approach that adopts a forward kinematics based layer to ensure the geometric validity of estimated poses. After applying standard convolutional and fully connected layers, the hand model pose parameters (mostly joint angles) are predicted. Then a new hand model layer maps the pose parameters to the hand joint locations via a forward kinematic process. The architecture of this framework is shown in Fig 48.

In [69], Ge et al. proposed a simple, yet effective approach for real-time hand pose estimation from single depth images using 3D CNNs. Their proposed 3D CNN taking a 3D volumetric representation of the hand depth image as input can capture the 3D spatial structure of the input and accurately regress full 3D hand pose in a single pass. The architecture of the proposed 3D CNN by this work is shown in Fig 49.

In [70], Spurr et al. proposed a method to learn a statistical hand model represented by a cross-modal trained latent space via a generative deep neural network. They derived an objective function from the variational lower bound of the VAE framework and jointly optimize the resulting cross-modal KL-divergence and the posterior reconstruction objective, naturally admitting a training regime that leads to a coherent latent space across multiple modalities such as RGB images, 2D keypoint detection or 3D hand configurations. Additionally, it grants a straightforward way of using semi-supervision. This latent space can be directly used to estimate 3D hand

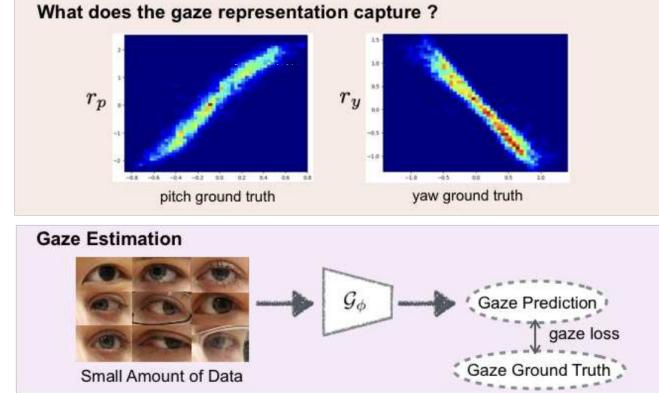
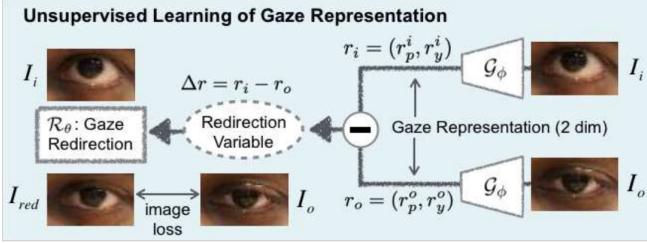


Fig. 46. The proposed framework for Unsupervised Representation Learning for Gaze Estimation. Courtesy of [61]

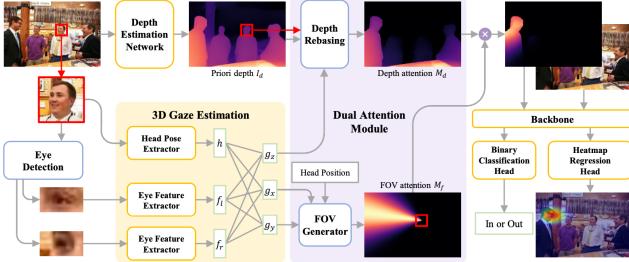


Fig. 47. The Architecture of Dual Attention Guided Gaze Target Detection. Courtesy of [62]

poses from RGB images, outperforming the state-of-the art in different settings. The high-level architecture of this framework is shown in Fig 50.

In [71], inspired by the point cloud autoencoder presented in self-organizing network (SO-Net), Chen et al. proposed SO-HandNet which aimed at making use of the unannotated data to obtain accurate 3D hand pose estimation in a semi-supervised manner. We exploit hand feature encoder (HFE) to extract multi-level features from hand point cloud and then fuse them to regress 3D hand pose by a hand pose estimator (HPE). We design a hand feature decoder (HFD) to recover the input point cloud from the encoded feature. The overview of the model architecture of this work is shown in Fig 51.

In [72], Moon et al. introduced a large-scale dataset, called InterHand2.6M, which contains 2.6M labeled single and interacting hand frames under various poses from multiple subjects. They also proposed a baseline network, InterNet, for 3D interacting hand pose estimation from a single RGB image. InterNet simultaneously performs 3D single and interacting hand pose estimation. Some of the sample frames from

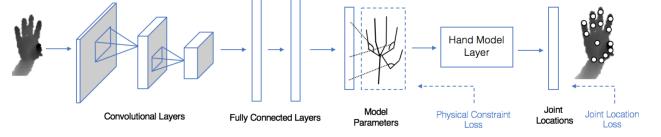


Fig. 48. The architecture of model based deep hand pose learning. Courtesy of [68]

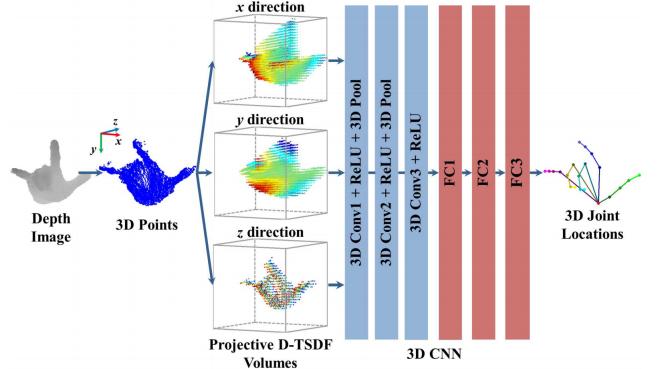


Fig. 49. The architecture of m3d CNN model for hand pose estimation. Courtesy of [69]

sequences with single hand are shown in Fig 52.

In [73], Caramalau et al. proposed a Bayesian approximation to a deep learning architecture for 3D hand pose estimation. Through this framework, they explored and analysed the two types of uncertainties that are influenced either by data or by the learning capability. Furthermore, they drew comparisons against the standard estimator over three popular benchmarks.

Some of the other works for hand tracking and pose estimation includes: Spatial attention deep net for hand pose estimation [74], Deepprior++ [75], Point-to-point regression pointnet for 3D hand pose estimation [76], Hand-transformer: non-autoregressive structured modeling for 3D hand pose estimation [77], 3D Hand Pose Estimation via aligned latent space injection and kinematic losses [78].

3) Human Pose Estimation and Tracking: In [79], Wei et al. showed a systematic design for how convolutional networks can be incorporated into the pose machine framework for learning image features and image-dependent spatial models for the task of pose estimation. They implicitly model long-range dependencies between variables in structured prediction tasks such as articulated pose estimation. Fig 53 shows the high level architecture of the proposed CPM framework.

In [80], Cao et al. proposed a real-time multi-person 2D pose estimation using part affinity fields. This approach uses a non-parametric representation, which they referred to as Part Affinity Fields (PAFs), to learn to associate body parts with individuals in the image. The architecture encodes global context, allowing a greedy bottom-up parsing step that maintains high accuracy while achieving real-time performance, irrespective of the number of people in the image. The architecture is designed to jointly learn part locations and their association via two branches of the same sequential prediction process. The overall pipeline of this framework is shown in Fig 54.

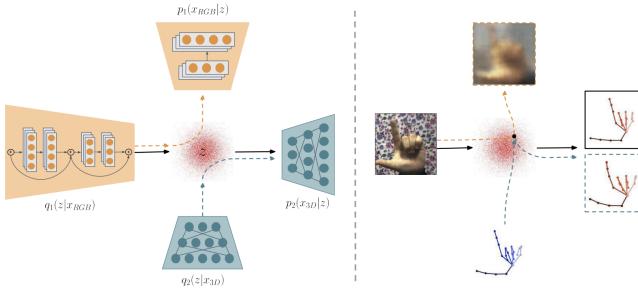


Fig. 50. Schematic overview of the cross-modal deep variational model. Left: a cross-modal latent space z is learned by training pairs of encoder and decoder q, p networks across multiple modalities (e.g., RGB images to 3D hand poses). Auxiliary encoder-decoder pairs help in regularizing the latent space. Right: The approach allows to embed input samples of one set of modalities (here: RGB, 3D) and to produce consistent and plausible posterior estimates in several different modalities (RGB, 2D and 3D). Courtesy of [70]

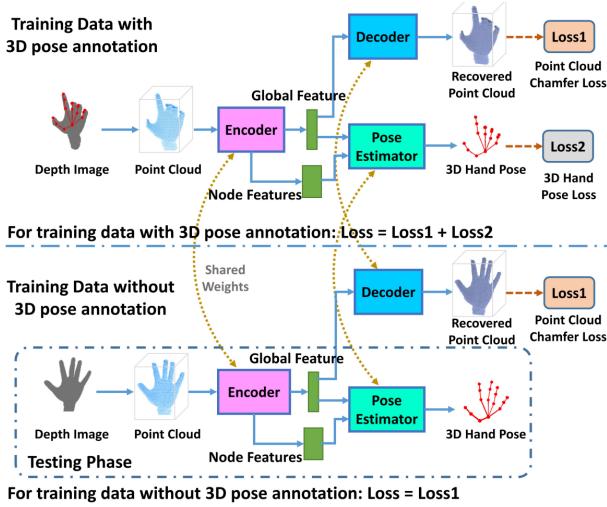


Fig. 51. Overview of the proposed SO-HandNet framework. Courtesy of [71]

In [81], Guler et al. proposed DensePose, which establishes dense correspondences between an RGB image and a surface-based representation of the human body. They gathered dense correspondences for 50K persons appearing in the COCO dataset by introducing an efficient annotation pipeline. The annotations of one sample image from this dataset is shown in Fig 55. They then used this dataset to train CNN-based systems that deliver dense correspondence ‘in the wild’, namely in the presence of background, occlusions and scale variations.

In [82], Pavllo et al. proposed a 3D human pose estimation in video with temporal convolutions and semi-supervised training. They demonstrated that 3D poses in video can be effectively estimated with a fully convolutional model based on dilated temporal convolutions over 2D keypoints. They also introduced back-projection, a simple and effective semi-supervised training method that leverages unlabeled video data. They started with predicted 2D keypoints for unlabeled video, then estimated 3D poses and finally back-project to the input 2D keypoints.

In [83], Xu et al. presented a statistical, articulated 3D human shape modeling pipeline, within a fully trainable, modular,

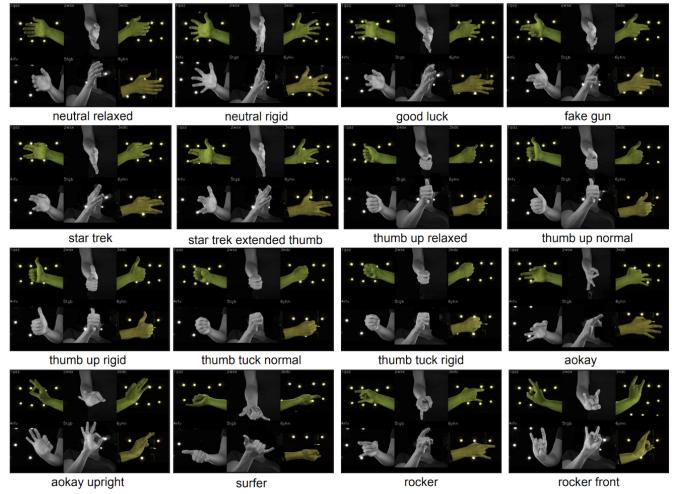


Fig. 52. Visualization of the single hand PP sequences from InterHand26M dataset. Courtesy of [72]

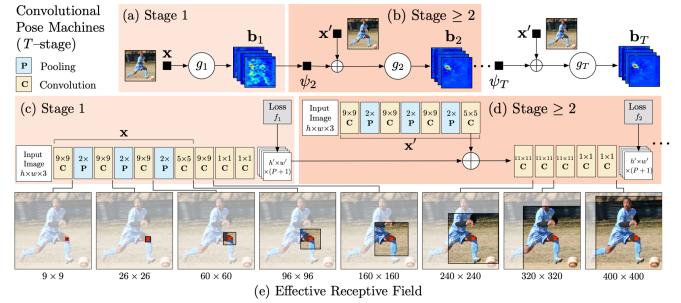


Fig. 53. Architecture and receptive fields of CPMs. Courtesy of [79]

deep learning framework. Given high-resolution complete 3D body scans of humans, captured in various poses, together with additional closeups of their head and facial expressions, as well as hand articulation, and given initial, artist designed, gender neutral rigged quad-meshes, they trained all model parameters including non-linear shape spaces based on variational auto-encoders, pose-space deformation correctives, skeleton joint center predictors, and blend skinning functions, in a single consistent learning loop. The models are simultaneously trained with all the 3D dynamic scan data (over 60, 000 diverse human configurations in our new dataset) in order to capture correlations and ensure consistency of various components. The high-level overview of this framework is shown in Fig 56.

In [84], Liu et al. proposed a Multi-View 3D Labeling and Keypoint Estimation for Transparent Objects, called KeyPose. They forwent using a depth sensor in favor of raw stereo input. They tried to address two problems: First, they established an easy method for capturing and labeling 3D keypoints on desktop objects with an RGB camera; Second, they developed a deep neural network, called KeyPose, that learns to accurately predict object poses using 3D keypoints, from stereo input, and works even for transparent objects. They also created a dataset of 15 clear objects in five classes, with 48K 3D-keypoint labeled images.

In [85], He et al. presented FFB6D, a Full Flow Bidirec-

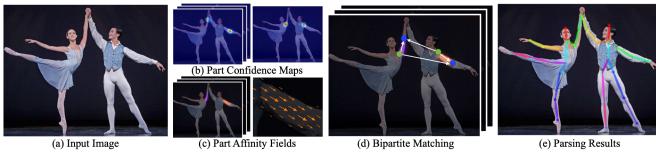


Fig. 54. Overall pipeline. Our method takes the entire image as the input for a two-branch CNN to jointly predict confidence maps for body part detection, shown in (b), and part affinity fields for parts association, shown in (c). The parsing step performs a set of bipartite matchings to associate body parts candidates (d). We finally assemble them into full body poses for all people in the image (e). Courtesy of [80]

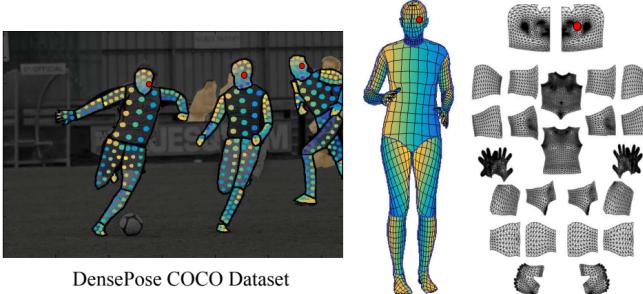


Fig. 55. DensePose-COCO Dataset annotations. Right: Partitioning and UV parametrization of the body surface. Courtesy of [81]

tional fusion network designed for 6D pose estimation from a single RGB-D image. Their key insight is that appearance information in the RGB image and geometry information from the depth image are two complementary data sources, and it still remains unknown how to fully leverage them. Towards this end, FFB6D is proposed, which learns to combine appearance and geometry information for representation learning as well as output representation selection. Specifically, at the representation learning stage, they built bidirectional fusion modules in the full flow of the two networks, where fusion is applied to each encoding and decoding layer. In this way, the two networks can leverage local and global complementary information from the other one to obtain better representations. The high-level overview of FFB6D framework is shown in Fig 57.

Some of the other popular frameworks for human pose estimation includes: regional multi-person pose estimation [86], simple baselines for human pose estimation and tracking [87], OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields [82], SimPoE: Simulated Character Control for 3D Human Pose Estimation [88].

E. Geometry Applications

Deep learning models developed for vision geometry are important for various AR applications (such as the ones in Games, Museums, Automotive, and Scene Understanding). There are various works developed in this direction. Here we cover some of the prominent works.

In [89], Ummenhofer et al. proposed a depth and motion Network for Learning Monocular Stereo, so called DeMoN. They formulated structure from motion as a learning problem. This network estimates not only depth and motion, but addi-

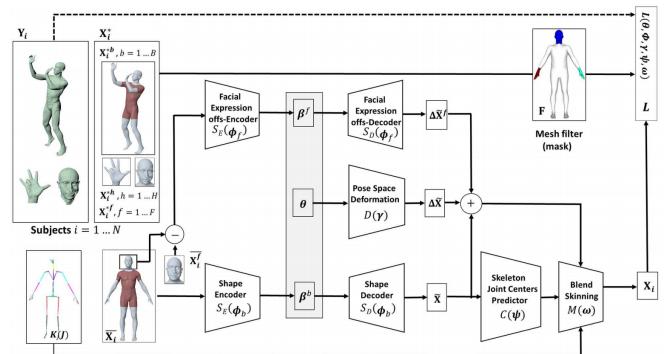


Fig. 56. Overview of our end-to-end statistical 3D articulated human shape model construction. Courtesy of [83]

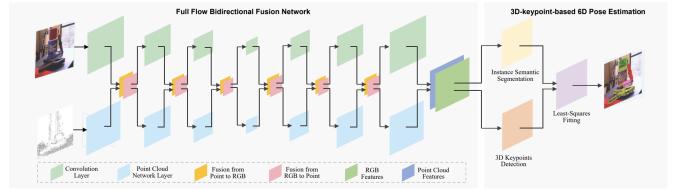


Fig. 57. The pipeline of FFB6D. A CNN and a point cloud network is utilized for representation learning of RGB image and point cloud respectively. In flow of the two networks, bidirectional fusion modules are added as communicate bridges. The extracted per-point features are then fed into an instance semantic segmentation and a 3D keypoint voting modules to obtain per-object 3D keypoints. Finally, the pose is recovered within a least-squares fitting algorithm. Courtesy of [85]

tionally surface normals, optical flow between the images and confidence of the matching. Fig 58 shows a sample result of the predicted depth map by DeMon.

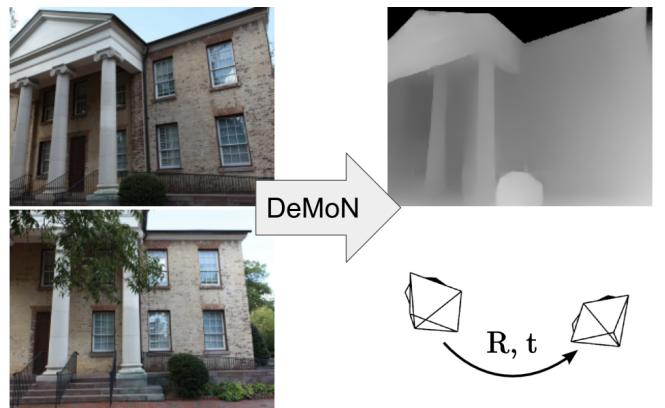


Fig. 58. Illustration of DeMoN. The input to the network is two successive images from a monocular camera. The network estimates the depth in the first image and the camera motion. Courtesy of [89]

In [90], Yin et al. proposed GeoNet, a jointly unsupervised learning framework for monocular depth, optical flow and egomotion estimation from videos. The three components are coupled by the nature of 3D scene geometry, jointly learned by our framework in an end-to-end manner. Fig 59 shows the overview of the GeoNet framework.

In [91], Gordon et al. present a novel method for simulta-

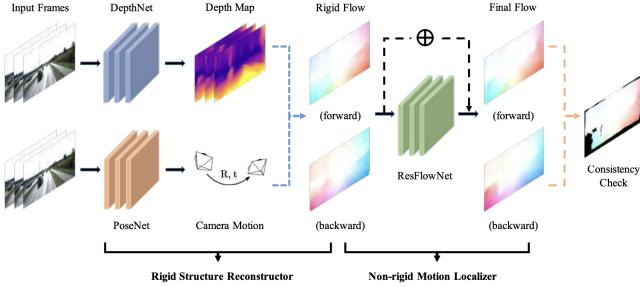


Fig. 59. The overview of GeoNet Framework. It consists of rigid structure reconstructor for estimating static scene geometry and non-rigid motion localizer for capturing dynamic objects. Courtesy of [90]

neously learning depth, ego motion, object motion, and camera intrinsics from monocular videos, using only consistency across neighboring video frames as a supervision signal. They addressed occlusions geometrically and differentiably, directly using the depth maps as predicted during training.

In [92], Guizilini et al. proposed a novel self-supervised monocular depth estimation method combining geometry with a new deep network, PackNet, learned only from unlabeled monocular videos. Their architecture leverages novel symmetrical packing and unpacking blocks to jointly learn to compress and decompress detail-preserving representations using 3D convolutions. The 3D inductive bias in PackNet enables it to scale with input resolution and number of parameters without overfitting, generalizing better on out-of-domain data.

In [93], Ranftl et al. introduced dense prediction transformers, an architecture that leverages vision transformers in place of convolutional networks as a backbone, for dense prediction tasks. They assemble tokens from various stages of the vision transformer into image-like representations at various resolutions and progressively combine them into full resolution predictions using a convolutional decoder. For monocular depth estimation, there is an improvement of up to 28% in relative performance when compared to a state-of-the-art fully convolutional network. Fig 60 shows the overview of the proposed dense prediction transformers framework.

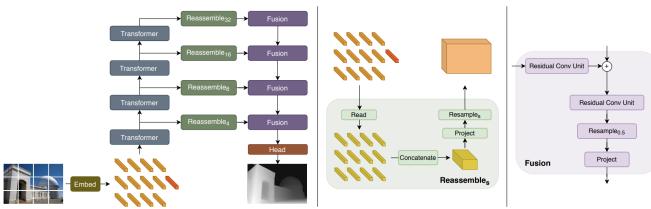


Fig. 60. The overview of dense prediction transformer framework. Courtesy of [93]

Some of the other representative works in this area includes: Unsupervised learning of depth and ego-motion from video [94], MegaDepth [95], and TransformerFusion [96].

F. Scene Understanding and Reconstruction

Simultaneous Localization and Mapping (SLAM) denotes the computational technique that creates and updates a map

of an unknown space where a robot agent is located, while simultaneously tracking the agent's location in it. It is a crucial step in many of the AR/MR, and also robotic applications.

In [97], Dai et al. introduced ScanNet, an RGB-D video dataset containing 2.5M views in 1513 scenes annotated with 3D camera poses, surface reconstructions, and semantic segmentations. To collect this data, they designed an easy-to-use and scalable RGB-D capture system that includes automated surface reconstruction and crowd-sourced semantic annotation. They showed that using this data helps achieve state-of-the-art performance on several 3D scene understanding tasks, including 3D object classification, semantic voxel labeling, and CAD model retrieval.

In [98], Zhang et al. developed an end-to-end system using a depth sensor to scan a scene on the fly. By proposing a Sigmoid-based Iterative Closest Point (S-ICP) method, they decouple the camera motion and the scene motion from the input sequence and segment the scene into static and dynamic parts accordingly. The static part is used to estimate the camera rigid motion, while for the dynamic part, graph node-based motion representation and model-to-depth fitting are applied to reconstruct the scene motions. With the camera and scene motions reconstructed, they further proposed a novel mixed voxel allocation scheme to handle static and dynamic scene parts with different mechanisms, which helps to gradually fuse a large scene with both static and dynamic objects.

In [99], Huang et al. proposed a computational framework to jointly parse a single RGB image and reconstruct a holistic 3D configuration composed by a set of CAD models using a stochastic grammar model. Specifically, they introduced a Holistic Scene Grammar (HSG) to represent the 3D scene structure, which characterizes a joint distribution over the functional and geometric space of indoor scenes. The proposed HSG captures three essential and often latent dimensions of the indoor scenes: i) latent human context, describing the affordance and the functionality of a room arrangement, ii) geometric constraints over the scene configurations, and iii) physical constraints that guarantee physically plausible parsing and reconstruction. They solved this joint parsing and reconstruction problem in an analysis-by-synthesis fashion, seeking to minimize the differences between the input image and the rendered images generated by our 3D representation, over the space of depth, surface normal, and object segmentation map. Fig 61 illustrates the overview of the proposed holistic 3D indoor scene parsing and reconstruction framework.

In [100], Shin et al. tackled the problem of automatically reconstructing a complete 3D model of a scene from a single RGB image. Their approach utilizes viewer-centered, multi-layer representation of scene geometry adapted from recent methods for single object shape completion. To improve the accuracy of view-centered representations for complex scenes, they introduced a novel “Epipolar Feature Transformer” that transfers convolutional network features from an input view to other virtual camera viewpoints, and thus better covers the 3D scene geometry. Unlike previous approaches that first detect and localize objects in 3D, and then infer object shape using category-specific models, their approach is fully convolutional, end-to-end differentiable, and avoids the resolution and memory limitations of voxel representations. Fig 62 illustrates the overview of the proposed framework.

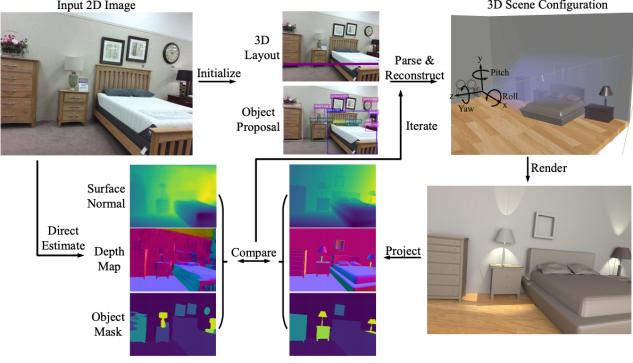


Fig. 61. Illustration of the proposed holistic 3D indoor scene parsing and reconstruction in an analysis-by synthesis fashion. Courtesy of [99]

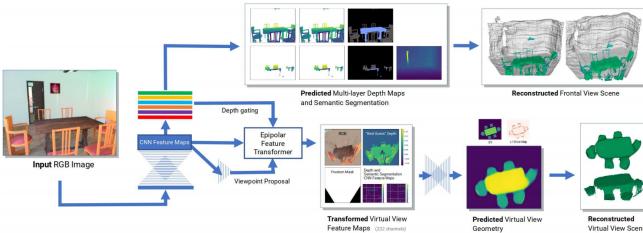


Fig. 62. Illustration of the proposed 3D scene reconstruction using Multi-layer Depth and Epipolar Transformers. Courtesy of [100]

In [101], Popov et al. proposed a coherent 3D scene reconstruction from a single RGB image, using encoder-decoder architectures, along with three extensions: (1) ray-traced skip connections that propagate local 2D information to the output 3D volume in a physically correct manner; (2) a hybrid 3D volume representation that enables building translation equivariant models, while at the same time encoding fine object details without an excessive memory footprint; (3) a reconstruction loss tailored to capture overall object geometry. They reconstruct all objects jointly in one pass, producing a coherent reconstruction, where all objects live in a single consistent 3D coordinate frame relative to the camera and they do not intersect in 3D space. Some of the sample reconstructed 3D scenes using this framework are shown in Fig 63.

In [102], Bovzivc et al. introduced TransformerFusion, a transformer-based 3D scene reconstruction approach. From an input monocular RGB video, the video frames are processed by a transformer network that fuses the observations into a volumetric feature grid representing the scene; this feature grid is then decoded into an implicit 3D scene representation. Key to their approach is the transformer architecture that enables the network to learn to attend to the most relevant image frames for each 3D location in the scene, supervised only by the scene reconstruction task. Features are fused in a coarse-to-fine fashion, storing fine-level features only where needed, requiring lower memory storage and enabling fusion at interactive rates. The feature grid is then decoded to a higher-resolution scene reconstruction, using an MLP-based surface occupancy prediction from interpolated coarse-to-fine 3D features. Fig 64 provides the overview of the proposed framework in TransformerFusion.



Fig. 63. Qualitative results of coherent 3D scene reconstruction on Pix3D dataset. Courtesy of [101]

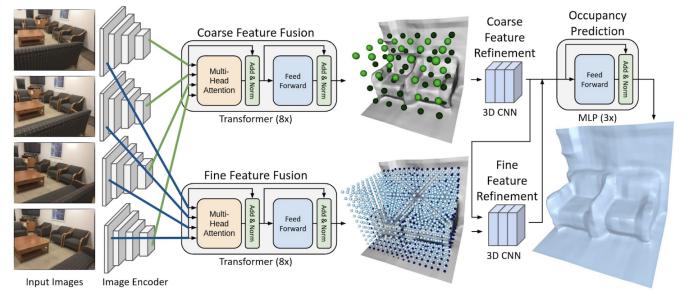


Fig. 64. The overview of TransformerFusion framework. Courtesy of [102]

Some of the other promising works in 3D scene reconstruction includes: CodeSlam [103], Moulding humans: Non-parametric 3d human shape estimation from single images [104], Atlas: end-to-end 3d scene reconstruction from posed images [105], From Points to Multi-Object 3D Reconstruction [106], and VolumeFusion [107].

IV. FUTURE DIRECTIONS

Although there has been a huge progress in AR domain in the past few years, several challenges lie ahead. We will next introduce some of the promising research directions that we believe will help in further advancing augmented reality algorithms.

A. AR *in-the-wild*

Many of the current models developed for AR applications work well only under constrained scenarios (such as simple background, or limited occlusion). Developing new models that perform well in general setting and complex environment is an important research area, which can further extend the application of AR models in different areas. In addition to algorithmic contributions, collecting more complex datasets (with more labeled data in the wild) would be helpful for this purpose.

B. See-through AR

When (AR) visual effect is overlaid on physical scene, ensuring the realistic feeling for the users/observers is crucial.

Even slightest artifact (due to various reasons such as: motion artifacts, quality inconsistency, imperfect segmentation and detection, etc.) could lead to a non-realistic experience for the user. Since human are the main end-user of many of the AR products, subjective tests/metrics could be very useful in assessing these models in early phase, but developing objective metrics to assess how realistic these AR effects/models are (in large-scale sense) is crucial in ensuring good user experience.

C. Realistic 3D Models

Many of today's AR models are developed for 2D images, but in order for AR to give people real-world-like feeling, it needs to work well in 3D setting too. Therefore developing AR models for 3D data is crucial (such as realistic human/clothes modeling and manipulation with fine 3D details and textures). There are already some works developed in this direction, but there is still big room for improvement.

D. Security and Privacy in AR setting

As augmented reality models become more widely used in people's daily life, security and privacy of AR systems are of great importance. While AR can offer several benefits and new opportunities, making sure users' privacy are taken into account, is very important for these models to become widely trusted. Hence, developing AR models which have minimal risk of identity thefts and adversarial attacks is imperative.

E. Remote Cooperative AR

In face-to-face collaboration, people use gesture, gaze and non-verbal cues to communicate in a clear fashion, and in many cases the surrounding environment and objects play a crucial role to providing context and meaning. Physical objects facilitate collaboration both by their appearance, their use as semantic representations, their spatial relationships, and their ability to help focus attention. AR system can be used to advance our remote cooperation and collaboration, by taking the surrounding environment of all parties (involved in a discussion or task) into account. However, that requires more powerful models which can process a lot more information and contexts. Co-located AR collaboration can blend the physical and virtual worlds so that real objects can be used to interact with three-dimensional digital content and increase shared understanding.

F. New Sensors for AR (smell, tactile, taste)

So far, the majority of AR systems are only based on data from visual and depth sensors. But there is no reason for AR to be limited to these sensors, and hopefully that in future there will be more advanced AR systems which can make use of other types of sensors too, such as smell, taste, tactile (for touch), and beyond.

G. AR Devices in Body

So far, the main interaction points with AR systems is through cellphones, laptops/PCs, and AR glasses. But developing displays and chips to enable easier interaction with AR systems could be another future direction. It is worth noting that, there are already some works moving in this direction,

such as Mojo Lens' revolutionary design that uses a tiny microLED display to share critical information, and smart sensors (powered by solid-state batteries) built into a Scleral lens that also corrects people's vision.

V. CONCLUSION

This work provides a detailed overview of augmented reality, its history, applications, and challenges. It introduces more than twenty applications of AR, and discusses about ten of them in detail. After that, it provides a survey of some of the recent deep learning based models developed for augmented reality applications, such as for clothing shopping, make-up try on, fitness and workout. Public datasets developed for those tasks are also mentioned in those sections, when available. Given AR's usefulness, it is continuously applied to new applications . Hence, this paper closes with a discussion about some of the challenges, and possible future directions of AR.

ACKNOWLEDGMENTS

We would like to thank Iasonas Kokkinos, Qi Pan, Lyric Kaplan, and Liz Markman for reviewing this work, and providing very helpful comments and suggestions.

REFERENCES

- [1] M. Billinghurst, A. Clark, and G. Lee, "A survey of augmented reality," 2015.
- [2] <https://www.theverge.com/2019/4/4/18294062/snapchat-landmarks-ar-lenses-filters-eiffel-tower-rainbows>.
- [3] <https://www.4danatomy.com/>.
- [4] <https://www.accuvein.com/>.
- [5] C. L. Loh, "Virtual fitting room using augmented reality," Ph.D. dissertation, UTAR, 2020.
- [6] A. d. F. S. Borges and C. H. Morimoto, "A virtual makeup augmented reality system," in *2019 21st Symposium on Virtual and Augmented Reality (SVR)*. IEEE, 2019, pp. 34–42.
- [7] <https://www.nissan-global.com/EN/TECHNOLOGY/OVERVIEW/i2v.html>.
- [8] <https://www.bbc.co.uk/taster/pilots/civilisations-ar>.
- [9] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3343–3351.
- [10] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104.
- [11] Q. Dong, S. Gong, and X. Zhu, "Multi-task curriculum transfer deep learning of clothing attributes," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 520–529.
- [12] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "Viton: An image-based virtual try-on network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7543–7552.
- [13] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5337–5345.
- [14] X. Han, X. Hu, W. Huang, and M. R. Scott, "Clothflow: A flow-based model for clothed person generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 471–10 480.

- [15] Z. Xie, X. Zhang, F. Zhao, H. Dong, M. C. Kampffmeyer, H. Yan, and X. Liang, "Was-vton: Warping architecture search for virtual try-on network," *arXiv preprint arXiv:2108.00386*, 2021.
- [16] Z. Guo, X. Zhang, H. Mu, W. Heng, Z. Liu, Y. Wei, and J. Sun, "Single path one-shot neural architecture search with uniform sampling," in *European Conference on Computer Vision*. Springer, 2020, pp. 544–560.
- [17] A. Neuberger, E. Borenstein, B. Hilleli, E. Oks, and S. Alpert, "Image based virtual try-on network from unpaired data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5184–5193.
- [18] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, "Towards photo-realistic virtual try-on by adaptively generating-preserving image content," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7850–7859.
- [19] K. Li, M. J. Chong, J. Zhang, and J. Liu, "Toward accurate and realistic outfits visualization with attention to details," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 546–15 555.
- [20] F. Zhao, Z. Xie, M. C. Kampffmeyer, H. Dong, S. Han, T. Zheng, T. Zhang, and X. Liang, "M3d-vton: A monocular-to-3d virtual try-on network," *ArXiv*, vol. abs/2108.05126, 2021.
- [21] A. Raj, P. Sangkloy, H. Chang, J. Hays, D. Ceylan, and J. Lu, "Swapnet: Image based garment transfer," in *European Conference on Computer Vision*. Springer, 2018, pp. 679–695.
- [22] I. Santesteban, M. A. Otaduy, and D. Casas, "Learning-based animation of clothing for virtual try-on," in *Computer Graphics Forum*, vol. 38, no. 2. Wiley Online Library, 2019, pp. 355–366.
- [23] E. Gundogdu, V. Constantin, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua, "Garnet: A two-stream network for fast and accurate 3d cloth draping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8739–8748.
- [24] V. Lazova, E. Insafutdinov, and G. Pons-Moll, "360-degree textures of people in clothing from a single image," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 643–653.
- [25] Z. Wu, G. Lin, Q. Tao, and J. Cai, "M2e-try on net: Fashion from model to everyone," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 293–301.
- [26] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin, "Fw-gan: Flow-navigated warping gan for video virtual try-on," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1161–1170.
- [27] H. Jae Lee, R. Lee, M. Kang, M. Cho, and G. Park, "La-viton: a network for looking-attractive virtual try-on," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
- [28] W.-L. Hsiao, I. Katsman, C.-Y. Wu, D. Parikh, and K. Grauman, "Fashion++: Minimal edits for outfit improvement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5047–5056.
- [29] C. Patel, Z. Liao, and G. Pons-Moll, "Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7365–7375.
- [30] W.-L. Hsiao and K. Grauman, "Vibe: Dressing for diverse body shapes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 059–11 069.
- [31] B. Ren, H. Tang, F. Meng, R. Ding, L. Shao, P. H. Torr, and N. Sebe, "Cloth interactive transformer for virtual try-on," *arXiv preprint arXiv:2104.05519*, 2021.
- [32] S. Choi, S. Park, M. Lee, and J. Choo, "Viton-hd: High-resolution virtual try-on via misalignment-aware normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 131–14 140.
- [33] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo, "Parser-free virtual try-on via distilling appearance flows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8485–8493.
- [34] F. Yang and G. Lin, "Ct-net: Complementary transferring network for garment transfer with arbitrary geometric changes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9899–9908.
- [35] S. Liu, X. Ou, R. Qian, W. Wang, and X. Cao, "Makeup like a superstar: Deep localized makeup transfer network," *arXiv preprint arXiv:1604.07102*, 2016.
- [36] T. Alashkar, S. Jiang, S. Wang, and Y. Fu, "Examples-rules guided deep neural network for makeup recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
- [37] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, and L. Lin, "Beautygan: Instance-level facial makeup transfer with deep generative adversarial network," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 645–653.
- [38] H. Chang, J. Lu, F. Yu, and A. Finkelstein, "Pairedcyclegan: Asymmetric style transfer for applying and removing makeup," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 40–48.
- [39] Q. Gu, G. Wang, M. T. Chiu, Y.-W. Tai, and C.-K. Tang, "Ladn: Local adversarial disentangling network for facial makeup and de-makeup," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 481–10 490.
- [40] W. Jiang, S. Liu, C. Gao, J. Cao, R. He, J. Feng, and S. Yan, "Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5194–5202.
- [41] T. Nguyen, A. T. Tran, and M. Hoai, "Lipstick ain't enough: Beyond color matching for in-the-wild makeup transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 305–13 314.
- [42] C. Cao, F. Lu, C. Li, S. Lin, and X. Shen, "Makeup removal via bidirectional tunable de-makeup network," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2750–2761, 2019.
- [43] X. Liu, R. Wang, C.-F. Chen, M. Yin, H. Peng, S. Ng, and X. Li, "Face beautification: Beyond makeup transfer," *arXiv preprint arXiv:1912.03630*, 2019.
- [44] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng, "Beautyglow: On-demand makeup transfer framework with reversible generative network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 042–10 050.
- [45] S. Velusamy, R. Parihar, R. Kini, and A. Rege, "Fabsoften: Face beautification via dynamic skin smoothing, guided feathering, and texture restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 530–531.
- [46] R. Kips, P. Gori, M. Perrot, and I. Bloch, "Ca-gan: Weakly supervised color aware gan for controllable makeup transfer," in *European Conference on Computer Vision*. Springer, 2020, pp. 280–296.
- [47] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [48] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857.
- [49] 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 pattern recognition*, 2018, pp. 8789–8797.
- [50] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.
- [51] 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*, 2019, pp. 4401–4410.
- [52] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE transactions on image processing*, vol. 28, no. 11, pp. 5464–5478, 2019.
- [53] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9899–9908.

- Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8188–8197.
- [54] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.
- [55] Z. Wu, D. Lischinski, and E. Shechtman, “Stylespace analysis: Disentangled controls for stylegan image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 863–12 872.
- [56] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, “Alias-free generative adversarial networks,” *arXiv preprint arXiv:2106.12423*, 2021.
- [57] K. Kraftka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, “Eye tracking for everyone,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2176–2184.
- [58] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “It’s written all over your face: Full-face appearance-based gaze estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 51–60.
- [59] T. Fischer, H. J. Chang, and Y. Demiris, “Rt-gene: Real-time eye gaze estimation in natural environments,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 334–352.
- [60] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, “Gaze360: Physically unconstrained gaze estimation in the wild,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6912–6921.
- [61] Y. Yu and J.-M. Odobez, “Unsupervised representation learning for gaze estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7314–7324.
- [62] Y. Fang, J. Tang, W. Shen, W. Shen, X. Gu, L. Song, and G. Zhai, “Dual attention guided gaze target detection in the wild,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 390–11 399.
- [63] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz, “Few-shot adaptive gaze estimation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9368–9377.
- [64] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, “Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 365–381.
- [65] S. Park, E. Aksan, X. Zhang, and O. Hilliges, “Towards end-to-end video-based eye-tracking,” in *European Conference on Computer Vision*. Springer, 2020, pp. 747–763.
- [66] R. Kothari, S. De Mello, U. Iqbal, W. Byeon, S. Park, and J. Kautz, “Weakly-supervised physically unconstrained gaze estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9980–9989.
- [67] M. Oberweger, P. Wohlhart, and V. Lepetit, “Hands deep in deep learning for hand pose estimation,” *arXiv preprint arXiv:1502.06807*, 2015.
- [68] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei, “Model-based deep hand pose estimation,” *arXiv preprint arXiv:1606.06854*, 2016.
- [69] L. Ge, H. Liang, J. Yuan, and D. Thalmann, “3d convolutional neural networks for efficient and robust hand pose estimation from single depth images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1991–2000.
- [70] A. Spurr, J. Song, S. Park, and O. Hilliges, “Cross-modal deep variational hand pose estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 89–98.
- [71] Y. Chen, Z. Tu, L. Ge, D. Zhang, R. Chen, and J. Yuan, “So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6961–6970.
- [72] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee, “Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 548–564.
- [73] R. Caramalau, B. Bhattacharai, and T.-K. Kim, “Active learning for bayesian 3d hand pose estimation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3419–3428.
- [74] Q. Ye, S. Yuan, and T.-K. Kim, “Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation,” in *European conference on computer vision*. Springer, 2016, pp. 346–361.
- [75] M. Oberweger and V. Lepetit, “Deepprior++: Improving fast and accurate 3d hand pose estimation,” in *Proceedings of the IEEE international conference on computer vision Workshops*, 2017, pp. 585–594.
- [76] L. Ge, Z. Ren, and J. Yuan, “Point-to-point regression pointnet for 3d hand pose estimation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 475–491.
- [77] L. Huang, J. Tan, J. Liu, and J. Yuan, “Hand-transformer: non-autoregressive structured modeling for 3d hand pose estimation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 17–33.
- [78] A. Stergioulas, T. Chatzis, D. Konstantinidis, K. Dimitropoulos, and P. Daras, “3d hand pose estimation via aligned latent space injection and kinematic losses,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1730–1739.
- [79] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.
- [80] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.
- [81] R. A. Güler, N. Neverova, and I. Kokkinos, “Densepose: Dense human pose estimation in the wild,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7297–7306.
- [82] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “Openpose: realtime multi-person 2d pose estimation using part affinity fields,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.
- [83] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, “Ghum & ghuml: Generative 3d human shape and articulated pose models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6184–6193.
- [84] X. Liu, R. Jonschkowski, A. Angelova, and K. Konolige, “Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 602–11 610.
- [85] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, “Ffb6d: A full flow bidirectional fusion network for 6d pose estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3003–3013.
- [86] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, “Rmpe: Regional multi-person pose estimation,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2334–2343.
- [87] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466–481.
- [88] Y. Yuan, S.-E. Wei, T. Simon, K. Kitani, and J. Saragih, “Simpoe: Simulated character control for 3d human pose estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7159–7169.
- [89] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, “Demon: Depth and motion network for learning monocular stereo,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5038–5047.
- [90] Z. Yin and J. Shi, “Geonet: Unsupervised learning of dense depth, optical flow and camera pose,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1983–1992.
- [91] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, “Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8977–8986.

- [92] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, “3d packing for self-supervised monocular depth estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2485–2494.
- [93] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision transformers for dense prediction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12179–12188.
- [94] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.
- [95] Z. Li and N. Snavely, “Megadepth: Learning single-view depth prediction from internet photos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2041–2050.
- [96] A. Božić, P. Palafox, J. Thies, A. Dai, and M. Nießner, “Transformerfusion: Monocular rgb scene reconstruction using transformers,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [97] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “Scannet: Richly-annotated 3d reconstructions of indoor scenes,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.
- [98] H. Zhang and F. Xu, “Mixedfusion: Real-time reconstruction of an indoor scene with dynamic objects,” *IEEE transactions on visualization and computer graphics*, vol. 24, no. 12, pp. 3137–3146, 2017.
- [99] S. Huang, S. Qi, Y. Zhu, Y. Xiao, Y. Xu, and S.-C. Zhu, “Holistic 3d scene parsing and reconstruction from a single rgb image,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 187–203.
- [100] D. Shin, Z. Ren, E. B. Sudderth, and C. C. Fowlkes, “3d scene reconstruction with multi-layer depth and epipolar transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2172–2182.
- [101] S. Popov, P. Bauszat, and V. Ferrari, “Corenet: Coherent 3d scene reconstruction from a single rgb image,” in *European Conference on Computer Vision*. Springer, 2020, pp. 366–383.
- [102] A. Božić, P. Palafox, J. Thies, A. Dai, and M. Nießner, “Transformerfusion: Monocular rgb scene reconstruction using transformers,” *arXiv preprint arXiv:2107.02191*, 2021.
- [103] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison, “Codeslam—learning a compact, optimisable representation for dense visual slam,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2560–2568.
- [104] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez, “Moulding humans: Non-parametric 3d human shape estimation from single images,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2232–2241.
- [105] Z. Murez, T. van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, “Atlas: End-to-end 3d scene reconstruction from posed images,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020, pp. 414–431.
- [106] F. Engelmann, K. Rematas, B. Leibe, and V. Ferrari, “From points to multi-object 3d reconstruction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4588–4597.
- [107] J. Choe, S. Im, F. Rameau, M. Kang, and I. S. Kweon, “Volumefusion: Deep depth fusion for 3d scene reconstruction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16086–16095.

Legal Disclaimer: This paper provides an overview of prominent machine learning models developed for augmented reality. It is not representative of Snap Inc’s practices, processes, techniques, and perspectives.