APSIPA Transactions on Signal and Information Processing, 2022, 11, e40

Overview Paper

DeepFake and its Enabling Techniques: A Review

Rachael Brooks[1], Yefeng Yuan[1], Yuhong Liu[1] and Haiquan Chen[2*]

[1] *Computer Science and Engineering, Santa Clara University, Santa Clara, CA, USA*

[2] *California State University, Sacramento, CA, USA*

ABSTRACT

Deepfake technology has been undoubtedly growing at a rapid pace since 2017. Particularly since using GAN architecture was popularized, research in this area has grown and seems to only be gaining momentum. One interesting area is animating images of full body humans using deep learning. This paper looks at the research done in this area and research that can influence it by looking at papers regarding human pose transfer, human motion transfer, and human motion generation. All of these types of papers have similar requirements, where a target pose must be abstracted to a skeleton and combined with appearance data from a source image to generate a result. The primary difference in the three types of research is whether or not there is motion in the result and whether that motion is given as an input or generated by the model. Overall, the research in this area is still new, and with the potential applications of this technology, both good and bad, there are many avenues of potential future research in this area in both creation and detection.

Keywords: Deep Learning, Image Animation, Human Pose Transfer, Human Motion Transfer, Deepfake

*Corresponding author: Yuhong Liu, yhliu@scu.edu

Received 24 April 2022; Revised 23 June 2022
ISSN 2048-7703; DOI 10.1561/116.00000024

1 Introduction

Deepfake technology has been undoubtedly growing at a rapid pace since 2017. There are various applications of deepfake technology in the entertainment industry, e.g., animating images or 3D models of actors, video dubbing of foreign films, and animating images of historical figures in a museum to provide a more interactive experience. There are also applications in online shopping, where customers are able to virtually try on clothing using appearance transfer and animation.

However, there has been no question about the danger that this technology poses as a tool of misinformation and disinformation. Many people first heard about deepfakes when they were used to steal celebrities' images and swap their faces onto pornography, which was posted on Reddit [30]. The next time deepfakes made it into the mainstream media is when MIT released a very convincing deepfake of Barack Obama, made from analyzing his speeches [65]. Through these events, many people saw the possibilities and risks of deepfakes and began researching their creation and detection. In the past few years there have been many different tools for creating fake media available to the public. Some are as common as face swapping on social media, like TikTok and Snapchat, where a person can put a different face on a body. Some go further – a creator on TikTok named deeptomcruise makes frighteningly realistic deepfakes of Tom Cruise [12]. As recently as March 2022, deepfakes have been used to spread disinformation in an international conflict, as a deepfake was posted on a Ukrainian news site of Ukrainian President Volodymyr Zelensky telling soldiers to surrender during the war with Russia [4]. There are also applications uploaded to GitHub for anyone to use [52, 61]. These are state-of-the-art models that anyone with access to Google Colab or a decent GPU can run with enough time. Most of these tools are developed based on Deep Neural Networks (DNNs).

As an emerging technique, *deepfake* does not have a generally accepted definition yet, though several have been proposed. For example, some representative definitions include:

- "manipulated digital media such as images or videos where the image or video of a person is replaced with another person's likeness" [5]
- "deepfakes . . . are created by techniques that can superimpose face images of a target person to a video of a source person to make a video of the target person doing or saying things the source person does . . . deepfakes are artificial intelligence-synthesized content that can also fall into two other categories, i.e., lip-sync and puppet-master" [51]

While these definitions cover some typical deepfake applications, they narrowly focus on manipulating human faces, while ignoring more complex manipulations on the human body, gesture, clothes, background, etc.

A more general definition is: "*believable media generated by a deep neural network*" [48]. Although this definition takes the emphasis off of faces and leaves room for advancements, it is too general and loses the focus of the human as a victim. In addition, the word "believable" is vague. If we were to animate an image of a historic figure, such as George Washington, it would not be "believable" as a real media taken of him, even though it may be created by the same technique as animating a realistic image of Joe Biden.

As a result, we claim that this paper mainly focuses on relevant techniques contributing to the creation of deepfakes with a human being as a target/victim. We define these techniques as synthetic image animation as follows.

Definition 1. *Synthetic image animation: Application of movement to a static source image through deep neural networks with the goal of creating synthetic visual media of a person.*

This definition distinguishes between manual image animation (where someone may use Photoshop or other software to animate images by hand) and image animation using deep neural networks. In addition, this definition also requires a static source image for appearance data. This requirement can help us focus more on technologies that can be leveraged to fabricate disinformation against certain human victims (i.e., by using the victim's image as the source image).

As for animation methods using DNNs, existing studies can be categorized into roughly three areas:

1. **Human Pose Transfer:** Taking the appearance of a person from a source image and a representation of a target pose and applying the source appearance to the target pose to create a new image.

2. **Human Motion Transfer:** Taking the appearance of a person from a source image or video and a representation of a target motion and applying the source appearance to the target motion to create a new video, typically with the goal of temporal coherence.

3. **Human Motion Generation:** Generating a video, typically with temporal coherence, of a person doing some movement, either with a source image for appearance data or generating the appearance as well.

Although we classify existing studies into these three categories based on their inputs and outputs, the techniques developed for each category are related. For example, the research that informs human pose transfer can be applicable to human motion transfer. In the most naïve approach, one could take the frames of a video and put them through the human pose transfer model and combine them at the end to create a video. The results from this approach can be of lower quality because there is nothing in the models enforcing temporal coherence, but it is technically a version of synthetic image animation. On a

higher level, the research done in human pose transfer in abstracting images, applying poses, and improving texture and clothing quality can be influential on human motion transfer and generation.

Also, on a technical level, the difference between human motion transfer and human motion generation is the work that goes into determining the next frame. For human motion transfer, this process is guided by a driving video, which should already have temporal coherence. Human motion generation involves training the model to determine the next frame and make sure the end result has temporal coherence. Therefore, this paper looks at papers in all three areas to present a fuller picture of the current literature.

Following the introduction, this paper begins with a breakdown of technologies used in synthetic image animation in Section 3. Section 4 has a deeper look into the different methods of classification of deepfake papers and models within human pose transfer, human motion transfer, and human motion generation. Then, there is a review of the methods for human pose transfer in Section 5. We then survey the synthetic image animation methods, including motion transfer and motion generation in Section 6. Though the focus of this paper is on the creation of media, Section 7 covers some deepfake detection methods for videos and images. Finally, Section 8 looks into the future of the field of synthetic image animation.

2 Related Works

Recent works surveying deepfake creation and detection tend to cover deepfakes more broadly. Papers [48, 51, 69] look at deepfakes as a whole, going over various types of deepfakes, the technology used, and methods to detect them. Both [51] and [48] provide accurate and in-depth descriptions of the specific technologies used for deepfake creation. Specifically, [48] has figures for all of the major deepfake creation models and discusses how technologies like GANs, CNNs, RNNs, etc. are combined in each model in clean, easy to read figures. While their contributions are notable, [48] focuses more on facial deepfakes and [51] has a strong focus on deepfake detection, rather than creation. While both reference some models that perform image animation, it is not a focus for either paper.

In [69], similar to [51], there is a strong focus on deepfake detection, and the paper discusses DNNs in reference to how they can help detection efforts, rather than creation. However, [69] does have a decent description of the different manipulations deepfakes can utilize, some of which are referenced earlier in this paper.

In [5], there is a strong focus around the technology used in deepfakes, going into the specifics about CNNs, RNNs, GANs, and LSTMs. Our paper is distinct from these in that it focuses on synthetic image animation and the

process taken to reach this point, with a focus on full body animation. As a result, this paper looks at topics such as pose transfer, motion transfer, and motion generation, as these three topics can have overlapping research, which is discussed further in Section 5.

3 Key Building Blocks of Existing Studies

In this section, we summarize the key components of synthetic image animation that will be used throughout this paper, including the inputs, networks, typical loss functions adopted, and data sets.

3.1 Inputs

We first introduce the typical inputs used by image animation papers. In the context of this paper, we adopt the following terms.

- **Source Image**: An image that provides appearance information for image animation or pose transfer.
- **Skeleton**: A representation of the outline of an image. Examples include segmentation maps, key-points, or 3D Models.
- **Target Image**: An image that has a target pose for animation.
- **Driving Video**: A video taken in as input that provides specific movements that a model will replicate with a different appearance. This does not include videos used to train a model.

3.2 Networks Used in Image Animation

The most common networks and architectures used in synthetic image animation tasks are defined here. Several of these networks can be combined to complete the tasks. The most popular seen for image animation are CNNs, GANs, U-Nets, and encoder-decoders. CNNs are central as they are the DNN used most often for image processing. U-Nets and encoder-decoders perform similar tasks, where they are applied to CNNs to apply appearance data and refine results. GANs have been central to the growth of image animation because the results they produce can be very high quality, as the deepfakes are directly compared to a ground truth image. RNNs and LSTMs are more commonly used in deepfake detection, but have applications in making fluid motion in an animation.

- **Encoder-Decoder**: A method used to compress and decompress data, typically used to apply and refine appearance data. A representation can be seen in Figure 1a.

- **U-Net**: An CNN-based encoder-decoder that includes skip connections between certain layers to improve the quality of the decompressed data, especially in situations with conditional appearance data [42, 56]. A representation can be seen in Figure 1b.

- **Generative Adversarial Network (GAN)**: A type of DNN with a generator and a discriminator, which generates media in an adversarial manner, where the generator synthesizes media and the discriminator determines if it is real or generated. A representation can be seen in Figure 1c.

- **Convolutional Neural Network (CNN)**: A type of DNN which is used to compress and analyze image data by converting small sections of the image into matrices and performing calculations on them.

- **Recurrent Neural Network (RNN)**: A type of DNN that is used for processing temporal data, such as videos, by keeping track of current behavior and expected behavior.

- **Long Short-term Memory(LSTM)**: A type of RNN that was proposed to overcome the vanishing gradient problem in RNNs. Its primary benefit is that it can both remember past data and forget it as new data comes in.

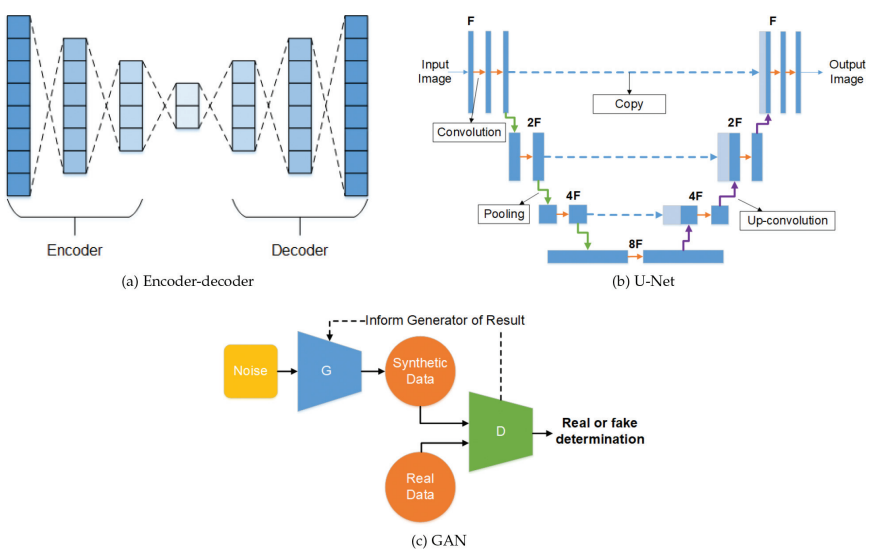


Figure 1: Most popular architectures for synthetic image animation.

3.3 Performance Metrics and Loss Functions

A critical part of creating a model to generate or animate images is picking or designing loss functions. The purpose of loss functions is to evaluate the distance between an output of a model and the expected output. Representations of the most common loss functions used in synthetic image animation are listed in Table 1.

- **Spatial Consistency**: Consistency of placement of objects from frame to frame.
- **Temporal Coherence**: Consistency of color and lighting from frame to frame.
- L_1 **Loss (and its derivatives)**: One of the most common loss functions seen in image animation, L_1 loss is the sum of all the absolute differences between the true and predicted values. It is used to minimize error.
- L_2 **Loss**: L_2 Loss is the same as L_1, but the values being summed are all squared. It is used less than L_1 loss, but it is seen in some papers.
- **Adversarial Loss**: Used for training GANs. There are typically two parts: one for generator loss and one for discriminator loss. Adversarial loss functions are much less standardized than L_1 or L_2 loss functions, and can differ greatly from each other depending on their specific applications.
- **Contextual Loss**: Used to train CNNs by measuring the similarity between the generated image and a target image [46]. Similar to adversarial loss, there are many ways to implement this type of loss function depending on the model the loss function is used for.
- **Perceptual Loss**: Perceptual loss is used to recover fine texture details and is a derivative of adversarial and contextual losses [33].
- **VGG Loss**: Based on the ReLU activation layers of a VGG network, VGG loss is defined as the euclidean distance between the feature representations of a reconstructed image and the reference image [33].
- **Feature Loss**: Used to determine what features are missing or present between two similar images.

3.4 Datasets

Dataset selection is critical for training models to perform well, according to the requirements of the model. The only consistent desirable requirement among datasets is that they have lots of data because, generally, the model

Table 1: Common loss functions in surveyed papers.

Loss	Function example	Citation
L_1	$\sum_{i=1}^{n} \lvert y_{true} - y_{predicted} \rvert$	Widely adopted
L_2	$\sum_{i=1}^{n} (y_{true} - y_{predicted})^2$	Widely adopted
Perceptual	$L_p(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{v=1}^{N} \lvert\lvert \mathbf{\Phi}^v(\mathbf{y}) - \mathbf{\Phi}^v(\hat{\mathbf{y}}) \rvert\rvert_2$	[50]
VGG	$\frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2$	[33]
Adversarial	$\frac{1}{2}\mathbb{E}_z[l(D(z,y) - 1)] + \frac{1}{2}\mathbb{E}_z[l(D(z,\hat{y}))] + \frac{1}{2}\mathbb{E}_z[l(D(G(z) - 1))]$, where $l(x) = x^2$	[50]
LSGAN (D)	$\min_D V_{LSGAN}(D) = \frac{1}{2}\mathbb{E}_{x \sim p_{data}(x)}[(D(x) - b)^2] + \frac{1}{2}\mathbb{E}_{z \sim p_z(z)}[(D(G(z)) - a)^2]$	[45]
LSGAN (G)	$\min_G V_{LSGAN}(G) = \frac{1}{2}\mathbb{E}_{z \sim p_z(z)}[(D(G(z)) - c)^2]$	[45]

can learn more and generate better results. However, the type of data can vary in resolution, type of media, and focus of the media (face, full body, etc.), and these differences can inform which datasets to choose. This section covers the more popular datasets used in research on human pose and motion transfer, however, many researchers created their own datasets from YouTube videos and may or may not have made the dataset public [9, 87].

- **DeepFashion [40]**: DeepFashion has over 800,000 high-resolution images with annotations and over 300,000 exact pairs. This dataset is popular for pose transfer research and is often used in conjunction with the Market-1501 dataset or other low-resolution datasets so researchers can compare the effectiveness of their models on high and low resolution data. This data set has been used by [10, 14, 21, 35, 42, 43, 50, 55, 78, 84, 86].

- **Market-1501 [81]**: A dataset with over 500,000 low resolution images, taken from cameras posted on a campus. The purpose of the dataset is for re-identification of people. However, it is commonly used in pose transfer with DeepFashion so models can compare their effectiveness on high and low resolution data. This data set has been used by [10, 14, 35, 42, 55].

- **VoxCeleb1 Dataset [49]**: This video dataset is a collection of 22,496 videos and 1251 speakers with their original audio. The videos are of celebrities giving interviews. While the dataset is targeted to those doing audio-visual deep learning research, it is used by papers researching pose

and motion transfer. This data set has been used by [24, 44, 62, 63, 73, 76].

- **VoxCeleb2 Dataset [11]**: This video dataset is a collection of 150,480 videos and 6112 speakers with their original audio. It is the second iteration of VoxCeleb1 and, like its predecessor, it can be used for papers researching audio-visual deep learning and pose and motion transfer. This data set has been used by [28, 44].

- **Tai-Chi Dataset [68]**: This dataset was gathered for training a motion generation model, but is useful for pose or motion transfer as well. It consists of more than 4500 videos gathered from YouTube, so the resolution is variable. This data set has been used by [68, 81].

- **Tai-Chi-HD Dataset [62]**: This dataset is similar to Tai-Chi, but it was not gathered by the same research group. This dataset is a collection of 280 high quality videos that are cropped to 256×256 pixels. The dataset was collected for motion transfer research. This data set has been used by [62–64].

- **Fashion Dataset [81]**: A dataset of videos where a single person models an outfit. The models are all different, as are the clothing types and textures. There are 600 total videos, split into training and testing, with each video containing roughly 350 frames. The videos are high resolution and have a static camera. This data set has been used by [39, 55, 62, 81].

- **iPER [38]**: This dataset contains 206 videos with one of thirty people performing some action. There is variability in the types of clothing and the height and weight of the people. There is a total of 241,564 frames in the dataset. This dataset was collected for "human motion imitation" and "human appearance transfer" [38], specifically with how clothing can impact the results. The dataset is used by [38, 39, 55, 75].

3.5 Other Background Terms

There are several commonly seen terms regarding deepfakes models in existing literature. We summarize them here as they can be helpful in the discussions later.

- **Generation**: Creating a new, realistic image or video. This manipulation is different from all the rest of the categories, as it does not require an original/driving image or video [69].

- **Face Swap**: Taking the facial features from one person and putting them on the body of a different person. If the two people share similar facial and body features, these can be extremely realistic [69].

- **Expression Transfer**: Changing the expression of a person in a target image to an expression represented in an image or some other medium, like key-points or segmentation maps [69].

- **Segmentation map-to-image**: Generating an image based on a segmentation map, or an image where different parts of the image are abstracted out in different shapes and colors and detail is added in [69].

- **Inpainting**: Filling in missing data from a picture by using training data and surrounding context as clues [69].

- **Object-agnostic**: A model that does not assume knowledge about the objects that will be animated [62].

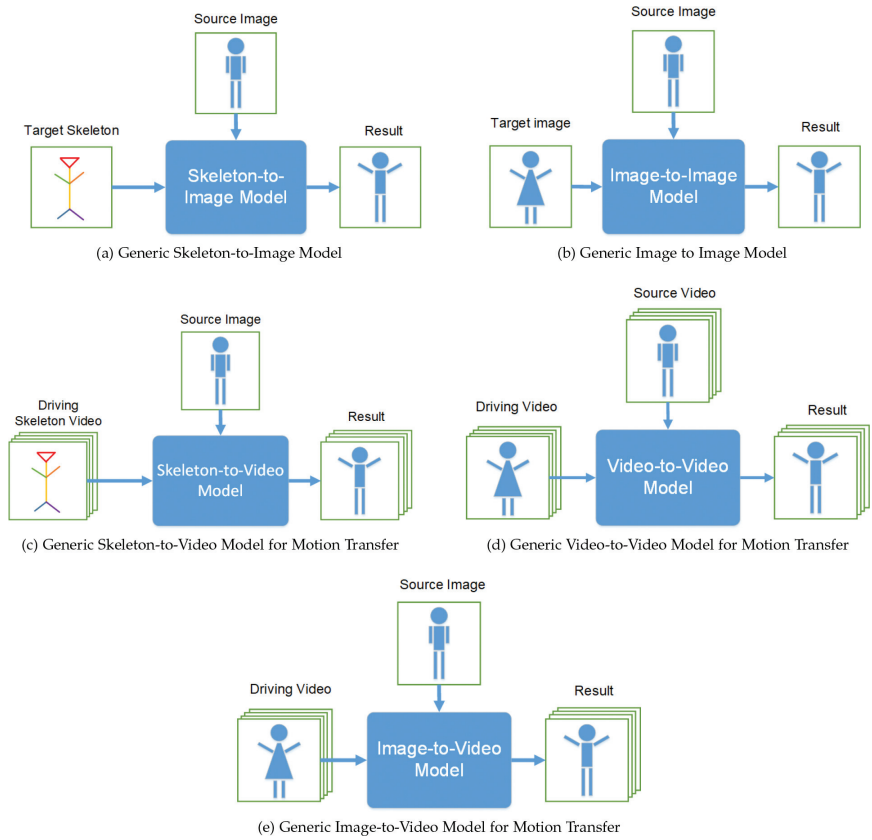
- **Object-specific**: A model that assumes knowledge about the specific object to animate [38].

4 Supporting Techniques

4.1 Types of Models

In Section 1, we provide definitions for the terms Human Pose Transfer, Human Motion Transfer, and Human Motion Generation. In the rest of the paper, we will be using those terms to classify and organize the literature we reviewed.

Within these three classifications, the models can be further categorized by the type of media used as sources and results, which are defined in Table 2. Skeleton-to-Image (Figure 2a) and Image-to-Image (Figure 2b) are both discussed with Human Pose Transfer in Section 5, as the resulting media is not a video. Human Motion Transfer and Human Motion generation are both discussed in Section 6, with Skeleton-to-Video (Figure 2c), Video-to-Video (Figure 2d), and Image-to-Video (Figure 2e) models, because their results are videos. It should be noted that these model classifications are generic; a specific model may have more or less number of sources or results depending on the goals of the researchers creating it. For example, several Skeleton-to-Image models require a skeleton for the source image and a target skeleton, or, a motion generation model may be discussed under Image-to-Video, but it does not have a driving video.



(a) Generic Skeleton-to-Image Model

(b) Generic Image to Image Model

(c) Generic Skeleton-to-Video Model for Motion Transfer

(d) Generic Video-to-Video Model for Motion Transfer

(e) Generic Image-to-Video Model for Motion Transfer

Figure 2: Summary of generic models.

4.2 Abstracting Images

The first step in determining how an image should move is determining which parts of the image should move and abstracting that to something deep learning models can understand and generate. In general, three approaches to this process are *image segmentation*, *key-point estimation*, and *3D pose estimation*.

Image segmentation is the partitioning of images or video frames into multiple segments and objects [47]. The benefit of this method is that it can be applied to any type of image, from images of cities [26], flowers [26], or people [82]. This is also used by the authors of [86] to differentiate between different types of clothing on a person. For example, a t-shirt may cut off about halfway down the upper arm, while a sweater may extend to the wrists. Other methods of image abstraction may not have the ability to map out clothing in this detail, which can make image segmentation a desirable method

Table 2: Models, definitions, and examples.

Model	Definition	Example	Figure
Skeleton-to-Image	A model that takes some skeleton of a pose or segmentation map and applies an appearance to it.	Pose Guided Person Image Generation [42]	Figure 2a
Image-to-Image	A model that takes a source image and a target image and creates a final image with the source pose and target appearance.	Human Appearance Transfer [82]	Figure 2b
Skeleton-to-Video	A model that takes a skeleton or segmentation map for the motion of a person and adds a target appearance to a final video.	High-Quality Video Generation from Static Structural Annotations [58]	Figure 2c
Video-to-Video	A model that takes a video of one person for the motion of another person and adds a target appearance to a final video.	Everybody Dance Now [9]	Figure 2d
Image-to-Video	A model that takes an image as the source and a video for the motion and applies the source appearance to the video motion.	Animating Arbitrary Objects via Deep Motion Transfer [60]	Figure 2e

for clothing and texture transfer. However, the increase in detail comes at a cost, as the time and work to create training data for image parsing is extensive [16].

The second type of image abstraction is key-points. Rather than the entire human body, key-points are typically a collection of some number of joints used as markers to determine how a body looks and where movements take place [7, 16]. Key-points are used in a large majority of the research referenced in this paper because they are lightweight and can be sufficient in full body person animation. An offshoot of key-points and image segmentation are semantic and co-part segmentation [16, 63], which define entire sections of movement, instead

of points of movement. This can provide more detail regarding movements, especially between varying body shapes [64].

The third option is to perform 3D pose estimation, which is the process of determining the position of a body in an image or video and representing it as a 3D model [50, 81, 82]. While 3D human pose estimation is more often used for tracking and identifying what pose a person is in by name, identifying how a person in a source image is posed is critical when determining how to apply a target image or driving video to achieve the target. Some of the body models often used are DeepPose [66] and DensePose [23]. In [29], the authors propose to learn a representation of 3D human dynamics over a temporal context of image features by predicting not only the current 3D human pose and shape but also changes in pose in the nearby past and future frames. The learned 3D dynamics knowledge can be transferred to static images by learning a "hallucinator." The authors of [31] propose an adversarial learning framework to discriminate between real human motions captured in-the-wild and those generated by the temporal pose and shape regression network. The benefits of these models are that they are detailed and could have more realistic results. The primary drawback is that they are resource intensive.

4.3 Clothing and Texture Transfer

Once a target from a source image has taken on a different pose, adjustments also need to be made for any clothing or hair. This practice is a subset of inpainting, where part of an image is removed, and a model is tasked with filling it in. If in pose transfer, for example, a source is in profile, but the target pose is face-forward, then the algorithm needs to generate the second half of the face, hair, any designs or textures on the clothing, etc. This is a complex task, but it is critical for the goal of creating deepfakes consistent with the source material.

While nearly all of the papers referenced here use clothing transfer, several papers focus explicitly on this task. For example, [86] utilizes an initial encoder to analyze the source pose and target pose, and a second encoder to analyze the textures from the source image and apply it to the target. Similarly, [21] utilizes GANs, CNNs, and an encoder-decoder architecture to evaluate the source image clothing and to apply inpainting to the new images. While the details differ between these papers, the general strategy is to determine the source image pose first, the target pose second, and then apply any required textural changes to the target image.

Unsupervised Image-to-Video Clothing Transfer [53] also deals with clothing transfer, but applies the techniques to animation specifically, instead of just static images. Specifically, they store memory before and after generating each frame to determine how the clothing should change from frame to frame to produce a coherent video. Image-to-Image clothing transfers do not need this capability, as the images do not need to maintain coherent textures and poses across multiple images.

5 Human Pose Transfer

In this section, we mainly focus on existing studies to transfer human poses. In particular, this section is split into four parts. The sections focus on two types of models that papers create: Skeleton-to-Image and Image-to-Image. The type of model included in those sections is expanded in Table 2.

5.1 Skeleton-to-Image

As shown in Figure 2a, the method of pose transfer uses a source image to get the necessary textures and a skeleton to assign a target pose. In this paper, "skeleton" is used as a general term for a computer representation of different poses. Skeletons can be in different forms, such as 3D poses [50] and 2D poses [7], two types of which are key-points [14, 85, 88] and segments [16, 23, 43, 86]. Often in transferring poses on images, a model will assign a skeleton to the source and target poses to simplify the process and identify which parts of the images should change.

Each of the forms of skeletons mentioned have different benefits and drawbacks, as mentioned in Section 4.2. The most commonly used is key-points because of their simplicity [7, 10, 34, 88]. Generally, the models in these papers took in a tuple of a source image, target skeleton, and source skeleton. In theory, these models could avoid inputting the skeleton of the source image using pose estimation. However, that is a more intensive process and is less necessary when generating images because datasets of still images are not difficult to come by.

A different type of Skeleton-to-Image, as defined for this paper, is referenced in [26]. This paper describes a model that is very similar to the other papers mentioned here, in that it can take a skeleton of a photo and add details, but it is different because part of the model is going from an image to a skeleton, which is more often seen in models that have visual media as the target or driving video input. Furthermore, since the images are not strictly of humans, the researchers use segmentation maps instead of key-points. As a final point, there is also functionality that generates photos that, instead of adding detail to a skeleton, it can change one feature of an image. For example, if an image input is in the day, the output could be the same scenery, but at night.

Pose Attention Transfer Networks (PATN) and Pose Attention Transfer Blocks (PATB) represent another way to perform pose transfer. Models based on this idea, instead of jumping from a starting pose to an ending pose, attempt to generate more realistic poses by generating the motion in between, thereby incrementally generating a more accurate picture. This concept is introduced in [88] which proposed utilizing cascading PATBs in a PATN to guide the pose transfer process. Some works that are based on this idea are [10, 78], and [34].

5.2 Image-to-Image

As shown in Figure 2b, the natural iteration from Skeleton-to-Image is to apply the pose from an image to the appearance from another image, also known as pose transfer and, when applied to humans, human pose transfer. Papers that implement Image-to-Image models are: [15, 35, 41, 43, 50, 59, 82, 84]. Generally, models implemented as Image-to-Image take in a target image and a source image and output a result, where the subject in the source image has the pose of the subject in the target image, much like in Figure 2b. There can be exceptions to this, such as [84], where the model requires that the source image is in a certain position, and there is no target image, as the model automatically outputs the subjects in two pre-trained poses.

In looking at the research done in this area, it is clear that the current trend is attempting to break down the process into smaller and smaller parts, with the goal of improved results. For example, if a model were to attempt to generate new images based on a pose all at once, it is likely that elements of the foreground and background will become mixed and the end result will have artifacts that indicate it is a fake.

Additionally, the generation process may go from one image to another all at once. A model could receive an input of an image of a person facing forward and a target pose of them facing away. In this system, there would be little process for determining what the back of a person's head may look like just from the front, likely resulting in lower quality outputs.

One way that researchers have attempted to break down the process is through the application of disentangling the foreground and background in images. If a model can correctly determine the foreground from the background in an image and disentangle the two parts, the resulting image is less likely to have artifacts that indicate it is a deepfake. There are a couple papers which discuss this technique. One example is in [7], which is focused around disentanglement, the act of separating a person, the background, and any items that person is holding. Another is in [43], where the researchers are specifically aiming to generate images in a disentangled manner.

6 Human Motion Transfer and Generation

There are two types of approaches for animating synthetic images to form videos, namely motion transfer and motion generation. Motion transfer and generation studies need to additionally take time coherence as an important criteria, which is not a requirement for pose transfer studies.

Most of the research found in the area focuses on motion transfer, though there are notable examples of human motion generation, such as [68] and [77]. The major difference between studies in these two categories is whether a

driving video is required. The human motion transfer models adopt a driving video, so that the person/object in the deepfake videos acts exactly in the same way as the one in the driving video. On the other hand, the human motion generation models do not use a driving video, but are trained on a variety of general motion videos so that any arbitrary motions can be created. Furthermore, general video generation models have been developed for some time. But most of them do not require a specific source image to create a deepfake for the specific person/object in the source image, which differentiates them from the two works discussed in this section.

Specifically, according to the different sources, we classify existing studies into three categories: Skeleton-to-Video, Video-to-Video and Image-to-Video, and discuss each category in detail. Understandably, it may seem odd to include a section on Video-to-Video models in a paper that is meant to focus on image animation. The reason this section is included is because several of the papers on Video-to-Video motion transfer utilize similar methods of motion and pose transfer, especially when broken down to the frame-by-frame level. The research done in [71] provides a decent example of this by breaking down the source movement video into frames and generating a new frame of the target person in the movement pose. Furthermore, we discuss Video-to-Video before Image-to-Video because order of discussion is based on the complexity of the models, not the complexity of the source material for them. Image-to-Video models can be more complex than Video-to-Video because Video-to-Video models can gather more data about the source person to apply to a result, while Image-to-Video models need to predict any missing data.

6.1 Skeleton-to-Video

As shown in Figure 2c, the idea of Skeleton-to-Video is that the models cut down on the work of extracting the initial motion by just taking in the skeletal motion. There are some models that use skeletal driving videos instead of driving videos of objects. An example of these videos could be when actors wear motion capture suits and videos of key-points are taken. Generally, the process of motion transfer is:

- **Step 1.** Take the driving video as input
- **Step 2.** Analyze the driving video for the motion
- **Step 3.** Extract the motion and apply it to some skeleton (typically a 3D model or key-points)
- **Step 4.** Apply the appearance from an input image to the skeleton

For Skeleton-to-Video models, steps 1, 2, and most of 3 are combined, though the model will still need to process the driving video.

One option is to use a segmented image as an input and to apply color, texture, and motion to it through DNNs [58]. In this case, the proposed task is pix2vid, which generates movement based on the structural annotation given. However, while there are some interesting applications and benefits to a system like this, it is harder to find datasets of image segments for training and testing. Furthermore, this system does not take in appearance information, which can be a significant negative for certain tasks.

A predecessor of pix2vid is vid2vid [74]. Instead of having a static skeleton used for the video synthesis, the skeleton is an image segmented video. This implementation means that the model does not have to dissect the skeleton from a video, and it also allows for different types of motion in non-human subjects. In the case of most of the papers in this review, the movement is limited to a singular object, usually a human. This paper stands out because it can represent the motion of many objects going by, for example, as if a spectator is looking out the window of a car.

In [55], the focus is on full body motion. The purpose of the paper is to generate images of models in different poses indicated by a skeleton of key-points, and to expand those generated images into videos. In order to create coherent animations, they generate the videos recurrently so that the model can extract correlations between adjacent frames. Specifically, this model extracts skeletons from videos of skeletons and cleans them up so that they align with required parameters. After that is done, the model generates clips using a source image, where the person in the source image does the same motions as the skeleton.

In another paper, [86], researchers were able to generate images from a segmentation map, and they were also able to create animations of garment transfer. Garment transfer is somewhat out of the realm of this paper, but the concept of expanding an existing Skeleton-to-Image model to an image animation one is adjacent to our scope.

6.2 Video-to-Video

As shown in Figure 2d, Video-to-Video models are those that take in videos for the source motion and the target appearance. Having the data of the target person's full appearance makes it easier for the model to generate a video of the target executing the motion of the source, when compared to Image-to-Video models.

The first and one of the most prevalent state-of-the-art Video-to-Video synthesis models is vid2vid [74]. This paper is referenced in Section 6.1 as well because it takes an image segmented video and adds texture and color. Because this model is considered state-of-the-art, the paper is included in the related works of [9, 28, 38, 62, 73, 81], and [1].

As mentioned, other Video-to-Video models focus on human motion transfer [1, 9, 19, 28, 32, 54, 71, 75, 87]. The most influential in this group is [9]. Paper [9] is one of the first papers on successful Video-to-Video motion transfer and introduces the researchers' method of motion transfer, "Everybody Dance Now," and a method to detect whether or not a video is fake. In order to ensure temporal coherence, this model generates two predicted outputs, one based on the previously generated frame and one based on the driving video frame, and puts them through a GAN. To add more realism to faces of the generated image, a FaceGAN was added which successfully added necessary detail. The final part of the paper was the creation and usage of a detection model by the researchers, which successfully identified the videos from the model.

Finally, an interesting exception comes up in [73], where few-shot vid2vid, a successor to vid2vid, is proposed. In many Video-to-Video models, the model has to be trained individually for each person the model should output. The goal of few-shot vid2vid is to be able to have multiple people as inputs and outputs with only one driving video and a few example images of each source. This paper stands out particularly because it does not fit in completely with any of the classifications established in this paper because the inputs are a driving video and several example images of each subject for the motion to be transferred to. It is included in this section because having to input more than one image is closer to a Video-to-Video model, which is taking in a sequence of images as frames, than it is an Image-to-Image model, which only takes in one image and has to generate any other appearance data from nothing else. The ability to get appearance data from only a few images and create multiple videos with only one model are both potentially huge areas of future research and could reduce the required size of datasets, but still produce high quality results.

6.3 Image-to-Video

The focus of this paper is the technology surrounding synthetic image animation. In this section, we look at papers that take a source image as input and add movement, either through motion transfer or motion generation. The question of how to animate images, or representations of objects, has existed since long before deep learning [8, 83]. With the revelation of GANs in 2014 [20], research into Deep Neural Networks and synthesizing images and movement received a massive boost.

The history of image animation using neural networks starts in 2016 with [77], which uses variational autoencoders and proposes CNNs to synthesize future frames of a video based on an image. Between 2016 and 2019, much of the research surrounded pose transfer and Video-to-Video applications. However, X2Face [76] was published in 2018, and research on other models,

like DwNet [81], MarioNETte [24], Liquid Warping GAN [38], and Monkey-Net [60] in 2019. Each of these papers proposes different models for synthetic image animation, though using a GAN is a common theme among them. Initial research into motion transfer and image animation tends to focus around humans.

X2Face [76] is an architecture with the ability to generate videos of faces from, in one iteration, a source frame and driving video. While the focus of this survey is on full body models, the work in X2Face is cited in multiple other papers in this area, such as [24, 62, 73], and others.

DwNet [81] utilizes DensePose [23], which focuses on human skeletons, and does not yet seem to have the capability for other types of focuses to be animated, to encode poses. However, DwNet's leading contribution to image animation is a warp module, which estimates the final location for source appearance placement, based on the target poses, and refines that estimate for the final output.

MarioNETte is another method proposed in [24]. This method proposes "a few-shot face reenactment framework." This means that while there is a driving video, the source image can actually be several images. Adding more data obviously means generating a more accurate animation. However, this has the obvious downside in that it is not as clearly image animation. In testing, however, the researchers also tested MarioNETte with single shot sources, which meets the required definition for an Image-to-Video model. This model performs better than some of the other original models, such as X2Face [76] and Monkey-Net [60], but compared to more advanced models, it is a lower performer.

Liquid Warping GAN is introduced in [38], which specifically focuses on body movement. The focus on humans can be seen in the skeleton used in processing the images, which focuses around human features, and the loss functions used, such as a face identity loss. This method works quite well on humans, but because it is object-specific, the applications are more limited.

Object-agnostic image animation is not a new idea but current research in this area, specifically Monkey-Net [60], can make for promising future research. The primary technology used in Monkey-Net is "using a set of sparse motion-specific key-points that were learned in an unsupervised way to describe relative pixel movements [60]." This technology is what enables Monkey-Net to reassign key-points to different images and objects that are inputted into the model, making it object-agnostic.

First Order Motion Model (FOMM) [62] also aimed to create an object-agnostic model by improving on the work done for Monkey-Net. FOMM improves on Monkey-Net by adding local affine transformations to model complex motions, an occlusion-aware generator to estimate objects not in the source image, and an adjusted equivariance loss function. The model was able to train key-points to identify movements for different types of pictures (human

faces, human full-bodies, animals, etc.), which is not seen in most models. Combined, these results outperform the original Monkey-Net, according to the given data. However, the work on FOMM was not final, and another paper has been published [64], making improvements on the original model. They do this by changing the segmentation methods in order to better apply motion between people of different body types or different clothing shapes.

Now that a few state-of-the-art technologies have been proposed, papers have been published that modify elements of previous work, such as in [44], in order to experiment with optimizing existing models.

It is also worth looking at papers that are outside of the norm of human motion transfer, like [68] and [77], which both use motion generation. These papers are still mentioned in this section because, while the original version of MoCoGAN [68] simply is trained on a dataset and then generates videos based on the dataset the model was trained on, [68] mentions a variant of MoCoGAN where an image is inputted into the model and the model generates motion based on that image. That is not to say that there is no source media; there is an entire dataset that trains the model so that it can generate the movement. However, motion generation for image animation is an interesting path that has been less researched than motion transfer.

7 Deepfake Detection

As more deepfake creation methods are introduced, it has never been easier to create a deepfake video using existing applications or following tutorials online. Due to the threats brought by deepfakes to personal privacy and social security, many detection methods have been developed along with the advancements of deepfake creation techniques. Previous detection methods relied on evaluating specified features from the inconsistencies of fake videos during their synthesis process. Nowadays, detection methods utilize deep learning technologies to extract inconsistent features automatically [51].

In this section, we divide the deepfake detection methods to two major categories, image manipulation detection and video manipulation detection. Most of these methods attempt to build a robust and accurate classifier to distinguish between real and fake contents.

7.1 Image Manipulation Detection

Image manipulation is fundamental to deepfakes, as a deepfake video is generated by a series of deefake images. Early applications such as face-swapping played a key role in promoting image manipulations, by splicing fake face regions to the original images. Therefore, using 3D head poses and a SVM classifier can reveal the errors of fake images [79]. With the utilization of deep

learning methods such as GANs, synthesized images can be easily created with outstanding quality. In order to accurately and efficiently detect fake images, a fake image detector needs to be forged. During the training process of image detectors, adversarial perturbations can be added to fool the system [18] to improve its robustness. Pairwise learning can be applied to increase the accuracy [25]. Deep learning approaches are also beneficial to improve the detectors. The DeepTag uses an encoder and decoder to recover embedded messages from facial images in order to prevent authentic images from manipulation [72]. CNNs and image segmentation are also used to examine the factors that affect detection accuracy [80].

7.2 Video Manipulation Detection

The development of image manipulation makes the video manipulation more accessible. However, most detection methods for image detection cannot be directly applied to video detection due to video compression [2] and temporal characteristics. In order to successfully detect deepfake videos, some methods evaluated the temporal dynamics of the videos, such as exploiting the dynamics of mouth shape with a spoken phoneme [3] and evaluating temporal dynamics with recurrent approaches [57, 67]. Other methods utilized CNNs to improve the robustness of the detecting classifier, such as extracting frame-level features [22], adopting optical flow fields [6], and detecting face warping artifacts [36]. Combining super resolution algorithms to deep learning can also improve the accuracy of the detector [27].

In addition, large-scale datasets with increasing number of videos have been created to evaluate the performance of DeepFake detection algorithms. The Celeb-DF dataset [37] contains 5,639 DeepFake videos generated from celebrities' videos on the internet. The DFDC dataset [13] created by Facebook has more than 100,000 video clips collected from 3,426 actors and actresses. These datasets contribute to the foundation of detection benchmarks.

8 Future of Field

We discuss the future of the field mainly from three aspects: further improvement, computational complexity reduction, and future applications.

8.1 Further Improvement

Looking at some current research can inform what to expect from future studies. First, further efforts towards improving how faces look in full body modeling for humans are important. Current research can do well when focused on just the face of a person, or just the body, but in a high-definition video, the face

tends to get less precision than required. There is a balance that needs to be found to find good results here.

Second, will future models be object-specific or object-agnostic? Research is likely to be prominently object-specific, as producing clean results is easier with object-specific modeling. However, although challenging, the research coming from object-agnostic models will be more interesting, as the media able to be produced by an object-agnostic model would have more capabilities.

Last but not least, there are a handful of papers that, instead of building entire models, focus on a detail of the model to improve, such as the texture of clothes, blurred edges, time inconsistency, etc. An example can be seen in [64], which improves upon FOMM by adjusting how the model abstracts objects to better pass appearance information. Although existing studies have achieved significant progress in making deepfakes realistic, it is still far away from deceiving human eyes. We consider it critical to further refine details of the outcomes, and as there are state of the art models out there already, future research will likely be focused on these details instead of the entire model.

8.2 Computational Complexity Reduction

The adoption of deep neural networks has significantly improved the quality of fake images/videos, however, the process typically requires very heavy computation and massive training data, making it less practical. Reducing computational complexity of image animation models is critical, as hackers have an incentive to reduce the complexity and make it easier to produce material and security researchers need to keep up in order to detect them.

The first place to look is datasets. Training and testing models on videos requires a non-negligible amount of computational power when compared to training on photos. While the results from Video-to-Video models seem to be better than Image-to-Video models, a researcher could use more data for training with the same amount of computational power for Image-to-Video, when compared to Video-to-Video. Image-to-Video also has an advantage in that there are more datasets of pictures than videos. Videos take up more data than images, and the creation of datasets is much more difficult and expensive, compared to photographs. Particularly, full body motion datasets are lacking, and these would be best for training and testing synthetic image animation models.

Another area researchers could look into to reduce computational complexity is few or one-shot papers. A few of these are referenced in our paper [24, 73] and demonstrate that this type of model has potential in creating quality deepfakes with less data in both Image-to-Video and Video-to-Video models. As these models take in less data they could work better with existing datasets.

8.3 Future Applications

Synthetic image animation has potential for future and current technology and development. Current development of this technology is happening in social media. Snapchat and TikTok are both major players in this game as part of creating fun filters for sharing content. One particular filter on TikTok is a Photo Animation filter, where users can apply the filter to drawings, photographs, and themselves. The filter will add eye movements and other effects to the photo to make it come alive.

Video games and movies also have potential with this technology, and research into using deepfakes in this industry is already taking place [8, 17, 70]. This technology could be used to insert players into the game with an avatar that has their likeness, such as with metaverse applications. Applications in movies could be used to create more dynamic action scenes, where a thorough image of an actor or actress could be used in place of a manually animated model or stunt double. This technology has also been used to bring historic photographs to life. Adding full body animation to those photographs could bring to life photos of dances, events, celebrities, or relatives, making the experience of learning about history more dynamic and even interactive.

9 Discussion and Conclusion

Before beginning this research, and certainly while working on it, the potential negative impact of deepfakes and image animation applications, particularly in their ability to mislead and generate disinformation, raised the question, "Why study this in the first place?" In most of the research published, the technology itself is neutral, with research teams motivated by the pursuit of knowledge and understanding the capacity of this technology. However, the consequences depend on who is leveraging it.

A question that better gets to the root of this problem is "Should this technology exist?" However, this question is still lacking, because the views on this issue will be different from person to person, institution to institution.

The real question is, "Will this technology exist?," to which the answer is yes. At this point, with or without academia, this technology is going to exist and grow. Considering concerns regarding how to detect deepfakes are growing, studying this technology, both its creation and detection, in the open is critical for researchers to develop detection techniques. In order to know how to detect deepfakes, we need to be able to know how they work, and the best way to do that is to publish research out in the open.

From designing new architectures that are more energy efficient to determining the best loss functions to produce higher quality outputs, human pose and motion transfer have many avenues of research yet to be explored. In this

paper we have focused on different types of architectures that can all lead to improvements in image animation research on full human bodies. In starting with pose transfer, we aimed to show the different parts of pose transfer and how they can relate to motion transfer. With motion transfer, there is a story in how each part of the process comes together. Skeleton-to-Image models do not have to be concerned with pose estimation in the driving video, and can focus on generating the appearance data on the result video correctly. Video-to-Video models do not have to be concerned with missing appearance data, and can focus on correctly extracting the motions from the source and applying them to the target. Image-to-Video models have to address all of these tasks as well as the ones mentioned in pose transfer. The task is difficult, but considering the pace at which architectures for this task have been developed and improved, that task will likely be accomplished in the coming years.

References

[1] K. Aberman, M. Shi, J. Liao, D. Lischinski, B. Chen, and D. Cohen-Or, "Deep Video-based Performance Cloning," *Computer Graphics Forum*, 38(2), 2019, 219–33.

[2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: A Compact Facial Video Forgery Detection Network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2018, 1–7.

[3] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting Deepfake Videos from Phoneme-viseme Mismatches," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, 660–1.

[4] B. Allyn, "Deepfake Video of Zelenskyy Could be 'Tip of the Iceberg' in Info War, Experts Warn," *NPR*, 2022, https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

[5] A. M. Almars, "Deepfakes Detection Techniques Using Deep Learning: A Survey," *Journal of Computer and Communications*, 9(5), 2021, 20–35.

[6] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[7] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing Images of Humans in Unseen Poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 8340–8.

[8] T. D. Bui *et al.*, "Fast and Realistic 2D Facial Animation based on Image Warping," in *2009 International Conference on Knowledge and Systems Engineering*, IEEE, 2009, 81–6.

[9] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody Dance Now," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, 5933–42.

[10] B. Chen, Y. Zhang, H. Tan, B. Yin, and X. Liu, "PMAN: Progressive Multi-Attention Network for Human Pose Transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[11] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *INTERSPEECH*, 2018.

[12] deeptomcruise, "deeptomcruise," May 2021, https://www.tiktok.com/@deeptomcruise/video/6965575763298962693.

[13] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The Deepfake Detection Challenge (DFDC) Preview Dataset," *arXiv preprint arXiv:1910.08854*, 2019.

[14] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated Warping-gan for Pose-guided Person Image Synthesis," *arXiv preprint arXiv:1810.11610*, 2018.

[15] P. Esser, E. Sutter, and B. Ommer, "A Variational U-net for Conditional Appearance and Shape Generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 8857–66.

[16] H.-S. Fang, G. Lu, X. Fang, J. Xie, Y.-W. Tai, and C. Lu, "Weakly and Semi Supervised Human Body Part Parsing via Pose-guided Knowledge Transfer," *arXiv preprint arXiv:1805.04310*, 2018.

[17] O. Gafni, L. Wolf, and Y. Taigman, "Vid2game: Controllable Characters Extracted from Real-world Videos," *arXiv preprint arXiv:1904.08379*, 2019.

[18] A. Gandhi and S. Jain, "Adversarial Perturbations Fool Deepfake Detectors," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, 1–8.

[19] T. Gomes, R. Martins, J. Ferreira, and E. Nascimento, "Do As I Do: Transferring Human Motion and Appearance between Monocular Videos with Spatial and Temporal Constraints," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, 3366–75.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, 27, 2014.

[21] A. Grigorev, A. Sevastopolsky, A. Vakhitov, and V. Lempitsky, "Coordinate-based Texture Inpainting for Pose-guided Image Generation" *arXiv preprint arXiv:1811.11459*, 2018.

[22] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *2018 15th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, IEEE, 2018, 1–6.

[23] 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, 7297–306.

[24] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, "Marionette: Few-shot Face Reenactment Preserving Identity of Unseen Targets," 34(7), 2020, 10893–900.

[25] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, "Deep Fake Image Detection based on Pairwise Learning," *Applied Sciences*, 10(1), 2020, 370.

[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image Translation with Conditional Adversarial Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 1125–34.

[27] N. S. Ivanov, A. V. Arzhskov, and V. G. Ivanenko, "Combining Deep Learning and Super-resolution Algorithms for Deep Fake Detection," in *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, IEEE, 2020, 326–8.

[28] S. Jeon, S. Nam, S. W. Oh, and S. J. Kim, "Cross-Identity Motion Transfer for Arbitrary Objects through Pose-Attentive Video Reassembling," in *European Conference on Computer Vision*, Springer, 2020, 292–308.

[29] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3D Human Dynamics from Video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 5614–23.

[30] L. Kelion, "Reddit Bans Deepfake Porn Videos," *BBC*, 2018, https://www.bbc.com/news/technology-42984127.

[31] M. Kocabas, N. Athanasiou, and M. J. Black, "Vibe: Video Inference for Human Body Pose and Shape Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 5253–63.

[32] M. R. Koujan, M. C. Doukas, A. Roussos, and S. Zafeiriou, "Head2head: Video-based Neural Head Synthesis," *arXiv preprint arXiv:2005.10954*, 2020.

[33] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, 4681–90.

[34] K. Li, J. Zhang, Y. Liu, Y.-K. Lai, and Q. Dai, "PoNA: Pose-Guided Non-Local Attention for Human Pose Transfer," *IEEE Transactions on Image Processing*, 29, 2020, 9584–99.

[35] Y. Li, C. Huang, and C. C. Loy, "Dense Intrinsic Appearance Flow for Human Pose Transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 3693–702.

[36] Y. Li and S. Lyu, "Exposing Deepfake Videos by Detecting Face Warping Artifacts," *arXiv preprint arXiv:1811.00656*, 2018.

[37] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 3207–16.

[38] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, 5904–13.

[39] W. Liu, Z. Piao, Z. Tu, W. Luo, L. Ma, and S. Gao, "Liquid Warping GAN with Attention: A Unified Framework for Human Image Synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[40] 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, 1096–104.

[41] F. Ma, G. Xia, and Q. Liu, "Spatial Consistency Constrained GAN for Human Motion Transfer," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[42] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. VAN GOOL, "Pose Guided Person Image Generation," *Advances in Neural Information Processing Systems*, 30, 2017.

[43] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled Person Image Generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 99–108.

[44] Y. S. Malik, N. Sabahat, and M. O. Moazzam, "Image Animations on Driving Videos with DeepFakes and Detecting DeepFakes Generated Animations," in *2020 IEEE 23rd International Multitopic Conference (INMIC)*, IEEE, 2020, 1–6.

[45] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least Squares Generative Adversarial Networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, 2794–802.

[46] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The Contextual Loss for Image Transformation with Non-aligned Data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, 768–83.

[47] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[48] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *ACM Computing Surveys (CSUR)*, 54(1), 2021, 1–41.

[49] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-scale Speaker Identification Dataset," in *INTERSPEECH*, 2017.

[50] N. Neverova, R. A. Guler, and I. Kokkinos, "Dense Pose Transfer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, 123–38.

[51] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep Learning for Deepfakes Creation and Detection: A Survey," *arXiv preprint arXiv:1909.11573v4*, 2022.

[52] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, *et al.*, "DeepFaceLab," 2020, https://github.com/iperov/DeepFaceLab.

[53] A. Pumarola, V. Goswami, F. Vicente, F. De la Torre, and F. Moreno-Noguer, "Unsupervised Image-to-video Clothing Transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, 0–0.

[54] J. Ren, M. Chai, S. Tulyakov, C. Fang, X. Shen, and J. Yang, "Human Motion Transfer from Poses in the Wild," *arXiv preprint arXiv:2004.03142*, 2020.

[55] Y. Ren, G. Li, S. Liu, and T. H. Li, "Deep Spatial Transformation for Pose-Guided Person Image Generation and Animation," *IEEE Transactions on Image Processing*, 29, 2020, 8622–35.

[56] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional Networks for Biomedical Image Segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, 2015, 234–41.

[57] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent-convolution Approach to Deepfake Detection-state-of-art Results on Faceforensics++," *arXiv preprint arXiv:1905.00582*, 2019.

[58] L. Sheng, J. Pan, J. Guo, J. Shao, and C. C. Loy, "High-Quality Video Generation from Static Structural Annotations," *International Journal of Computer Vision*, 128, 2020, 2552–69.

[59] A. Siarohin, S. Lathuilière, E. Sangineto, and N. Sebe, "Appearance and Pose-conditioned Human Image Generation Using Deformable Gans," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[60] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Animating Arbitrary Objects via Deep Motion Transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 2377–86.

[61] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First Order Motion Model for Image Animation," 2019, https://github.com/AliaksandrSiarohin/first-order-model.

[62] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order Motion Model for Image Animation," *Advances in Neural Information Processing Systems*, 32, 2019, 7137–47.

[63] A. Siarohin, S. Roy, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Motion-supervised Co-Part Segmentation," *arXiv preprint arXiv:2004.03234*, 2020.

[64] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion Representations for Articulated Animation," *arXiv preprint arXiv:2104.11280*, 2021.

[65] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning Lip Sync from Audio," *ACM Transactions on Graphics (ToG)*, 36(4), 2017, 1–13.

[66] A. Toshev and C. Szegedy, "Deeppose: Human Pose Estimation via Deep Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, 1653–60.

[67] L. Trinh, M. Tsang, S. Rambhatla, and Y. Liu, "Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, 1973–83.

[68] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing Motion and Content for Video Generation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 1526–35.

[69] L. Verdoliva, "Media Forensics and Deepfakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 2020, 910–32.

[70] C. Wang, C. Xu, and D. Tao, "Self-supervised Pose Adaptation for Cross-Domain Image Animation," *IEEE Transactions on Artificial Intelligence*, 1(1), 2020, 34–46.

[71] H. Wang, M. Huang, D. Wu, Y. Li, and W. Zhang, "Supervised Video-to-Video Synthesis for Single Human Pose Transfer," *IEEE Access*, 9, 2021, 17544–56.

[72] R. Wang, F. Juefei-Xu, Q. Guo, Y. Huang, L. Ma, Y. Liu, and L. Wang, "DeepTag: Robust Image Tagging for DeepFake Provenance," *arXiv preprint arXiv:2009.09869*, 2020.

[73] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, "Few-shot Video-to-video Synthesis," *arXiv preprint arXiv:1910.12713*, 2019.

[74] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video Synthesis," *arXiv preprint arXiv:1808.06601*, 2018.

[75] D. Wei, X. Xu, H. Shen, and K. Huang, "GAC-GAN: A General Method for Appearance-controllable Human Video Motion Transfer," *IEEE Transactions on Multimedia*, 2020.

[76] O. Wiles, A. Koepke, and A. Zisserman, "X2face: A Network for Controlling Face Generation Using Images, Audio, and Pose Codes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, 670–86.

[77] T. Xue, J. Wu, K. L. Bouman, and W. T. Freeman, "Visual Dynamics: Probabilistic Future Frame Synthesis via Cross Convolutional Networks," *arXiv preprint arXiv:1607.02586*, 2016.

[78] L. Yang, P. Wang, C. Liu, Z. Gao, P. Ren, X. Zhang, S. Wang, S. Ma, X. Hua, and W. Gao, "Towards Fine-grained Human Pose Transfer with Detail Replenishing Network," *IEEE Transactions on Image Processing*, 30, 2021, 2422–35.

[79] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 8261–5.

[80] C.-M. Yu, C.-T. Chang, and Y.-W. Ti, "Detecting Deepfake-forged Contents with Separable Convolutional Neural Network and Image Segmentation," *arXiv preprint arXiv:1912.12184*, 2019.

[81] P. Zablotskaia, A. Siarohin, B. Zhao, and L. Sigal, "DwNet: Dense Warp-based Network for Pose-guided Human Video Generation," *arXiv preprint arXiv:1910.09139*, 2019.

[82] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu, "Human Appearance Transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 5391–9.

[83] D. Zhang, Z. Miao, and M. Wang, "Images Based Human Volumetric Model Reconstruction and Animation," in *2010 International Conference on Image Analysis and Signal Processing*, IEEE, 2010, 210–3.

[84] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, Z. Jie, and J. Feng, "Multi-view Image Generation from a Single-view," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, 383–91.

[85] W. Zhao, Q. Xie, Y. Ma, Y. Liu, and S. Xiong, "Pose Guided Person Image Generation Based on Pose Skeleton Sequence and 3D Convolution," in *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, 1561–5.

[86] H. Zheng, L. Chen, C. Xu, and J. Luo, "Pose Flow Learning From Person Images for Pose Guided Synthesis," *IEEE Transactions on Image Processing*, 30, 2021, 1898–909.

[87] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. Berg, "Dance Dance Generation: Motion Transfer for Internet Videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[88] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive Pose Attention Transfer for Person Image Generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 2347–56.