XPose: Realism and Stability in Human Video Animations through ControlNet Integration of DensePose and DWPose

Ziheng Jiang[1], Chentao Zhang[2,*]

{858693822@qq.com[1], 23198478@studentmail.ul.ie[2]}

School of Communication and Information Engineering, Shanghai University, Shanghai, China[1]
Department of Science & Engineering, University of Limerick, Limerick, Ireland[2]
*corresponding author

Abstract. Enhancing the realism and stability of human animations in video generation remains a significant challenge, particularly in maintaining consistent facial expressions, body proportions, and clothing. This paper introduces a new framework, which addresses these challenges by integrating Dense Human Pose (DensePose) and Dynamic Warping Pose (DWPose) models within a ControlNet structure, called XPose. XPose combines mid-sample and down-sample ControlNet components to effectively guide the video generation process, while also exploring the impact of Lineart overlay on video quality. The framework dynamically adjusts the contributions of DensePose and DWPose through a weighted concatenation approach, with a 1.5:1 ratio identified as optimal based on quantitative evaluation using Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS) metrics. Evaluated on the TikTok dancing video dataset, XPose demonstrates significant improvements in the consistency and stability of generated animations, particularly in maintaining facial expressions and body sizes. Experimental results indicate that XPose outperforms existing methods such as MagicAnimate, achieving up to a 23.5% improvement in video fidelity under specific configurations. The findings suggest that XPose offers a robust solution for high-quality human video animations, with future research planned to explore the use of ControlNet images for further refinement.

Keywords: Video Generation, Human Animation, DensePose, ControlNet.

1 Introduction

Many methods and models have been developed to animate static images using sequential motion signals, such as video, depth, and pose data. These techniques have significant potential applications across various domains, including the movie industry, social media, and entertainment. This growing interest has led to numerous explorations and the creation of various approaches to bring static images to life.

Among these methods, low-cost data-driven animation frameworks have garnered the most attention, as the abundance of video data makes them both feasible and efficient. Two primary models are predominantly used for this purpose: Generative Adversarial Networks (GANs) [1] based and diffusion-based frameworks. GAN-based methods typically employ a warping function to deform the reference image into the target pose and use GANs to extrapolate the

CONF-MLA 2024, November 21, Adana, Turkey

DOI 10.4108/eai.21-11-2024.2354584

missing or occluded body parts [2]. While these methods can generate visually plausible animations, they often struggle with motion transfer capability, resulting in unrealistic details in occluded regions and limited generalization ability for cross-identity scenarios. On the other hand, diffusion-based methods [3] harness appearance and pose conditions to generate the target image based on pre-trained diffusion models. These methods process videos in a frame-by-frame manner and stack results along the temporal dimension, which can lead to flickering and temporal inconsistency in the animations.

This issue arises from the lack of effective temporal modeling and poor preservation of reference identity in existing methods. To address these challenges, a model named MagicAnimate [4] was introduced, a novel diffusion-based framework designed to enhance temporal consistency and preserve the reference image's fidelity. MagicAnimate incorporates a video diffusion model with temporal attention blocks to encode temporal information and a novel appearance encoder to retain intricate details of the reference image. Additionally, a simple video fusion technique is employed to encourage smooth transitions for long video animations. The method achieves state-of-the-art performance on benchmarks such as the TikTok dancing dataset, surpassing the strongest baseline by over 38% in terms of video fidelity. Empirical results demonstrate that MagicAnimate significantly improves the quality of human image animations, offering robust generalization across various identities and motion sequences.

Despite advancements in animation generation, facial expressions and hand gestures produced by existing methods often lack stability and can appear unrealistic, occasionally distorting the intended portrayal. To address these challenges and enhance the realism of animations, we developed a new model called XPose. XPose integrates Distillation for Whole-body Pose Estimators (DWPose) [5] and Dense Human Pose Estimation in The Wild (DensePose) [6]. This integration enables XPose to capture comprehensive body pose data and contours more effectively, resulting in animations that are both more accurate and stable. By combining DWPose and DensePose, XPose achieves more lifelike and natural animations. To further optimize the contributions of DensePose and DWPose within the XPose framework, we utilize a weighted concatenation approach. This allows us to dynamically adjust the influence of each model, making XPose more adaptable to various input conditions and enhancing its performance across diverse scenarios. Additionally, we implemented an innovative image preprocessing technique within XPose, called Lineart from OpenPose. This technique generates line art versions of input images by emphasizing contours and boundaries, providing precise structural constraints. The incorporation of Lineart is particularly effective in reducing distortions related to clothing and body boundaries, thereby improving the overall quality and realism of the generated animations. We believe these enhancements make XPose a significant advancement in producing more authentic and meticulously controlled animations.

2 Literature Review

2.1 Data driven animation

Nowadays, research in image animation has primarily focused on the human body or face, utilizing diverse training data and domain-specific knowledge such as key points, semantic parsing, and statistical parametric models [7-9]. This has led to various animation techniques, categorized into implicit and explicit methods. Implicit Animation Methods transform the source image to the target motion signal by deforming the reference image in sub-expression space [10] or manipulating the latent space of a generative model [11]. Explicit Animation Methods: These methods warp the source image to the target pose using techniques like 2D optical flow [12], 3D deformation fields [7], or direct face swapping [8]. Recent approaches [9] also explore deforming points in 3D neural representations for improved temporal and multi-view consistency.

2.2 Diffusion model

Recent progress in diffusion models has led to significant improvements in text-to-image generation, controllable image generation, and video generation. A common approach involves generating 2D optical flow with diffusion models and using frame-warping techniques to animate reference images. Many frameworks use Stable Diffusion as their backbone and employ ControlNet [13] to condition animations on OpenPose keypoint sequences. Reference image conditioning often utilizes CLIP to encode images into a semantic text token space, guiding the generation process through cross-attention. However, these methods typically process each frame independently, neglecting temporal information and leading to flickering animations. MagicAnimate addresses these issues by incorporating advanced temporal modeling and robust appearance encoding to produce high-quality, temporally consistent human image animations

2.3 ControlNet

ControlNet enhances text-to-image diffusion models by introducing spatial conditioning controls. It leverages the robust encoding layers of pretrained models like Stable Diffusion, connecting them with trainable copies using zero convolution layers. This design ensures that the training process starts with minimal noise, preserving the pretrained model's strengths. ControlNet supports various conditioning inputs, allowing users to guide image generation with additional images specifying the desired composition. This flexibility makes ControlNet suitable for a wide range of applications, from artistic creation to practical design tasks.

3 Methodology

3.1 Data set description and preprocessing

To evaluate the performance of Xpose, we utilized a dataset consisting of 350 dancing videos sourced from TikTok [14]. Each video was truncated to a duration of 1-3 seconds. The frame dimensions of all videos were standardized to 512x512 pixels for input. Preprocessing was performed using DensePose and DWPose, resulting in the generation of corresponding guide videos. For a fair comparison of results, we employed the same demo video as a test set to evaluate the Magic Animate model.

3.2 Proposed approach

Given a reference image and two motion sequences and, each containing N frames. Our objective is to generate a video with high quality, consistent clothes, body and face through the inference of image and motion sequences based on pretrained models. Existing work with DensePose lacks fine details with features, including fingers, facial expressions and clothes. To address this, we propose XPose based on the combination of DWPose and DensePose. In addition, existing video of the face and hands is very unstable. We argue that these can be solved by blending the original and line drawings together. Therefore, we introduce a novel approach that combines original picture with its line sketch.

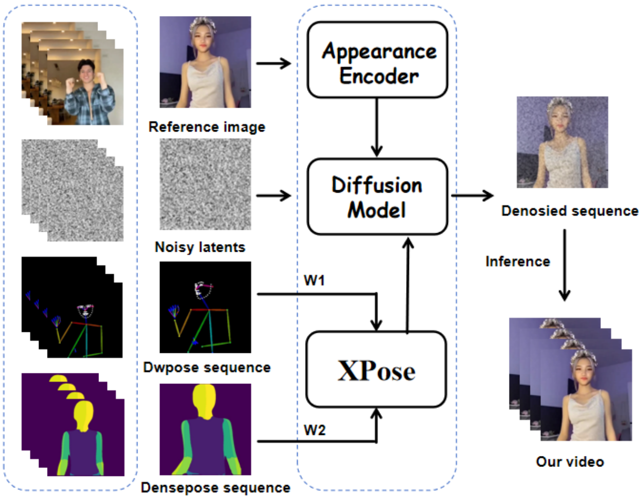

Fig. 1. XPose pipeline.

As illustrated in our inference flow (Figure 1), the Appearance Encoder extracts high-dimensional information from the reference image and inputs it into the diffusion model. The Xpose ControlNet then integrates the sampled information from DWPose and DensePose, applying different weights w1 and w2 respectively. Subsequently, a sequence of noisy latents is generated for each frame. The diffusion model then processes this sequence to produce a denoised output. Finally, this denoised sequence undergoes further inference to generate the final video.

XPose. DWPose is a framework aimed at improving the accuracy and efficiency of whole-body pose estimation, which involves localizing key points for the human body, hands, face, and feet [5]. It provides skeletal information, informing the model of the relative positions of various human joints. It also offers facial and hand skeletal information, which serves as additional reference points for facial expression generation, thus reducing facial distortion. The OpenPose model in ControlNet 1.1 is compatible with DWPose input, allowing us to utilize its pre-trained model directly, facilitating program improvements. However, when used independently, ControlNet cannot guide the precise contours of the human figure, leading to some unpredictable distortions.

DensePose aims to create detailed correspondences between Red, Green, Blue (RGB) images and a 3D model of the human body. It marks specific contours of the human body, clearly delineating the edges of the figure in video frames, which significantly aids ControlNet in understanding the figure's proportion within the frame. However, DensePose only provides contour information and lacks details on finger joints and facial expressions. Therefore, we build a novel network XPose, combining DensePose and DWPose within ControlNet to achieve better results than using a single pose model alone, as depicted in Figure 1. In addition, in ControlNet, we have two sampling methods, Down block resolution sampling and Mid-block resolutions sampling. The Down block resolution sampling is responsible for controlling the resolution of the subsample blocks. Its adjustments affect the details of how the network processes images during the forward process. The Mid-block resolutions sampling will affect how mid-tier features are represented, which in turn will affect the quality and detail of the resulting images.

Set Down block resolution samples for ControlNet of DWPose to ds1 and its weight to w1, and Down block resolution samples for DensePose to ds2 and its weight to w2. Then ds is the weighted sum of ds1 and ds2, and then a weight normalization is performed, and the total Down block resolution samples is:

$$ds = \frac{w_1 * ds_1 + w_2 * ds_2}{w_1 + w_2} \tag{1}$$

In addition to Down block resolution samples, Mid-block resolution samples should also be weighted and added. Set the mid-block resolution sample of DWPose as dm1. If mid-block resolution sample of DensePose is equal to dm2 and its weight is w2, then

$$dm = w_1 * dm_1 + w_2 * dm_2 \tag{2}$$

We have different settings to implement the XPose in video generating tasks. With suitable sample steps, we can get more features with more computational resources. Besides, the distribution of weights in XPose is also very significant since with the best weights, the video presents greater stability and robustness. More details are given in the experiment section.

Appearance encoder with line sketch. We use line drawings to enhance the edge information of the people in the image and highlight the details in the original image by overlaying the line drawings with the original reference image. This can be done by marking some of the edges of the artwork that are not obvious, so that ControlNet can better match the outline information of the artwork with that of the reference video. We generate the corresponding line diagram using the five preprocessors built into the stable diffusion ControlNet plug-in Figure 2.

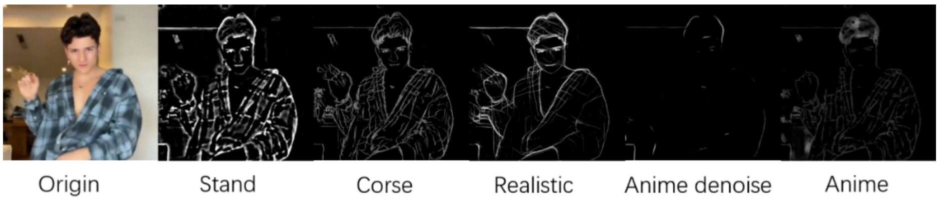

Fig. 2. Comparison of five different Lineart preprocessors used in ControlNet.

As illustrated Figure 2, the complexity of the line diagram generated by different preprocessors varies significantly. For our purpose, the line diagram used for superimposing should not be too complicated, as excessive detail can obscure the original colors of the image. Consequently, we selected Anime Denoise as our method, which prevents excessive line diagrams compared with the original image, as depicted in Figure 3.

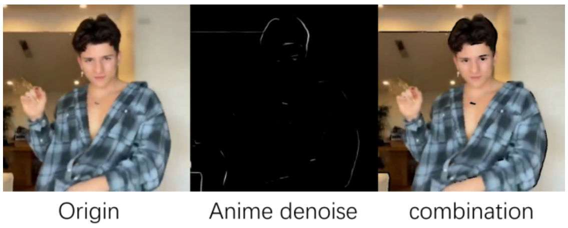

Fig. 3. Performance results after image overlay.

Set linear image is A, origin image is O. combination is C, threshold is t, the default is to perform operations on each pixel.

$$C = \begin{cases} O & if \quad A \leq t \\ (0,0,0) & if \quad A > t \end{cases} \tag{3}$$

We aim for the lineart to enhance the edge information of the reference image input, thereby increasing the stability of the human body in the resulting output. Additionally, the generated clothing should exhibit greater stability, minimizing errors in clothing generation.

3.3 Implementation details

We deployed the Xpose project on servers leased from the AutoDL platform, utilizing a single A100-PCIE-40GB Graphics Processing Unit (GPU) for computational power. The random seed was set to 1, with a sampling step of 25 and a guidance scale of 7.5. It is imperative that the DWPose video and DensePose video originate from the same source video. The optimal weight allocation between DWPose and DensePose was determined to be 1.5 to 1.

4 Result and Discussion

We provide quantitative comparisons against latest method MagicAnimate based on Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet Inception Distance (FID). Moreover, we test weight allocation method by setting different ratio while keeping sample step size 25 and the random seed 1. At last, we provide our findings about the stability of Lineart.

4.1 Evaluation metrics

We employ two evaluation metrics, LPIPS and FID, on the TikTok dataset to assess the performance of our model on single-frame image quality. In LIPIS, we feed two images as inputs into the neural network for feature extraction, and the output of each layer was normalized after activation, and then L2 distance was calculated after weighted multiplication of the w layer. In FID, for the generated image set and the real image set, their feature representations are computed through the Inception V3 model, respectively. The LPIPS value is expected to be lower, which indicates a higher perceptual similarity between the generated image and the original image. Similarly, a lower FID value signifies that the statistical distribution of the generated images closely aligns with that of the real images, reflecting higher generation quality.

4.2 Results of XPose

Table 1 and Figure 4 presents quantitative results between our model and MagicAnimate on TikTok datasets. Our method surpasses MagicAnimate in terms of the average score of evaluation metrics. Notably, our new model XPose improves against MagicAnimate by 9.1% with no line sketch and 23.5% in Deep Line and 10.6% in Simple Line. The observed improvement of a single-frame quality is attributed to the fact that MagicAnimate primarily captures the general outline of the character, neglecting finer details such as the hands and face. In contrast, DWPose effectively captures these critical details through localizing key points for the human body, improving the accuracy of human pose estimation and enhancing the overall quality of the animation. Additionally, by incorporating line drafts, more detailed information is integrated into the artwork, resulting in increased stability of pose generation during video generation.

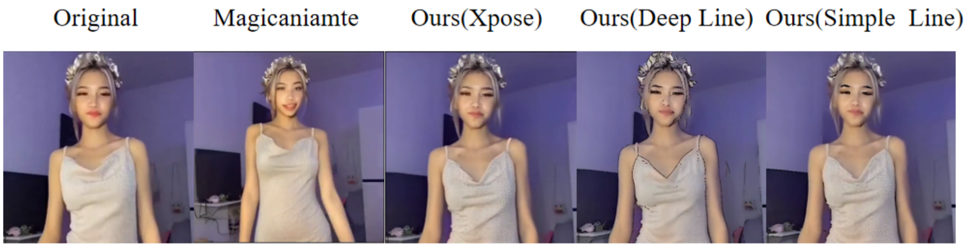

Fig. 4. Comparisons of different models.

Table 1. Quantitative results of different methods.

Evaluation Metrics	MagicAnimate	Ours (Xpose)	Ours (Deep Line)	Ours (Simple Line)
FID	110.2	**100.2**	**84.3**	**98.5**

4.3 Results of weight allocation method

At the same time, we tested the different weight ratio allocation numbers of our model XPose by changing the proportion of DWPose and DensePose,which is shown in Table 2 and Figure 5. After conducting several experiments, our model achieves the best FID and LPIPS scores with a weight ratio of 1.5:1, specifically 58.32 and 0.2532, respectively. The reason why this proportion proved to be the most effective is that while both DWPose and DensePose contribute to generating consistent videos, DWPose excels in handling the finer details of hands and faces, whereas DensePose does not effectively capture these features. Additionally, DensePose's control signals lack background information, which MagicAnimate fails to address, resulting in its inability to accurately learn the dynamic backgrounds presented in the TikTok dataset. However, an excessive increase in the weight assigned to DensePose negatively impacts the performance of our model. This is primarily due to the constraints imposed by its approach to human body shape outlining, which restricts the flexibility of the generated video. In contrast, DWPose demonstrates better performance in handling this aspect. Ultimately, we observe instability in the Lineart method, as indicated by its relatively high LPIPS value. This instability may be attributed to the LPIPS metric capturing numerous details like traced black lines of human body introduced by the Lineart process, which can obscure the original single-frame image and negatively impact the overall quality of the generated video.

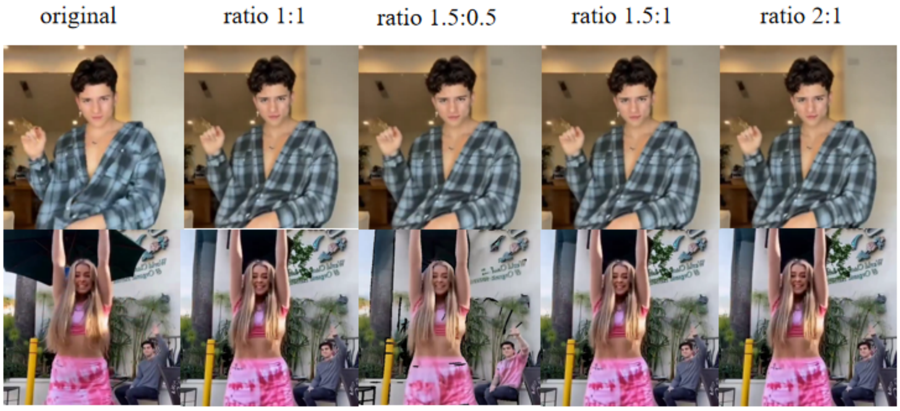

Fig. 5. Comparisons of different weights.

Table 2. Quantitative results of different weights .

Evaluation Metrics	1:1	1.5:0.5	1.5:1	2:1	1.5:1(Lineart)
FID	62.05	65.95	58.32	60.08	67.95
LIPIS	0.2535	0.2539	0.2532	0.2533	0.3097

5 Conclusion

We present XPose, a framework that enhances video generation through the integration of DensePose and DWPose. By combining the mid-sample and down-sample components of ControlNet, XPose establishes a unified guiding mechanism for video creation. Additionally, we investigate the influence of Lineart overlay on video quality. XPose effectively utilizes the strengths of both DensePose and DWPose, dynamically adjusting their contributions within ControlNet to optimize the generated output. Our quantitative analysis, using metrics such as FID and LPIPS, reveals that a weight ratio of 1.5:1 yields the most favorable results. While the use of line sketching offers some benefits in character body generation, it also introduces challenges such as instability and unwanted black edges. Experimental results confirm that XPose significantly enhances the consistency of facial expressions and body proportions, delivering robust and stable performance. In future research, we intend to explore the potential of using ControlNet images for guiding video generation. This study primarily focused on using a guide video as input for ControlNet, and further investigation could reveal additional improvements by using image-based guidance.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Güler R A Neverova N Kokkinos I 2018 Densepose: Dense human pose estimation in the wild Proceedings of the IEEE conference on computer vision and pattern recognition pp 7297-7306

[2] Chan C Ginosar S Zhou T et al. 2019 Everybody dance now Proceedings of the IEEE/CVF international conference on computer vision pp 5933-5942

[3] Wu J Z Ge Y Wang X et al. 2023 Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation Proceedings of the IEEE/CVF International Conference on Computer Vision pp 7623-7633

[4] Xu Z Zhang J Liew J H et al. 2024 Magicanimate: Temporally consistent human image animation using diffusion model Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp1481-1490

[5] Yang Z Zeng A Yuan C et al. 2023 Effective whole-body pose estimation with two-stages distillation Proceedings of the IEEE/CVF International Conference on Computer Vision pp 4210-4220

[6] Cao Z Simon T Wei S E et al. 2017 Realtime multi-person 2d pose estimation using part affinity fields Proceedings of the IEEE conference on computer vision and pattern recognition pp 7291-7299

[7] Cao C Hou Q Zhou K 2014 Displaced dynamic expression regression for real-time facial tracking and animation ACM Transactions on graphics vol 33 no 4 pp 1-10

[8] Nirkin Y Keller Y Hassner T 2019 Fsgan: Subject agnostic face swapping and reenactment Proceedings of the IEEE/CVF international conference on computer vision pp 7184-7193

[9] Xu H Song G Jiang Z et al. 2023 Omniavatar: Geometry-guided controllable 3d head synthesis Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 12814-12824

[10] Thies J Zollhofer M Stamminger M et al. 2016 Face2face: Real-time face capture and reenactment of rgb videos Proceedings of the IEEE conference on computer vision and pattern recognition pp 2387-2395

[11] Oorloff T Yacoob Y 2023 Robust one-shot face video re-enactment using hybrid latent spaces of stylegan2 Proceedings of the IEEE/CVF International Conference on Computer Vision pp 20947-20957

[12] Siarohin A Lathuilière S Tulyakov S et al. 2019 First order motion model for image animation Advances in neural information processing systems p 32

[13] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023 Adding conditional control to text-to-image diffusion models. In ICCV

[14] Jafarian Y Park H S 2021 Learning high fidelity depths of dressed humans by watching social media dance videos Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 12753-12762