Beyond Perspectives: Enhancing Pose Estimation via Viewpoint Transformation

Jung Won Yoon, Hyun Jun Yook, Jae Eun Seo, Jae Hun Choi, and Youn Kyu Lee*
Department of Computer Engineering
Hongik University
Seoul, Republic of Korea
pattyyoon@g.hongik.ac.kr, hyunjunyook@g.hongik.ac.kr, itws009@g.hongik.ac.kr,
chlwogns316@g.hongik.ac.kr, younkyul@hongik.ac.kr

***Abstract*—A number of approaches have been proposed to address the occlusion and truncation issues in pose estimation, but they typically require additional devices or specific recording environments. In this paper, we propose a novel pose estimation mechanism that utilizes GAN-based viewpoint transformation. Our mechanism complements missing pose information without requiring additional devices or a pre-aligned camera setups. It achieves this by transforming a supplementary viewpoint video to the target viewpoint video utilizing GAN and using it to complement missing keypoints in the target viewpoint video. The evaluation results confirm that our mechanism effectively complements missing pose information and provides reliable viewpoint transformation performance.**

***Index Terms*—human pose estimation, deep learning, viewpoint transformation, keypoint complementation**

I. Introduction

Human pose estimation (HPE), which involves determining the pose of a human by estimating the positions of body keypoints, has been extensively employed to assess the postural correspondence in videos captured by mobile devices, such as indoor workout analysis and health condition assessment [1] [2] [3]. However, HPE is susceptible to challenges such as "occlusion," which occurs when a body part is obstructed by an object or a person, and "truncation," which happens when body parts are outside the recording boundaries [1]. As a result, inaccurate HPE can occur, particularly in mobile videos recorded under constrained environments with limited shooting angles and focal lengths [4].

Although a number of methods have been proposed to address these challenges, they require additional devices such as inertial measurement units (IMUs) [5] or depth sensors [6]. Furthermore, multi-view pose estimation techniques have also been proposed [7], but their applicability is limited to specific recording environments due to the requirement of a pre-aligned set of cameras, fixed shooting positions, and synchronization between devices [8].

We propose a novel pose estimation mechanism that utilizes a generative adversarial network (GAN) to convert viewpoints of different videos and complement the keypoints extracted from each viewpoint. Specifically, the target (viewpoint) video and the supplementary (viewpoint) video are simultaneously captured using different cameras. The supplementary video is then transformed to align with the viewpoint of the target video through image-to-image translation using GAN. Keypoints are extracted from both the transformed video and the target video. For any missing keypoints in the target video, they are complemented using the keypoints from the transformed video. Our mechanism enhances the overall performance of pose estimation by enabling more accurate keypoint estimation through the complementation of missing keypoints, without requiring additional devices or a pre-aligned camera setups.

The contributions of this paper are as follows: (1) Proposal of a novel pose estimation mechanism that utilizes GAN-based viewpoint transformation to enable complementation of missing keypoints; (2) Proposal of a novel multi-view pose estimation mechanism that can be applied without additional devices or pre-aligned camera setups; (3) Case study using real-world datasets and validation of the proposed mechanism's effectiveness.

This paper is organized as follows: Section 2 provides an overview of related works. Section 3 presents the mechanism proposed in this paper. Section 4 presents the experimental results of our mechanism. Finally, Section 5 concludes the paper.

II. Related Work

A. Image-to-Image Translation

Image-to-image translation enables the transformation of one image domain to another by training neural network models such as CNN and GAN with the mapping information between the input and output images [9]. Pix2pix utilizes conditional GAN to learn mapping information between images and generate an image by transforming specific aspects of the input image to a desired condition [10]. It is widely utilized in diverse generative tasks, including map generation, aerial photo generation, and image colorization. Unlike pix2pix, which relies on a set of image pairs with similarities, CycleGAN operates with unpaired training data. It enables image domain transformation by learning the mapping information between two unpaired domains: the source domain and the target domain [11]. UGATIT operates with unpaired training data, which incorporates attention feature maps and a learnable

*Corresponding Author

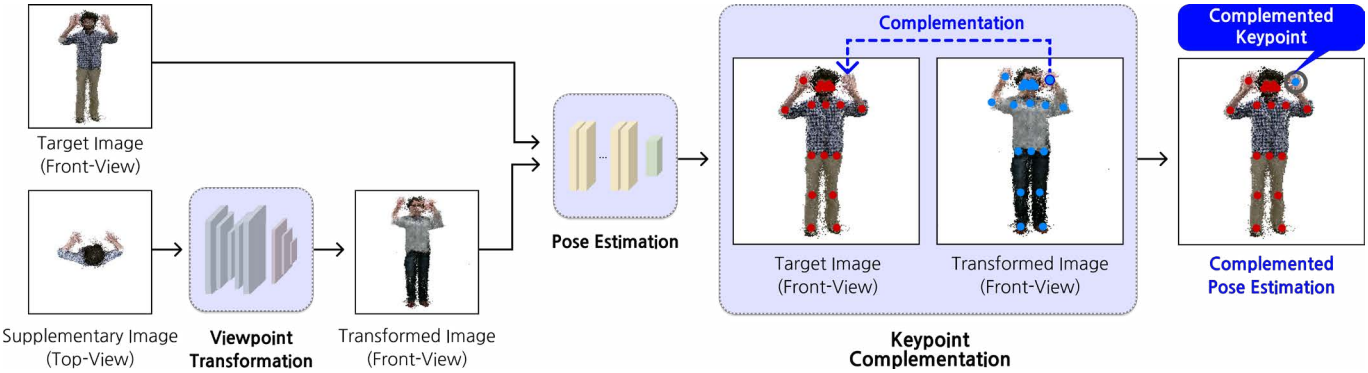

Fig. 1. Overview of Our Mechanism

normalization function to guide the translation process by distinguishing important regions from minor ones [12].

B. Human Pose Estimation

Human pose estimation is a technique that analyzes the pose of a person in videos or images by identifying the locations of joints or body parts as keypoints [1]. OpenPose, a popular real-time multi-person pose estimation mechanism, is a CNN-based 2D method for estimating the poses of multiple individuals [13]. AdaFuse is a multi-view fusion method that leverages adaptive weights to combine information from multiple viewpoints captured by a set of pre-aligned cameras [8]. DECA incorporates a capsule encoder to model each joint of the input image as a capsule object and utilizes a multi-task decoder for pose estimation, thereby demonstrating enhanced performance, especially in top-view scenarios [14]. ViTPose incorporates a ViT encoder and a lightweight decoder, enabling fast inference and delivering superior performance compared to other pose estimation methods [15].

Overall, existing pose estimation methods generally fail to successfully address "occlusion" and "truncation" and essentially rely on specialized recording environments, such as additional devices or pre-aligned camera setups. Hence, it is required to design a new pose estimation mechanism that can be applied in general recording environments and is capable of effectively addressing these challenges.

III. Our Method

In this paper, we propose a novel human pose estimation (HPE) mechanism that complements missing keypoints through viewpoint transformation by incorporating GAN-based image-to-image translation. Our proposed mechanism addresses common challenges in HPE in general recording environments, including occlusion and truncation.

Fig. 1 presents an overview of the proposed mechanism, which utilizes a frame-wise segmented image of a video as input. During the training process, the viewpoint transformation module is trained to convert a supplementary image into a transformed image with the same viewpoint as the target image. During the HPE process, the trained viewpoint transformation module converts a supplementary image into a transformed image. Subsequently, the pose estimation module estimates the keypoints of both the target image and the transformed image. By utilizing the keypoints estimated from the transformed image, the keypoint complementation module fills in the missing keypoints of the target image. The detailed operation of each module is as follows.

A. Viewpoint Transformation

The viewpoint transformation module converts a supplementary image to have the same viewpoint as the target image. It incorporates *pix2pix* [10], which learns the mapping information between image pairs and generates an image that reflects the intended aspect of the input image. By training the viewpoint transformation module with a paired dataset of target images and supplementary images, it can effectively convert a supplementary image into a transformed image with the target viewpoint.

B. Pose Estimation

The pose estimation module estimates the keypoints of the human instance detected within the target image and the transformed image, respectively. The pose estimation module incorporates *ViTPose* [15], a transformer-based HPE model, which extracts a feature map, performs upsampling of the feature map, and regresses the heatmap to determine a total of 17 keypoints.

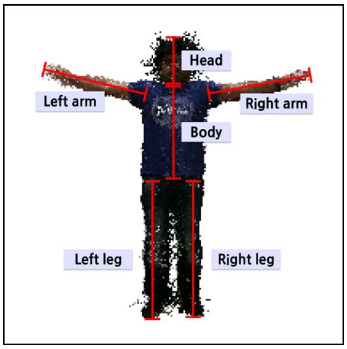

Fig. 2. Example of Measuring Lengths of Body Parts

Algorithm 1: *Keypoint complementation*

Input: $TG \Leftarrow$ a list of keypoints for the target image
$TF \Leftarrow$ a list of keypoints for the transformed image
$\delta \Leftarrow$ a confidence score threshold
$k \Leftarrow$ a number of all keypoints
Output: $Comp \Leftarrow$ a list of complemented keypoint coordinates

1 Let (C, S) be a keypoint where C represents the estimated coordinates and S represents the confidence score
2 Let N be a list of non-missing keypoint indexes
3 Let n be a number of non-missing keypoints
4 Let s be a sum of translation vector of non-missing keypoints
5 Let v be an average translation vector

6 **for** $i \leftarrow 1$ *to* k **do**
7 **if** $TG[i].S > \delta$ **then**
8 *append* i to N
9 $n \leftarrow n + 1$

10 **for** $i \leftarrow 1$ *to* k **do**
11 **if** $i \in N$ **then**
12 *add* $(TG[i].C - TF[i].C)$ to s

13 $v = \frac{s}{n}$

14 **for** $i \leftarrow 1$ *to* k **do**
15 **if** $i \in N$ **then**
16 *append* $TG[i].C$ to $Comp$
17 **else**
18 *append* $(v + TF[i].C)$ to $Comp$

19 **return** $Comp$

C. Keypoint Complementation

The keypoint complementation module complements the missing keypoints from the target image using the keypoints from the transformed image. It assumes that the transformed image is generated with the same proportional lengths for all body parts as the target image. Specifically, the lengths of body parts (head, left arm, right arm, body, left leg, and right leg) are measured as depicted in Fig. 2. The keypoint complementation module verifies whether the proportions between the lengths of these body parts are identical between the target image and the transformed image. The detailed mechanisms of the keypoint complementation module are depicted in Algorithm 1.

In Algorithm 1, TG and TF refer to the list of keypoints for the target image and the transformed image, respectively. Each keypoint is represented as a tuple: (C, S), where C represents the estimated coordinates and S represents the confidence score. Additionally, δ denotes a pre-defined threshold for the confidence score, and k represents the number of all keypoints.

Algorithm 1 iterates the following operation for k times. In each iteration, keypoints from the target image are classified as non-missing keypoints if their confidence score exceeds a pre-defined threshold (=δ). Then the index of each non-missing keypoint is appended to a list (=N), and the number of non-missing keypoints (=n) is incremented (lines 6-9).

Algorithm 1 then iterates the following operation for k times. If the index i is contained in N, the algorithm calculates the cumulative sum of the translation vectors (=s) based on the difference between the coordinates of the i-th keypoint in TG and TF. Afterwards, s is divided by the number of non-missing keypoints (=n) to compute the average translation vector (=v) that will be used for keypoint complementation (lines 10-13).

Finally, Algorithm 1 iterates the following operation for k times. If the index i is contained in N, the algorithm appends the coordinates of the target image keypoint to $Comp$. Otherwise, it appends the coordinates obtained by adding the average translation vector (=v) to the transformed image keypoint to $Comp$. Ultimately, the algorithm returns the list of complemented keypoints (=$Comp$) (lines 14-19).

IV. Evaluation

We evaluate the effectiveness of our proposed mechanism by addressing the following two research questions:

- **RQ#1**: How effectively does our proposed mechanism improve pose estimation?
- **RQ#2**: How effectively does our proposed mechanism work for viewpoint transformation?

In our evaluations, we selected the front-view as the target viewpoint and the top-view as the supplementary viewpoint. Since the top-view provides a completely different perspective from the front-view, it can include additional information that complements potential occlusions and truncations in the front-view video. Since our mechanism uses frame-wise segmented images of a video as input, as described in Section 3, the evaluations were performed using an image dataset.

A. Experimental Setting

In our evaluations, we selected the PanopTOP31K dataset [16], which comprises images of entire-body poses from multiple viewpoints for 23 different subjects (front-view: 25,604 train images and 8,044 test images; top-view: 25,604 train images and 8,044 test images). The viewpoint transformation module was implemented based on *pix2pix* [10], an image-to-image translation model, and trained using the PanopTOP31K dataset. The pose estimation module was implemented based on *ViTPose* [15].

The environment and hyperparameters used in our evaluations are as follows: (1) *pix2pix*: NVIDIA Geforce RTX 3070 GPU, Python 3.9.16, PyTorch 2.0.1+cu117, plateau lr_policy, 1 batch size, 149 epochs; (2) *ViTPose*: ViTPose-25-S, NVIDIA Geforce RTX 3070 GPU, Python 3.9.16, PyTorch 2.0.1+cu117;

For RQ#1, to validate the effectiveness of our proposed mechanism, we first generated a masked image dataset as follows. As shown in Fig. 3, we selected eight masking regions: (a) left hand, (b) right hand, (c) both hands, (d) left arm, (e) right arm, (f) both arms, (g) left leg, and (h) right leg. From the front-view test images, we randomly selected 20 images and masked the eight regions, resulting in a total of 160 masked images. For our evaluation metric, we defined the percentage of correct keypoints (PCK) as follows. We set

TABLE I
PCK PERFORMANCE COMPARISON BETWEEN MASKED IMAGES AND OUR MECHANISM

Target	Left Hand	Right Hand	Both Hands	Left Arm	Right Arm	Both Arms	Left Leg	Right Leg	Mean
Masked	96.58%	96.32%	91.32%	91.32%	91.05%	81.58%	95.26%	93.42%	92.11%
Our Mechanism	**97.89%**	**96.58%**	**96.84%**	**96.84%**	**96.84%**	**94.47%**	**97.37%**	**96.84%**	**96.71%**

the keypoints estimated from the unmasked target image as the ground truth. Then, we measure the PCK by comparing the keypoints estimated from the masked image and the keypoints complemented by our mechanism on the masked image.

For RQ#2, to assess the viewpoint transformation performance of our proposed mechanism, we utilized 8,044 top-view test images. To evaluate the similarity between a front-view test image, which corresponds to a top-view test image pair, and the transformed (front-view) images obtained by applying our viewpoint transformation module to the top-view test image, we utilized the perceptual similarity metric LPIPS [17] and the structural similarity metric SSIM [18].

B. Experimental Results

(RQ#1) Effectiveness of our mechanism in pose estimation: Table I presents the PCK performance comparison between the masked images and their complemented images using our mechanism for each masking region. For the masked image, the PCK for each region is as follows: left hand = 96.58%, right hand = 96.32%, both hands = 91.32%, left arm = 91.32%, right arm = 91.05%, both arms = 81.58%, left leg = 95.26%, and right leg = 93.42%. For the complemented image, the PCK for each region is as follows: left hand = 97.89%, right hand = 96.58%, both hands = 96.84%, left arm = 96.84%, right arm = 96.84%, both arms = 94.47%, left leg = 97.37%, and right leg = 96.84%. The results demonstrate that the complementation achieved through our mechanism successfully enhanced the performance, resulting in an average increase of 4.60 percentage points.

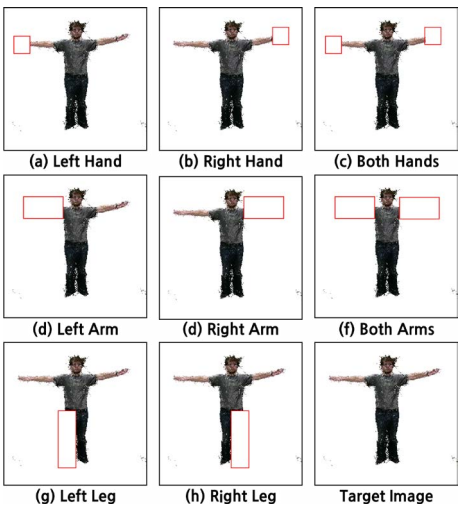

Fig. 3. Examples of Masked Images

TABLE II
PERFORMANCE COMPARISON OF VIEWPOINT TRANSFORMATION BETWEEN THE EXISTING METHOD AND OUR MECHANISM

Methods	LPIPS	SSIM
DiOr [23]	0.176	0.806
Our Mechanism	**0.127**	**0.829**

(RQ#2) Effectiveness of our mechanism in viewpoint transformation: Table II presents the performance comparison of viewpoint transformation between the existing method (*DiOr* [19]) and our mechanism. The evaluation is based on the perceptual similarity and structural similarity between each front-view test image and the transformed image obtained through our mechanism. Note that the transformed image was derived by applying our mechanism to transform the corresponding top-view test image into the front-view image. When comparing the performance of our mechanism (LPIPS = 0.127 and SSIM = 0.829) with that of the state-of-the-art pose transfer method, *DiOr* [19] (LPIPS = 0.176 and SSIM = 0.806), our mechanism exhibited superior performance. Specifically, our mechanism achieved a lower LPIPS score by 0.049 and a higher SSIM score by 0.023. Considering the characteristics of the evaluation metrics (i.e., lower values indicating better performance for LPIPS, and higher values indicating better performance for SSIM), the results confirm that our mechanism provides effective viewpoint transformation performance.

Fig. 4 shows the representative images transformed by our mechanism. Specifically, the first column displays the target image, the second column displays the supplementary image paired with the target image, and the last column displays the transformed image obtained by applying our mechanism to the supplementary image.

The results of the comparison for each pose are as follows. For case (a), when examining the distinctive features of the target image, such as the pose of raising both hands, the position of hands on the sides of the head, and the angles between the shoulders and elbows, the transformed image was generated to closely resemble the pose of the target image. For case (b), when examining the distinctive features of the target image, such as the pose of straightened arms to both sides and the position of hands raised to shoulder height, the transformed image was generated to closely resemble the pose of the target image. For case (c), when examining the distinctive features of the target image, such as the pose of both hands placed on the waist, the position of the hands on the waist, the angle of the shoulders, and the angle of the elbows bent towards the body, the transformed image was generated to closely resemble the

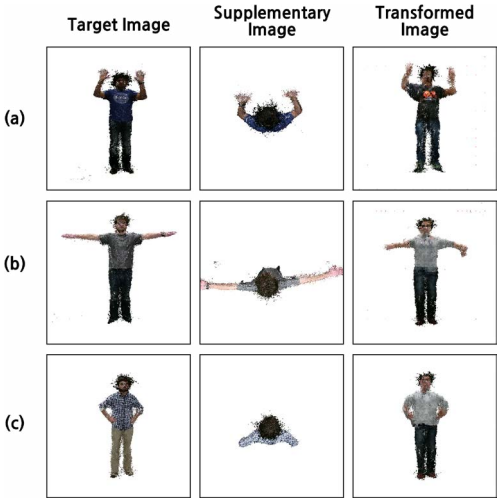

Fig. 4. Representative Examples of Viewpoint Transformation via Our Mechanism

pose of the target image.

V. Conclusion

In this paper, we propose a novel human pose estimation (HPE) mechanism that employs GAN-based viewpoint transformation to complement missing keypoints in the target viewpoint image. The results of keypoint complementation using our mechanism on the masked image confirm the effectiveness of our proposed approach in complementing the missing keypoints in the target viewpoint images. Furthermore, the results of evaluating the similarity between the viewpoint-transformed images generated by our mechanism and the target images verify that our mechanism provides valid viewpoint transformation performance. Our future work will focus on validating the effectiveness of our mechanism on high-resolution datasets containing various poses. Additionally, we will explore the potential of our mechanism to be applied in multi-person scenarios, as well as its compatibility with advanced image reconstruction methods [20].

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-00165648).

References

[1] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Comput. Surv.*, Jun. 2023.

[2] A. Nagarkoti, R. Teotia, A. K. Mahale, and P. K. Das, "Realtime indoor workout analysis using machine learning & computer vision," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, pp. 1440–1443, IEEE, Jul. 2019.

[3] J. Stenum, K. M. Cherry-Allen, C. O. Pyles, R. D. Reetzke, M. F. Vignos, and R. T. Roemmich, "Applications of pose estimation in human health and performance across the lifespan," *Sensors*, vol. 21, p. 7315, Nov. 2021.

[4] J. Zhang, D. Zhang, X. Xu, F. Jia, Y. Liu, X. Liu, J. Ren, and Y. Zhang, "Mobipose: Real-time multi-person pose estimation on mobile devices," in *Proc. 18th Conf. Embedded Netw. Sensor Syst.*, pp. 136–149, Nov. 2020.

[5] R. Kianifar, V. Joukov, A. Lee, S. Raina, and D. Kulić, "Inertial measurement unit-based pose estimation: Analyzing and reducing sensitivity to sensor placement and body measures," *J. Rehabil. Assist. Technol. Eng.*, vol. 6, p. 2055668318813455, Jan. 2019.

[6] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake, "Efficient human pose estimation from single depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 2821–2840, Dec. 2013.

[7] S. Ershadi-Nasab, E. Noury, S. Kasaei, and E. Sanaei, "Multiple human 3d pose estimation from multiview images," *Multimedia Tools and Appl.*, vol. 77, no. 12, pp. 15573–15601, 2018.

[8] Z. Zhang, C. Wang, W. Qiu, W. Qin, and W. Zeng, "Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild," *Int. J. Comput. Vis.*, vol. 129, no. 3, pp. 703–718, 2021.

[9] A. Alotaibi, "Deep generative adversarial networks for image-to-image translation: A review," *Symmetry*, vol. 12, p. 1705, Oct. 2020.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5967–597, Jul. 2017.

[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 2223–2232, Oct. 2017.

[12] J. Kim, M. Kim, H. Kang, and K. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Apr. 2020.

[13] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7291–7299, Jul. 2017.

[14] N. Garau, N. Bisagno, P. Bródka, and N. Conci, "Deca: Deep viewpoint-equivariant human pose estimation using capsule autoencoders," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 11677–11686, Oct. 2021.

[15] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 38571–38584, 2022.

[16] N. Garau, G. Martinelli, P. Bródka, N. Bisagno, and N. Conci, "Panoptop: a framework for generating viewpoint-invariant human pose estimation datasets," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, pp. 234–242, Oct. 2021.

[17] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 586–595, Jun. 2018.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, Apr. 2004.

[19] A. Cui, D. McKee, and S. Lazebnik, "Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 14638–14647, Oct. 2021.

[20] H. Yoo, P. M. Hong, T. Kim, J. W. Yoon, and Y. K. Lee, "Defending against adversarial fingerprint attacks based on deep image prior," *IEEE Access*, vol. 11, pp. 78713–78725, Jul. 2023.