

A ROBUST FRAMEWORK OF CHROMOSOME STRAIGHTENING WITH ViT-PATCH GAN

Sifan Song^{1,3*}, Jinfeng Wang^{1,3*}, Fengrui Cheng^{1*}, Qirui Cao¹, Yihan Zuo¹, Yongteng Lei¹, Ruomai Yang¹, Chunxiao Yang², Frans Coenen³, Jia Meng¹, Kang Dang^{4†}, Jionglong Su^{1†}

¹ Xi'an Jiaotong-Liverpool University, Suzhou, China

² Suzhou Sano Precision Medicine Ltd., Suzhou, China

³ University of Liverpool, Liverpool, UK

⁴ VoxelCloud, Inc., Los Angeles, USA

ABSTRACT

Chromosomes carry the genetic information of humans. They exhibit non-rigid and non-articulated nature with varying degrees of curvature. Chromosome straightening is an important step for subsequent karyotype construction, pathological diagnosis and cytogenetic map development. However, robust chromosome straightening remains challenging, due to the unavailability of training images, distorted chromosome details and shapes after straightening, as well as poor generalization capability. In this paper, we propose a novel architecture, ViT-Patch GAN, consisting of a self-learned motion transformation generator and a Vision Transformer-based patch (ViT-Patch) discriminator. The generator learns the motion representation of chromosomes for straightening. With the help of the ViT-Patch discriminator, the straightened chromosomes retain more shape and banding pattern details. The experimental results show that the proposed method achieves better performance on Fréchet Inception Distance (FID), Learned Perceptual Image Patch Similarity (LPIPS) and downstream chromosome classification accuracy, and shows excellent generalization capability on a large dataset.

Index Terms— Chromosome Straightening, Vision Transformer, Generative Adversarial Networks.

1. INTRODUCTION

In a normal human cell, there are 22 pairs of autosomes (Type 1-22) and one pair of heterosomes (Type X & Y in male and two copies of Type X in female). By karyotype analysis, chromosome aberrations can be detected in the diagnosis of many genetic diseases, such as the Klinefelter syndrome [1] and specific cancers [2]. The banding patterns of chromosomes (unique light and dark stained bands) provide important evidence for the development of cytogenetic maps. Due to their non-rigid nature, condensed chromosomes exhibit varying degrees of curvature under the microscope. Therefore, chromosome straightening is an important upstream task for chromosome classification [3] and the subsequent karyotype construction and cytogenetic map development [4].

Chromosome straightening has been studied for a long time and its development may be described in three stages. (i) Printed images of bent chromosome are physically cut and rearranged for straightening [5, 6]. (ii) Geometric algorithms are extensively designed based on chromosome micrographs for automatic straightening, which consist of two main categories: extraction of the medial axis [7, 8, 9, 10] and finding bending points [11, 12, 13]. However, these geometric methods may extract inaccurate medial axes when chromosomes have large widths. The methods of bending point localization often use a cut line to separate chromosome arms, leading to discontinuous banding patterns. (iii) A Single Chromosome Straightening Framework (SCSF) [14] using conditional generative adversarial networks (cGAN) is proposed to straighten chromosomes with uninterrupted chromosome banding patterns. However, it requires a large number of input image pairs for training a model for each chromosome. ChrSNet [15] is proposed as a two-module framework with self-attention mechanism, but some edge details are not well-preserved. In addition, there are many studies on image deformation, such as medical image registration [16, 17]. The image registration technique requires two real-world chromosomes with the identical shape details but different curvatures, which is almost impossible to obtain. Thus it also fails in this chromosome straightening task. Recently, First Order Motion Model (FOMM) [18] and PCA-based Motion Estimation Model (PMEM) [19] are proposed by learning key point-based and region-based representations for motion transfer, respectively. However, they still require to train on sufficient image pairs for robust performance.

Challenges remain in the three stages of chromosome straightening. (1) *Lack of training images*. It is almost impossible to take micrographs of two chromosomes with identical stained banding pattern but different curvatures due to the diversity of random mutations, chromosome condensation and laboratory conditions. Thus, it is challenging to train a robust deep learning-based model for straightening. (2) *Distorted chromosome details and shapes after straightening*. The condensed chromosomes are non-rigid with varying degrees of curvature. Straightened chromosome results require a high degree of preservation of consistent shape and details in the source image. Image registration and motion transfer methods tend to generate distorted results with driving image

* Contributed Equally

† Corresponding Authors. kdang@voxelcloud.net.cn; Jionglong.Su@xjtlu.edu.cn

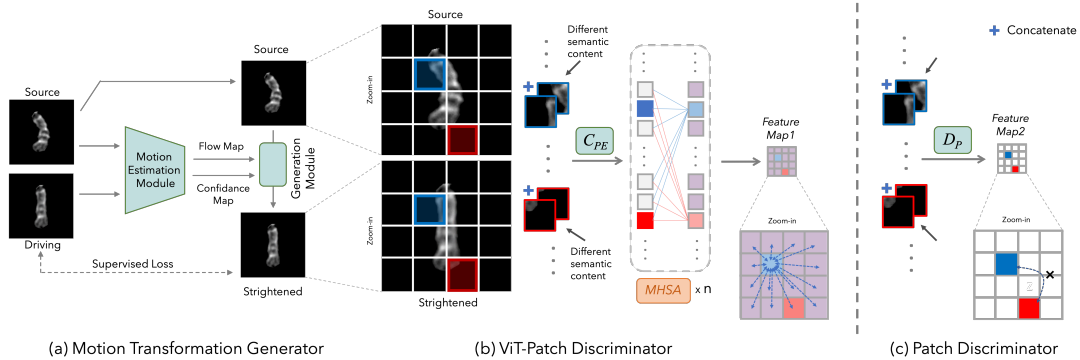


Fig. 1. Overview of the architectures of the proposed framework, including (a) motion transformation generator and (b) our proposed ViT-Patch discriminator; (c) gives the structure comparison of a basic patch discriminator.

shapes. (3) *Poor generalization capability.* The recent cGAN-based chromosome straightening framework [14] only learns mapping dependencies for each specific banding patterns, making it very time-consuming for large-scale applications.

The main contributions of this research address the above problems and are as follows:

- We propose a robust cGAN-based framework for chromosome straightening on a small dataset, by transferring the chromosome straightening task to the motion transformation task of non-rigid objects.
- We propose a novel architecture, ViT-Patch discriminator, to improve the detail preservation ability of our framework. Compared with existing methods, straightened chromosomes retain more shape and banding pattern details of the corresponding source images.
- Different from SCSF [14], our trained model shows excellent generalization capability and has the ability to be applied to a large chromosome dataset for straightening.

2. METHODOLOGY

2.1. Network Architecture

Fig. 1 presents an overview of the proposed ViT-Patch GAN architecture. It is comprised of two parts, a motion transformation generator (Fig. 1-a) and a ViT-Patch discriminator (Fig. 1-b). Compared to SCSF [14], ViT-Patch GAN requires a small dataset size for training (642 images). It straightens chromosomes with highly preserved details, and has high generalization capacity.

Generalized Framework of Chromosome Straightening. We apply PMEM [19] as the generator part since it learns motion representations through self-learned region estimation. We can consider the chromosome straightening task as a motion transformation task for non-rigid objects. In the training stage, the generator requires a *training source* and *training driving* images with the same chromosome but different curvatures. As shown in Fig. 1-a, using the Motion Estimation Module, the flow and confidence maps containing a combination of region and background transformations are then fed into the Generation Module to synthesize the straightened chromosome. Subsequently, the straightened

chromosome is supervised with the *training driving* image by a supervised reconstruction loss (\mathcal{L}_1), consistent with PMEM [19]. One of the challenges of this task is the lack of *training driving* data. However, PMEM was trained on sufficient image pairs (video clips containing only the same object) [19]. On a small dataset, PMEM may inadequately transfer the shape of the driving image to the source image, leading to inaccurate straightening results.

To alleviate this drawback, we propose a ViT-Patch discriminator (Fig. 1-b) that encodes not only local details but also global feature connection. The final loss is the standard adversarial loss of the generator and discriminator in the cGAN-based framework [20]. During testing, the converged model straighten chromosomes by feeding *test source* and *test driving* images. It is worth noting that the *test driving* can be a different straight chromosome to the *test source* since the motion representation has been learned.

ViT-Patch Discriminator. Fig. 1-b and Fig. 1-c give the comparison between our proposed ViT-Patch discriminator and a basic patch discriminator. In a basic patch GAN [20], the semantic content of the corresponding patches at the same position between the source and generated images is generally the same. However, in this task, curvatures of a chromosome change after straightening. As a result, patches at the same position before and after straightening may contain significantly different chromosome patterns (*e.g.* the concatenated patches with black arrows in Fig. 1-b and Fig. 1-c).

In training a basic patch GAN, two patches at the same position are concatenated and fed into the discriminator network (D_P) that contains several consecutive convolutional blocks to output the *Feature Map2* for adversarial training (Fig. 1-c). Although adjacent patches overlap according to the receptive field, long-distance patches are independent and not informatively linked. This deficiency may lead to inaccuracy and limited performance in the chromosome straightening task. To address this, for the proposed ViT-Patch discriminator, patches at the same position are fed into a convolutional patch embedding layer (C_{PE}), and then processed by N Multi-Head Self-Attention (MHSA) Blocks as ViT [21] (4-16 used for ablation experiments in Section 3.3). Afterwards, the encoded features contain long-range dependencies across

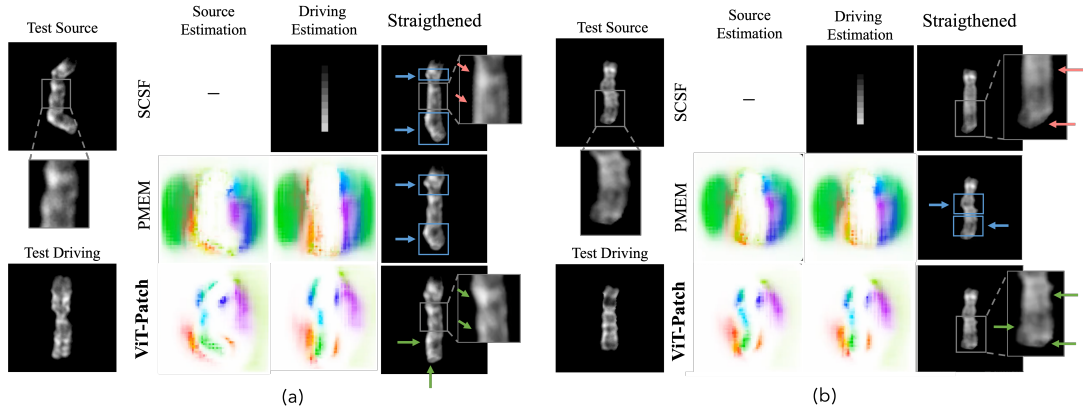


Fig. 2. Visual results of chromosome straightening generated by different methods.

the image to compensate for information loss. Thus the *Feature Map1* contains not only the local semantic content but also the information connectivity of the entire chromosome.

2.2. SL-matching Scheme

In addition to proposing a ViT-Patch discriminator to improve generalization, empirically, we found a large shape difference between the *test source* and *test driving* chromosome, leading to significant distortion. To alleviate this problem, we propose a Size-LPIPS matching scheme (SL-matching) to select a *test driving* with a similar size and shape for each *test source* image. The SL-matching scheme is comprised of two phases. In Phase 1 we perform a line-by-line scan of the chromosome image and record the midpoint of each line containing non-zero pixels. The resulting set of midpoints is first smoothed using a moving average algorithm, and its length calculated. Next, we perform a column-by-column scan, where we take the x -axis coordinates of all non-zero pixel columns as the width. The top three candidates whose lengths and widths are most similar to the corresponding *test source* are selected. In Phase 2 we calculate the perceptual score between a *test source* and each candidate chromosome. This score is the average result generated by LPIPS with AlexNet and VGG backbones [22]. Finally, the selected image has the largest size and perceptual similarity.

3. EXPERIMENTS

3.1. Experiment Setup

Datasets and Implementation Details. A total of 16696 chromosome micrographs were used for the evaluation as provided by [14]. All data had been desensitized, and patients’ personal information has been removed. Since 642 out of the 16696 chromosomes were straightened by SCSF [14], these 642 chromosomes were used as the dataset (D_{train}) for training and testing the ViT-Patch GAN in a 4:1 ratio. In the training stage, the original chromosomes and corresponding straightened results in D_{train} were utilized as *training*

Table 1. Comparison with existing studies, the best and second best results highlighted in bold and underlines. The DCA result is the average score of each experiment with cross-validation.

DCA (%)	FID↓	LPIPS _A ↓	LPIPS _V ↓	DCA _{R34}	DCA _{R50}	DCA _{D169}
Original	-	-	-	91.72	85.63	83.59
SCSF [14]	54.90	0.1272	0.0974	90.63	86.87	85.31
FOMM [18]	53.71	0.1310	0.0982	92.19	85.16	88.28
PMEM [19]	43.80	0.1196	0.0897	92.03	86.09	84.69
ViT-Patch	42.21	0.1160	0.0874	93.91	90.16	89.06

↓ represents that lower results are better.

source and *training driving* images, respectively. Our clinical experts select 1200 real-world straight chromosomes of different lengths from all chromosome types as the dataset, $D_{driving}$. In the testing stage, $D_{driving}$ was used to select *test driving* for each *test source* utilizing the SL-matching scheme. A large dataset was employed in this study to assess generalization capacity: the remaining 14854 chromosome images provided by [14] (D_{large}).

All experiments were implemented using one NVIDIA GeForce RTX 2080Ti GPU and coded using PyTorch. The Adam optimizer and a batch size of 1 were used. The total training epoch was 50 and the ratio of the training and test sets was 4:1 (5-fold cross-validation). The initial learning rates of the generator and discriminator were $5e^{-5}$ and $1e^{-5}$, respectively. We used MultiStepLR for the generator and discriminator with milestones 30 and 45. After completing cross validation, all 642 chromosomes of D_{train} were straightened. Subsequently, we performed downstream classification experiments (7 chromosome types) using initial learning rate of $4e^{-5}$ and ReduceLRonPlateau with patience 5 and early stopping protocol with patience 20. All input images were preprocessed to png format (256×256).

Evaluation Metrics. To quantitatively assess the quality of straightened chromosomes, we used the following three evaluation metrics: (i) Fréchet Inception Distance (FID) (ii) Learned Perceptual Image Patch Similarity (LPIPS) and (iii) Downstream Classification Accuracy (DCA). FID [23] assesses the quality of generated images by comparing the distribution between the real and generated images. LPIPS [22] estimates the perceptual similarity. Its results are closer to human judgment than many traditional metrics, such as \mathcal{L}_2 ,

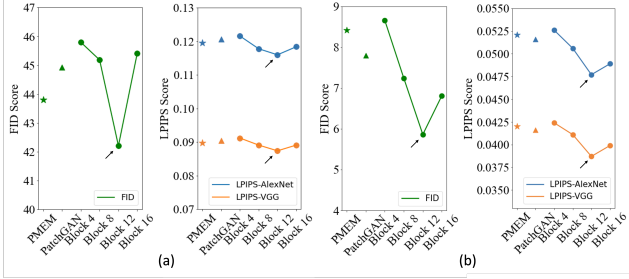


Fig. 3. Comparison of straightening performance (a) of ablation study and (b) on a large-scale dataset (D_{large}).

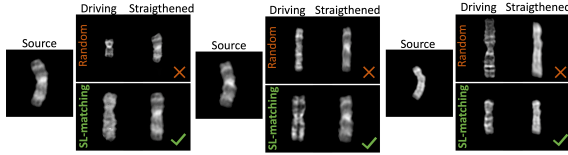


Fig. 4. Comparison of straightening performance based on different *test driving* images selected by random and our proposed SL-matching schemes.

PSNR and SSIM, especially for deformed objects (between original and straightened chromosomes) [22]. $LPIPS_A$ and $LPIPS_V$ represent the scores generated by AlexNet and VGG backbones. Since chromosome classification is an important downstream task of chromosome straightening for subsequent karyotype construction and pathological diagnosis, the DCA can be used to assess the performance.

3.2. Comparison with State-of-the-Art Methods

Table 1 gives the quantitative assessment of chromosome straightening on D_{train} . We compare the ViT-Patch GAN with state-of-the-art studies, SCSF [14], FOMM [18] and PMEM [19]. These results are the average scores using 5-fold cross-validation on D_{train} . The FID value of ViT-Patch GAN decreases from 54.90 to 42.21, implying that the straightening quality of our method is closest to the real-world chromosome images. Moreover, ViT-Patch GAN achieves the best LPIPS scores on both AlexNet and VGG backbones, demonstrating the clear advantages of our proposed method in this task.

We perform downstream classification experiments using 642 chromosomes straightened by these state-of-the-art methods in exactly the same configuration. Each DCA is the average best result generated by 4-fold cross-validation classification experiments with commonly used models, ResNet34 (R34), ResNet50 (R50) and DenseNet169 (D169) [24, 25]. Although there is overfitting with the increasing depth of classification networks, we observe that ViT-Patch GAN significantly outperforms other methods by a large margin on DCA_{R34} , DCA_{R50} and DCA_{D169} . The proposed method achieves gains of 2.19%, 4.53% and 5.47% on DCA compared to the original bent chromosomes (baselines). Such results highlight that our proposed ViT-Patch GAN has an excellent reconstruction quality of straightened chromosomes.

Fig. 2 gives the comparison between two straightened chromosomes. Compared to SCSF and PMEM, ViT-Patch

GAN achieves the most accurate straightening details (green arrows) with fewer reconstruction errors. These errors are generally of two types: (i) bending is not well recovered (blue arrows with boxes), and (ii) the chromosome shape and details are not well preserved (red arrows). This inadequate straightening is mainly caused by inaccurate motion representation. For SCSF, only a backbone responsible for straightening results in a lack of features. The estimated regions of PMEM are wide and mostly located on the background, resulting in the curvature being almost not straightened (blue arrows). In contrast, ViT-Patch GAN estimates more meaningful and accurate motion representation for straightening, resulting in highly reliable straightening results.

3.3. Ablation Study and Cross-Dataset Experiments

We conduct ablation experiments to compare the performance of different architectures. We also implement PatchGAN, which uses a basic patch discriminator with PMEM as the generator. Blocks 4-16 are experiments for our proposed ViT-Patch discriminator with a series number of MHSA blocks. Fig. 3-a shows that the performance on all metrics increases from Blocks 4 to 12, and Blocks 12 of our method achieves the best results (black arrows). In contrast, Blocks 16 may lead to overfitting, resulting in decreased performance.

We further conduct experiments with the exactly same ViT-Patch GAN model (Blocks 12), but only with different *test driving* images picked by random selection and the proposed SL-matching scheme. Fig. 4 demonstrates that the SL-matching scheme is important to drive source images for generating accurate straightening results.

We perform cross-dataset experiments to evaluate the generalization capability of ViT-Patch GAN (Fig. 3-b). The models (those used in Section 3.3) are trained on D_{train} . The performance relationship between Fig. 3-b and Fig. 3-a is consistent. Our ViT-Patch GAN with Blocks 12 outperforms the others on all metrics, suggesting its robustness in the cross-dataset setting. Although our proposed chromosome straightening framework is trained on only 642 chromosomes (D_{train}), the ViT-Patch GAN learns the global motion representation rather than only mapping dependencies of specific patterns (*i.e.* SCSF). The converged model can further straighten chromosomes of all types on the large dataset, D_{large} . Thus, it may be further utilized for large-scale applications of chromosome straightening.

4. CONCLUSIONS

In this paper we propose a novel robust chromosome straightening framework, which includes a generator for learning chromosome motion representation and a ViT-Patch discriminator for generating more realistic straightened results. The ViT-Patch discriminator encodes both the local detail and long-range dependency. Qualitative and quantitative results demonstrate that the efficacy of our proposed method in retaining more chromosome shape and banding pattern details. Our framework also has excellent generalization capability in chromosome straightening for a large size dataset.

5. COMPLIANCE WITH ETHICAL STANDARDS

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee.

6. ACKNOWLEDGMENTS

This work was supported by the Key Program Special Fund in XJTLU (KSF-A-22).

7. REFERENCES

- [1] Anders Bojesen and Claus H Gravholt, "Klinefelter syndrome in clinical practice," *Nature Clinical Practice Urology*, vol. 4, no. 4, pp. 192–204, 2007.
- [2] Donna G Albertson, Colin Collins, Frank McCormick, and Joe W Gray, "Chromosome aberrations in solid tumors," *Nature genetics*, vol. 34, no. 4, pp. 369–376, 2003.
- [3] Wenbo Zhang, Sifan Song, Tianming Bai, Yanxin Zhao, Fei Ma, Jionglong Su, and Limin Yu, "Chromosome classification with convolutional neural network based deep learning," in *2018 11th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)*. IEEE, 2018, pp. 1–5.
- [4] Gleb N Artemov, Vladimir N Stegny, Maria V Sharakhova, and Igor V Sharakhov, "The development of cytogenetic maps for malaria mosquitoes," *Insects*, vol. 9, no. 3, pp. 121, 2018.
- [5] Stegniĭ, "Systemic reorganization of the architecture of polytene chromosomes in onto-and phylogenesis of malaria mosquitoes. structural features regional of chromosomal adhesion to the nuclear membrane," .
- [6] VN Stegnii and VM Kabanova, "Chromosome analysis of anopheles atroparvus and anopheles maculipennis (diptera, culicidae)," *Zoologicheskii zhurnal*, 1978.
- [7] Tanvi Arora, Renu Dhir, and Manish Mahajan, "An algorithm to straighten the bent human chromosomes," in *2017 Fourth International Conference on Image Information Processing (ICIIP)*. IEEE, 2017, pp. 1–6.
- [8] Sahar Jahani and Seyed Kamaledin Setarehdan, "Centromere and length detection in artificially straightened highly curved human chromosomes," *International journal of Biological engineering*, vol. 2, no. 5, pp. 56–61, 2012.
- [9] Devaraj Somasundaram and VR Vijay Kumar, "Straightening of highly curved human chromosome for cytogenetic analysis," *Measurement*, vol. 47, pp. 880–892, 2014.
- [10] Ning Xie, Xu Li, Kang Li, Yang Yang, and Heng Tao Shen, "Statistical karyotype analysis using cnn and geometric optimization," *IEEE Access*, vol. 7, pp. 179445–179453, 2019.
- [11] Mehrgan Javan Roshtkhari and Seyed Kamaledin Setarehdan, "A novel algorithm for straightening highly curved images of human chromosome," *Pattern recognition letters*, vol. 29, no. 9, pp. 1208–1217, 2008.
- [12] Monika Sharma, Oindrila Saha, Anand Sriraman, Ramya Hebbalaguppe, Lovekesh Vig, and Shirish Karande, "Crowdsourcing for chromosome segmentation and deep classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 34–41.
- [13] Jiping Zhang, Wenjing Hu, Shuyuan Li, Yaofeng Wen, Yong Bao, Hefeng Huang, Chenming Xu, and Dahong Qian, "Chromosome classification and straightening based on an interleaved and multi-task network," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 8, pp. 3240–3251, 2021.
- [14] Sifan Song, Daiyun Huang, Yalun Hu, Chunxiao Yang, Jia Meng, Fei Ma, Frans Coenen, Jiaming Zhang, and Jionglong Su, "A novel application of image-to-image translation: Chromosome straightening framework by learning from a single image," in *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2021, pp. 1–9.
- [15] Sunyi Zheng, Jingxiong Li, Zhongyi Shui, Chenglu Zhu, Yunlong Zhang, Pingyi Chen, and Lin Yang, "Chrsnet: Chromosome straightening using self-attention guided networks," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV*. Springer, 2022, pp. 119–128.
- [16] Fakhre Alam and Sami Ur Rahman, "Challenges and solutions in multimodal medical image subregion detection and registration," *Journal of medical imaging and radiation sciences*, vol. 50, no. 1, pp. 24–30, 2019.
- [17] Nitin Arora, Mamta Martolia, and Alaknanda Ashok, "A comparative study of the image registration process on the multimodal medical images," *Asia-pacific Journal of Convergent Research Interchange*, vol. 3, no. 1, pp. 1–17, 2017.
- [18] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe, "First order motion model for image animation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [19] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov, "Motion representations for articulated animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13653–13662.
- [20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
- [21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [22] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.