Learn to Walk Across Ages: Cross-age Gait Analysis with Spatio-temporally Augmented Representation

Submitted to
Graduate School of Information Science and Technology
Osaka University

September 2022

Yiyi ZHANG

ABSTRACT

Video is becoming an important carrier of information in this digital era under the development of communication technology and widespread availability, where surveillance video is one of the most important types. Due to the widely deployed surveillance system, large-scale videos are captured everyday, making the computer-aided surveillance analysis system necessary.

Gait is a very important subject in surveillance video due to its noncontact, noninvasive and can be perceivable at a long distance. Psychological experiments have shown that gait includes not only identity information, but is also discriminative in a variety of human attributes such as age and gender. Among all the human attributes, since age is inevitable in person's life span, it is one of the crucial factors in gait analysis.

In this thesis, we propose the task of cross-age gait video translation, which aims to realize a function of age progression/regression on gait. The translated gait videos attempted to preserve the identity of the original subject while keeping the realism of the generation quality. Cross-age gait video translation can not only improve the fundamental understanding of age features in a subject, but also support a range of applications such as criminal investigation and fugitive research. Moreover, cross-age gait translation can be a promising solution for age-invariant gait recognition.

In Chapter 1, we first review the background and motivation of the proposed cross-age gait video translation task. we then summarize two aspects in realizing this: gait video translator with spatio-temporally augmented network design, and the spatio-temporally augmented high fidelity input.

In Chapter 2, to learn an effective gait video translator across ages, we proposed spatio-temporally augmented multi-age group gait video translation framework, which aims to ensure three aspects: aging effect, individuality preservation, and gait realism. Specifically, we build our framework on a multi-domain image translation model. Because the existing multi-domain image translation model was originally designed for a still image, we extend it to gait video by introduc-

ing a motion-augmented network architecture with three streams, where gait period, period-normalized phase-synchronized gait video, and its frame difference sequence are each input to one stream. We also designed a discriminator with a slow-fast path to learn spatio-temporally augmented gait representation for cross-age gait video translation. Our framework quantitatively and qualitatively outperforms state-of-the-art age progression/regression methods on the largest gait database with age, OULP-Age, with respect to both age group classification and identity recognition.

In Chapter 3, since appearance-based gait analysis approaches usually use silhouette or silhouette-based template as input, silhouette quality plays an important role in gait analysis. To learn a spatio-temporally augmented high fidelity input for the proposed cross-age gait video translation task, we studied natural image matting task and achieved competitive results on widely acknowledged matting benchmarks. To capture global contextual information from a whole image without degrading the image quality, an end-to-end three-branch image matting framework is proposed, which can exploit unknown-relevant global contextual information condensed from the high-resolution image. We then proposed a matting-oriented contextual aggregation can cope with such a situation by making use of all the pixels in the deformed foreground/background where foreground/background pixels are dominant. The proposed method can estimate alpha matte and background simultaneously while keeping the matting equation, which can improve the foreground extraction performance qualitatively.

In Chapter 4, We further provided a discussion on how the high fidelity input may influence the gait video translator across ages. We first designed a scheme to automatically estimate the trimap for the proposed matting method, so that the matting can be adopted without user interaction. Specifically, we adopted a inpainting method to predict the background and finetuned the proposed matting method on OULP-Age. We then provided thorough experiments on the largest gait database with age information, OULP-Age, to reveal how the input quality of silhouette affect the performance of age progression/regression task on age group classification and cross-age gait recognition.

Finally, conclusions are drawn and future work is discussed in Chapter 5.

Key words: spatio-temporal, video, gait, age progression and regression, matting

List of Publications

Publication List Relevant with PhD Thesis

A. Jounral Papers

- Learn to Walk Across Ages: Motion Augmented Multi-age Group Gait Video Translation. Y. Zhang, Y. Makihara, D. Muramatsu, J. Zhang, L. Niu, L. Zhang, Y. Yagi , IEEE Access, pp. 1-9, 2021. DOI: 10.1109/access.2021.30616842.
  (Corresponding to Chapter 2 in the PhD thesis)

- Natural Image Matting with Attended Global Context. Y. Zhang, L. Niu, Y. Makihara, J. Zhang, W. Zhao, Y. Yagi, L. Zhang, Journal of Computer Science and Technology. Accepted
  (Corresponding to Chapter 3 in the PhD thesis)

Publication List Irrelevant with PhD Thesis

A. Jounral Papers

- Hallucinating Uncertain Motion and Future for Static Image Action Recognition. L. Niu, S. Huang, X. Zhao, L. Kang, Y. Zhang, L. Zhang, Computer Vision and Image Understanding. Volume 215, pp. 103337, 2022. DOI: 10.1016/j.cviu.2021.103337

B. International Conference/Workshop (Full Paper Reviewed)

- Warp Gait Across Ages: Cross-age Gait Video Translation with Part-aware Flow Warping. Y. Zhang, L. Niu, Y. Makihara, H. Yan, Y. Yagi, L. Zhang, in submission

- Exploiting Motion Information from Unlabeled Videos for Static Image Action Recognition. Y. Zhang, L. Niu, Z. Pan, M. Luo, J. Zhang, D.

Cheng, L. Zhang. Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34, 07 (AAAI 2020), pp. 12918-12925, DOI: 10.1609/aaai.v34i07.6990. New York, NY, USA, Feb. 2020

- Spatio-Temporal Attention-Based Neural Network for Credit Card Fraud Detection. D. Cheng, S. Xiang, C. Shang, Y. Zhang, F. Yang, L. Zhang. Proceedings of the AAAI Conference on Artificial Intelligence. Volume 34, 01 (AAAI 2020). pp. 362-369, DOI: 10.1609/aaai.v34i01.5371. New York, NY, USA, Feb. 2020

- Contagious Chain Risk Rating for Networked-guarantee Loans. D. Cheng, Z. Niu, Y. Zhang, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2020). pp. 2715-2723. DOI: 10.1145/3394486.3403322. Virtual Event / San Diego, CA, USA, Aug. 2020

- A Dynamic Default Prediction Framework for Networked-guarantee Loans. D. Cheng, Y. Zhang, F. Yang, Y. Tu, Z. Niu, L. Zhang. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM 2019). pp. 2547-2555. DOI: 10.1145/3357384.3357804. Beijing, China, Nov. 2019

- Exploiting Motion Information from Unlabeled Videos for Static Image Action Recognition. Y. Zhang, L. Niu, Z. Pan, M. Luo, J. Zhang, D. Cheng, L. Zhang. Symposium on Brain-inspired Computing and Intelligence, Shanghai Jiao Tong University. Poster. Shanghai, China. Nov. 2019.

- Evaluate the Government's Action on Air Pollution. Y. Zhang. Shanghai Jiao Tong University "Qian Xue-Sen Cup" Student Science and Technology Innovation Fair, Poster. Shanghai, China. Apr. 2017.

C. Awards

- Finalist (top 8%). World Artificial Intelligence Conference (WAIC 2018) - MedAI Challenge. AION: Intelligent Pathological Diagnosis System. Aug.

2018.

- First Prize. Shanghai Jiao Tong University Qian “Xue-Sen Cup” Student Science and Technology Innovation Fair. Apr. 2017.

Acknowledgements

This thesis has been completed based on Agreement on the Double-Degree Program for PhD between the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University and the Graduate School of Information Science and Technology, Osaka University. I deeply appreciate the program giving me a chance to get PhD degree.

Beside, this thesis would not have been possible without the support of countless people.

First and foremost, many thanks to the referee of this thesis: Prof. Yasushi Yagi, Prof. Yasushi Makihara, Prof. Hajime Nagahara and Prof. Tomoya Nakamura for their invaluable feedback.

I am deeply grateful to my advisor Prof. Yasushi Yagi for taking me as the double degree student in Osaka University. Although my one-year stay in Japan is short, the research environment of the lab and his attitude towards research really impressed me. I would like to express my sincere gratitude to Prof. Yasushi Makihara for his guidance and support throughout my PhD study. The weekly meeting schedule is really helpful in keeping my pace and monitor my research progress. Prof. Tatsuhiro Tsuchiya is the contact person of my double degree program, I could not have undertaken this journey without his effort.

I am indebted to my advisor Prof. Liqing Zhang in Shanghai Jiao Tong University. I joined the BCMI lab back in my third year bachelor. In the seven years, he encouraged me to step out of my comfort zone and challenge myself. I also benefited greatly from Associate Prof. Li Niu for his wealth of knowledge and meticulous comment. His ingenuity of thought and hardworking are something I will always keep aspiring to.

I'm grateful to Prof. Daigo Muramatsu who provided helpful suggestions on my project. As my neighbor in the lab, his humor also created a relaxed atmosphere to help me adjust to the new environment. Special thanks to Project Associate Prof. Kota Aoki, Project Associate Prof. Md Atiqur Rahman Ahad for taking care of me during my stay, Project Assistant Prof. Xiang Li, Dr. All Shehata Hassanein

Allam for helpful discussions.

Assistant Prof. Shuqiong Wu and Project Assistant Prof. Chi Xu are two excellent women researchers I met in my PhD journey. Although female researchers are the minority in the computer science community, thank you for setting up the role model and inspiring me to pursue my academic career after graduation.

I also appreciate the help of secretaries in Yagilab Ms. Nobue Yuasa, Ms. Megumi Kaneshiro to me with lots of administrative procedures. Your smile is like sunshine that enriched my life in Japan. Also, I would thank Ms. Jun Ma, Ms. Yan Chen, Ms. Chao Wei in SEIEE Graduate Education Office, Shanghai Jiao Tong University, for their support on all the procedures in the SJTU side.

I am fortunate to have met many friends during my PhD study to work hard and play hard together. Dr. Ruochen Liao, Dr. Chengju Zhou, Mr. Yang Yu, Mr. Akos Godo, Ms. Margaret Dy Manalo, Mr. Alsherfawi Aljazaerly Mohamad Ammar Ayman in Yagi lab. Dr. Haohua Zhao, Dr. Dawei Cheng, Dr. Jianfu Zhang, Dr. Yi Tu, Dr. Zhangxuan Gu, Mr. Ziqi Pan, Mr. Junjie Chen, Mr. Jiangtong Li, Mr. Wentao Wang, Mr. Jiahao Chang, Mr. Weijie Zhao, Mr. Yuxuan Duan, Mr. Songyu Ke, Mr. Zhaohe Liao in BCMI.

Lastly, my parents Jun Zhang and Ying Wang deserve endless gratitude. Thank you for providing a lovely family environment and supporting me in all aspects to pursue my dream.

Table of Contents

Chapter I
Introduction

1.1 Background and Motivation

In this digital era, video has become an important source under the development of communication technology and widespread availability of mobile devices. One of the main factor that makes a video different from image is the temporal information [1]. Analyzing spatio-temporal representation of video offer the promise of understanding low-level motion [2, 3], multi-frame dependencies [4] and temporal structure [5] that go beyond what can be discerned from a single image and achieved remarkable success in video-oriented tasks such as action recognition [1, 6], video question answering [7], etc.

Surveillance video is an important type of video. In the past twenty years, more surveillance cameras have been deployed to keep up with the growing demand to monitor public and private areas for security reasons [8]. Nowadays, the surveillance system is becoming an essential component of the smart city to be deployed in a wide range of scenarios, such as shopping malls for customer behavior analysis, highways for monitoring traffic load, and streets for crime reduction. However, the analysis of large-scale captured videos manually can be tedious work. Hence, an automatic surveillance system that can analyze the captured videos is necessary.

Computer vision is the core technique that aims to equip the automated surveillance system with human-like vision system that can understand images and videos as humans do. Thanks to the rapidly growing computing power, publicly

available large-scale dataset (i.e., ImageNet [9], MSCOCO [10]) and the effective architecture such as convolutional neural network, automated surveillance system nowadays succeed in a wide range of tasks, such as object detection [11,12], object tracking [13, 14], anomaly detection [15], etc.

Among the above mentioned tasks, computer vision target on human is one the most promising topics which can be beneficial to individuals and the society as a whole. For example, pedestrian detection is crucial for autonomous vehicles to eliminate crashes. Face detection is one of the fundamental techniques used in digital cameras for auto-focus. Temperature measurement has been widely adopted for automatic access control during COVID-19 situation so as to reduce the infection risk.

Person identification and human attribute estimation are two fundamental technique in computer-aided human analysis. Biometric sources such as DNA, fingerprint, face and gait are linked to distinct individuals. They are believed to contain a considerable amount of personal information for person identification, which has been widely addressed in the literature [16–19]. Human attribute (e.g., age, gender, and ethnicity) can also be estimated from biometric data [20–24], which is believed to improve the person identification accuracy [25, 26].

Unlike other biometric sources such as DNA, fingerprint and face, gait is a unique biometric feature that is noncontact, noninvasive, and can be perceivable even at a long distance from a camera. With the increasing demand of visual surveillance and the rising concern of data collection privacy, gait recognition with no concern of data privacy is extremely useful in the contexts of surveillance systems. Gait-based features can be divided into appearance-based and model-based, where appearance-based is dominant. Meanwhile, gait silhouette (i.e., a temporal sequence of gait) is a spatio-temporal source, and hence a variety of spatio-temporal gait representations have been proposed in gait analysis community [27–36].

In addition, psychological experiments showed that gait is also discriminative in a variety of attributes such as age and gender [37–39] through the composite of appearance and motion. Among them, since age change is inevitable during a

person's life span, it is one of the crucial factors in gait analysis.

Gait-based age estimation can not only improve the fundamental understanding of age features in a subject, but can also support a range of applications such as age-based person retrieval and people counting in a wide area (e.g., a shopping mall), fugitive research, time-lapsed identity verification in criminal investigation, and transfer onto digital avatars to help generalize cross-age virtual characters in virtual reality. However, gait analysis is sensitive to appearance changes due to covariants, where age is one of them. The aging process of the subject is likely to affect the gait pattern, which makes gait-based person identification difficult. Thereafter, gait video translation across ages (Fig. 1.1-1), i.e., age progression/regression on gait while preserving the identity and realism of the generated videos is a promising solution for age-invariant gait recognition.

Although the property (i.e., age group classification and cross-age recognition performance) of the generated cross-age gait silhouette sequences is the main concern in the age-invariant gait recognition, the generation quality will also be emphasized under some scenarios. For example, in healthcare field application to remind a person to pay more effort to keep youthfulness in gait by watching his/her age-progressed gait video with function, the poor generation quality of the generated gait video might not be convincing to persuade the user.

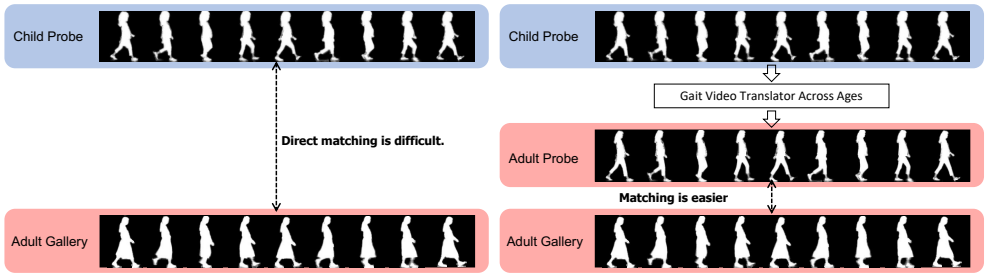


Fig. 1.1-1. (a). Traditional way of age-invariant gait recognition. (b). Overview of the proposed cross-age gait video translation for age-invarant gait recognition.

To realize the cross-age gait video translation, we investigate two necessary aspects: (1) an effective gait video translator across ages, which is studied in Chapter 2. (2) since appearance-based gait analysis approaches usually use silhouette

or silhouette-based template as input, hence silhouette quality plays an important role in gait analysis. We therefore studied a high fidelity input which enables the translator to learn spatio-temporal gait representation in Chapter 3

1.2 Cross-age Gait Video Analysis

While the face image translate across ages can exploit more spatial information such as wrinkles, muscles, hence it has been widely addressed in the literature [40–51], gait video does not contain such color and texture information. Cross-age gait video analysis need to focus on both the spatial and temporal aspects of the aging pattern, where geometric transformation of the body shape is the main concern.

To model the spatial aspect of gait, a widely acknowledged strategy is to convert the gait silhouette into a single template [52–57]. Static template-based gait representation has been proved feasible to preserve the discriminative age pattern with appearance [58]. For example, we can conclude from static gait template children has larger head-to-body ratio, senior has smaller stride. Static template-based gait recognition methods have the advantage of low requirement of the captured quality, robustness toward noise, and attractive small storage and calculation cost, making it suitable for real-world surveillance systems.

Some commonly used template-based gait representation include: Gait Energy Image (GEI) [54], which is calculated by averaging each frame of a normalized gait cycle. The higher intensity of the pixel in GEI stands for the higher frequency human walking occurs in this position. Gait Entropy Image (GEnI) [55] describes the uncertainty of each pixel in the gait silhouette with Shannon entropy. Gait flow image (GFI) [53] refers to the average $n-1$ binary flow images in one gait cycle with n frames. Chrono-gait image (CGI) [56] tend to compress the gait silhouette sequences into one multichannel image, which is expected to ensure a higher variance than intensity in grayscale image so as to preserve temporal information.

Although template-based gait representation encode information as abundant

as possible, temporal information are likely to be omitted in the compression process.

To address this issue, researchers model the temporal aspect of gait directly from gait silhouette sequences with 3D convolution [3] and LSTM network [59]. For example, Feng *et al.* [60] proposed a CNN-based pose estimation method to transform the gait feature from one view to another. Gait sequence is modeled with LSTM recurrent neural network, where each frame is transformed to a joint heapmap as input. Liao *et al.* [61] proposed a framework to extract spatio-temporal features from pose instead of image, which can effectively improve the gait recognition performance. Battistone *et al.* [62] introduced a model to learn long short-term dependencies of structured data and temporal information with graph structure. Zhang *et al.* [63] introduced a gait feature with disentangled pose and appearance information obtained in RGB sequence. The disentangled pose and appearance information in each frame is integrated with a LSTM-based framework to model the critical temporal change.

Modeling the temporal information in gait is feasible to alleviate the influence of covariates such as appearance change, clothing, and carrier. However, when the gait silhouette sequence contains discontinuous frames (i.e., occlusion in the captured video) or a video with a different frame rate, the performance of gait recognition models with temporal information will largely degrade.

Recently, gait representation learned from unordered sets does not rely on the specific order of the gait silhouette sequence has aroused increasing interest. Compared to phase-synchronised gait silhouette sequence, methods that leverage gait representation from an unordered set have a simpler preprocessing step. Therefore, being efficient in real-time prediction.

Set-based approaches have been successfully in fields such as point cloud [64–66], content recommendation [67], and image caption [68]. To the best of our knowledge, GaitSet [69] introduced by Chao *et al.* is the only work to exploit set-based representation with gait silhouette sequence. Gait is a periodic motion with a unique pose in each gait cycle phase. Given an unordered set of frames, they can be easily rearranged to a correct order by observation, making

the given order unnecessary. GaitSet therefore leveraged all silhouettes in one or more sequences of a person as a set of unordered frames, and introduced a global-local fused deep network with set pooling operation to fuse features in different scales.

The aforementioned approaches can be summarized according to the spatial and temporal aspect in Fig. 1.2-2. GEINet [70] takes in the input of GEI, where both the temporal and spatial aspect of the representation learned is low due to the compression of input. Optical flow map [71] make use of the optical flow therefore is the highest in temporal aspect. However, spatial information such as body shape is not considered in this architecture. GaitPart [72] takes in a silhouette sequence and exploits the micro-motion in different parts. GaitSet [69] learn from unordered sequence with a novel set pooling. Therefore, it obtains less temporal but better spatial gait representation. Our model make use of gait period, motion augmented block and slow-fast discriminator to exploit both spatial and temporal aspect of gait representation.

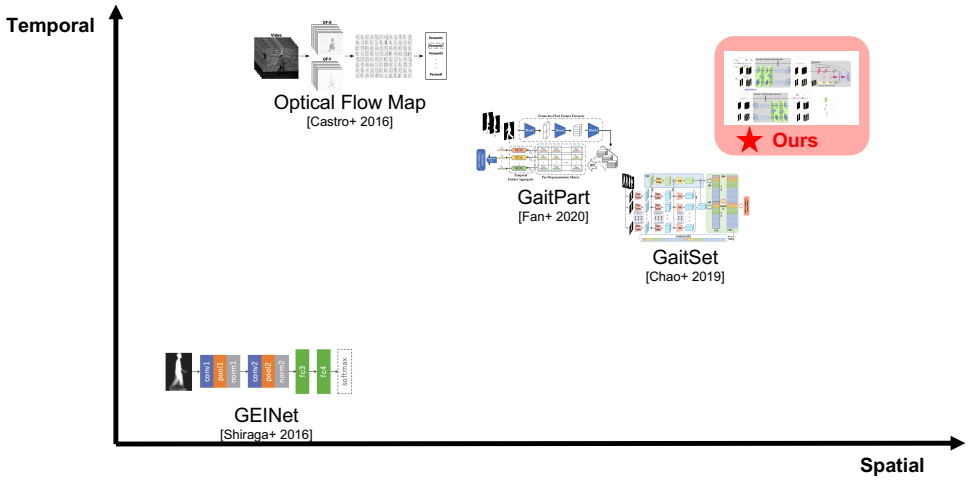


Fig. 1.2-2. Summary of Existing Gait Representation in Network Architecture. Our proposed gait representation preserves the most information in both spatial and temporal aspect.

In the first study, we proposed the task of cross-age gait video translation for the first time, where the realistic generation quality and identity preservation

are of equal importance. To realize this, a novel spatio-temporal augmented gait representation is necessary. Specifically, we introduced a framework with motion augmented block in the generator to obtain temporal information from gait video. Meanwhile, a novel discriminator with a slow path and a fast path is proposed to capture spatial-related information and temporal-related information respectively.

1.3 Inputs for Gait Video Analysis

Silhouette extraction from the captured gait video is a necessary preprocessing step in appearance-based gait analysis framework. We summerize the existing gait representation in the input stage in Fig. 1.3-3. In existing works, gait silhouettes are extracted by segmentation supported by straightforward background subtraction technique [73]. Although recently many excellent segmentation networks have been proposed [74–76], we argue that binary segmentation masks are not suitable for silhouette extraction in video-based gait recognition methods due to limitations in spatial and temporal resolution.

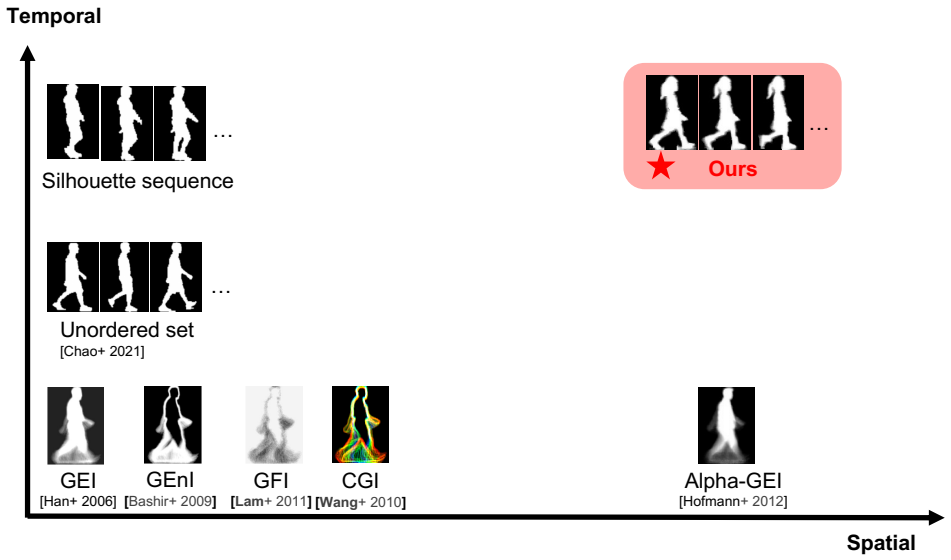


Fig. 1.3-3. Summary of Existing Gait Representation in Input. Our method learn directly from alpha matte gait silhouette sequence, which preserves the most information in both spatial and temporal information in input.

(a) (b) (c) (d) (e) (f)

Fig. 1.3-4. Comparison between segmentation and matting silhouette extraction results. (a),(d) Original image. (b),(e) Silhouette by segmentation. (c),(f) Silhouette by matting. We also zoom in some details (red box) for comparison.

For limitation in spatial aspect, captured gait silhouette is in low resolution, hence each pixel in the captured silhouette represents a group of pixels in a higher resolution with an average value. The detailed information in the higher resolution get washed out since the binary segmentation mask classifies each pixel into foreground or background. To address this problem, alpha matte indicates the opacity of the foreground, which tends to preserve the informative identity characteristics in the gait silhouette of higher resolution. For example, details of shoes and hairstyle can be clearly captured by matting in Fig. 1.3-4.

For limitation in temporal aspect, captured gait silhouette is usually processed into a different frame rate for analysis (i.e., period normalize [58] to obtain the same number of frames for each gait silhouette sequence). Thus, the pixel value describe the motion pattern of several frames of this specific area, not likely to be accurately represented by a binary segmentation mask due to the presence of motion (i.e., binary segmentation mask will probably lead to discontinuity such as leg area in Fig. 1.3-4). Alternatively, the continuous value in alpha matte tend to provide a better representation of the motion pattern.

In the second study, we therefore investigate the spatio-temporally augmented fidelity input for cross-age gait video translation. To obtain such gait representation, we train with a novel network architecture. To make the most use of foreground/background information in an alpha matte estimation network under the limitation of memory usage, we introduced a network that can learn fine-

grained contextual information from the whole input and estimate alpha matte and background simultaneously while keeping the definition of matting equation. In the discussion chapter, we provide a scheme that trimap can be automatically estimated in the matting method introduced in the second study. We further provide thorough experiments on the largest gait database with age to reveal how accuracy of silhouette affect the performance of age progression/regression on age group classification and cross-age gait recognition.

1.4 Organization of the Thesis

The thesis is organized as follows: Chapter 1 is the overall introduction of age aspect in gait analysis. Chapter 2 to 4 are studies conducted during doctoral program.

In Chapter 2, we proposed multi-age group gait video translation for the first time, which aims to preserve the identity while realize the function of age progression/regression on gait. Specifically, we build our framework on a multi-domain image translation model. Because the existing multi-domain image translation model was originally designed for a still image, we extend it to gait video by introducing a motion-augmented network architecture with three streams, where gait period, period-normalized phase-synchronized gait video, and its frame difference sequence are each input to one stream. We then train the network to ensure three aspects: aging effect (using an age group classification loss), individuality preservation (using a reconstruction loss), and gait realism (using an adversarial loss). Our framework quantitatively and qualitatively outperforms state-of-the-art age progression/regression methods on the largest gait database, OULP-Age, with respect to both age group classification and identity recognition.

In Chapter 3, we studied natural image matting task and achieved competitive results on widely acknowledged matting benchmarks. To capture global contextual information from a whole image without degrading the image quality, an end-to-end three-branch image matting framework is proposed, which can exploit unknown-relevant global contextual information condensed from the high-

resolution image. We then proposed a matting-oriented contextual aggregation can cope with such a situation by making use of all the pixels in the deformed foreground/background where foreground/background pixels are dominant. The proposed method can estimate alpha matte and background simultaneously while keeping the matting equation, which can improve the foreground extraction performance qualitatively.

In Chapter 4, We further provided a discussion on how the high fidelity input may influence the gait video translator across ages. We first designed a scheme to automatically estimate the trimap for the proposed matting method, so that the matting can be adopted without user interaction. Specifically, we adopted a inpainting method to predict the background and finetune the proposed matting method on OULP-Age. We then provided thorough experiments on the largest gait database with age information, OULP-Age, to reveal how the input quality of silhouette affect the performance of age progression/regression task on age group classification and cross-age gait recognition.

Finally, conclusions are drawn and future work is discussed in Chapter 5.

Chapter II
Multi-age Group Gait Video Translation

2.1 Introduction

Gait refers to a person's walking style and is considered as one of behavioral biometrics, which contains a variety of attributes of the person such as age, gender, health status, and identity. Among the attributes, since age changes in gait is inevitable during a person's life span, it is considered as one of covariates for gait-based person identification a.k.a. gait recognition. Viewed from another perspective, the gait provides a cue to estimate an age or age group and hence gait-based age estimation also enjoy a rich body of literature [39, 77–81].

Taking the above-mentioned fact into consideration, if we realize a function of age progression/regression on gait, i.e., translation of the gait from one age to another by preserving the identity (see Fig. 2.1-1), we may be able to employ it for many applications. For example, assuming that a criminal investigator tries to find a perpetrator by gait recognition after a long time (e.g., 10 years), he/she can mitigate intra-subject variations between a matching pair of an enrollment and a query by progressing the age of the enrolled gait by 10 years with the function. As another application in the healthcare field, a person may pay more efforts to keep youthfulness in gait by watching his/her age-progressed gait video with the function.

Although age progression/regression has been extensively studied in the face analysis community [47, 48], the study [58] is the only one on gait age progression/regression, to the best of our knowledge. In their work, the authors progress a

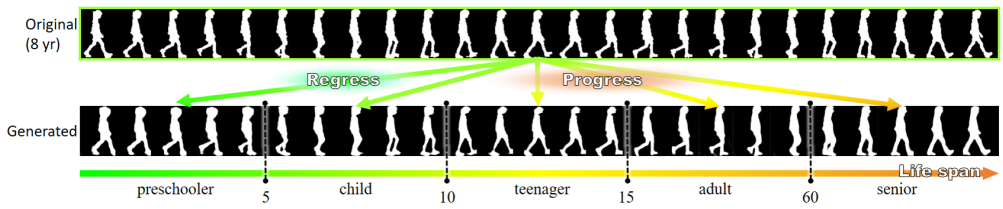


Fig. 2.1-1. Given one's current gait video, we attempt to translate into his/her gait video in different age groups of life.

gait energy image (GEI) [82] (also called an averaged silhouette [83]), which is the most widely used gait template in the gait recognition community, using a subject-independent geometric warping field. Taking into account a variety of potential applications, it is naturally preferable to translate not a static gait template such as a GEI but a gait video (e.g., recent gait recognition no longer relies on a static gait template but employs a gait video as input instead). In addition, since one process of gait age progression/regression may differ from the process in another individual, it is preferable to translate a gait video in a subject-dependent way, unlike [58], which adopts a subject-independent method of translation. Moreover, most gait analysis studies overlook an important component in gait, i.e., the gait period (or gait cycle), or they regard it just as a normalization factor. In fact, static gait templates such as GEI [82], frequency-domain feature [57], chrono-gait image [56], are all period-normalized templates (e.g., in case of GEI, gait silhouettes are aggregated over a gait period and then divided by the gait period to obtain an average), and hence the gait period is "washed out"in those gait templates. In addition, the gait period has not yet been explicitly exploited, even in recent deep learning-based gait analysis [69, 72, 84]. This is partly because the gait period is unstable due to intra-subject variations (e.g., changes in walking speed) in the context of gait-based person identification. The gait period is, however, naturally quite important information in the context of age. For example, a child's gait period is usually shorter than that of an adult or elderly person, in other words, children's gaits have higher cadence (or frequency). The gait period is therefore considered to be one of key components for representing the characteristics of each age group.

In this work, we therefore propose a method of gait video translation among

multiple age groups. For better representation of the age groups' characteristics, we convert a gait video into the gait period and a period-normalized (or rate-normalized) gait video with a fixed number of frames, and then translate both of them in an original age group to those in a target age group. As such, we can reflect aging effect not only in gait sequence of one gait period but also in the gait period (see the supplementary videos to recognize the difference in gait periods among age groups). In addition, we make the gait video translation subject-dependent by introducing StarGAN, a kind of conditional GAN framework, which was originally designed for generic static image translation. The main contributions of this work include the followings.

1. Gait video translation among multiple age groups

We propose a method of gait video translation from one age group to another for the first time, unlike the existing work [58] translates just a static gait template. This is beneficial for potential applications such as cross-age gait recognition using not a static gait template but a gait video as an input.

2. Gait-oriented motion-augmented video translation

In order to better translate all contents in a cyclic gait video, we design a video translation framework with three streams: a period-normalized gait silhouette sequence, a period-normalized frame difference sequence, and a gait period. While the first one mainly encodes body shape aspects in a gait video, the latter two encode motion aspects in it. This technically differentiates the proposed method from other work on still image-based face age regression/progression and general image/video translation.

3. State-of-the-art performance

Our method outperforms modified versions of state-of-the-art of facial age progression/regression qualitatively and quantitatively in both person identification and age group classification tasks with the OULP-Age, the largest publicly available gait database with age information.

2.2 Related Work

2.2.1 Face Age Progression/regression

Extensive studies on face aging can be mainly divided into three groups: physical model-based approaches, prototype-based approaches, and deep learning-based approaches [40].

The physical model-based approaches correlate biological and physical mechanism (i.e., craniofacial growth, skin, and winkles) with human age using models such as an and-or graph, a concatenational graph evolution aging, a craniofacial growth. Although those models are carefully designed, they rely heavily on the imperfect human knowledge [48] and require sufficient training samples with aging sequence over a long age span for each individual, which are almost impossible to be collected for gait.

In the prototype-based approaches, averaged faces are created as the prototype for each group, the difference between each prototype is regarded as the transition pattern. To offset the missing personal characteristic in the transition pattern due to the averaged faces, Shu *et al.* proposed novel aging dictionary learning methods to better preserve personality [44, 45]. However, it still relies on dataset (i.e. CACD [85]) with short period paired data, which is unlikely to be acquired for gait. Meanwhile, to obtain progressed and regressed facial image, the prototype-based approaches need to be trained twice [46].

Recently, the deep learning-based approaches to facial age progression/regression have achieved the state-of-the-art performance in age group classification and identity preservation with no paired data. Particularly, Zhang *et al.* [46] proposed a conditional adversarial autoencoder where each facial image corresponds to a point on the manifold. Translated aging facial images are obtained through stepping along the aging axis on the manifold. Other GAN-based approaches [47–49] adopt a pretrained classifier and a conditional GAN architecture to generate faces conditioned on age during progression and regression to preserve the identity. Although deep learning-based approaches to face aging can generate

good simulation results, a gait is not a static image but a video (image sequence), and hence we consider to directly handle the gait video as an input/output and to better handle motion and appearance aspect.

2.2.2 Gait-based Age Analysis

Prior studies in gait-based age group classification have demonstrated the fact that gait contains discriminative aging patterns in a long-elapsed time period. For example, Mannami have adopted a frequency domain feature [86] to classify four age groups: children (under 15 years old), adult males, adult females, and the elderly (over 65 years old). Chuen *et al.* have investigated correlation of image-based gait features (i.e. stride length, body length, head-to-body ratio) to distinguish between children and adults [87]. These internally contained discriminative aging pattern in gait video make the task of multi-age group gait translation practical. Actually, Xu *et al.* have conducted the first gait age translation among multi-age groups [58]. However, similar to most of the current analysis in gait that utilize image-based features such as frequency domain feature [57], chrono-gait image [56], gait flow image [53], Gabor GEIs [52], their work rely on the GEI, which also falls into the category of image-based gait features.

Although the static image-based gait feature including the GEI had been considered simple yet effective representations for gait analysis, they have several problems such as highly compressed motion information and entanglement of appearance (body shape) and motion information. They have been therefore gradually replaced with spatio-temporal or more disentangled representations. For example, Chao *et al.* have introduced GaitSet [69], a framework that make use of gait sequences to achieve the state-of-the-art performance in gait recognition, outperforming methods that utilize image-based features by a large margin. In this paper, we also leverage gait videos to obtain gait aging pattern from both appearance and motion.

2.2.3 Video-to-video Translation

Video-to-video translation addressed in [88–91] has aroused attention from researchers recently. Given a video to drive a motion, these methods transfer the driving motion to an input static image with a different content, and generate a video with the different content and the same driving motion. Unlike the previous work tries transferring the same motion, our task aim to learn the aging pattern automatically across ages without the driving motion.

As an extension of image-to-image translation, prior works in video-to-video translation [88,89] require paired data. Bansal *et al.* proposed RecycleGAN [90] to first facilitate unpaired video-to-video translation. Meanwhile, these approaches have limitations: they rely on the variants of CycleGAN [92], and hence they are only capable of learning the relations between two different domains at a time, which lack scalability in handling multiple domains.

2.3 The Proposed Method

2.3.1 Overview

Our goal is to translate a gait video among multiple age groups. In order to efficiently handle multi-age group translation, we build our method upon StarGAN [93] because it can translate images among multiple domains just by a single unified model (i.e., with less network parameters compared with pairwise translation models). The StarGAN is a kind of a conditional GAN, and hence it has a generator and a discriminator as shown in the overview in Fig. 2.3-2. In addition, for better treatment of all the components appeared in the gait video, we take a triplet of period-normalized gait silhouette sequence, frame difference sequence, and a gait period as an input for the generator/discriminator. We will describe preprocessing, the generator/discriminator, and loss functions, in the following subsections.

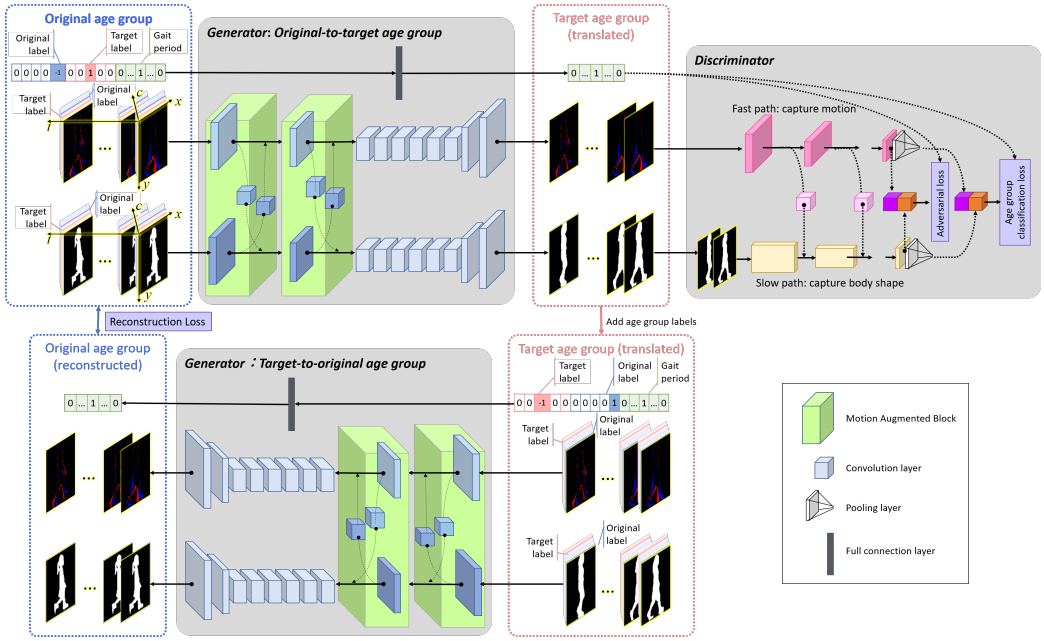


Fig. 2.3-2. Overview of the proposed framework, consisting of a three-stream generator/discriminator. We assign positive and negative values in frame difference to red and blue just for visualization purpose. Given an original triplet of period-normalized gait silhouette sequence, frame difference sequence, and a gait period, as well as an original and target age groups' labels, the generator translates the triplet to the target age group. The translated triplet of the target age group is further translated back to that of the original age group to preserve individuality by a reconstruction loss, while the triplet is also fed into the discriminator to ensure reality and characteristics of the target age group by adversarial and age group classification losses, respectively.

2.3.2 Preprocessing

We briefly describe preprocessing to prepare input data, i.e., a triplet of period-normalized gait silhouette sequence, frame difference sequence, and a gait period, for our generator/discriminator.

Given a gait video, gait silhouettes were first extracted by graph-cut segmentation supported by background subtraction [73], since color and texture information are relevant with neither gait nor body shape. We then obtained size-normalized and registered silhouette sequences in size 88 by 128 pixels. One gait period (or cycle) was detected by maximizing auto-correlation along the temporal axis [57]. We morphed the sequence to produce a period-normalized phase-synchronized gait silhouette sequence per subject with N_{img} frames [58], where N_{img} is experimentally to 25. Examples of the resultant sequence is shown in Fig. 2.3-3. We also extract frame difference sequence, which is also used in a motion-oriented gait recognition method [94].

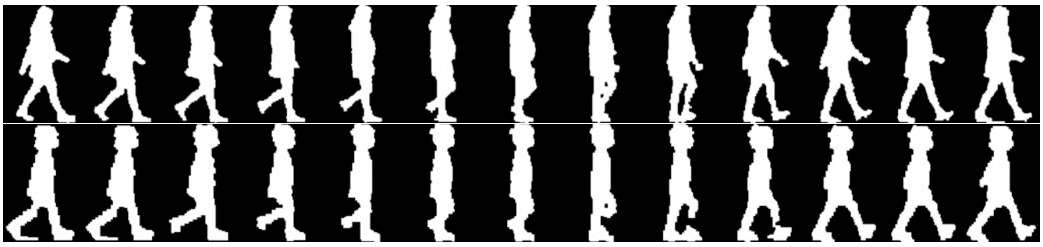


Fig. 2.3-3. Period-normalized phase-synchronized silhouette sequences (top: an adult, bottom: a child). Thirteen frames from a half gait period are shown due to space limitation.

The image size of each frame in the gait silhouette sequence and the frame difference sequence was finally converted to 128 by 128 pixels by adding zero-padded columns at left and right sides, and then frames in each sequence are stacked over channel dimensions, for the convenience of treatment in a network architecture.

As for the gait period [frame], we represent it an indicator vector rather than an integer scalar for more flexible non-linear translation. We therefore first experimentally set minimum and maximum gait periods as $P_{\text{min}} = 20$ and $P_{\text{max}} = 41$, respectively, and then set $N_{\text{period}} (= P_{\text{max}} - P_{\text{min}} + 1)$-dimensional one-hot vector

$\vec{x}_{\text{period}} \in \mathbb{R}^{N_{\text{period}}}$ whose i-th component represent the gait period $(P_{\min} + i - 1)$. Specifically, given an input gait period P, the $(P - P_{\min} + 1)$-th component of the vector $\vec{x}_{\text{period}}$ is 1 and the others are zero-padded.

2.3.3 Generator with Motion Augmented Block

We feed the generator the triplet described in Section 2.3.1 to augment the generator with more motion information.

At the beginning, since we build our framework upon StarGAN [93], i.e., image translation framework among multiple domains, we also prepare indicators for multiple domains, (i.e., age groups).

As for the gait period, we prepare N_{age}-dimensional one-hot vector, where N_{age} is the number of age groups. In this vector, the element for the target age group is set to one and the others are set to zero. In addition, we prepare an indicator vector for an original (or source) age group with the same dimension. We then concatenate the two indicator vectors to the gait period indicator vector $\vec{x}_{\text{period}}$ in channel dimension, which results in $(N_{\text{period}} + 2N_{\text{age}})$ channels in total (see Fig. 2.3-2).

As for the gait silhouette sequence, we prepare an indicator set of N_{age} matrices with 128×128 pixels, where a matrix for a target age is set to all ones and the other matrices are all zero-padded. We also prepare the indicator set of matrices for an original age group. We then concatenate both of them for each frame of the gait silhouette sequence in channel dimension, which results in $(2N_{\text{age}} + 1)$ channels in total. The input structure of the silhouette sequence is consequently a 4th-order tensor of $128 \times 128 \times (2N_{\text{age}} + 1) \times N_{\text{img}}$ (see Fig. 2.3-2). Similarly, we concatenate the indicator set of matrices for each frame of the frame difference sequence with $(N_{\text{img}} - 1)$ frames, which results in a $128 \times 128 \times (2N_{\text{age}} + 1) \times (N_{\text{img}} - 1)$ structure.

Once the generator takes an input triplet of period-normalized gait silhouette sequence, frame difference sequence, and a gait period at the original age group, it translates them to a triplet at the specified target age group. As for the gait period, we simply employ 3 full connection layers to output the N_{period}-dimensional gait

period indicator vector.

As for the period-normalized silhouette sequence and frame difference sequence, we extend network structure used in CycleGAN [92] by using 3D convolution to handle not a still image but a video (i.e., sequence).

Moreover, we introduce a motion augmented block (MAB) so that the model can share and exchange information about the body shape-dominant silhouette sequence and motion-dominant frame difference sequence in the intermediate layers. This is because the body shape and motion have some correlation, for example, a certain body shape (e.g., a fat body shape) may limit the gait motion (e.g., range of joint motion).

More specifically, let the feature maps of the gait sequence and the frame difference sequence at the i-th stage be C_i and M_i, respectively. We first apply three-dimensional convolution filters f^c_{conv} (*resp.*, f^m_{conv}) for C_i (*resp.*, M_i) to obtain intermediate feature maps C'_i (*resp.*, intermediate difference map M'_i) as

$$C'_i = f^c_{\mathrm{conv}}(C_i) \tag{2.3.1}$$

$$M'_i = f^m_{\mathrm{conv}}(M_i). \tag{2.3.2}$$

We further apply three-dimensional convolution filters $f^{c\to m}_{\mathrm{conv}}$ (*resp.*, $f^{m\to c}_{\mathrm{conv}}$) to the intermediate feature maps C'_i (*resp.*, intermediate difference map M'_i), and obtain the feature maps at the second stage by exchanging the appearance and motion information as

$$C_{i+1} = C'_i + f^{m\to c}_{\mathrm{conv}}(M'_i) \tag{2.3.3}$$

$$M_{i+1} = M'_i + f^{c\to m}_{\mathrm{conv}}(f_{\mathrm{diff}}(C'_i)). \tag{2.3.4}$$

Note that the three-dimensional convolution filters have the same structure (i.e., kernel size 1, stride 1, dilation 1) yet have different weights. After passing two MABs, we apply some more convolution and deconvolution layers to generate a silhouette sequence and a frame difference sequence of the target age group.

2.3.4 Discriminator with SlowFast Path

We adopted a discriminator with a SlowFast path inspired by [95] as a primitive analogy to human visual system to better discriminate the generated progressed/regressed sequences. While sparsely sampled translated silhouette sequence (i.e., every five frames) are fed into the slow path to capture more body shape-relevant features with high channel capacity, the translated frame difference sequence with full frames are fed into the fast path to capture more motion-relevant features with low channel capacity.

Moreover, we fuse features from the fast path to the slow path by using lateral connection, which has demonstrated its effectiveness in optical flow-based two-stream network [96, 97]. Finally, output from the both paths as well as the translated gait period indicator vector are fed into an adversarial loss and age group classification loss through several layers, as described in the following subsection.

2.3.5 Loss Functions

As for the generator, we compute reconstruction losses with L1-norm between original and reconstructed silhouette sequences/frame differences and with a cross entropy between original and reconstructed gait period to preserve individuality as demonstrated in [47, 48]. Specifically, similar to [92, 93, 98], we apply the generator twice: original age group to target one; target age group to original one, to get reconstruction. Let $\vec{x}_{\text{sil}}$, $\vec{x}_{\text{diff}}$, and $\vec{x}_{\text{period}}$ are the silhouette sequence, the frame difference sequence, and the gait period of the original age group c_{org}, respectively, and $G_s(\vec{x}_s; c_{\text{trg}})$ is a generator to the target age group c_{trg} for a component $s \in \{\text{sil}, \text{diff}, \text{period}\}$. The reconstruction loss is then computed as

$$L_{\text{rec}} = \sum_{s \in \{\text{sil},\text{diff},\text{period}\}} \lambda_s \mathbb{E}_{\vec{x}_s, c_{\text{org}}, c_{\text{trg}}} [d_s(\vec{x}_s, c_{\text{org}}, c_{\text{trg}})] \tag{2.3.5}$$

$$
\begin{aligned}
&d_s(\vec{x}_s, c_{\text{org}}, c_{\text{trg}}) = \\
&\begin{cases}
||\vec{x}_s - G_s(G_s(\vec{x}_s; c_{\text{trg}}); c_{\text{org}})||_1, & \text{if } s \in \{\text{sil}, \text{diff}\} \\
f_{\text{ce}}(\vec{x}_s, G_s(G_s(\vec{x}_s; c_{\text{trg}}); c_{\text{org}}), & \text{if } s \in \{\text{period}\},
\end{cases}
\end{aligned}
\tag{2.3.6}
$$

where f_{ce} stands for cross entropy loss, weights λ_{sil}, λ_{diff}, and λ_{period} are set to 1, 0.01, and 1, respectively.

As for the discriminator, we apply two loss functions: an adversarial loss and a age group classification loss.

The adversarial loss is introduced to make translated gait videos realistic. More specifically, we choose Wasserstein GAN [99] to make generated sequences indistinguishable from real sequences and stabilize training as

$$
\begin{aligned}
L_{\text{adv}} = &\lambda_{\text{adv}} \mathbb{E}_{\vec{x}}[D(\vec{x})] - \lambda_{\text{adv}} \mathbb{E}_{\vec{x},c}[D(G(\vec{x}, c))] \\
&- \lambda_{\text{gp}} \mathbb{E}_{\tilde{\vec{x}}}[(||\nabla_{\tilde{\vec{x}}} D(\tilde{\vec{x}})||_2 - 1)^2],
\end{aligned}
\tag{2.3.7}
$$

where $\vec{x}$ denotes a concatenated vector of $\vec{x}_{\text{sil}}$, $\vec{x}_{\text{diff}}$, and $\vec{x}_{\text{period}}$, $\tilde{\vec{x}}$ denotes a uniformly sample straight line of the concatenated vector between real and fake ones, and c is an age group label.

The age group classification loss is introduced to preserve a age group-specific property in gait. Unlike previous studies [47, 48] use a pre-trained age group classifier, we train the age group classifier through optimization of both generator and discriminator in an end-to-end manner similarly to [93]. Specifically, the age group classification loss poses constraints on real videos to optimize discriminator, and pose constraints on fake videos to optimize generator as follows

$$
\begin{aligned}
L_{\text{cls}}^{\text{real}} &= \mathbb{E}_{\vec{x}, c_{\text{org}}}[-\log D(c_{\text{org}}|\vec{x})] \\
L_{\text{cls}}^{\text{fake}} &= \mathbb{E}_{\vec{x}, c_{\text{trg}}}[-\log D(c_{\text{trg}}|G(\vec{x}, c_{\text{trg}}))].
\end{aligned}
\tag{2.3.8}
$$

Finally, the full loss function composed of the losses introduced above are

defined for each of the discriminator and the generator as

$$\begin{aligned} L_D &= -L_{\text{adv}} + \lambda_{\text{cls}} L_{\text{real}} \\ L_G &= L_{\text{adv}} + \lambda_{\text{cls}} L_{\text{fake}} + \lambda_{\text{rec}} L_{\text{rec}}. \end{aligned} \tag{2.3.9}$$

2.4 Experiments

2.4.1 Datasets

We trained and evaluated our model on OULP-Age dataset [100], which is the largest gait database with age information in the world. We used a subset of 26,159 subjects with ages ranging from 2 to 90 years old. We randomly chose 2,616 subjects (roughly 10% of the entire dataset for quantitative and qualitative evaluation, and used the remaining for training.

We then set the age groups. Gait reflects human's physical growing process, where the difference between neighboring age group is not necessarily the same [101, 102]. Generally, young children and teenagers (i.e., under 20 years old) grow more rapidly than adults, since the body of adults have grown into a mature physical state [103]. Therefore, instead of dividing ages with a uniform interval, we defined five age groups: [0, 5], [6, 10], [11, 15], [16, 60], and over 60 similarly to [104]. Examples of the dataset are shown in Fig. 2.3-3, and its statistics are shown in Table 2.4.1.

Table 2.4.1 Statistics of subset of OULP-Age.

Age group	[0,5]	[6,10]	[11,15]	[16,60]	Over 60	Total
#Training	639	4,148	2,961	15,108	687	23,543
#Test	71	462	329	1,678	76	2,616

2.4.2 Experimental Setup

We use Adam optimizer [105] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for optimization. The learning rate starts from 0.0001 and fix for 100 epochs, then decays by

0.1 every 100 epochs, and stays at 0.000001 finally. The batch size is set to 8. Similar to [99], we perform one generator update after five discriminator updates. We compared the proposed method with the state-of-the-arts on deep generative adversarial facial aging including CAAE [46], IPCGAN [47], and S2GAN [48] qualitatively and quantitatively. For qualitative evaluation, we visualize the generated gait sequence in subsection 2.4.3. For quantitative evaluation, we conducted experiment on age group classification and cross-age gait recognition (i.e., identity recognition) in subsections 2.4.4 and 2.4.5, respectively.

2.4.3 Qualitative Visualization

I checked the generation results of 2616 subjects in the test set and the randomly sampled generation results for visualization in the training phase for every 1000 epoch, and visualized typical examples of generated gait videos from different methods given target age group for comparison in Figures 2.4-4, 2.4-5 and 2.4-12. CAAE [46] projects the encoded vector to a latent manifold which is constrained by a simple uniform distribution, and does not well preserve individuality (e.g., hairstyle in Fig. 2.4-12 and woman's skirt in Fig. 2.4-13 cannot be reflected in translated target age group gait video by CAAE. Even the gender of the translated subject is hard to distinguish in CAAE results). While IPCGAN [47] adopts a pre-trained classifier to better preserve individuality, it produces artifacts in arm in some cases. S2GAN [48] is a state-of-the-art facial aging method and introduced a well-designed S2 module which well captures age group-specific characteristics (e.g., senior people tend to be fatter). However, it fails to preserve individuality (e.g., the woman in Fig. 2.4-5 is slimmer than general one, but the generated video by S2GAN ignores this characteristic). On the other hand, our method well captures both age-specific characteristics and individuality, yielding the best qualitative result.

We also visualize the age progressed and regressed gait video for multiple age groups with our method in Figures 2.4-7 to 2.4-11. We can see that our method successfully realizes realistic body shapes such as body-length, head-to-body ratio, leg length for kid (0–5 years old), child (6–10 years old), and teenager (11–15

Fig. 2.4-4. Input: Original gait video (1st row, age group [0-5]). Output: translated ones to age group [11-15] by CAAE, IPCGAN, S2GAN, and our method (from the 2nd to 5th row)t.

Fig. 2.4-5. Input: Original gait video (1st row, age group [16-60]). Output: translated ones to age group [6-10] by CAAE, IPCGAN, S2GAN, and our method (from the 2nd to 5th row).

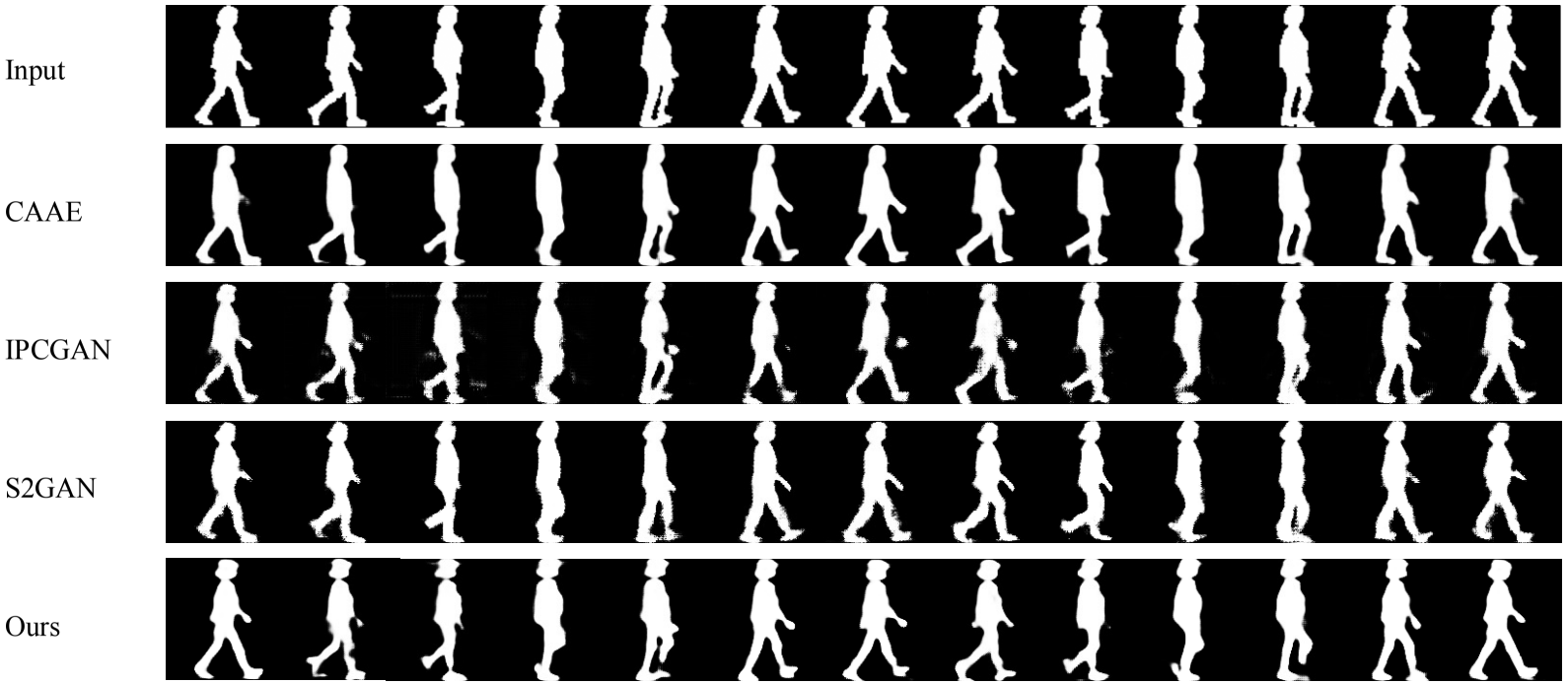


Fig. 2.4-6. Input: Original gait video (1st row, age group Over 60). Output: translated ones to age group [16-60] by CAAE, IPCGAN, S2GAN, and our method (from the 2nd to 5th row).

years old), respectively. Meanwhile, our method generates smaller stride length in the gait video of senior age group (over 60 years old), which corresponds to the physical phenomenon that people tend to walk slower when aged.

Since such self-check may not be convincing enough, we'll conduct subjective tests in the future to provide the qualitative evaluation of whether the generated videos keep age group/identity characteristic from human perspective.

2.4.4 Age Group Classification

Age group classification on real gait sequences can be widely deployed in real-world applications such as market research to find the potential consumer of the product, automatic age-dependent crowd counting, etc. However, the age group classification experiment conducted in this thesis is based on simulated data, which is regarded as a measure to evaluate the generation quality of aging patterns. In other words, such experiment is to evaluate to what extent do the generated cross-age gait silhouette sequence simulate the real characteristics of the intended age group.

Specifically, we first designed an age group classifier using a modified ResNet-18 architecture [106] by following [93], and then trained it with real gait silhouette sequence of OULP-Age (the same training and test set split). The input dimension of first convolution layer of ResNet-18 is modified to $N_{\text{img}} = 25$ in order to handle not a single still image but a gait silhouette sequence. By classifying the generated gait silhouette sequences from different methods using the same pre-trained ResNet-18 classifier for an unseen test set, we can check to which extent each method generates age progressed/regressed video that can present the characteristics of the intended age group.

Experimental results on age group classification accuracies among benchmarks are shown in Table 2.4.2. We also report two more properties: whether the multi-age group gait sequence generation model is trainable in an end-to-end manner, and the number of network parameters.

As a result, we can see that the proposed method outperforms the other benchmarks and that it yielded the best accuracies for all of the age groups. Besides, un-

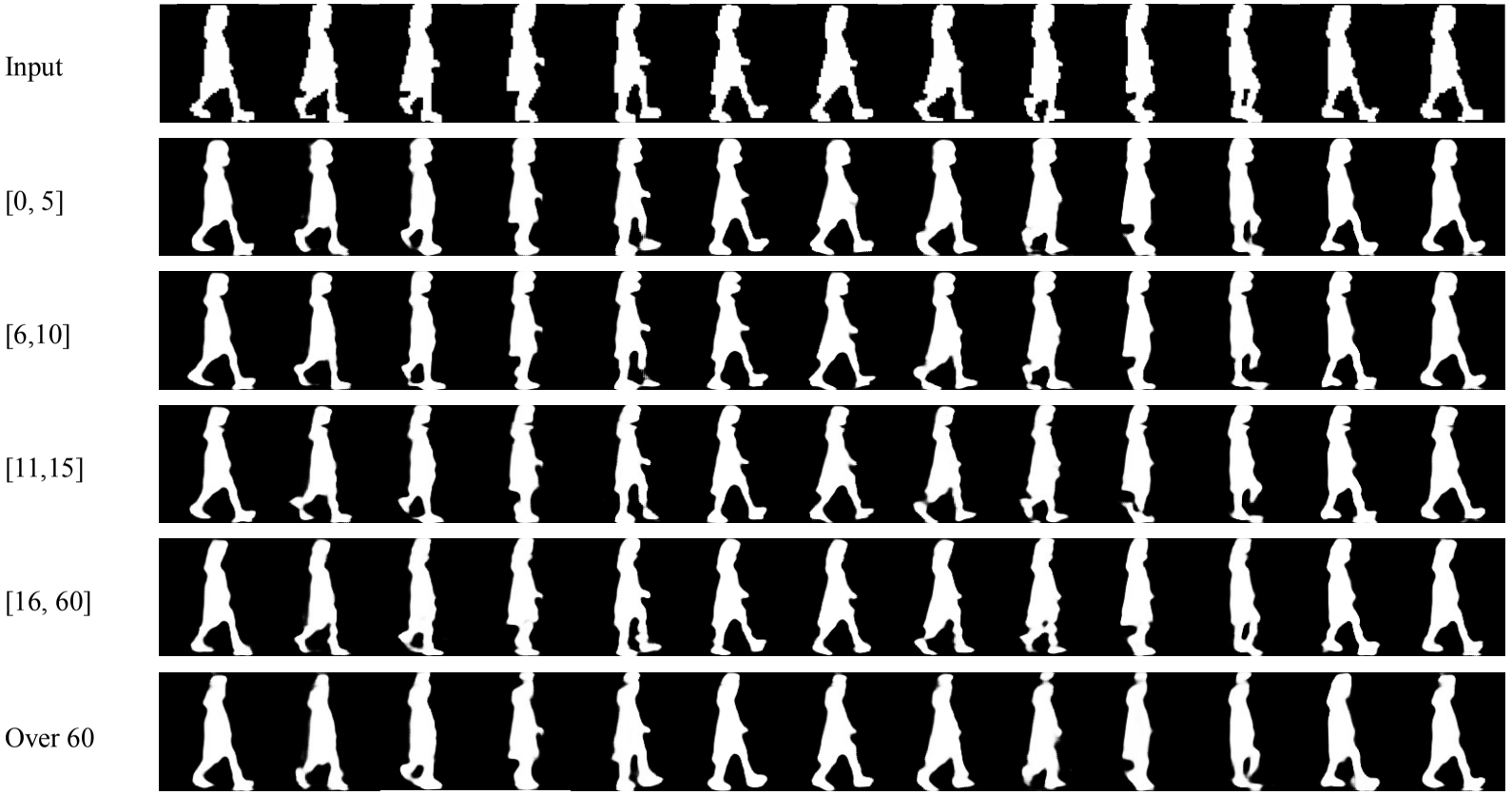


Fig. 2.4-7. Input: Original gait video (1st row, age group [0,5]). Output: translated ones to age groups [0, 5], [6, 10], [11, 15], [16, 60], over 60 by our method (from the 2nd to 6th row).

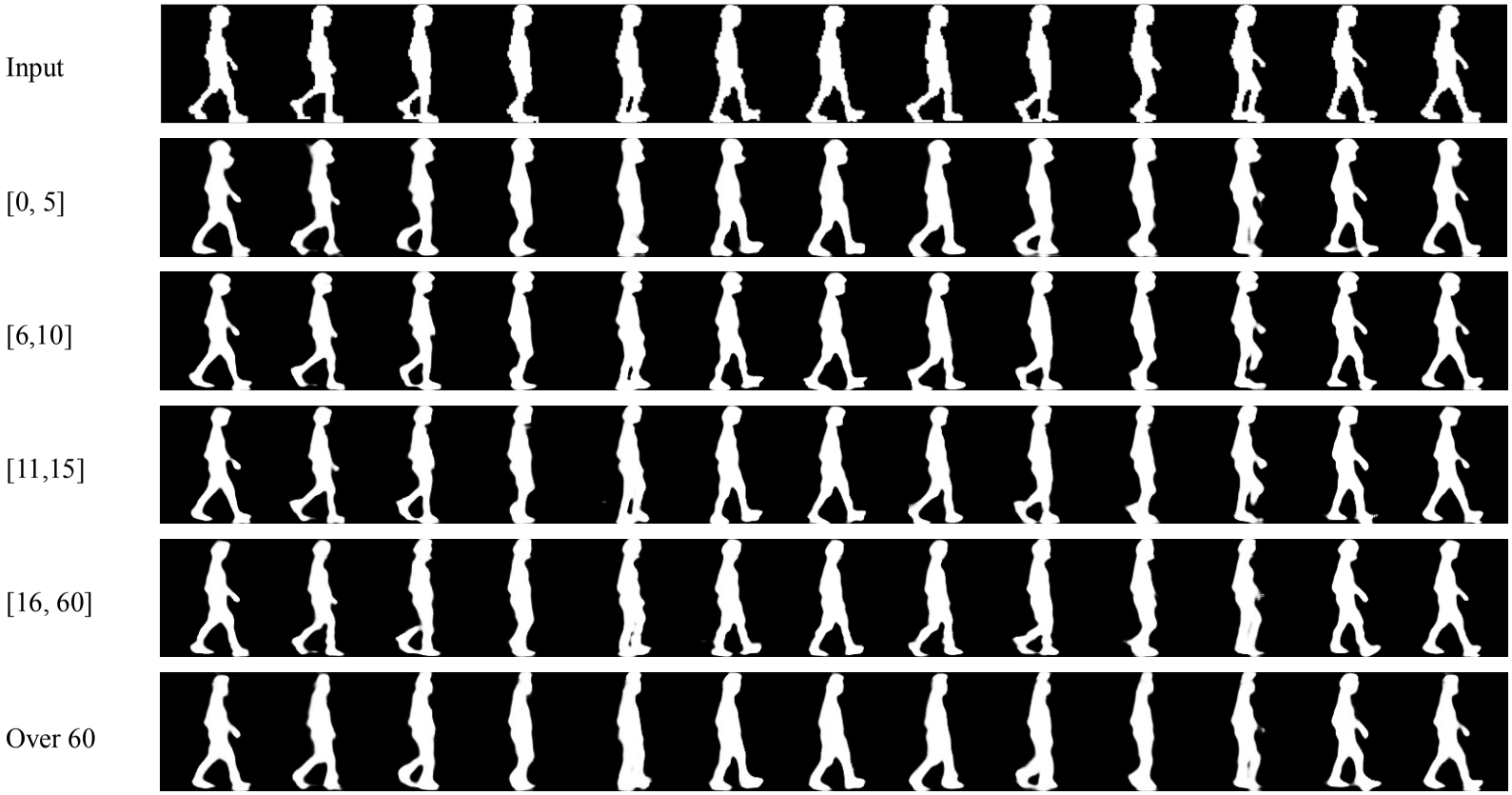


Fig. 2.4-8. Input: Original gait video (1st row, age group [6,10]). Output: translated ones to age groups [0, 5], [6, 10], [11, 15], [16, 60], over 60 by our method (from the 2nd to 6th row).

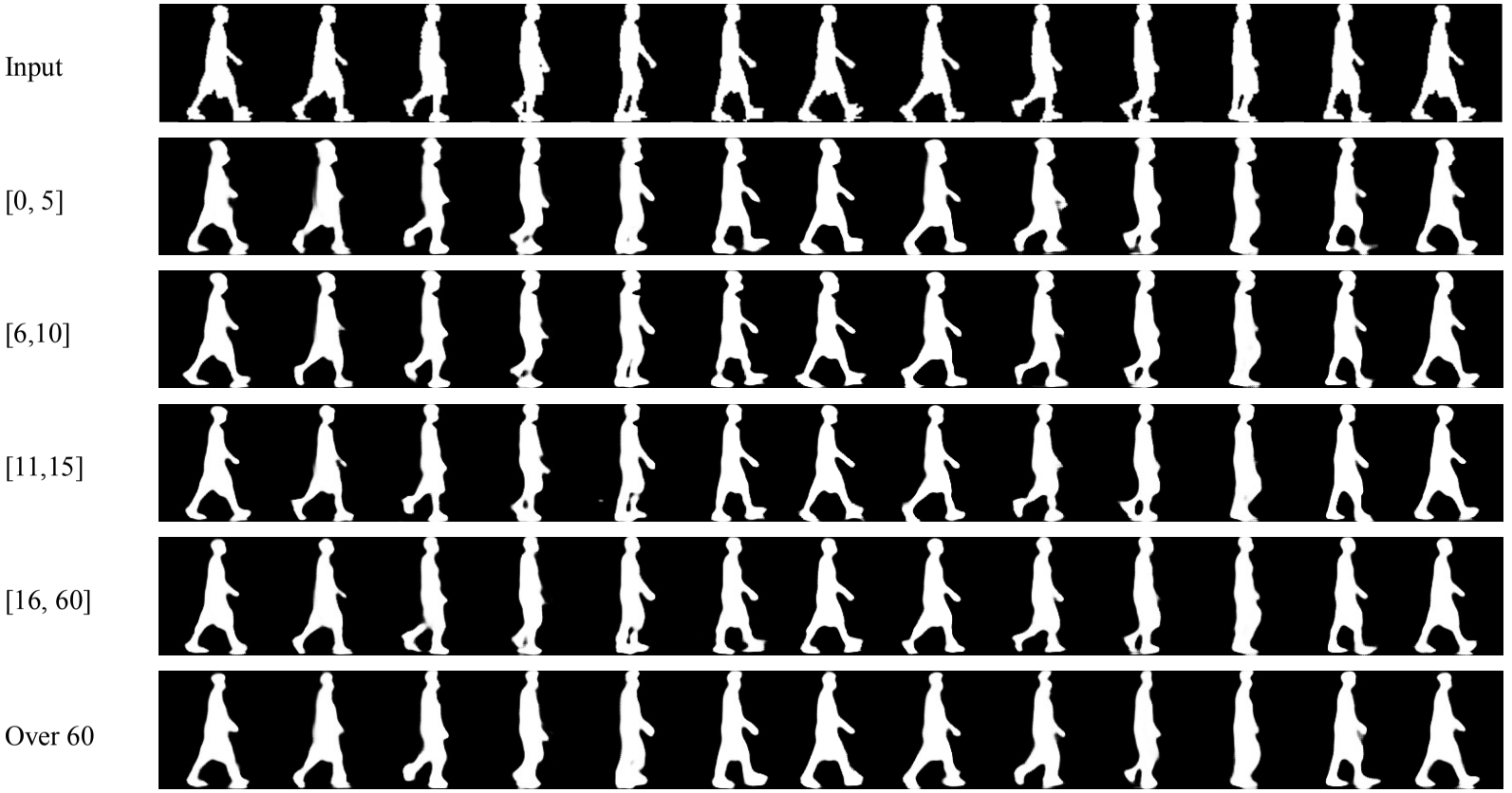


Fig. 2.4-9. Input: Original gait video (1st row, age group [11,15]). Output: translated ones to age groups [0, 5], [6, 10], [11, 15], [16, 60], over 60 by our method (from the 2nd to 6th row).

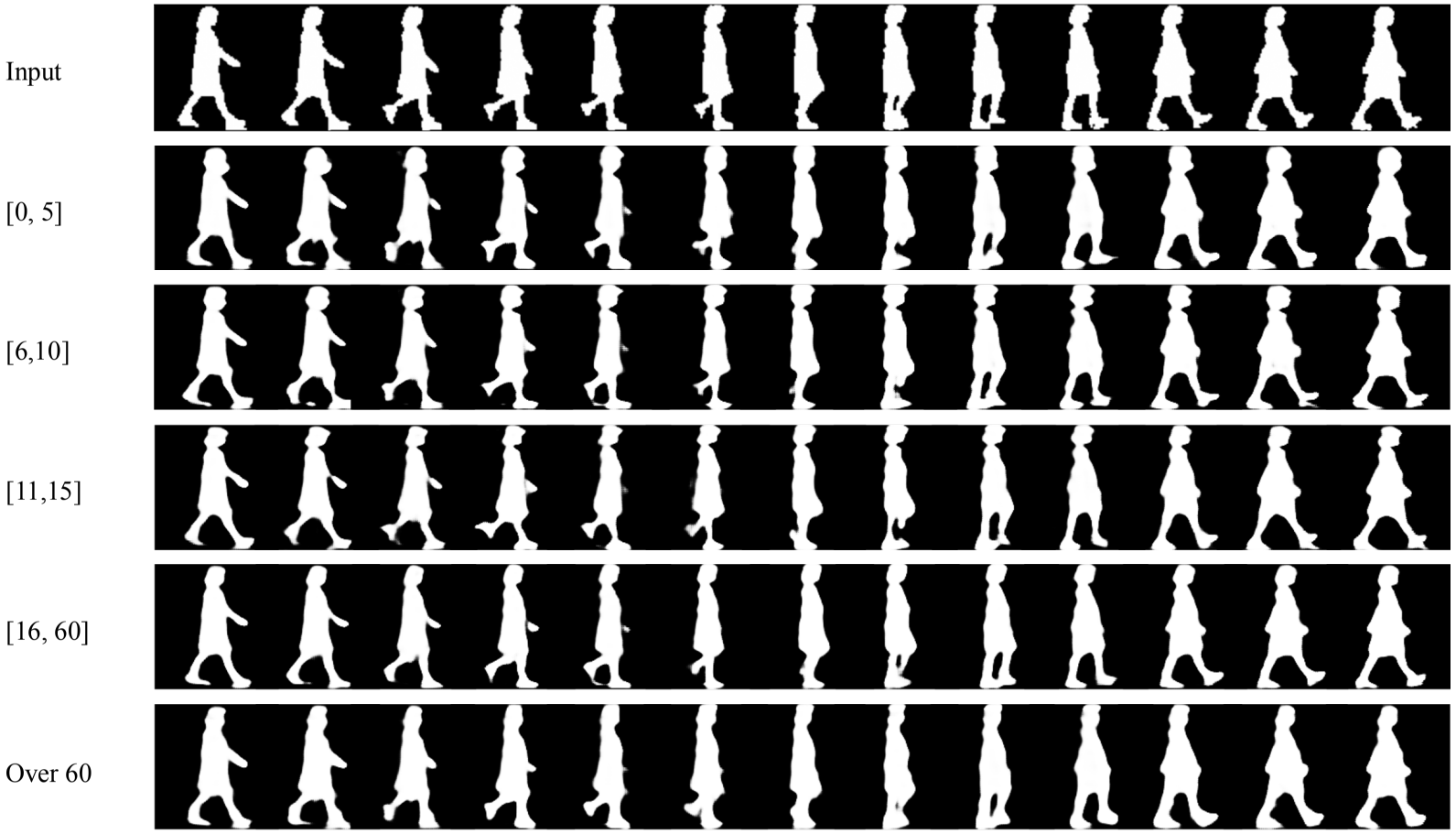


Fig. 2.4-10. Input: Original gait video (1st row, age group [16,60]). Output: translated ones to age groups [0, 5], [6, 10], [11, 15], [16, 60], over 60 by our method (from the 2nd to 6th row).

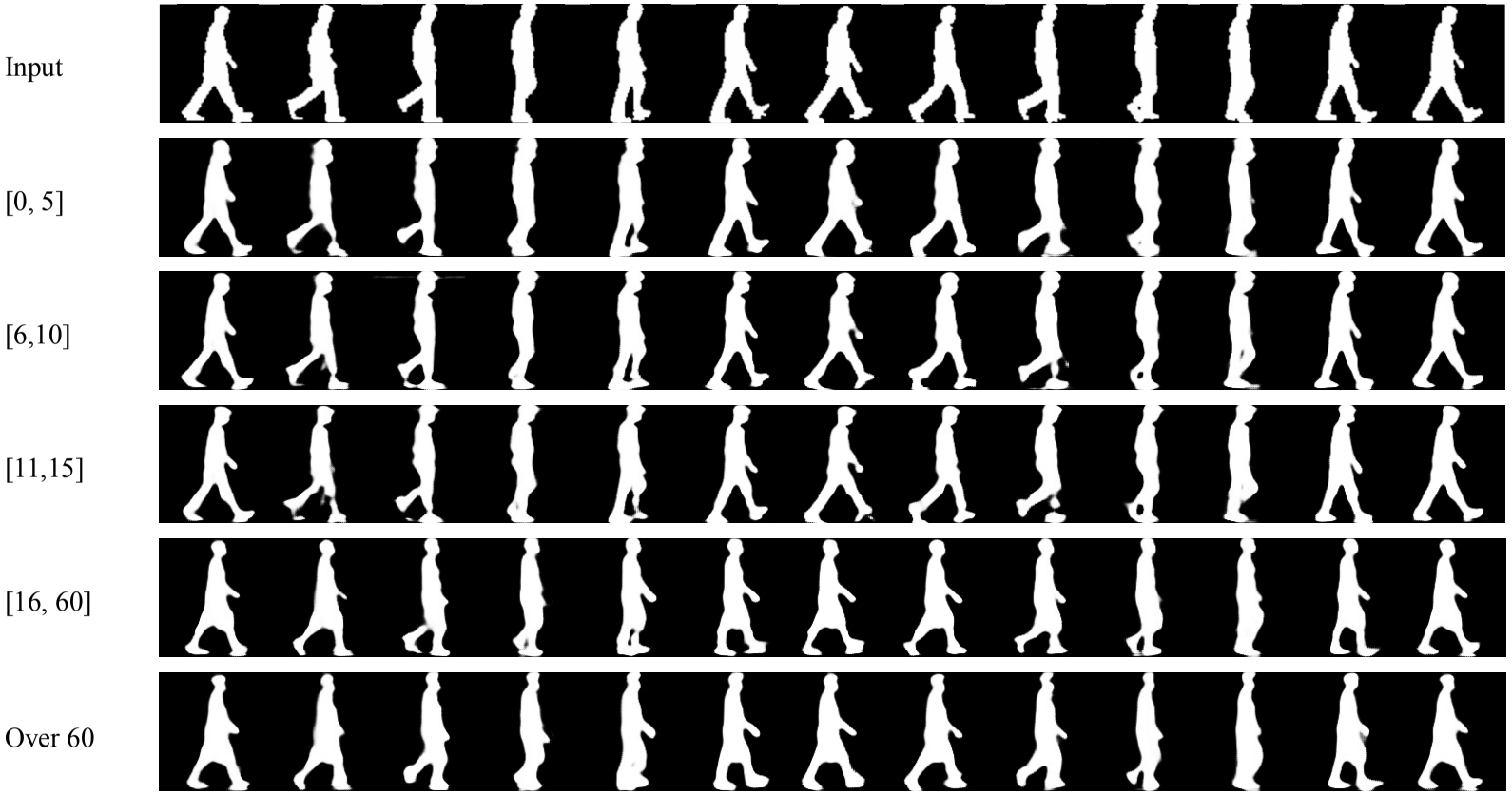
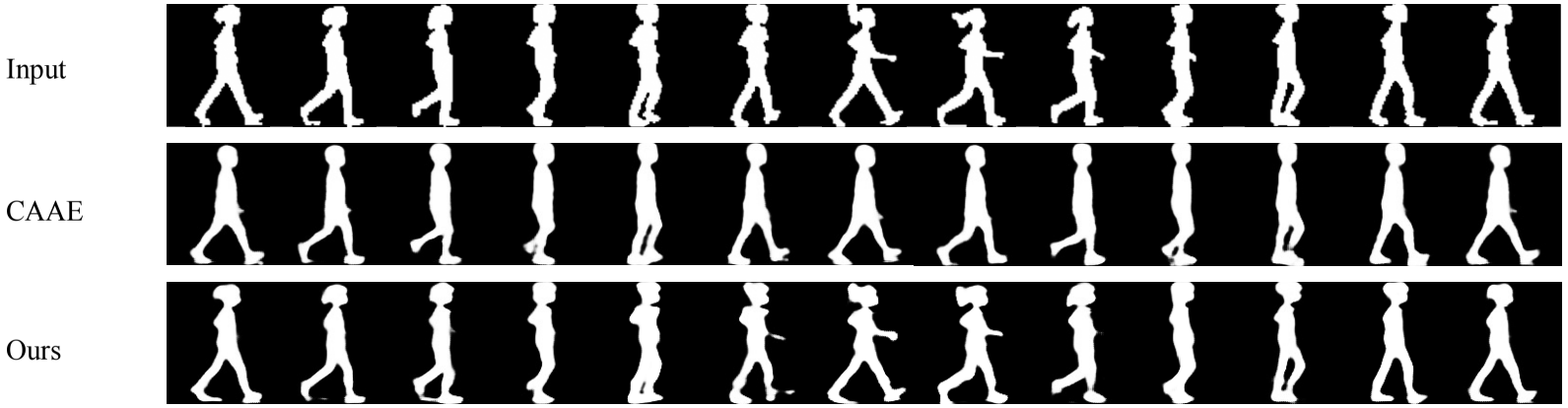
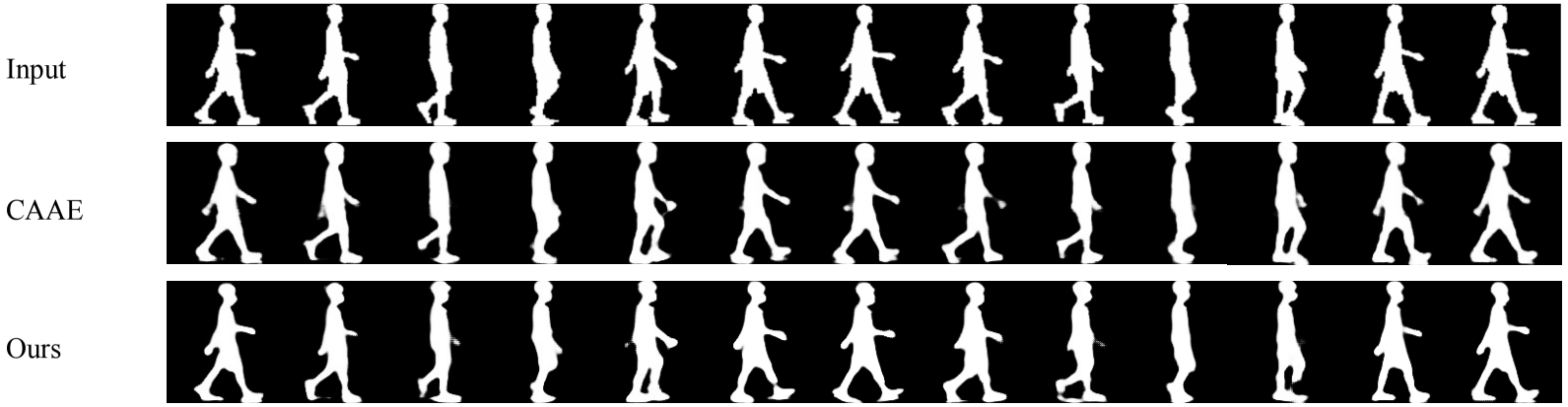


Fig. 2.4-11. Input: Original gait video (1st row, age group Over 60). Output: translated ones to age groups [0, 5], [6, 10], [11, 15], [16, 60], over 60 by our method (from the 2nd to 6th row).

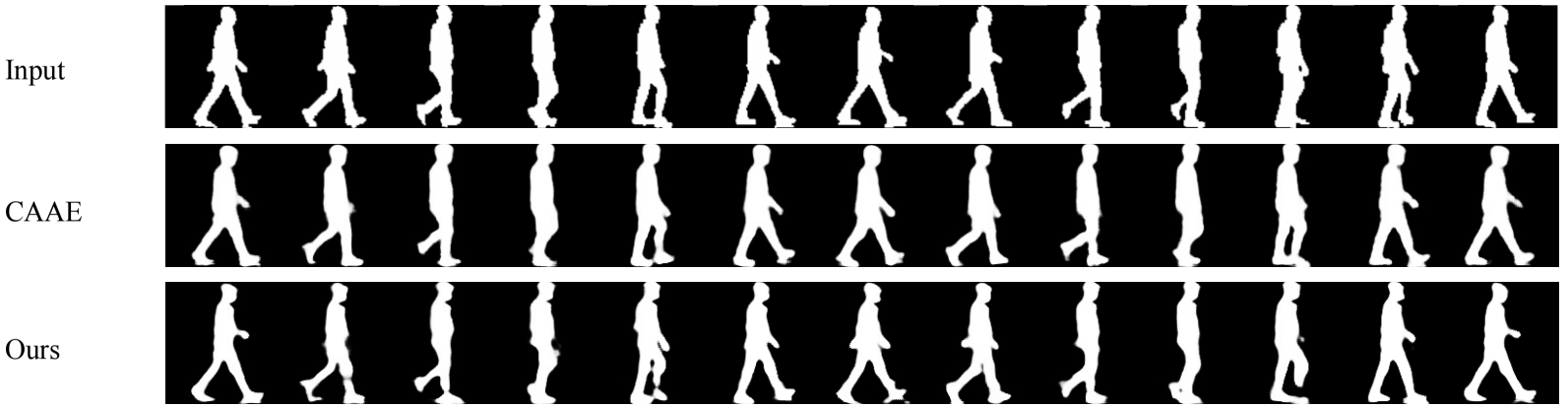


(a) Subject A. Female in age group [11-15]

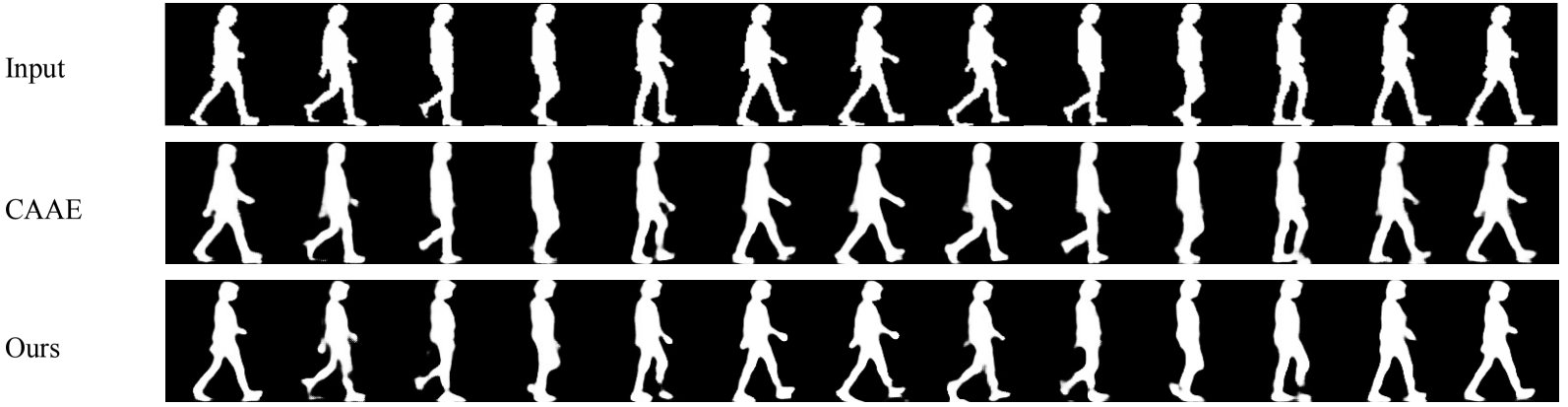


(b) Subject B. Male in age group [11-15]

Fig. 2.4-12. Input: Original gait video in age group [11-15]. Output: translated ones to age group [6-10] by CAAE and our method in (a) and (b).



(a) Subject A. Male in age group [16-60]



(b) Subject B. Female in age group [16-60]

Fig. 2.4-13. Input: Original gait video in age group [16-60]. Output: translated ones to age group [11-15] by CAAE and our method in (a) and (b).

Table 2.4.2 Age group classification accuracy for benchmarks. E2E indicates end-to-end training. Bold font indicates the best accuracy.

Method	Age group (Ground Truth)					Avg	E2E	#Param
	[0, 5]	[6, 10]	[11, 15]	[16, 60]	Over 60			
CAAE	10.53	58.75	35.09	92.69	16.73	42.76	√	44.4M
IPCGAN	12.16	65.63	42.24	81.50	16.17	43.54		36.4M
S2GAN	11.72	62.52	46.73	93.86	19.85	46.93		36.7M
Ours	**22.74**	**72.90**	**84.17**	**98.01**	**49.54**	**65.47**	√	39.3M

like IPCGAN [47] and S2GAN [48] requires a pre-trained age group classifier to train the multi-age group translation model, the proposed method does not require it, i.e., it can train the model in an end-to-end manner, which can save training time and efforts for the pre-training. Moreover, we notice that the number of network parameters for the proposed method is comparable to the benchmarks, which shows a good scalability of the proposed method.

2.4.5 Cross-age Gait Recognition

We conducted cross-age gait recognition experiments to evaluate the preservation of individuality using the age progressed/regressed generated gait sequence in addition to real ones. The dataset for the cross-age gait recognition experiment composed of three subsets: a training set, a gallery set, and a probe set. The training set contains $23,543$ subjects, the gallery and probe sets form a test set composed of the other $2,616$ subjects that are disjoint from the training set. The real silhouette sequences are assigned to the gallery, while the generated sequences are assigned to the probe. The training and test phase is illustrated in Fig. 2.4.5.

We first train the proposed cross-age gait silhouette translation framework to generate five gait sequences per subject, which correspond to the five age groups. Since the real cross-age gait database is not available, we designed the following simulated experimental setting. In the training phase, we train GaitSet [69], which is a state-of-the-art network structure in gait recognition for gait silhouette sequences (as opposed to static gait templates), with multiple real and generated gait silhouette sequences from training set. Cross entropy loss and triplet loss are

imposed to keep the identity across ages during training. In the test phase, the real silhouette sequences of the test set are assigned to the gallery, while the corresponding generated sequences are assigned to the probe. The trained GaitSet model takes in one input from probe, and compare the final representation of the input subject with subjects in gallery.

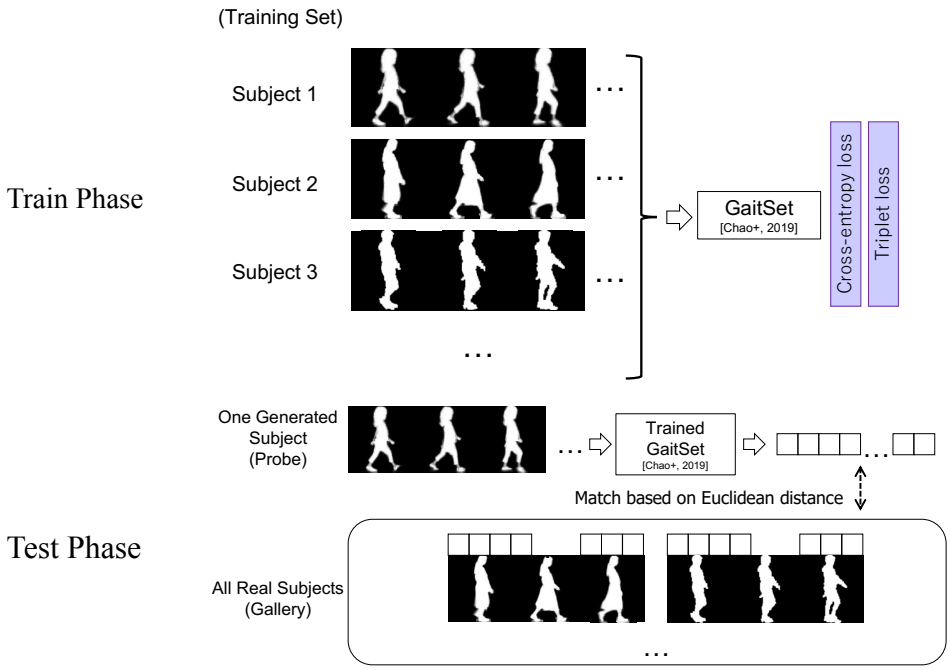


Fig. 2.4-14. Train and test phase of GaitSet [69].

In an identification scenario, we matched a probe to all the subjects in gallery and evaluated rank-1 identification rate based on dissimilarities (i.e., euclidean distance between the final representations in the trained GaitSet network). We computed the standard deviation (uncertainty) sFRR of false rejection rate (FRR) pFRR in case of a single attempt per subject according to [107, 108], which is represented as:

$$\sigma_{FRR} = \sqrt{\frac{p_{FRR}(1-p_{FRR})}{n-1}} \tag{2.4.1}$$

where n is the number of subjects and it is $2,616$ in our case. The standard deviation of true acceptance rate and rank-1 identification rate can be computed in the

same manner. Take results for the age group $[0,5]$ as an example, in the form of rank-1 identification ($\pm$the standard deviation), the following results are obtained. CAAE: 84.9%($\pm$0.70%), IPCGAN: 99.6%($\pm$0.12%), S2GAN: 98.8%($\pm$0.21%), Proposed: 99.7%($\pm$0.11%). We therefore confirmed that there is still statistical significant difference between the proposed method and the benchmark method even though the absolute difference of the rank-1 identification rate is less than 1% (between the proposed method and IPCGAN).

In a verification scenario, an input pair of a probe and a gallery is accepted as the same subject (i.e., positive sample pair) if the dissimilarity measure between them is below an acceptance threshold, and is rejected otherwise (i.e., negative sample pair or different subject pair). We computed the equal error rate (EER) of the false acceptance rate and false rejection rate as a typical performance measure. As reference, we also confirmed the statistical significant difference in terms of EER.

Table 2.4.3 Rank-1 identification rates [%] and EER [%] ($\pm$standard deviation [%]) for each age group in probe. Bold font indicates the best performances.

Measure	Method	Age group					Avg.
		$[0,5]$	$[6,10]$	$[11,15]$	$[16,60]$	Over 60	
Rank-1	CAAE	84.9 ($\pm$0.70)	84.2 ($\pm$0.70)	84.3 ($\pm$0.71)	85.5 ($\pm$0.71)	85.2 ($\pm$0.69)	84.80
	IPCGAN	99.6 ($\pm$0.12)	**99.8** ($\pm$**0.09**)	**99.9** ($\pm$**0.06**)	99.2 ($\pm$0.17)	99.3 ($\pm$0.16)	99.56
	S2GAN	98.8 ($\pm$0.21)	97.9 ($\pm$0.28)	98.2 ($\pm$0.26)	98.8 ($\pm$0.21)	99.0 ($\pm$0.19)	98.54
	Ours	**99.7** ($\pm$**0.11**)	99.5 ($\pm$0.14)	99.3 ($\pm$0.16)	**99.8** ($\pm$**0.09**)	**99.7** ($\pm$**0.11**)	**99.60**
EER	CAAE	0.93 ($\pm$0.19)	0.97 ($\pm$0.19)	0.96 ($\pm$0.19)	0.96 ($\pm$0.19)	0.91 ($\pm$0.19)	0.94
	IPCGAN	0.73 ($\pm$0.17)	0.70 ($\pm$0.16)	0.96 ($\pm$0.19)	0.93 ($\pm$0.19)	0.84 ($\pm$0.18)	0.83
	S2GAN	0.78 ($\pm$0.17)	0.88 ($\pm$0.18)	0.96 ($\pm$0.19)	0.70 ($\pm$0.16)	0.64 ($\pm$0.16)	0.79
	Ours	**0.24** ($\pm$**0.10**)	**0.20** ($\pm$**0.09**)	**0.31** ($\pm$**0.11**)	**0.32** ($\pm$**0.11**)	**0.34** ($\pm$**0.11**)	**0.28**

Experimental results of the cross-age gait recognition are summarized in Table 2.4.3. As a result, we can see that the proposed method yielded the best accu-

racy for all age groups, which indicates the superiority of the proposed method in terms of individuality preservation.

Under the real-world application such as time-lapsed identity verification in criminal investigation, the proposed method achieves an EER of 0.28%, which indicate the criminal investigator can narrow down the search scope of 1000 people to 2.8. Since the manual check of large-scale surveillance videos can be a tedious work, the proposed cross-age gait video translation method can be leveraged as a way of screening to largely reduce the human labor of the criminal investigator. Although such simulated cross-age gait recognition experiment may not be ideal for evaluation of identity preserve across ages and might not guarantee the performance on real data, the GaitSet model learns well during training provide evidence that identity is discriminative in the generated cross-age gait silhouette sequences.

2.4.6 Ablation Study

We made ablation studies on individual modules and report the accuracies of age group classification in Table 2.4.4.

First, in order to validate the effectiveness of the gait period, we removed the period stream in both generator and discriminator. As a result, it turns out that the accuracy decreases by approx. 15% without the gait period, which indicates that the gait period is essential for age analysis in gait.

Second, in order to validate the effectiveness of the SlowFast path, we replaced it with two extended one-stream discriminators derived from StarGAN [93], i.e., we used the one-stream discriminator for both gait silhouette sequence stream and frame difference sequence stream. As a result, it turns out that the accuracy decreases by approx. 2% without the SlowFast path. Moreover, the SlowFast path is more efficient than the one-stream discriminator [93] w.r.t. the number of network parameters (i.e., the number of parameters with the one-stream discriminator is more than twice). This is because the SlowFast path can save the number of parameters by limiting the temporal resolution in the slow path (i.e., sampling by every five frames) as well as by limiting the channel capacity in the fast path.

Finally, we removed the MAB to validate the effectiveness of interaction between the gait silhouette sequence and the frame difference sequence. As a result, it turns out that the accuracy decreases by approx. 7% without the MAB. This indicates that interaction between motion and body shape by the MAB is essential to age group classification.

This ablation study consequently demonstrates the benefits of the MAB, the SlowFast path, and the gait period.

Table 2.4.4 Ablation study of different modules in our network on average age-group classification accuracy [%]. We show the results for the final proposed method (top row) and compare them with the results obtained when individual modules are removed to validate their effectiveness (second to bottom rows). #Params: number of network parameters in millions.

Modules			Accuracy	#Params in million	
MAB	SlowFast	Period		Generator	Discriminator
√	√	√	**65.47**	20.01	19.35
√	√		50.34	20.00	19.30
√		√	63.21	20.01	44.83
	√	√	58.05	16.60	19.35

2.4.7 Failure Mode Analysis

We list a typical false example in Fig. 2.4-15, where the first row is the input of age group [0, 5] years old, the second row is the output of translated ones to age group [16, 60] years old. Young children usually have larger head-to-body ratio compare to adults. When translating from age group [0,5] to [16, 60], our method can successfully generate smaller head-to-body ratio (small head for adult), but will sometimes lead to artifacts in head (red circle). Meanwhile, since our method is image-based one, generated arm and legs might not be consistent across frames (green circle). We will solve it by introducing a model-based method, which enables us to prevent from such artifacts and inconsistent body shape, in future work.

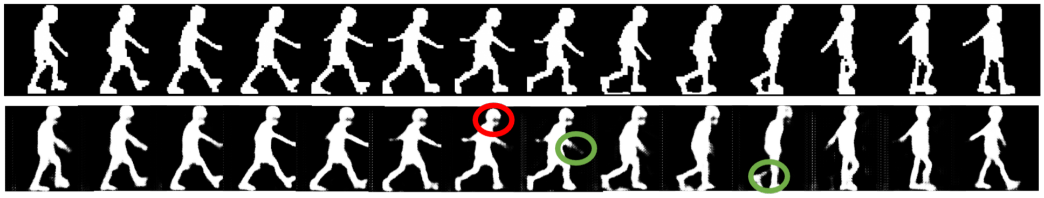


Fig. 2.4-15. False example. Input: 1st row, original gait video of age group [0, 5] years old. Output: 2nd row, translated gait video of age group [16, 60] years old.

2.5 Summary

In this chapter, we introduced an end-to-end motion augmented multi-age group gait video translation framework, which exploits both motion and body shape information from gait sequences. Specifically, we proposed three-stream generator/discriminator with the gait period, period-normalized gait silhouette sequence, and the frame difference sequences in conjunction with motion augmented block and the SlowFast path. Experiments on OULP-Age demonstrated the superiority of the proposed method quantitatively and qualitatively among other state-of-the-art methods.

Chapter III
General Image Matting with simultaneous Alpha and Background Output

3.1 Introduction

Image matting, which refers to the problem that accurately extracts the foreground opacity from an image, also known as alpha matte, is a fundamental operation for various tasks during the post-production stage of feature films, such as background replacement, layer separation and color correction.

Mathematically, the color mixtures in soft transitions between foreground and background are typically represented with the composition equation

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i, \tag{3.1.1}$$

where $\alpha_i \in [0, 1]$ stands for the weight or the opacity of the foreground color at pixel i, I_i is the original image at pixel i, F_i and B_i denote the foreground and background at pixel i respectively. Since neither the foreground and background colors nor the alpha mattes are known, estimating the value of alpha matte is a highly ill-posed problem. To alleviate the difficulty of this problem, typically an image is accompanied with a user input *trimap*, which is a rough segmentation of the input image into foreground, background, and unknown region with undetermined

opacity.

Existing image matting methods can be categorized into traditional methods and learning-based methods. Traditional methods can be further divided into propagation-based methods [109–112], sampling-based methods [113–115], and a hybrid of both [116–118]. In propagation-based methods, the unknown alpha values are propagated from the known region according to the matting equation (3.1.1). Nevertheless, propagation-based methods rely heavily on the continuity of image. In sampling-based methods, for each pixel in the unknown region, the best foreground and background color pair is sampled to compute the alpha matte. However, due to the absence of high-level semantic information, sampling-based methods perform badly when the foreground and background color distributions have a large overlap. In addition, the above traditional matting methods are very inefficient, which usually cost minutes to process a single image.

Recently, learning-based matting methods [119–129] have achieved rapid progress by estimating the alpha values directly through a convolutional encoder-decoder neural network. Different from traditional methods that might take several minutes to process a single image even with GPUs, learning-based methods can output results in real-time, and avoid "smearing" or "chunky" artifacts by utilizing high-level semantic information. Nevertheless, existing learning-based methods suffer from deficiency of global contextual information which has been proved essential in matting performance [127]. This is because a matting image may be up to several megapixels, which is too big for learning-based network to capture global contextual information from the original matting image due to the limit size of the receptive field for convolution layers. Other existing methods [129, 130] crop the unknown regions during training so as to reduce the input size and preserve the quality of input. However, cropping the unknown regions will ignore global contextual information in the regions which are not cropped.

In this work, we introduce an end-to-end image matting network, which can seek for unknown-relevant global contextual information from the whole image. Our network consists of two stages as shown in Fig. 3.1-1. In the first stage, we use a deformable sampling layer to downsize the original large image into two de-

formed compact images, magnifying the foreground and background separately. In this way, we still obtain desired small input for the network while maintaining the foreground and background information maximally. In the second stage, we adopt a contextual attention (CA) layer [131] to extract foreground and background information which is related to the unknown region in the crop. However, contextual aggregation for image inpainting cannot be applied in a straightforward way. Key differences from the inpainting-oriented CA layer are as follows: (1) We need to aggregate both foreground and background cues in the unknown areas which is vital for alpha matte estimation. More specifically, a composite color in the input image and foreground/background information gives a strong cue for alpha matte by taking account of the matting equation. (2) The existence of the definite foreground/background is not guaranteed, e.g., no definite foreground in case of a semi-transparent foreground object, and hence no patches may be available either for foreground or background. To cope with this, we employ deformed sampling technique and use all the pixels in the deformed foreground/background, where foreground/background pixels are dominant. Moreover, we would claim a framework of simultaneous alpha matte, background and foreground estimation as a contribution. Unlike most of the existing matting studies that do not estimate background or foreground, and hence a foreground image directly extracted from an input image is always contaminated with a background color, our method can estimate purified foreground as well as purified background, which should be preferable for matting applications (superimposition of foreground objects on another background image). Even though a few studies try estimating foreground and background as well as an alpha matte, the outputs are estimated independently. Hence matting equation is not guaranteed to be held in those methods. Therefore, the proposed method, which satisfies the matting equation in the estimation process, could generate a more reasonable foreground/background.

Our contributions can be summarized as follows: (1) To capture global contextual information from a large input image without degrading the image quality, we propose an end-to-end three-branch image matting framework, which can exploit unknown-relevant global contextual information condensed from the

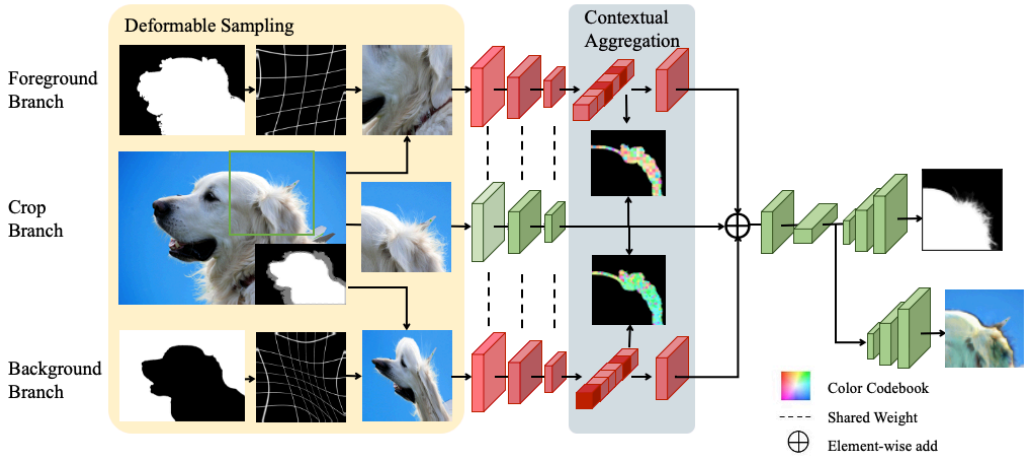


Fig. 3.1-1. Natural Image Matting with Attended Global Context framework.

whole large image. (2) We propose a matting-oriented contextual aggregation supported by deformed sampling. Unlike the original contextual aggregation designed for image inpainting, which cannot consider a situation of no definite foreground/background region in image matting, our matting-oriented contextual aggregation can cope with such a situation by making use of all the pixels in the deformed foreground/background where foreground/background pixels are dominant. (3) Our method can estimate alpha matte and background simultaneously while keeping the matting equation, which can improve the foreground extraction performance qualitatively. (4) Our method achieves competitive results on both Composition-1k and the alphamatting.com benchmark.

3.2 Related Work

Both traditional methods and learning-based methods for image matting have been widely investigated in computer graphics and vision community. Traditional methods can be further categorized into sampling-based methods and propagation-based methods. In this section, we provide a brief review of some representative examples.

Sampling-based Methods: Sampling-based methods [111, 116–118] first gather a set of samples from known foreground and background color regions, and then choose the best foreground-background color pair from them for each pixel in the unknown region. The sampling-based methods differ in building the sample set and selecting good samples within the set. Bayesian Matting [111] models foreground and background colors as mixtures of Gaussians and samples from the foreground and background pair with the maximum likelihood. Instead of constraining the sampling set to the boundaries of trimap, comprehensive sampling [118] models sampling distance as a function of the proximity of the unknown pixel to the known region. Shared Matting [117] recognizes that true samples may be located farther away from an unknown region, and therefore collects samples by shooting rays in different directions. However, the ray-based sampling strategy can find spatially distant samples, but most of them are nearby. Thus the true samples can still be missing. To decrease the probability of missing a true sample, He *et al.* propose the global matting method [116] to search in a global sample set that contains all available samples, which can decrease the probability of missing a true sample. However, the increasing number of samples makes the subsequent sample selection more expensive.

In summary, the methods in this category take advantage of natural image statistics to solve the ill-posed matting problem. These methods can perform well when a strong correlation between the unknown pixels and known ones exists, i.e., the input image contains smooth regions and the trimap is well-defined. However, local smoothness assumption might not hold especially in a complex scene (e.g., a highly textured and cluttered scene), leading to the fundamental limitation for sampling-based methods. In this paper, we propose a learning-based matting method with attended global contextual information to improve efficiency, which shares a similar idea with sampling-based methods when gathering global information from the whole image.

Propagation-based Methods: Propagation-based methods [109, 110, 112, 132, 133] formulate the estimation of alpha matte as optimization problem to propagate alpha matte values from known foreground and background region to un-

known region based on appearance similarity. The propagation-based methods differ in the way to construct the correlation between pixels and to propagate the alpha matte values. Poisson matting [109] formulates the matting problem as solving Poisson equations with the matte gradient field under the assumption that intensity changes in the foreground and background are locally smooth. Closed form matting [110] proposes a closed-form solution by minimizing a quadratic cost function based on alpha matte under the constraint of local smoothness. Non-local matting [132] is the first work to apply the non-local principle in the matting problem to obtain the sparsity in matte representation and produce good graph clusters. To be faster and more accurate with sparse user label, *K*NN matting [133] employs the non-local principle by using K nearest neighbors to match non-local neighborhoods. Authors in Information flow matting [112] propose a color-mixture flow that utilizes the local and non-local affinities of colors and spatial smoothness.

In summary, the propagation-based methods are likely to be more robust than sampling-based methods when dealing with complex images. However, as alpha matte is estimated in a propagation fashion, from known pixels to unknown ones, small errors could be propagated and accumulated to produce more significant errors. Besides, since affinities are defined only in local windows, they cannot consider the global context at a far distance and hence may result in poor performance when the global context plays an important role. Instead of defining affinities between neighboring pixels to employ local image statistics, the matting-oriented contextual aggregation layer we introduce in this paper can enhance features in the unknown region with relevant foreground/background features in a differentiable way.

Learning-based Methods: Modern deep learning-based methods directly learn an alpha matte from an image and its trimap. Early methods in this category [119, 134] rely on traditional matting methods to approximate the actual matting with the goal of improving overall matte quality. With the invention of the encoder-decoder learning framework, learning-based methods have demonstrated their capability in learning high-level semantics that can avoid "smearing" or high-frequency "chunky"artifacts [123]. Lutz *et al.* [120] improve the decoder

structure by adding atrous spatial pyramid pooling [135] to the decoder to resample the features at several scales. Zhang *et al.* [125] leverage two decoder branches for foreground and background respectively, followed by a fusion module to integrate the raw alpha result. Lu *et al.* [128] introduce an index-guided encoder-decoder framework, where indices are self-learned and used to guide the pooling and upsampling operators. Usually, images are cropped for input during training due to speed and memory concerns in learning-based methods, meanwhile, the architecture of deep convolutional neural network can be ineffective to learn spatial distant information [136, 137], leading to the lack of global contextual information.

In summary, learning-based methods have the most impressive performance among the three categories since the availability of large-scale datasets. However, they might be unable to capture spatial distant information in large images due to the limited size of the receptive field for convolution layers. To tackle this problem, we introduce an end-to-end matting system that collects global contextual information related to unknown regions, which is consistent with global sampling-based methods.

3.3 Natural Image Matting with Attended Global Context

3.3.1 Overview

Our goal is to predict an alpha matte as well as a pure background given an input composite image and a trimap. Note that a foreground is computed based on the basic matting equation (3.1.1) given the input image I, the predicted background $\hat{B}$, and the predicted alpha matte $\hat{\alpha}$. The reason why we estimate the background instead of the foreground is that the ground-truth background is usually available for an entire image while the ground-truth foreground is valid only in a region where the alpha matte value is non-zero in the training phase.

Moreover, we utilize condensed global context information from foreground and background for better prediction, because (1) contextual information in spa-

tially distant area is essential to alpha matte estimation performance [127]; (2) the matting image of the original size is too large for the receptive field of learning-based network.

For this purpose, we design an end-to-end natural image matting with condensed global context with three encoder branches and two decoder branches shown in Fig. 3.1-1. Among the three encoder branches, the middle encoder branch processes the image crop while the top (resp., bottom) encoder branch processes the condensed foreground (resp., background).

In the first stage, inspired by [138], we use a deformable sampling layer to condense the original large image into two smaller images that reflect the foreground and the background respectively. By taking the background as an example, we use the background mask as the attention map to generate a distorted coordinate system (visualized in the grid map in Fig. 3.1-1), and then condense the original large image according to the distorted coordinate system. As shown in Fig. 3.1-1, background pixels are more densely sampled than other pixels, producing a deformed condensed background.

In the second stage, we use the contextual attention layer [131] to seek foreground and background information that is related to spatially distant unknown region in the cropped area. In this way, we can preserve the context information of foreground and background to the utmost separately. Again, by taking background as an example, we extract patches from a background feature map, compare features of the unknown region with background patches all over the whole image, and reconstruct features of the unknown region with relevant background patches, generating background-aware features of the unknown region.

Finally, we aggregate the original crop feature map, foreground-aware feature map, and background-aware feature map by element-wise summation, then feed the aggregated feature map into the remaining network to produce the alpha matte. In the following, we will fully describe the first stage in Subsection 3.3.2 and the second stage in Subsection 3.3.3.

3.3.2 Foreground and Background Deformable Sampling

Most convolutional neural networks only allow constrained input size considering computational and memory efficiency as well as the benefit of effective batch training. However, matting images are usually as large as megabytes in the natural image matting task. Previous methods generally use uniform downsampling and cropping to satisfy the desired input size, but this will cause significant information loss. To preserve the global foreground and background information to the utmost separately, we adopt a deformable sampling layer inspired by [138], and append it to the start of our network to improve downsampling when obtaining information from foreground and background. Different from using saliency estimator to get a saliency map, we simply use foreground and background binary masks as two saliency maps, because we want to maintain foreground and background information in two downsized images separately.

Given the trimap of the original large image $\boldsymbol{I}$, we denote the foreground (resp., background) binary mask as $\boldsymbol{M}^{\mathrm{f}}$ (resp., $\boldsymbol{M}^{\mathrm{b}}$). In practice, as the foreground region of many images is very small, we also include unknown region in the foreground mask $\boldsymbol{M}^{\mathrm{f}}$.

Specifically, the deformable sampling layer f takes in the foreground mask $\boldsymbol{M}^{\mathrm{f}}$ (resp., the background mask $\boldsymbol{M}^{\mathrm{b}}$) and image $\boldsymbol{I}$ to get a deformed compact foreground $\hat{\boldsymbol{I}}^{\mathrm{f}}$ (resp., $\hat{\boldsymbol{I}}^{\mathrm{b}}$) as follows,

$$\hat{\boldsymbol{I}}^{\mathrm{f}} = f(\boldsymbol{I}, \boldsymbol{M}^{\mathrm{f}}), \quad \hat{\boldsymbol{I}}^{\mathrm{b}} = f(\boldsymbol{I}, \boldsymbol{M}^{\mathrm{b}}). \tag{3.3.1}$$

Next, we choose background as an example to introduce deformable sampling layer and foreground can be processed similarly. To fill in the pixel (x, y) in deformed compact background $\hat{\boldsymbol{I}}^{\mathrm{b}}$, we sample a pixel from original large image $\boldsymbol{I}$ with the sampling position x (resp., y) determined by function $\boldsymbol{u}^{\mathrm{b}}(\cdot,\cdot)$ (resp., $\boldsymbol{v}^{\mathrm{b}}(\cdot,\cdot)$). This process can be written as

$$\hat{\boldsymbol{I}}^{\mathrm{b}}(x, y) = \boldsymbol{I}[\boldsymbol{u}^{\mathrm{b}}(x, y), \boldsymbol{v}^{\mathrm{b}}(x, y)]. \tag{3.3.2}$$

Note that all coordinates are normalized within the range $[0, 1]$. Similar to [138],

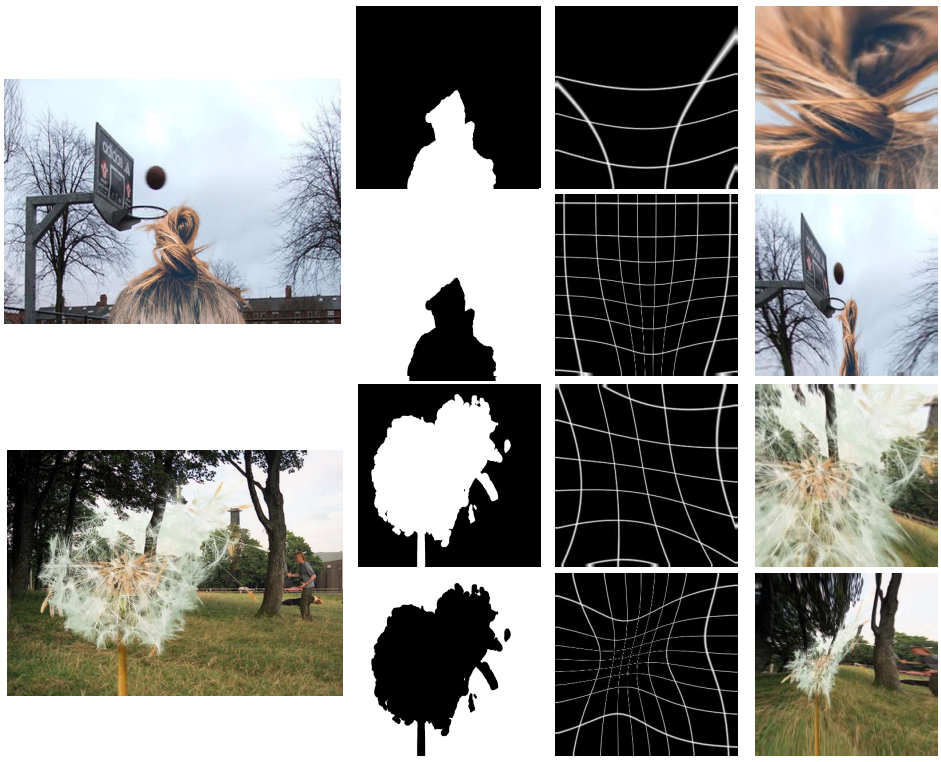
the functions $\boldsymbol{u}^{\mathrm{b}}(\cdot,\cdot)$ and $\boldsymbol{v}^{\mathrm{b}}(\cdot,\cdot)$ can be formulated as

$$\boldsymbol{u}^{\mathrm{b}}(x,y) = \frac{\sum\limits_{x',y'} \boldsymbol{m}^{\mathrm{b}}(x',y')k((x,y),(x',y'))x'}{\sum\limits_{x',y'} \boldsymbol{m}^{\mathrm{b}}(x',y')k((x,y),(x',y'))}, \tag{3.3.3}$$

$$\boldsymbol{v}^{\mathrm{b}}(x,y) = \frac{\sum\limits_{x',y'} \boldsymbol{m}^{\mathrm{b}}(x',y')k((x,y),(x',y'))y'}{\sum\limits_{x',y'} \boldsymbol{m}^{\mathrm{b}}(x',y')k((x,y),(x',y'))}, \tag{3.3.4}$$

in which $\boldsymbol{m}^{\mathrm{b}}(x',y')$ is the (x',y')-th element in $\boldsymbol{M}^{\mathrm{b}}$ and distance kernel $k((x,y),(x',y')) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{(\mathrm{x}-\mathrm{x}')^2+(\mathrm{y}-\mathrm{y}')^2}{2\sigma^2}\right\}$ with σ set as one third of the width of the background mask. Intuitively, for a target pixel (x,y) in deformed compact background $\hat{\boldsymbol{I}}^{\mathrm{b}}$, we sample a "near-target" and "near-background" pixel from original image $\boldsymbol{I}$. "Near-target" is realized by the distance kernel $k((x,y),(x',y'))$, which expects the sampled pixel to be spatially close to the target pixel to avoid the convergence of all sampled pixels. "Near-background" is realized by background mask $\boldsymbol{m}^{\mathrm{b}}(x',y')$, which encourages the sampled pixel to be from the background region. In this way, deformed compact background $\hat{\boldsymbol{I}}^{\mathrm{b}}$ is more densely sampled from the background region in $\boldsymbol{I}$, which can zoom and exaggerate the background region in the downsized image.

Another interpretation is that deformable sampling layer distorts the coordinate system with $\boldsymbol{m}^{\mathrm{b}}(x',y')$ and $k[(x,y),(x',y')]$. The foreground or background would be magnified when sampling based on the distorted coordinate system. Several examples are provided in Fig. 3.3-2 together with the distorted coordinate system. In the deformed grid, expanded cells correspond to the regions to magnify. By virtue of the deformable sampling layer, we can obtain deformed compact background and foreground, which can maintain the background and foreground information to the utmost while elastically preserving the image structure. As the CA layer, which will be mentioned in the next subsection, can automatically align the features with learned global information without spatial constrains, features do not need to be aligned.



(a) (b) (c) (d)

Fig. 3.3-2. Distorted coordinate system. (a) Original image. (b) Foreground and background binary mask based on trimap. (c) Visualization of distorted coordinate system. (d) Deformed condensed foreground and background image.

3.3.3 Unknown-related Contextual Information Aggregation

A convolutional neural network processes image features with local convolutional kernel layers, but such local operators are not effective to obtain features from distant spatial locations. Meanwhile, it has been proved in [116] that spatially distant pixels from global image can also contribute to the alpha matte estimation in a local region. To solve this problem, we enhance features in the unknown region of the crop by exploiting global contextual information with the CA layer. The core idea of the CA layer is to use foreground/background patches as convolutional filters to represent unknown region with relevant foreground/background features.

The CA layer is differentiable and thus can be integrated into an end-to-end network. Besides, the CA layer is fully-convolutional, and thus supports input crops of arbitrary shape. After passing the crop, deformed compact foreground, deformed compact background through the encoder, we can obtain the crop feature map, the foreground feature map, and the background feature map, respectively. In spirit to the sampling-based matting methods, we consider the problem as matching crop features in the unknown region with foreground and background patches, which can be achieved by the CA layer elegantly. In the following, we use background as an example to describe the CA layer and foreground can be handled in a similar way.

First, follow [131], we extract 3×3 patches from background feature map and apply them as convolutional filters to the crop features in the unknown region. With proper normalization, this convolutional operation is equivalent to computing the cosine similarity between the unknown patch and background patch:

$$\boldsymbol{s}^{\mathrm{b}}_{x,y,x',y'} = \cos(\boldsymbol{F}^{\mathrm{b}}_{x',y'}, \boldsymbol{F}^{\mathrm{u}}_{x,y}), \tag{3.3.5}$$

where $\boldsymbol{F}^{\mathrm{u}}_{x,y}$ is the 3×3 patch centered around pixel (x,y) in unknown region and $\boldsymbol{F}^{\mathrm{b}}_{x',y'}$ is the 3×3 patch centered around pixel (x',y') in the background. After applying the convolutional operation, we can obtain the similarity score map with the number of channels being the number of background patches. Then, we apply

channel-wise softmax to normalize the similarity scores for each pixel, leading to the attention score map $\boldsymbol{S}^{\mathrm{b}}$, in which each pixel-wise attention vector $\boldsymbol{s}^{\mathrm{b}}_{x,y}$ represents the attention scores over all background patches *w.r.t.* the unknown patch centered around pixel (x,y). We visualize attention score map based on a color codebook, as illustrated in Fig. 3.1-1. Specifically, we use color to indicate the relative location (e.g., green on top-right, pink on bottom-left) of the most relevant background patch (the largest attention score in $\boldsymbol{s}^{\mathrm{b}}_{x,y}$) for each pixel (x,y) in the unknown region (e.g., green on top-right, pink on bottom-left). For example, top right area of dog ear in the background attention map is green, which means that it obtains information from top-right of the deformed background (i.e., the sky).

Then, we apply deconvolutional operation with background patches as deconvolutional filters to attention score map $\boldsymbol{S}^{\mathrm{b}}$, to reconstruct unknown region. This is equivalent to representing each pixel in the unknown region with relevant background patches. We refer to such reconstructed features of unknown region as background-aware features. Similarly, we can obtain foreground-aware features by representing each pixel in unknown region with relevant foreground patches. Please refer to [131] for more details of the CA layer.

For the unknown region in the crop, we aggregate original crop features, foreground-aware features, and background-aware features by element-wise summation. For known region in the crop, we keep using the original crop features. Then, we input the aggregated crop feature map into the remaining network and obtain the alpha matte. To be specific, the above procedure of contextual aggregation is described in Algorithm 1

3.3.4 Training Losses

In this paper, we leverage three losses: alpha matte prediction loss, background reconstruction loss, and foreground reconstruction loss. Alpha matte prediction loss is the most commonly used loss in deep image matting task [123, 128–130, 139]. We only apply alpha matte prediction loss among other losses proposed in the deep image matting task according to [130]. We apply background recon-

Algorithm 1: Procedure of Contextual Aggregation.

Input : Original image I_{u}, deformed foreground/background image $\hat{I}_{\mathrm{d}}$, a set of pixels in unknown area $\boldsymbol{U}$, a hyper parameter for softmax λ

Output: Generated foreground/background image by contextual aggregation I_{out}

for each pixel (x,y) in $\boldsymbol{U}$ **do**
  $counter(x,y) \leftarrow 0$
  $I_{\mathrm{out}}(x,y) \leftarrow 0$
for each pixel (x,y) in $\boldsymbol{U}$ **do**
  // Set a 3×3 patch around (x,y)
  $P_{\mathrm{u}}(x,y) \leftarrow \{(x+i,y+j) | -1 \leq i \leq 1, -1 \leq j \leq 1\}$
  for each pixel (x',y') in the deformed foreground/background image **do**
    // Set a 3×3 patch around (x',y') as
    $P_{\mathrm{d}}(x',y') \leftarrow \{(x'+i,y'+j) | -1 \leq i \leq 1, -1 \leq j \leq 1\}$
    // Compute cosine similarity between the patches $P_{\mathrm{u}}(x,y)$ and $P_{\mathrm{d}}(x',y')$
    $s(x',y';x,y) \leftarrow \cos[P_{\mathrm{d}}(x',y'), P_{\mathrm{u}}(x,y); I_{\mathrm{u}}, \hat{I}_{\mathrm{d}}]$
  // Compute softmax over (x',y')
  $prob(x',y';x,y) \leftarrow \mathrm{softmax}_{x',y'}[\lambda s(x,y;x,y)]$
  // Update an image and counter
  for each pixel $(x+i,y+j)$ in patch $P_{\mathrm{u}}(x,y)$ **do**
    $I_{\mathrm{out}}(x+i,y+j) \leftarrow I_{\mathrm{out}}(x+i,y+j) + \sum_{x',y'} prob(x',y';x,y)\hat{I}_{\mathrm{d}}(x+i,y+j)$
    $counter(x+i,y+j) \leftarrow counter(x+i,y+j) + 1$
  for each pixel (x,y) in $\boldsymbol{U}$ **do**
    $I_{\mathrm{out}}(x,y) \leftarrow I_{\mathrm{out}}(x,y) / counter(x+i,y+j)$

struction loss and foreground reconstruction loss to obtain better background as well as more purified foreground, which contributes to better qualitative performance in composition.

The alpha matte prediction loss is summation of an absolute difference between the ground-truth alpha matte value α_i and the predicted alpha matte value $\hat{\alpha}_i$ over all pixels, which is defined as

$$L_\alpha = \sum_i |\hat{\alpha}_i - \alpha_i|. \tag{3.3.6}$$

The background reconstruction loss is summation of a weighted L_1 norm between predicted background $\hat{B}_i$ and ground-truth background B_i over all pixels, which is defined as

$$L_{\mathrm{bg}} = \sum_i w_{\mathrm{bg},i} |\hat{B}_i - B_i|, \tag{3.3.7}$$

where $w_{\mathrm{bg},i}$ is a weight at the i-th pixel. As one of the simplest ways, one may set the weight to a contribution ratio of a background in a composite image, i.e., $(1-\alpha)$ as evident from the basic matting equation (3.1.1). The weight gets, however, small in a foreground/background blending region, and hence we may fail in accurately estimating a purified background/foreground in that region. We therefore aim at increasing the weight near the foreground/background blending region. Specifically, we dilate the weight by using the eroded alpha $\alpha_i^{\mathrm{erode}}$ and set the resultant weight as $w_{\mathrm{bg},i} = 1 - \alpha_i^{\mathrm{erode}}$, which enables us to consequently extract more purified foreground as shown in Fig. 3.4-3. The erosion kernel size k is randomly chosen from a set $\{k | 1 \leq k \leq 30\}$.

The foreground reconstruction loss is also summation of a weighted L_1 norm between an estimated foreground $\hat{F}_i$ and ground-truth foreground F_i over all pixels as

$$L_{\mathrm{fg}} = \sum_i w_{\mathrm{fg},i} |\hat{F}_i - F_i|, \tag{3.3.8}$$

where $w_{\mathrm{fg},i}$ is a weight at the i-th pixel. We simply set $w_{\mathrm{fg},i} = \alpha$ unlike the background weight $w_{\mathrm{bg},i}$ uses the eroded alpha. This is because the ground-truth background is defined over an entire image, while the ground-truth foreground is valid

only for region where the alpha value is positive. The estimated foreground $\hat{F}_i$ is computed from an input image I_i, ground-truth alpha α_i, and predicted background $\hat{B}_i$ based on the matting composition equation in (3.1.1) as

$$\hat{F}_i = \frac{I_i - (1-\alpha_i)\hat{B}_i}{\max(\alpha_i, \varepsilon)}, \tag{3.3.9}$$

where ε is a small positive value to avoid zero division. We set ε as 0.003 in the experiment since the minimum non-zero value of the ground-truth alpha matte is larger than 0.003.

The overall loss of our method is finally defined as

$$L = L_\alpha + \lambda_l (L_{\text{fg}} + L_{\text{bg}}), \tag{3.3.10}$$

in which λ_l is set to 0.1 in our experiments.

3.4 Experiments

3.4.1 Datasets

We train and evaluate our model on the Composition-1k dataset [123], the largest public natural image matting dataset, and the alphamatting.com benchmark [140]. The Composition-1k dataset contains a training set of 431 unique foregrounds. The test set contains 1,000 images, generated by 50 unique foreground objects, each composite onto 20 background images. The alphamatting.com benchmark contains 8 test images, each of which has 3 trimaps: small, large and user.

3.4.2 Implementation Details

Following [129, 130], we randomly composite two foregrounds with probability of 0.5 respectively to generate a new foreground as well as its alpha matte. During training, we randomly use crops of different sizes 480×480, $512 \times$

512, 640×640 from the original images to learn more scale-robust model. The size of deformable downsampled foreground and background is fixed as 512×512.

Similar to [126, 128–130], we randomly select background images at each iteration and composite the training image on the fly, which enables the network to see completely new input image throughout the training process. Random flipping and rotation are applied to all the cropped and resized training images for variation. Also, we create a batch with a single foreground composited on 8 randomly selected backgrounds to enhance the robustness of alpha matte estimation to diverse background images. We train the model with 100,000 iterations. We warmup the learning rate to 10^{-4} in 6,000 iterations, and then apply cosine decay to the learning rate. We use Adam optimizer [105] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for optimization.

3.4.3 Evaluation Metrics

In this paper, four widely acknowledged metrics are used for evaluation: SAD (sum of absolution difference), MSE (mean square error), gradient, and connectivity. They are proposed in [140] to reflect the visual quality of the alpha matte. For all the four metrics, the lower value of the metrics indicate the better predicted alpha mattes. We report the average over all images in the test set.

3.4.4 Comparison with the State-of-the-art

We compare our method with the state-of-the-art baselines on both Composition-1k dataset and alphamatting.com benchmark. We report the comparison result with traditional matting methods [110, 116–119, 133, 141] and learning-based matting methods [123, 125–130, 139, 142, 143] in Table 3.4.1. Note that we did not report the comparison result with [144], since it leverages class information as foreground semantics, which is unavailable in public datasets. From Table 3.4.1, we can observe that learning-based methods are significantly better than traditional methods. Among learning-based methods, our method achieves

Table 3.4.1 Quantitative Results of Composition-1k against the State-Of-The-Art Methods

Methods	SAD ↓	MSE ↓	Gradient ↓	Connectivity ↓
Shared Matting	128.9	0.091	126.5	135.3
Learning-based Matting	113.9	0.048	91.6	122.2
Comprehensive Sampling	143.8	0.071	102.2	142.7
Global Matting	133.6	0.068	97.6	133.3
Closed-form Matting	168.1	0.091	126.9	167.9
KNN Matting	175.4	0.103	124.1	176.4
DCNN Matting	161.4	0.087	115.1	161.9
Deep Image Matting	54.6	0.017	36.7	55.3
Fusion Matting	49.0	0.020	34.3	50.6
IndexNet Matting	45.8	0.013	25.9	43.7
Ada Matting	41.7	0.010	16.8	-
Samplenet	40.35	0.0099	-	-
ATNet	40.5	0.013	21.5	39.4
GCA Matting	35.28	0.0091	16.92	32.53
Context-aware Matting	35.8	0.0082	17.3	33.2
HDMatt	33.5	0.0073	14.5	29.9
A2U Matting	32.15	0.0082	16.39	29.25
Ours	**30.72**	**0.0070**	**13.59**	**29.73**

Note: ↓ means the lower is the better.

the best results on Composition-1k dataset, which demonstrates the effectiveness of seeking for unknown-related global contextual information from the whole image.

We report SAD, MSE, gradient and connectivity of different methods on the alphamatting.com benchmark in Tables 3.4.2 to 3.4.5. We compare the proposed method with top six deep learning-based methods including our baseline method, i.e., Natural Image Matting via Guided Contextual Attention (GCA Matting). As a result, we observe that our method achieves competitive results among the state-of-the-art matting methods, especially in SAD and MSE metric.

3.4.5 Foreground Extraction

Extracting a purified foreground from an input composite image is one of the main applications of matting (e.g. video production, graphics, and consumer ap-

Table 3.4.2 SAD Comparison with Top 6 State-Of-The-Art Baselines on the Alphamatting.com Online.

SAD	overall	AVG			Troll			Doll			Donkey			Elephant			Plant			Pineapple			Plastic bag			Net		
		S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U
Ours	**10.6**	13.9	**8.6**	**9.3**	9	9.6	10.2	4.9	**4.7**	5.6	**2.9**	3.1	2.7	1.1	1.1	**1.3**	5.9	6.2	**7.2**	2.8	2.9	3.3	**16.3**	**17**	**15.9**	19.9	20.9	22.8
HDMatt [139]	11	13	9.8	10.1	9.5	10	10.7	4.7	4.8	5.8	2.9	**3**	**2.6**	1.1	1.2	**1.3**	5.2	**5.9**	6.7	2.4	**2.6**	**3.1**	17.3	17.3	17	21.5	22.4	23.2
AdaMatting [126]	13.5	12	12.5	16	10.2	11.1	10.8	4.9	5.4	6.6	3.6	3.4	3.4	**0.9**	**0.9**	1.8	**4.7**	6.8	9.3	**2.2**	**2.6**	3.3	19.2	19.8	18.7	**17.8**	**19.1**	**18.6**
A2U Matting [143]	13.6	12.8	10.5	17.5	9.3	9.7	10.9	4.8	4.9	5.3	3	3.1	2.8	1	1.1	1.4	5.1	6.7	8.5	2.5	3	5.9	17.3	18.4	18.1	20.6	20.6	27.3
SampleNet [129]	14	**11.4**	14.1	16.5	9.1	9.7	**9.8**	**4.3**	4.8	**5.1**	3.4	3.7	3.2	**0.9**	1.1	2	5.1	6.8	9.7	2.5	4	3.7	18.6	19.3	19.1	20	21.6	23.2
GCA Matting [130]	15.3	16.4	12.6	17	**8.8**	**9.5**	11.1	4.9	4.8	5.8	3.4	3.7	3.2	1.1	1.2	**1.3**	5.7	6.9	7.6	2.8	3.1	4.5	18.3	19.2	18.5	20.8	21.7	24.7

Note: S, L, U stand for small, large, user trimap respectively.The lowest errors are in bold.

Table 3.4.3 MSE Comparison with Top 6 State-Of-The-Art Baselines on the Alphamatting.com Online.

MSE	overall	AVG			Troll			Doll			Donkey			Elephant			Plant			Pineapple			Plastic bag			Net		
		S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U
Ours	**11**	14.8	10.4	**8**	**0.3**	**0.3**	0.4	0.2	**0.2**	0.3	**0.1**	**0.1**	**0.1**	**0**	**0**	**0**	0.5	0.5	**0.6**	0.2	**0.2**	**0.2**	**0.9**	**0.9**	**0.9**	0.8	0.8	0.8
HDMatt [139]	11.2	13.9	**9.8**	10	**0.3**	**0.3**	0.4	0.2	**0.2**	0.3	**0.1**	**0.1**	**0.1**	**0**	**0**	**0**	**0.4**	**0.4**	0.6	**0.1**	**0.2**	**0.2**	**0.9**	**0.9**	**0.9**	0.8	0.8	0.8
AdaMatting [126]	14.2	11.5	13.1	18	**0.3**	0.4	0.4	0.2	**0.2**	0.3	0.2	0.2	0.2	**0**	**0**	0.1	**0.4**	0.6	1	**0.1**	**0.2**	0.3	1.1	1.2	1.1	**0.6**	**0.6**	**0.6**
SampleNet [129]	14.6	**10.6**	15.1	18	**0.3**	**0.3**	**0.3**	**0.1**	**0.2**	**0.2**	0.2	0.2	0.2	**0**	**0**	0.1	**0.4**	0.6	1.2	**0.1**	0.3	0.3	1.1	1.1	1.2	0.7	0.8	0.8
A2U Matting [143]	16.1	13.8	12.6	21.9	**0.3**	**0.3**	0.4	0.2	**0.2**	0.3	0.2	0.2	**0.1**	**0**	**0**	0.1	**0.4**	0.6	0.8	0.2	**0.2**	0.8	1	1	1	0.7	0.7	1.2
GCA Matting [130]	16.3	16.3	14.9	17.6	**0.3**	**0.3**	0.4	0.2	**0.2**	0.3	0.2	0.2	0.2	**0**	**0**	0.1	0.5	0.6	0.8	0.2	**0.2**	0.5	1	1.1	1.1	0.7	0.8	0.9

Note: S, L, U stand for small, large, user trimap respectively.The lowest errors are in bold.

Table 3.4.4 Gradient Comparison with Top 6 State-Of-The-Art Baselines on the Alphamatting.com Online.

Gradient	overall	AVG			Troll			Doll			Donkey			Elephant			Plant			Pineapple			Plastic bag			Net		
		S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U
HDMatt [139]	**9**	9.8	**7.4**	**9.9**	0.2	0.2	**0.2**	**0.1**	**0.1**	0.3	**0.1**	**0.1**	**0.2**	0.2	0.2	**0.3**	**1.1**	**1.2**	1.6	0.6	**0.6**	**0.9**	**0.5**	**0.5**	**0.6**	**0.3**	0.4	**0.4**
Ours	11.9	14.3	10.6	10.8	0.2	**0.1**	**0.2**	0.2	**0.1**	0.3	**0.1**	0.2	**0.2**	0.2	0.2	**0.3**	1.3	1.4	**1.5**	0.7	0.8	**0.9**	0.6	0.8	**0.6**	0.4	0.4	0.5
A2U Matting [143]	12.3	11.4	8.5	16.9	0.2	0.2	**0.2**	**0.1**	**0.1**	**0.2**	**0.1**	0.2	**0.2**	0.2	0.2	0.4	**1.1**	1.3	1.9	0.6	0.7	1.7	0.6	0.6	**0.6**	**0.3**	**0.3**	**0.4**
AdaMatting [126]	13.7	**9.6**	11.4	20	0.2	0.2	**0.2**	**0.1**	**0.1**	0.4	0.2	0.2	**0.2**	**0.1**	**0.1**	**0.3**	**1.1**	1.4	2.3	**0.4**	**0.6**	**0.9**	0.9	1	0.9	**0.3**	0.4	**0.4**
GCA Matting [130]	14.1	14.1	12.6	15.6	**0.1**	**0.1**	**0.2**	**0.1**	**0.1**	0.3	0.2	0.2	**0.2**	0.2	0.2	**0.3**	1.3	1.6	1.9	0.7	0.8	1.4	0.6	0.7	**0.6**	0.4	0.4	**0.4**
SampleNet [129]	15.5	10.9	13.4	22.3	**0.1**	**0.1**	**0.2**	**0.1**	**0.1**	**0.2**	0.2	0.3	0.3	**0.1**	0.2	0.5	**1.1**	1.5	2.7	0.6	0.9	1	0.8	0.9	0.9	0.4	0.4	**0.4**

Note: S, L, U stand for small, large, user trimap respectively.The lowest errors are in bold.

Table 3.4.5 Connectivity Comparison with Top 6 State-Of-The-Art Baselines on the Alphamatting.com Online.

Connectivity	overall	AVG			Troll			Doll			Donkey			Elephant			Plant			Pineapple			Plastic bag			Net		
		S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U	S	L	U
GCA Matting [130]	**23.1**	26.4	**20.9**	**22**	1.1	1.1	1	0.2	0.2	0.2	**0.2**	**0.2**	**0.2**	**0**	**0**	**0**	**0.1**	**0.1**	**0.1**	0.1	0.1	0.1	1.1	1.3	1.3	1.9	1.5	1.6
AdaMatting [126]	23.6	**21**	26.1	23.8	1.1	1.1	1.1	**0.1**	0.2	0.2	**0.2**	**0.2**	**0.2**	**0**	**0**	**0**	**0.1**	**0.1**	**0.1**	**0**	**0**	0.1	6.8	13.3	1.4	**1.3**	**1.3**	**1.3**
SampleNet [129]	27.1	28.7	24.3	28.5	0.9	0.9	**0.8**	**0.1**	**0.1**	**0.1**	**0.2**	**0.2**	**0.2**	**0**	**0**	**0**	**0.1**	**0.1**	0.2	**0**	0.1	0.2	1.5	1.5	1.8	3.8	3.9	3.8
A2U Matting [143]	28	30.7	28.7	24.9	**0.8**	**0.8**	**0.8**	0.2	0.2	**0.1**	**0.2**	**0.2**	**0.2**	**0**	**0**	**0**	**0.1**	0.2	0.3	0.1	0.1	0.1	1	**0.9**	**1.1**	4.8	4.6	4.5
HDMatt [139]	31.7	36.7	28	30.3	1.5	1.3	1.3	0.3	0.3	0.3	0.3	0.3	0.3	**0**	**0**	**0**	**0.1**	**0.1**	**0.1**	**0**	**0**	**0**	**0.9**	**0.9**	1.2	2.4	2.2	2.3
Ours	35.2	40	29.5	36.9	1.7	1.7	1.6	0.2	0.2	0.2	0.3	0.3	0.3	**0**	**0**	**0**	**0.1**	0.2	0.2	0.1	0.1	0.1	1.6	1.7	1.8	3.2	2.9	2.7

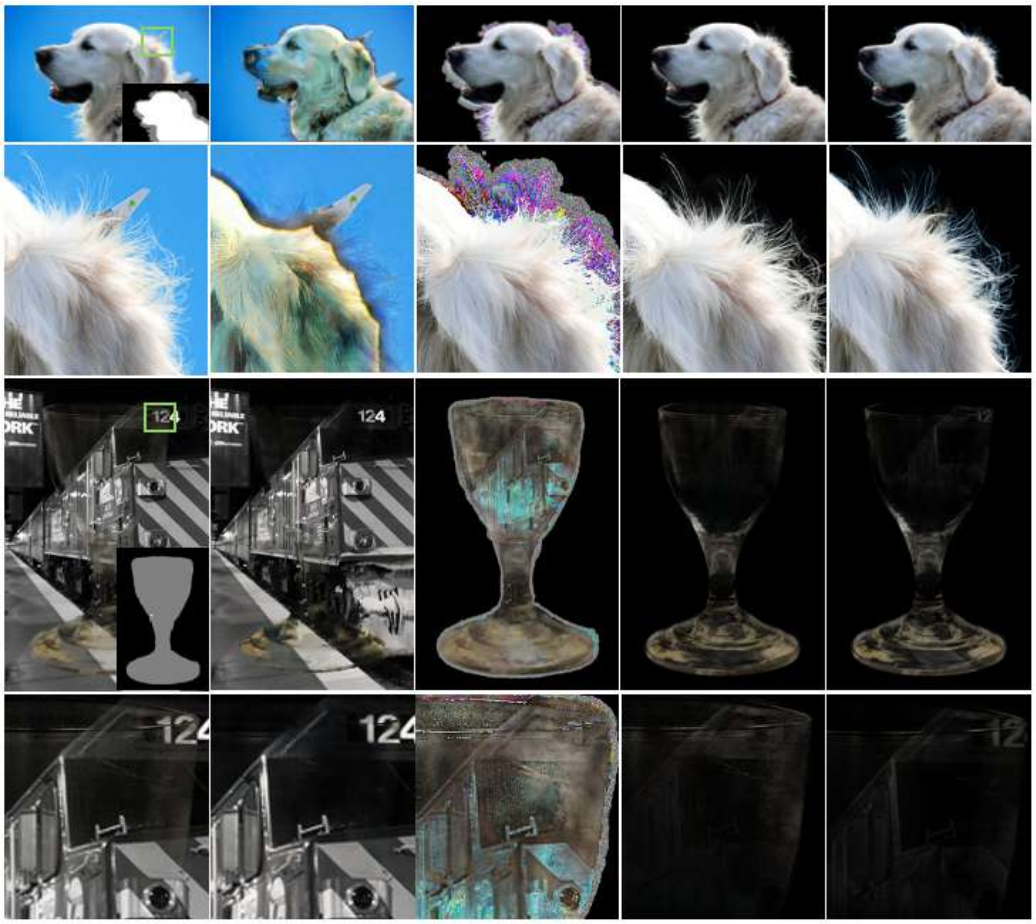
Note: S, L, U stand for small, large, user trimap respectively.The lowest errors are in bold.

plication). Therefore, we conduct foreground extraction application to compare purified foreground with traditional foreground in Fig. 3.4-3 to evaluate the effectiveness of background prediction. In the traditional foreground, we extract foreground by multiplying predicted alpha matte with the input composite image. In the purified foreground, we first compute foreground from matting equation (3.1.1) by the image, predicted alpha matte and predicted background as computed foreground, then multiply alpha matte with the computed foreground as the purified foreground (i.e., more purified foreground than the input composite image itself). We can tell from the result that background prediction can successfully filter background scene to achieve better visualization performance, i.e., the blue-ish sky near the of dog's fur in the first example will not be included with background prediction in purified foreground.

3.4.6 Qualitative Analyses

We visualize the results of different methods on both Composition-1k test set and alphamatting.com benchmark. Since the official code of Deep Image Matting [123] has not yet been released by the original authors, we choose the most frequently used PyTorch implementation released by the third party①. For the accuracy difference, we modify the last convolution layer of the encoder part according to another third party implementation② and confirm that we obtain the same accuracy as the original paper's one. For GCA Matting [130], IndexNet Matting [128], and ContextNet Matting [127], we generate visualized results using their released official codes and show them in Fig. 3.4-4. Deep image matting [123] is one of the most significant studies in learning-based image matting methods. GCA Matting [130], IndexNet Matting [128] ,and ContextNet Matting [127] are recently proposed state-of-the-art methods. GCA matting and IndexNet matting perform better than DIM especially in semi-opaque images such as "Jellyfish" in Fig. 3.4-4. However, due to the local operator in the architecture, it cannot obtain global information from the whole image, which leads to noisy prediction in tiny area such as "Hair" in Fig. 3.4-4. ContextNet considers global context information, there-

① https://github.com/foamliu/Deep-Image-Matting-PyTorch, Dec. 2021
② https://github.com/huochaitiantang/pytorch-deep-image-matting, Dec. 2021



(a) (b) (c) (d) (e)

Fig. 3.4-3. Effectiveness of background prediction. (a) Image and corresponding trimap. (b) Predicted background. (c) Computed foreground. (d) Purified foreground. (e) Traditional foreground

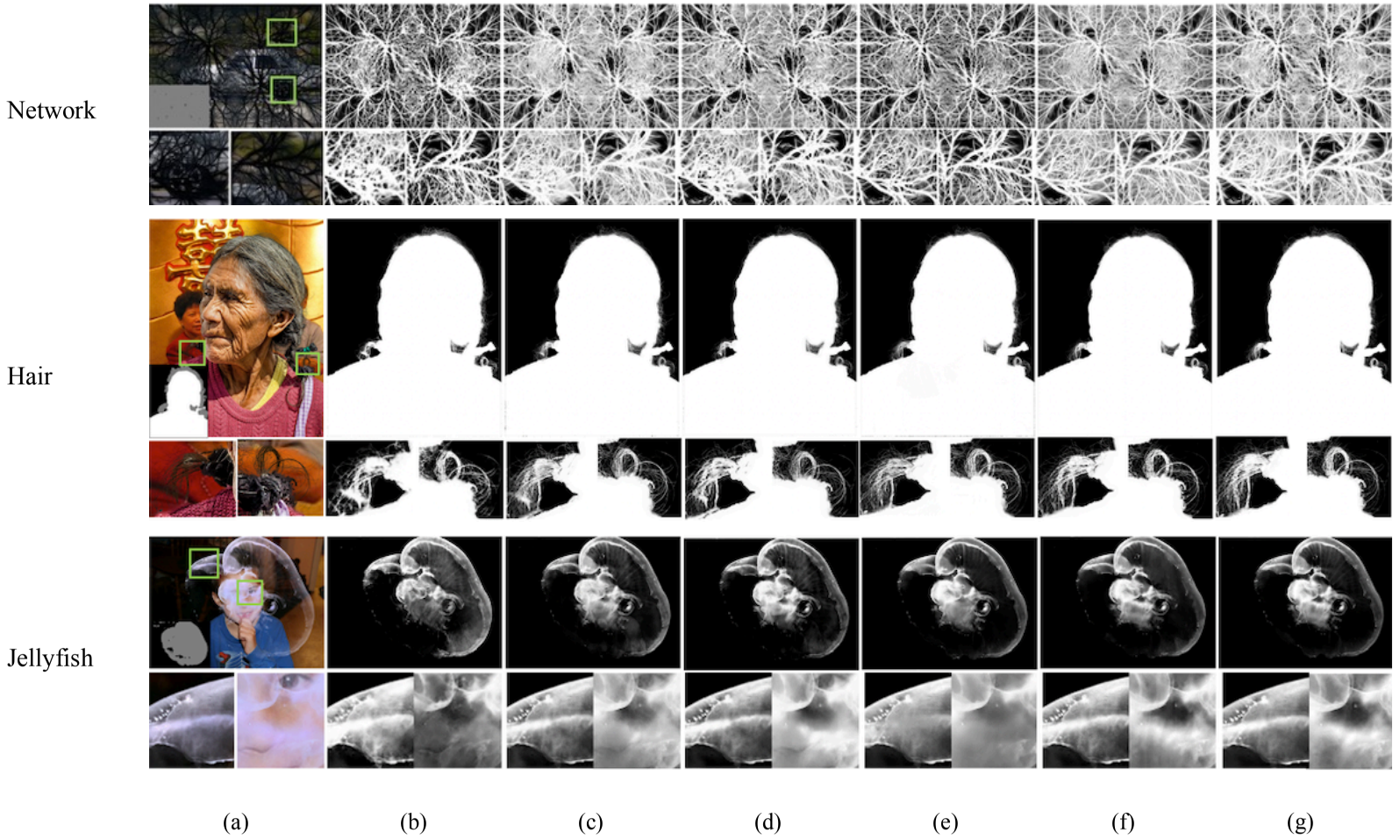


Fig. 3.4-4. The visualization comparison against the state-of-the-art methods on the Composition-1k dataset. (a) Image and corresponding trimap. (b) Deep Image Matting [123]. (c) GCA Matting [130]. (d) IndexNet Matting [128]. (e) ContextNet Matting [127]. (f) Our approach. (g) Ground Truth. We also zoom in some details (green box) for comparison.

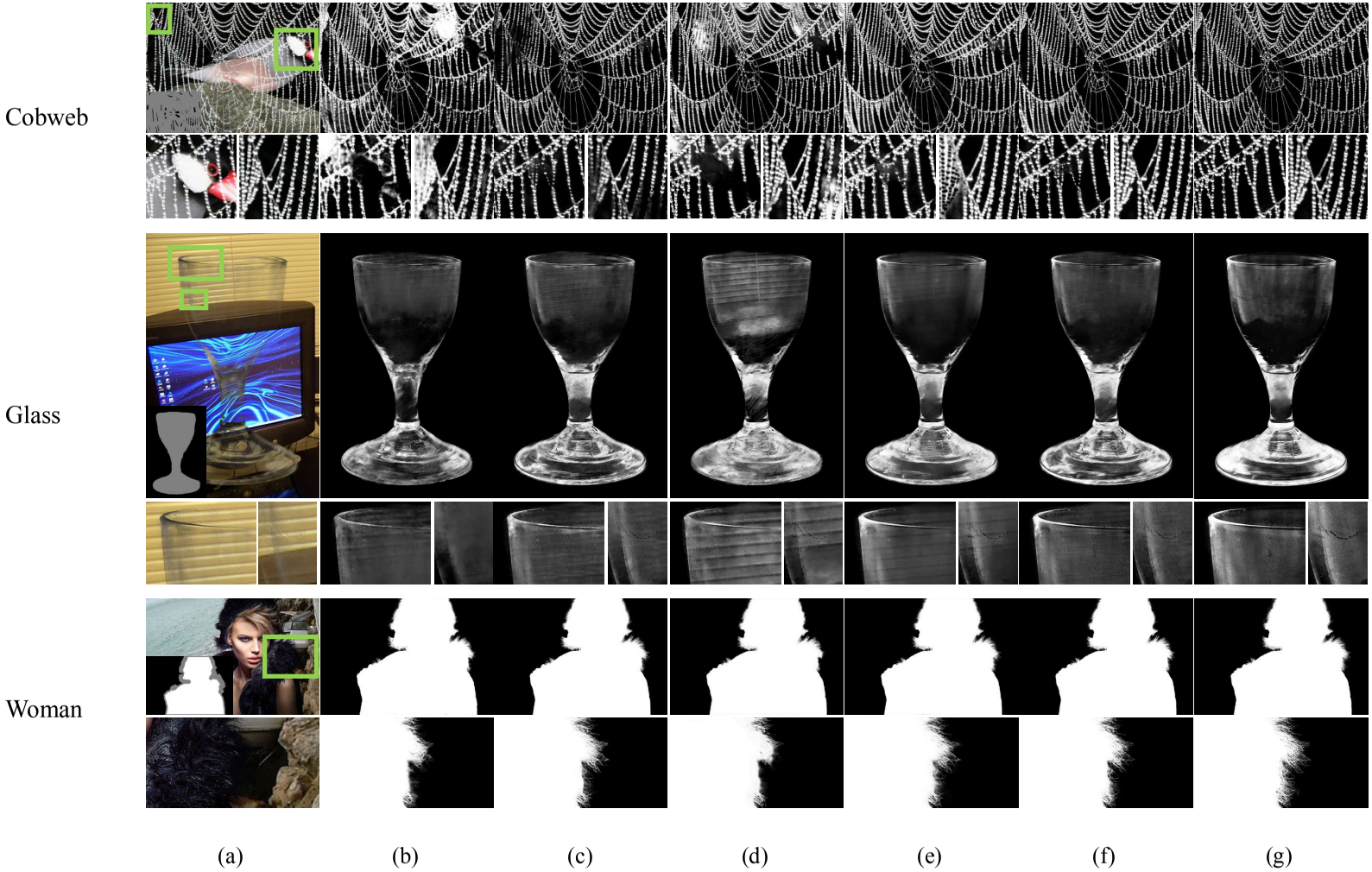


Fig. 3.4-5. The visualization comparison against the state-of-the-art methods on the Composition-1k dataset. (a) Image and corresponding trimap. (b) Deep Image Matting [123]. (c) GCA Matting [130]. (d) IndexNet Matting [128]. (e) ContextNet Matting [127]. (f) Our approach. (g) Ground Truth. We also zoom in some details (green box) for comparison.

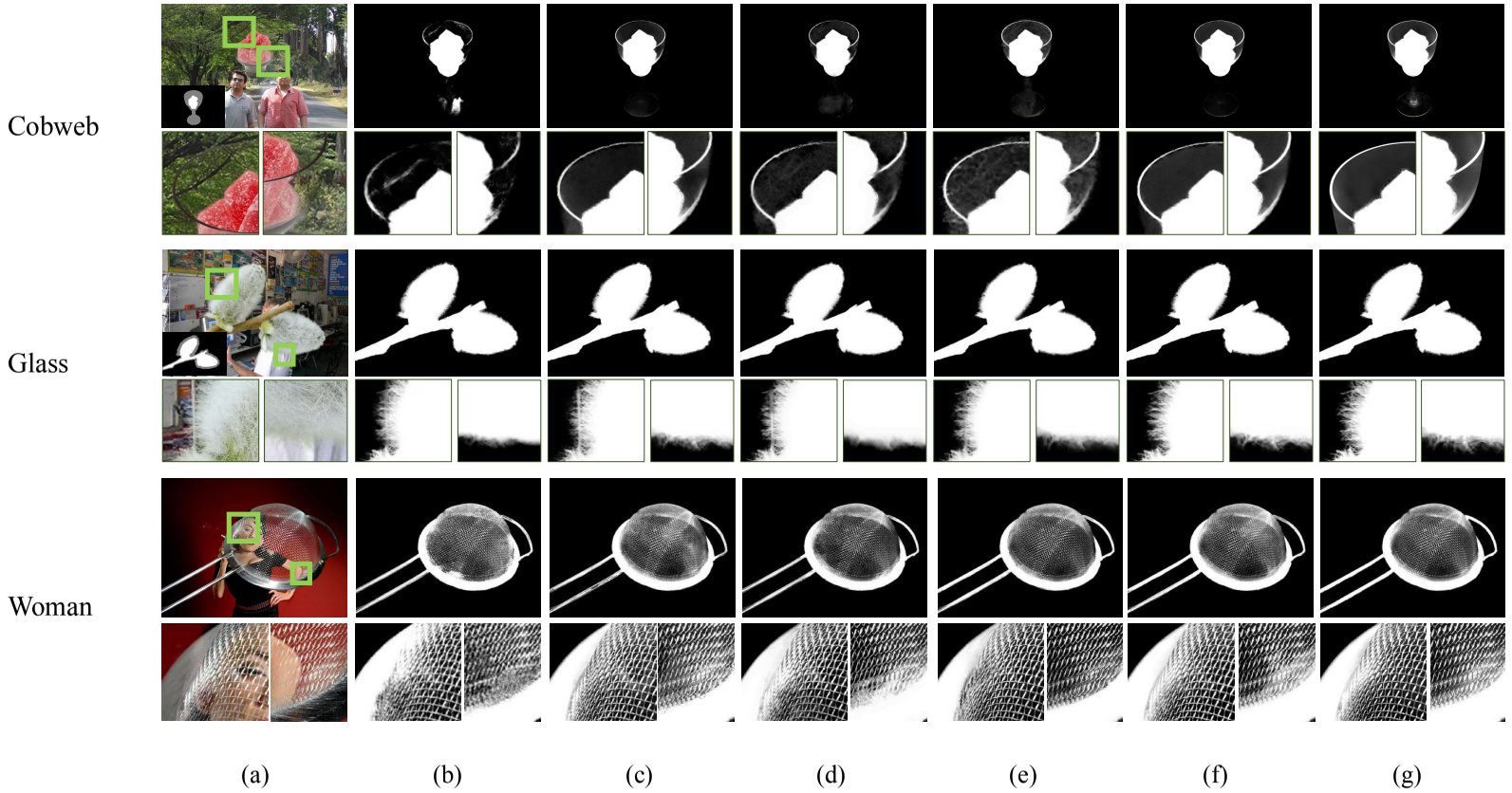


Fig. 3.4-6. The visualization comparison against the state-of-the-art methods on the Composition-1k dataset. (a) Image and corresponding trimap. (b) Deep Image Matting [123]. (c) GCA Matting [130]. (d) IndexNet Matting [128]. (e) ContextNet Matting [127]. (f) Our approach. (g) Ground Truth. We also zoom in some details (green box) for comparison.

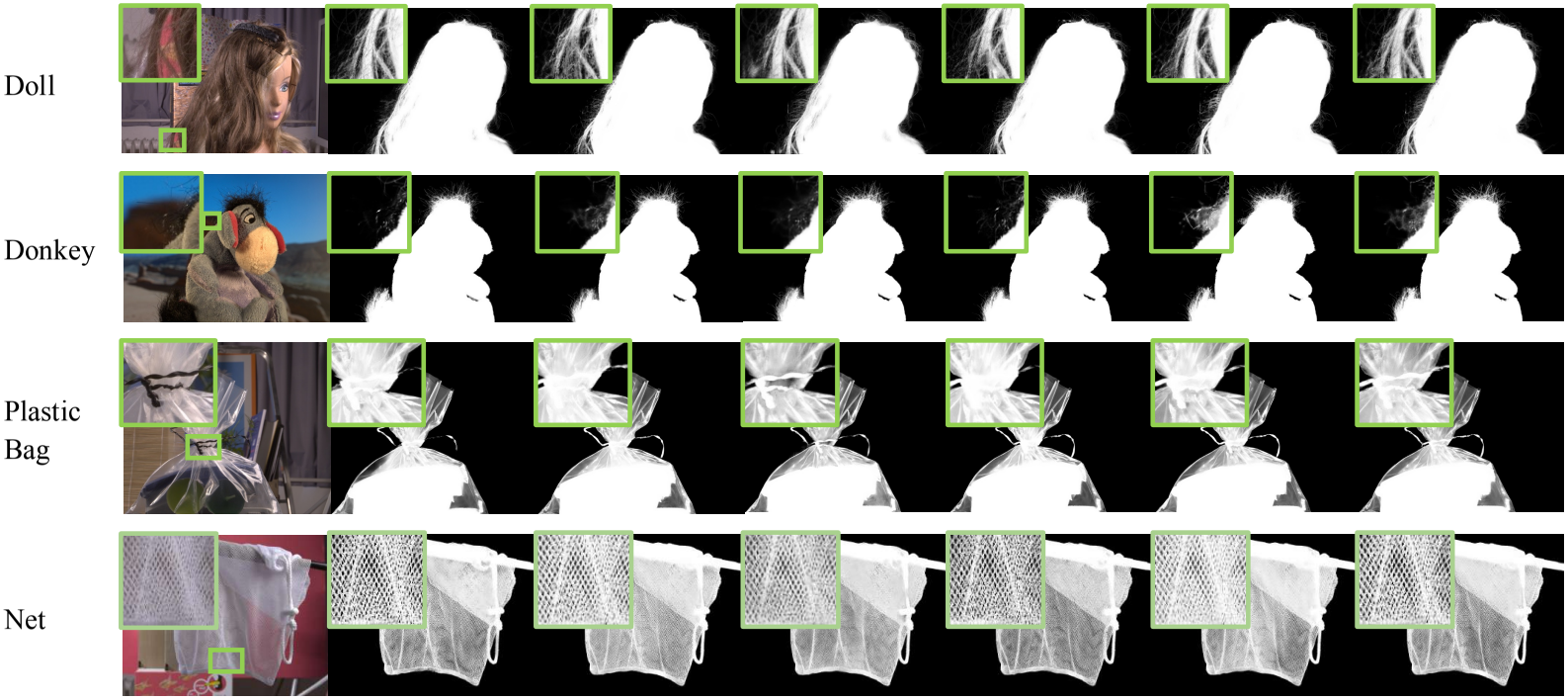


Fig. 3.4-7. The visualization comparison against the state-of-the-art methods on alphamatting.com benchmark. (a) Image and corresponding trimap. (b) SampleNet Matting [129]. (c) GCA Matting [130]. (d) Ada Matting [126]. (e) A2U Matting [143]. (f) HDMatt [139]. (g) Our approach. We also zoom in some details (green box) for comparison.

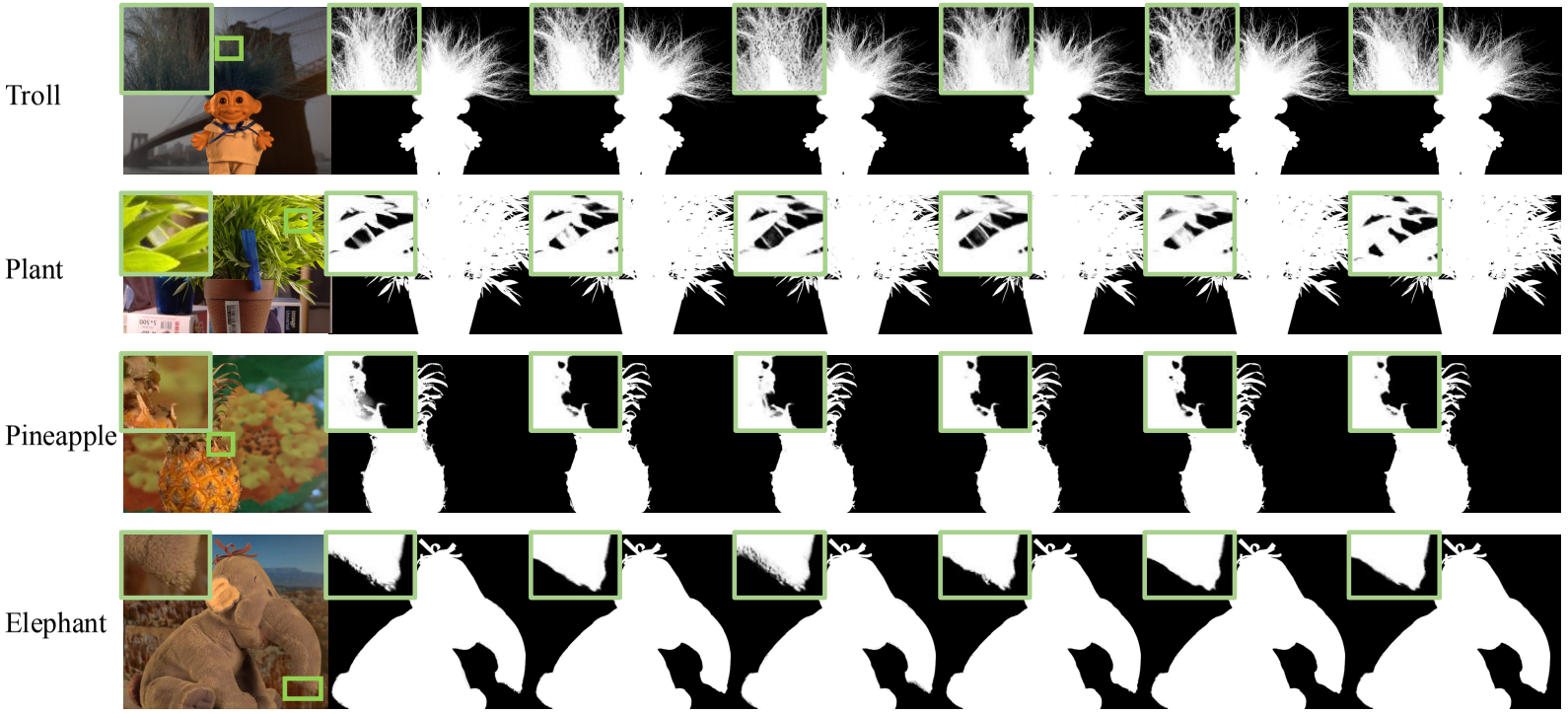


Fig. 3.4-8. The visualization comparison against the state-of-the-art methods on alphamatting.com benchmark. (a) Image and corresponding trimap. (b) SampleNet Matting [129]. (c) GCA Matting [130]. (d) Ada Matting [126]. (e) A2U Matting [143]. (f) HDMatt [139]. (g) Our approach. We also zoom in some details (green box) for comparison.

fore performs better in fine structures such as “Network” and “Hair”. However, it cannot separate the boy’s face in the background from semi-opaque jellyfish in the foreground. Our method considers condensed global context in the whole image and predicts alpha matte and background simultaneously, which combines the advantage of both traditional methods and learning-based methods, and separate background object from foregrounds, yielding the closest qualitative results compared with the ground-truth.

In Fig. 3.4-8, we demonstrate some qualitative comparisons of our method and the top-performance deep learning-based matting methods on the alphamatting.com benchmark. We can observe that our model is good at handling the images with fine structures (e.g., doll), similar foreground and background color (e.g., donkey’s hair), and highly transparent object (e.g., plastic bag), which have been puzzling the image matting community for a long time.

3.4.7 Ablation Study

We conduct ablation studies on each module and show the results in Table 3.4.6 and list results of GCA Matting as the baseline. First, to evaluate the effectiveness of the condensed foreground that is generated by the deformable sampling layer, we only use deformed foreground as input for foreground encoder branch without using background encoder branch, which is referred to as “+ condensed f”. We also evaluate the effectiveness of condensed background in the same way as “+ condensed b”. The results show that either condensed foreground information or condensed background information can boost the matting performance.

In addition, considering that uniform downsampling can also preserve the global contextual information of the whole large image and fit the input image size of network, we uniformly downsample the foreground (resp., background) image (the complementary background (resp., foreground) region to 512×512 as the input for foreground (resp., background) encoder branch, which is referred to as “+ downsampled f, g”. The obtained results are much worse than those of our full method, which demonstrates that the deformable sampling layer can preserve more information than uniform downsampling. We also uniformly downsample

the whole image to 512×512 as input for an extra branch other than the crop branch, which is referred to as "+ downsampled whole". The obtained results are worse than "+ downsampled f, g", which indicates the advantage of using two separate foreground and background branches. Finally, our full method outperforms all the above special cases and significantly improves our baseline GCA Matting.

Table 3.4.6 Ablation Study of Different Modules

Methods	SAD ↓	MSE ↓	Gradient ↓	Connectivity ↓
Baseline	35.28	0.0091	16.92	32.53
+ condensed f	33.09	0.0072	15.44	29.41
+ condensed b	32.45	0.0074	14.99	29.45
+ downsampled whole	35.10	0.0082	16.78	30.92
+ downsampled f,g	33.73	0.0079	15.96	30.74
+ masked f, b	31.14	**0.0067**	14.65	30.27
Ours w/o. CA	37.23	0.0090	17.37	33.05
Ours	**30.72**	0.0070	**13.59**	**29.73**

Note: f and b represent foreground and background respectively. ↓ means the lower is the better.

Moreover, we mask out regions that do not belong to foreground/background in condensed background/foreground, which is referred to as "+ masked f, b" in Fig. 3.4-9. The obtained results are comparable to those of our full method, which indicates masked regions will not invalidate the performance of our purposed method. We also conduct additional ablation study on the proposed method without the CA module to validate its effectiveness. Specifically, we remove the CA layer and instead add deformed foreground/background feature map and crop feature map in an element-wise way, which is referred to as "Ours w/o. CA". The obtained results are worse than those of our full method and baseline. This might because without CA layer, the foreground feature map, background feature map, and crop feature map are spatially misaligned, thus having negative impact to the performance.

Fig. 3.4-9. Example image of experiment conducted on masked condensed deformed FG/BG. Regions that do not belong to FG/BG in the deformed image are masked out as black.

3.4.8 Visualization of Our Method on Natural Images

We evaluate our model on natural images gathered from Internet with user defined trimap in Fig. 3.4-10. We visualize the predicted alpha mattes, composite on the white background, composite on a different background, distorted coordinate system in grid, deformed condensed foreground and background obtained in the deformable sampling stage, and attention score map with the color of each pixel in unknown region indicating the knowledge learned through unknown-related contextual information aggregation stage.

The attention score maps in the last column indicate the relative direction (e.g., green for top-right, pink for bottom-left) of most interested foreground/background patch for each pixel in unknown area according to the color codebook at the bottom-right corner. For example, we picked up a certain pixel in the model's widespread hair in the unknown area in Fig. 3.4-11, and illustrated the corresponding displacement vector (purple in pseudo color). As a result, we can see that the certain pixel selected the lower-left hair pixel in the definite foreground.

Also, as shown in Fig. 3.4-10, the visualized color maps imply that global contextual information exists in both foreground and background, which can also be confirmed in Table 3.4.6 that using both condensed foreground and background can achieve the best matting performance among other special cases.

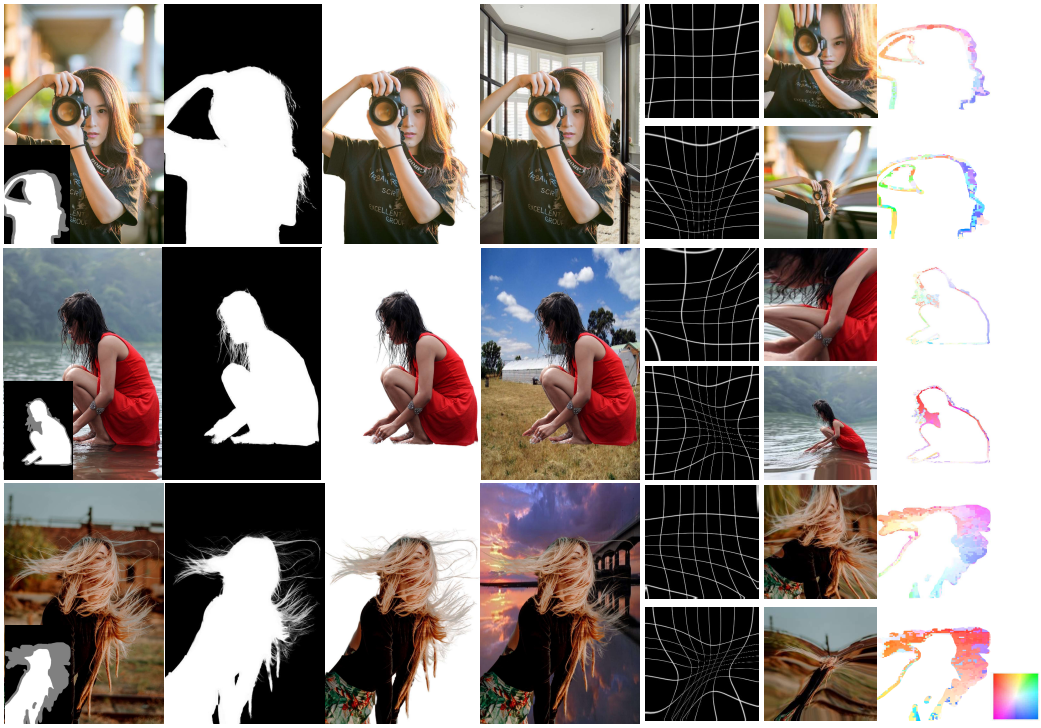
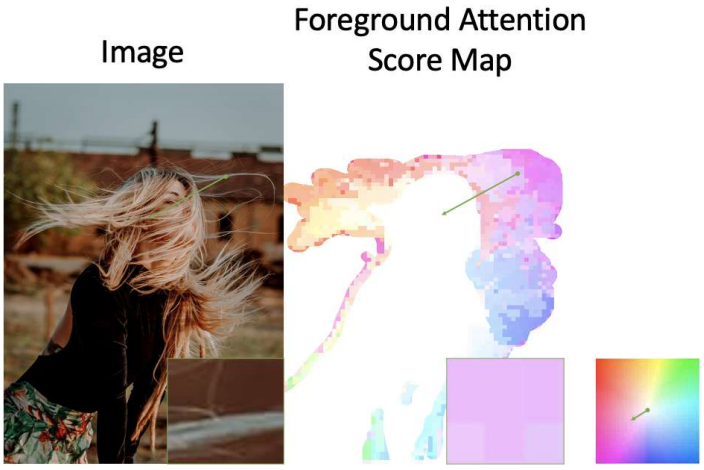


Fig. 3.4-10. Qualitative results of natural image from Internet with user defined trimap input. (a) Image and corresponding trimap. (b) Predicted alpha mattes. (c) Composite on white background. (d) Composite on a different background. (e) Distorted coordinate system. (f) Deformed condensed foreground and background. (g) Visualization of attention score map based on color codebook at the bottom-right.

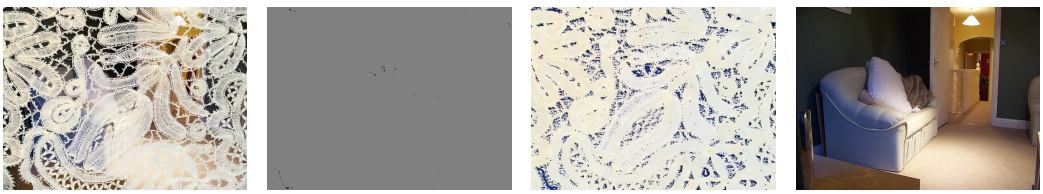


(a) (b)

Fig. 3.4-11. A detailed example of visualization. (a) Image. (b) Foreground Attention Score Map.

3.4.9 Failure Mode Analysis

We list a typical false example in Fig. 3.4-12, where background is small and scattered across the image. In such extreme cases, background deformable sampling cannot generate deformed condensed background. We will solve it by tackling these special issue separately in future work.



(a) (b) (c) (d)

Fig. 3.4-12. False example. (a) Image. (b) Trimap. (c) Foreground. (d) Background.

3.5 Summary

We proposed an end-to-end natural image matting network, which tends to seek unknown-relevant global contextual information from a large input image

without degrading the image quality. Specifically, we proposed a matting-oriented contextual aggregation supported by deformed sampling. As image inpainting-oriented contextual aggregation cannot consider a situation of no definite foreground/background region in image matting, our method tackles such situation by making use of all the pixels in the deformed foreground/background where foreground/background pixels are dominant. Although a small portion of background/foreground pixels is included in the deformed foreground/background image, we proved that it does not significantly affect the result. Besides, our method can estimate alpha matte and background simultaneously while keeping the matting equation, which can improve the foreground extraction performance qualitatively. Experiments on the Composition-1k dataset and alphamatting.com benchmark demonstrated the effectiveness of our method.

Chapter IV
Discussion

In the previous chapters, we proposed the task of cross-age gait video translation for the first time. The task is to generate age-progressed/regressed gait videos while preserving the input subject's identity as well, which can be employed in applications such as time-lapsed gait recognition and might shed light on cognitive reasoning to disentangle the age pattern from the silhouette.

We then studied two necessary aspects to realize the proposed cross-age gait video translation task. One is an effective gait video translator across ages, where the spatio-temporally augmented gait representation is the main concern. The other one is high fidelity input for the cross-age gait video translation. Since appearance-based gait analysis approaches usually use silhouette or silhouette-based template as input, the silhouette quality plays an important role in the proposed cross-age gait video translation task. However, we argue that the existing silhouette extraction with segmentation masks is not suitable for the proposed cross-age gait video translation due to the limitations in spatial and temporal resolution. We therefore conducted the second study to obtain a spatio-temporally augmented high fidelity input with matting framework.

In this chapter, we provided a discussion on how the high fidelity input may influence the gait video translator across ages. We first designed a scheme to automatically estimate the trimap for the proposed matting method, so that the matting framework introduced in the second study can be adopted without user interaction. Specifically, we adopted a inpainting method to predict the background and finetuned the proposed matting method on OULP-Age. We then provided thorough experiments on the largest gait database with age information, OULP-Age,

to reveal how the input quality of silhouette affect the performance of age progression/regression task on age group classification and cross-age gait recognition.

4.1 Finetune the Matting Model on OULP Dataset

Alpha matte represents the opacity of the foreground, which is expected to capture fine-grained details in spatial aspect and motion pattern in temporal aspect. To obtain a more accurate alpha matte, we finetune the natural image matting with attended global context framework proposed in the second study on the largest gait database with age information OULP-Age [100].

Since the proposed matting framework requires ground-truth background, which is not available in OULP-Age, we regard the binary mask area as the missing area and apply image matting technique to fill in the holes after removing the subject in RGB sequences. We adopt the state-of-the-art inpainting method Edge Connect [145] to generate a coarse background in Fig. 4.1-2 (b). Edge connect model first hallucinate edges in the missing regions as a priori with an edge generator, and an image completion network combines the predicted edges with color and texture information of the rest of the image to fill in the missing regions. Results in Fig. 4.1-2 (b) shown that areas apart from edge have visually consistent pixel intensities, however, the transition near the edge area is not very smooth. Therefore, we refine the coarse inpainting result with edge map generated by canny operation in Fig. 4.1-2 (c).

To smoothen the boundary area of the coarse matting, we designed a diffusion process of weight attenuation by edge (see Fig. 4.1-1). Specifically, to update the weight of pixel of interest (POI)-centered patch, we first define a neighbor patch of size k for the POI. The diffusion weight of neighbor patch is initialized as edge map in Fig. 4.1-2 (c) (i.e., edge pixels in the corresponding edge map has high initial weight of 1, whereas non-edge area has low initial weight of 0). POI is regarded as the start node of the diffusion process. In each diffusion step, the diffusion algorithm calculates a new weight to the pixels that are one step to the current node by equation 4.1.1. The diffusion weight of the pixels will be updated

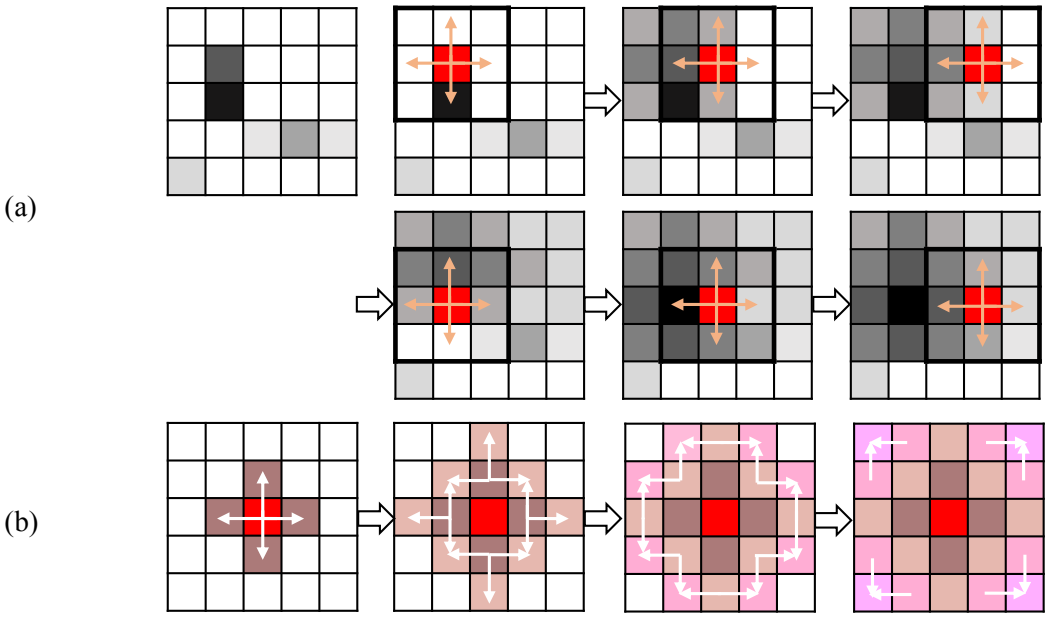


Fig. 4.1-1. (a) Sliding Neighbor Patch in Edge Weight Attenuation. (b) Edge Weight Attenuation Process in a Single Neighbor Patch.

if the new diffusion weight exceeds the previous value.

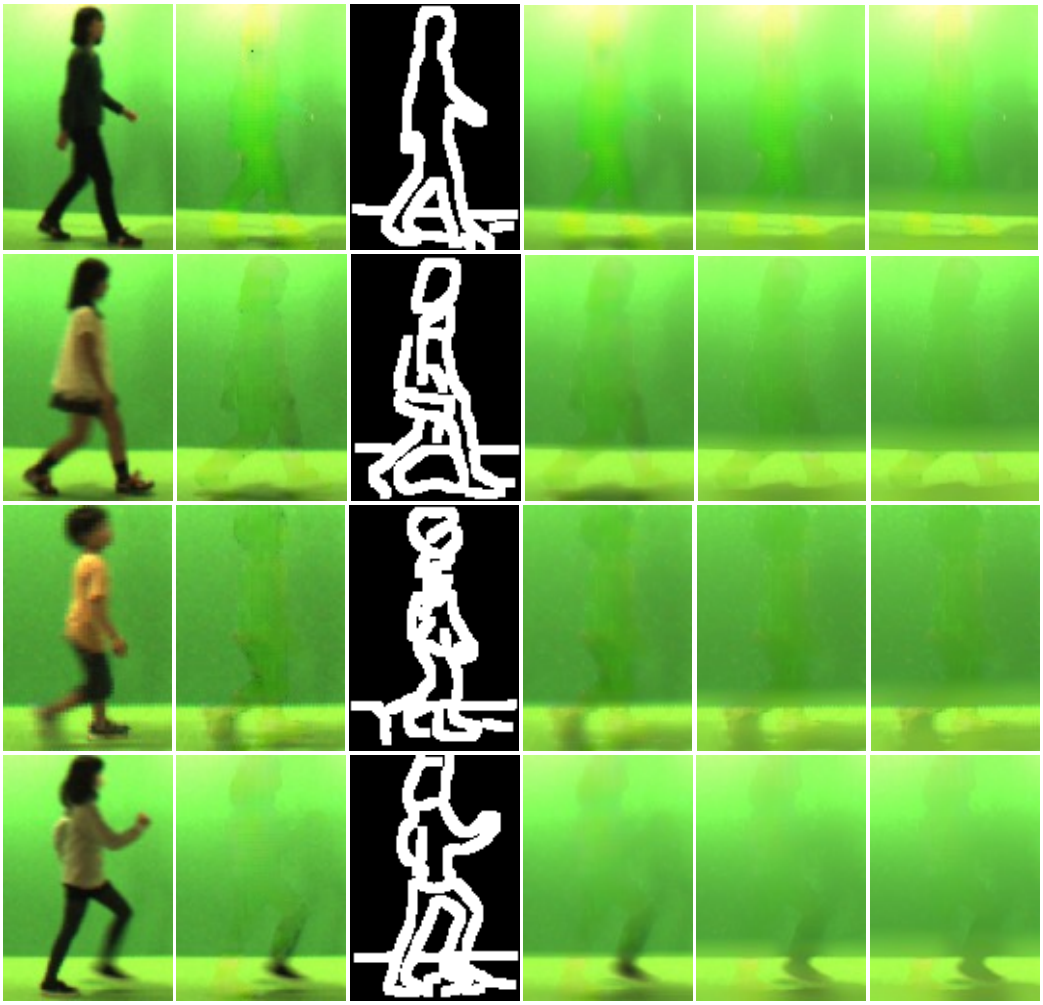
$$v_{i-1,j} \leftarrow \max(v_{i-1,j}, \exp(-\frac{v_{i,j}}{2*\sigma^2})), \tag{4.1.1}$$

$$v_{i+1,j} \leftarrow \max(v_{i+1,j}, \exp(-\frac{v_{i,j}}{2*\sigma^2})), \tag{4.1.2}$$

$$v_{i,j+1} \leftarrow \max(v_{i,j+1}, \exp(-\frac{v_{i,j}}{2*\sigma^2})), \tag{4.1.3}$$

$$v_{i,j-1} \leftarrow \max(v_{i,j-1}, \exp(-\frac{v_{i,j}}{2*\sigma^2})), \tag{4.1.4}$$

where $v_{i,j}$ stands for the diffusion weight value of pixel position (i, j) in the last step. Similarly, $v_{i-1,j}, v_{i+1,j}, v_{i,j-1}, v_{i,j+1}$ represent the weight value of pixel position $(i-1, j), (i+1, j), (i, j-1), (i, j+1)$ in the last step. σ is set to $0.15*k+0.35$, where k stands for the window size of the neighbor patch. The refined background in Fig. 4.1-2 (d)-(f) is obtained by multiplying the updated diffusion weight with coarse inpainting result.



(a) (b) (c) (d) (e) (f)

Fig. 4.1-2. Background prediction for OULP-Age database. (a) Image. (b) Coarse inpainting result. (c) Edge map. (d) Refined result with neighbor patch size 7. (e) Refined result with neighbor patch size 21. (f) Refined result with neighbor patch size 31.

4.2 Age Group Classification with Matting

Follow the first study, we conducted age group classification experiment in order to evaluate the generation quality of aging patterns. This experiment to check whether the age progressed/regressed image truly presented the characteristics of the intended age group. In other words, to what extent do the cross-age group classification performance of the generated gait silhouette video simulate the real ones, and whether the matting input can narrow down the gap.

Specifically, we first designed an age group classifier using a modified ResNet-18 architecture [106], and then trained two classifiers with binary and alpha matte of real gait silhouette sequence of OULP-Age dataset repectively. The input dimension of first convolution layer of ResNet-18 is modified to $N_{\text{img}} = 25$ in order to handle not a single still image but a gait silhouette sequence. By classifying the generated gait silhouette sequences from different methods using the same pre-trained ResNet-18 classifier for an unseen test set, we can check to which extent each method generates age progressed/regressed video that can present the characteristics of the intended age group.

Experimental results on age group classification accuracies among benchmarks are shown in Table 4.2.1. We also report the confusion matrix of the proposed method with and without matting input, respectively in Tables Tables 4.2.2 and 4.2.3.

Table 4.2.1 Age group classification accuracy[%] for benchmarks. E2E indicates end-to-end training. Bold font indicates the best accuracy.

Method	Silhouette Type	Age group (Ground Truth)					Avg
		[0, 5]	[6, 10]	[11, 15]	[16, 60]	Over 60	
Pretrained Clf	binary	73.24	78.57	55.62	96.13	26.32	85.28
Ours	binary	22.74	72.90	**84.17**	**98.01**	49.54	65.47
Pretrained Clf	alpha	81.43	77.16	67.18	93.76	36.49	85.51
Ours(Matting)	alpha	**51.13**	**74.14**	71.77	94.16	**51.05**	**68.45**

As a result, we can tell from the result that the proposed method with matting input outperforms the proposed method without matting. We further compare the confusion matrix between with and without matting, which shows the predicted

Table 4.2.2 Confusion matrix of ours method with matting.

Confusion Matrix		Predicted Class				
		[0,5]	[6,10]	[11,15]	[16,60]	Over 60
Actual Class	[0,5]	51.13	40.42	4.13	4.05	0.27
	[6,10]	1.32	74.14	19.04	4.98	0.51
	[11,15]	0.31	10.55	71.77	17.37	0.00
	[16,60]	0.62	2.26	2.69	94.16	0.27
	Over 60	0.14	0.35	3.82	44.64	51.05

Table 4.2.3 Confusion matrix of ours method without matting.

Confusion Matrix		Predicted Class				
		[0,5]	[6,10]	[11,15]	[16,60]	Over 60
Actual Class	[0,5]	22.74	34.01	2.92	38.79	1.54
	[6,10]	0.84	72.90	12.87	12.59	0.81
	[11,15]	0.00	3.20	84.17	12.08	0.55
	[16,60]	0.30	0.43	0.80	98.01	0.46
	Over 60	0.84	2.64	3.09	43.89	49.54

misclassified class will be closer to the actual class with matting input. Take results for age group [0,5] as an example, generated subjects with target age group [0, 5] are unlikely to be misclassified as age group [16, 60] with matting (possibility of 4.05%) compared to those without matting (possibility of 38.79%). Such improvement can benefit real-world applications, such as person search by age query and people counting by age group.

4.3 Cross-age Gait Recognition with Matting

We conducted cross-age gait recognition experiments to evaluate the preservation of individuality using the age progressed/regressed generated gait sequence in addition to real ones. Since the real cross-age gait database is not available, we designed the following simulated experimental setting. The dataset for the cross-age gait recognition experiment composed of three subsets: a training set, a gallery set, and a probe set. The training set contains 23,543 subjects, while the gallery

and probe sets form a set set composed of the other 2,616 subjects that are disjoint from the training set. Since each subject in OULP-Age has a single gait sequence, we generated five gait sequences per subject, which correspond to the five age groups, i.e., each subject has 6 gait sequences in total (one real and five generated).

To keep the same experimental setting with the first study, we employed Gaitset [69], which is a state-of-the-art network structure in gait recognition for gait silhouette sequences (as opposed to static gait templates) and adopted the code from the official implementation③.In both training and testing phases, the input sequences are prepossessed into size of 64×64 for GaitSet requirement. In the training phase, a batch with size of $p \times k$ is sampled from training set, where p stands for the number of persons and k stands for the number of training samples each person has in the batch. In the thesis, p is set to 8, whereas k is set to 16. Cross entropy loss and triplet loss are imposed to keep the identity across ages during training. In the test phase, the real silhouette sequences are assigned to the gallery, while the generated sequences are assigned to the probe.

In an identification scenario, we matched a probe to all the subjects in gallery and evaluated rank-1 identification rate based on dissimilarities (i.e., L2 norm between the final representations in the trained GaitSet network). We computed the standard deviation (uncertainty) sFRR of false rejection rate (FRR) pFRR in case of a single attempt per subject according to [107, 108], which is represented as:

$$\sigma_{FRR} = \sqrt{\frac{p_{FRR}(1-p_{FRR})}{n-1}} \tag{4.3.1}$$

where n is the number of subjects and it is 2,616 in our case. The standard deviation of true acceptance rate and rank-1 identification rate can be computed in the same manner. The average rank-1 identification ($\pm$the standard deviation)increases from 99.6% without matting input to 99.86% with matting input.

In a verification scenario, an input pair of a probe and a gallery is accepted as the same subject (i.e., positive sample pair) if the dissimilarity measure between them is below an acceptance threshold, and is rejected otherwise (i.e., negative sample pair or different subject pair). We computed the equal error rate (EER)

③ https://github.com/AbnerHqC/GaitSet, April 2022

of the false acceptance rate and false rejection rate as a typical performance measure. As reference, we also confirmed the statistical significant difference in terms of EER. The average EER decreases from 0.28% without matting to 0.21% with matting. We therefore confirmed that there is still statistical significant difference between matting and without matting input, even though the absolute difference of the rank-1 identification rate is less than 1%.

Table 4.3.1 Rank-1 identification rates [%] and EER [%] (±standard deviation [%]) for each age group in probe. Bold font indicates the best performances.

Measure	Method	Age group					Avg.
		[0,5]	[6,10]	[11,15]	[16,60]	Over 60	
Rank-1	Ours	**99.7** (±**0.11**)	99.5 (±0.14)	99.3 (±0.16)	99.8 (±0.09)	99.7 (±0.11)	99.60
	Ours(Matting)	**99.7** (±**0.11**)	**99.9** (±**0.06**)	**99.9** (±**0.06**)	**99.9** (±**0.06**)	**99.9** (±**0.06**)	**99.86**
EER	Ours	0.24 (±0.10)	0.20 (±0.09)	0.31 (±0.11)	0.32 (±0.11)	0.34 (±**0.11**)	0.28
	Ours(Matting)	**0.20** (±**0.09**)	**0.10** (±**0.06**)	**0.22** (±**0.09**)	**0.20** (±**0.09**)	**0.32** (±**0.11**)	**0.21**

Experimental results of the cross-age gait recognition are summarized in Table 4.3.1. As a result, we can see that the proposed method with matting yielded the best accuracy for all age groups, which indicates the superiority of the matting input in terms of individuality preservation.

4.4 Qualitative Visualization

We visualize the generated gait videos of our proposed method with binary and alpha matte from matting as input for comparison in Figures 4.4-4 and 4.4-5. While the result of binary input can reflect the identity and detail of the subject (i.e., we can easily infer the generated sequence of subject A is female, subject B is male), the proposed method with matting input can further reflect the subject A's hairstyle and subject B's belongings such as hat. For better comparison, we also animate the generated gait videos in the target age group with the proposed method given with or without matting input in Fig. 4.4-3.

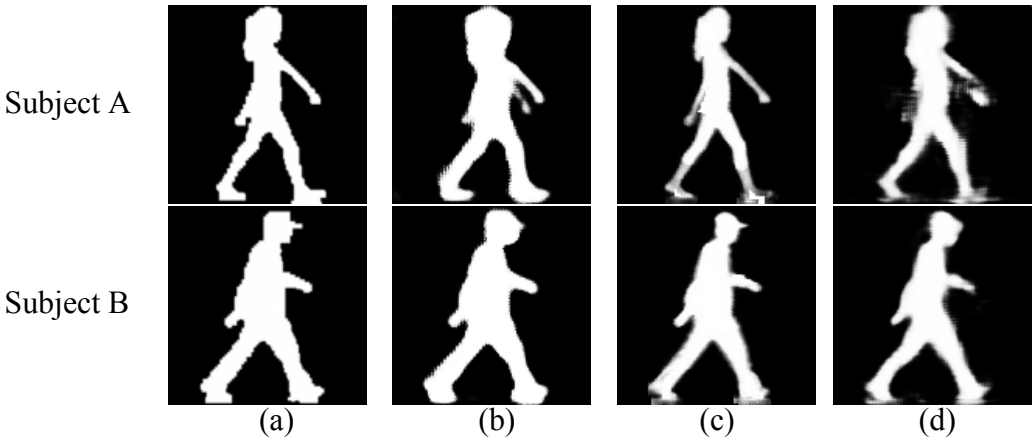


Fig. 4.4-3. Animated generated gait videos of our proposed method. Subject A, original gait video of age group [6, 10] years old, translated to target age group [0, 5] years old. Subject B, original gait video of age group Over 60 years old, translated to target age group [6, 10] years old.(a) Input: original gait video in binary. (b) Output: translated to target age group by our proposed method with binary input. (c) Input: original gait video in alpha matte. (d) Output: translated to target age group by our proposed method with alpha matte input by matting.

4.5 Correlation Between Generation Quality and Cross-age Group Classification

Our observation shows that the generation quality is not correlated with the mis-age group classification. We list typical examples of good quality in success mode, good quality in failure mode, poor quality in success mode, and poor quality in failure mode respectively in Fig. 4.5-6. When translating from age group [16, 60] to age group [0, 5] in the fourth example (Fig. 4.5-6 (g), (h)), artifacts appear in the arm region and the translated gait silhouette video is misclassified as an age group different from the target. Meanwhile, another subject also translating from age group [16, 60] to age group [0,5] in the third example (Fig. 4.5-6 (e), (f)) has artifacts in the head and leg regions. However, the generated gait silhouette sequence has been correctly classified to the target age group. Moreover, good generation quality may not guarantee the translated gait video be correctly classified to the target age group. The translated gait silhouette video of the second example (Fig. 4.5-6 (c), (d)) has good generation quality. Still, it is misclassified

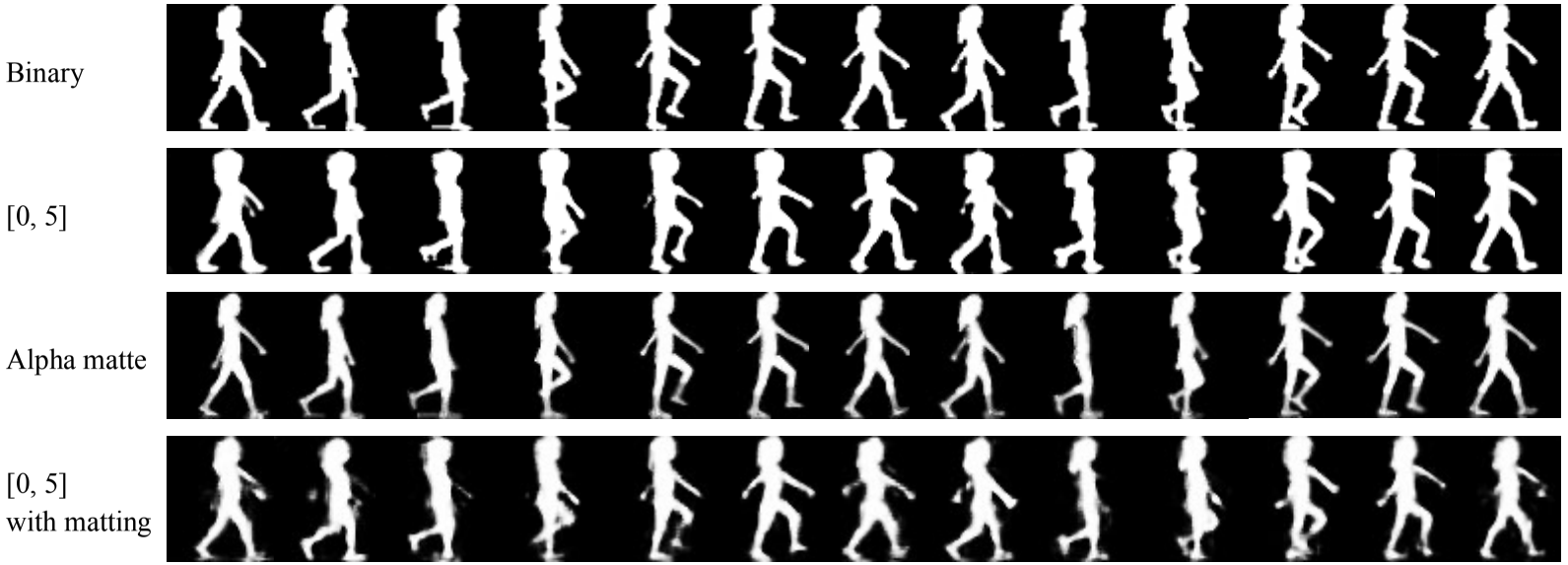


Fig. 4.4-4. Input: Binary sequence gait video (1st row, age group [6,10]). Output: Translated to age groups [0, 5] with proposed method (2nd row). Alpha matte result by matting model (3nd row). Translated to age groups [0, 5] with proposed method given matting input (4nd row).

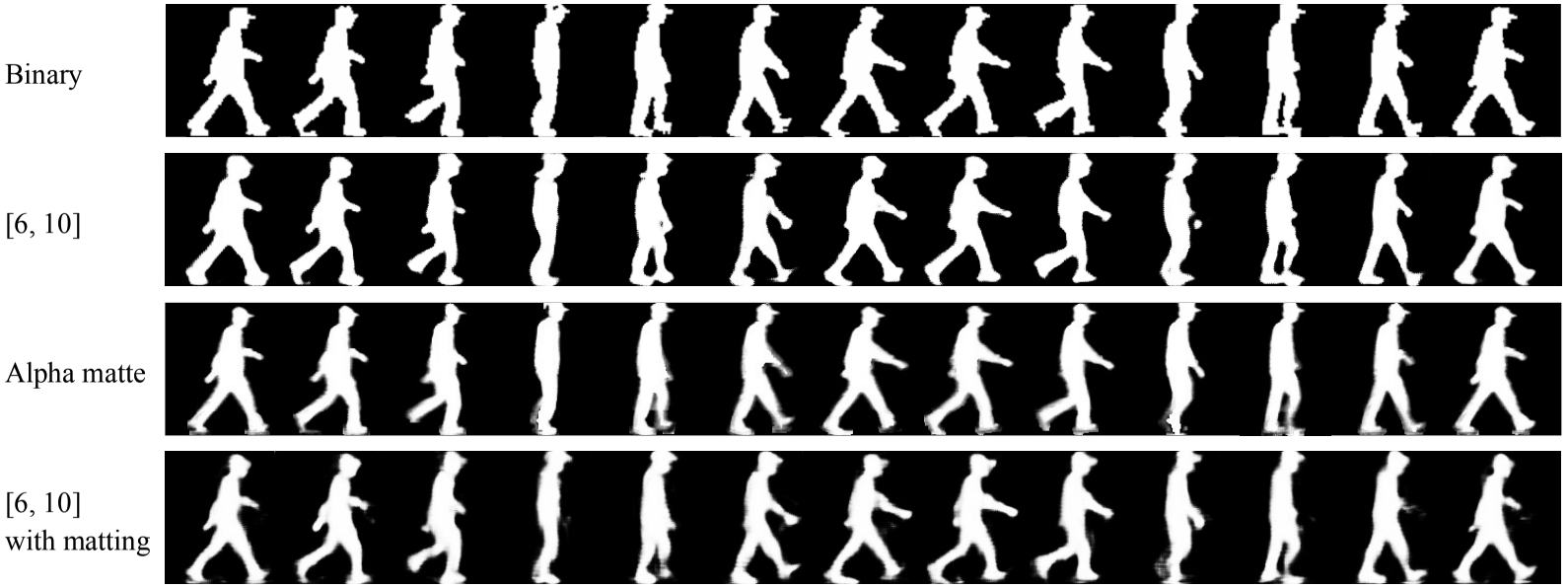


Fig. 4.4-5. Input: Binary sequence gait video (1st row, age group Over 60). Output: Translated to age groups [6, 10] with proposed method (2nd row). Alpha matte result by matting model (3nd row). Translated to age groups [6, 10] with proposed method given matting input (4nd row).

to age group [6,10] instead of the target age group Over 60.

Although the property and the quality of the generated gait silhouette sequence are two different aspects, good generation quality can be necessary for some applications. For example, in healthcare field application to remind a person to pay more effort to keep youthfulness in gait by watching his/her age-progressed gait video with function, the poor generation quality of the generated gait video might not be convincing to persuade the user.

(a)

(b)

Good generation quality in success mode. (a) Input of a subject in age group Over 60. (b) Translation to target age group [6, 10] successfully but in poor generation quality.

(c)

(d)

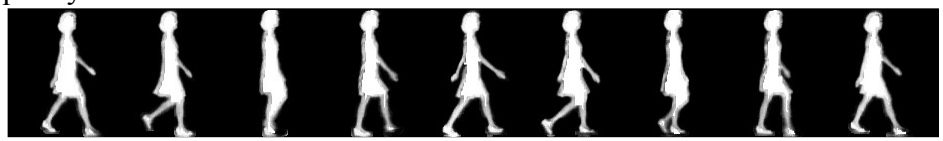
Good generation quality in failure mode. (c) Input of a subject in age group [0,5]. (d) Translation to target age group Over 60 but is misclassified as [6, 10].

(e)

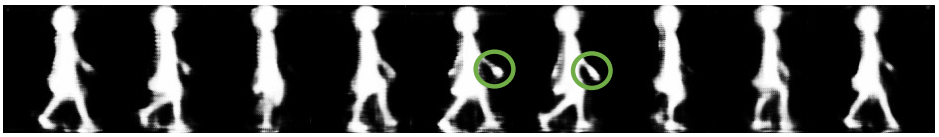
(f)

Poor generation quality in success mode. (e) Input of a subject in age group [16,60]. (f) Translation to target age group [0, 5] successfully but in poor generation quality.

(g)



(h)



Poor generation quality in failure mode. (g) Input of a subject in age group [16, 60]. (h) Translation to target age group [0, 5] but is misclassified as [6, 10].

Fig. 4.5-6. Correlation between generation quality and cross-age group classification.

Chapter V
Conclusion

This thesis focus on the age aspect of gait, which is yet to be thoroughly addressed in the literature. More specifically, we studied the gait-based video age progression and regression task, which is to generate gait videos of the target age while preserving the input subject's identity at the same time. Person recognition based on gait is an important research area in computer vision and pattern recognition due to its unobtrusive, noninvasive and non-perceivable nature. However, gait recognition can be challenging since gait can be easily affected by covariates such as age changes, which are inevitable during a person's life span.

Age progression/regression is an essential task in biometric recognition and computer vision with a wide range of applications, such as security, entertainment, etc. Age progression/regression has demonstrated its effectiveness in face analysis. However, it has not yet been extended to the gait community due to no paired dataset and lack of color and texture in gait sequences.

In the study of multi-age group gait video translation, we proposed the task of age-controllable gait video translation for the first time to learn the spatio-temporally augmented gait representation. Unlike the existing work that translates a static gait template, our work generates the age-progressed/regressed gait video for each input subject. The proposed framework consists of a motion augmented generator and a discriminator with SlowFast path to exploit both the temporal and spatial aspects of gait representation. The framework ensures the aging effect, individuality preservation, and gait realism from three inputs: gait period, period-normalized phase-synchronized gait video, and its frame difference sequence. This study is also the first to make use of the gait period, which have

been overlooked in most gait analysis, for age pattern representation and shown its effectiveness.

Since binary segmentation masks are not suitable for silhouette extraction due to the limit on spatial and temporal resolution. To obtain a spatio-temproally augmented high fidelity input to improve the cross-age recognition performance as well as the generation quality, we conducted the second study of a general matting framework that achieved competitive results on widely acknowledged matting benchmarks.

In the study of general image matting with simultaneous alpha and background output, we proposed an end-to-end three-branch image matting framework, which tends to seek unknown-relevant global contextual information from the whole large image and extract better foreground with predicted alpha matte. Specifically, the proposed framework is composed of a deformable sampling layer, which can obtain deformed condensed foreground and background, and a contextual attention layer to locate information from condensed foreground and background that are relevant to the unknown region. The proposed method can estimate alpha matte and background simultaneously while keeping the matting equation, which can improve the foreground extraction performance qualitatively. Experiments on two widely acknowledged benchmark datasets on matting have demonstrated the effectiveness of our proposed method.

In the discussion chapter, we explored how the quality of silhouette affects the performance of generated age progressed and regressed gait video of the target age group. Following the previous experiment settings, we conducted the age group classification and cross-age gait recognition experiments by utilizing matting results from the second study as input. Quantitative results show that the improvement from binary to matting input is not marginal. Meanwhile, from the qualitative result, although some detailed information such as gender can be observed given binary input, the same method with matting input can further reflect identity information such as hairstyle and belongings.

In future work, despite the progress in the gait-based video age progression/regression task, spatial inconsistency before and after translation remain challenging. Although alpha mattes from matting can preserve more fine-grained detail in contour, and describe better transparent areas (i.e., motion) within a frame,

inconsistency during the translation might largely degrade the qualitative quality of generated sequences. We will solve it by including a module with spatial transformation capability to enable warping from the original video to the generated video of the target age group to ensure a smooth transition between frames in future work.

In addition, we will conduct cross-dataset analysis to evaluate the generalization capability of the proposed method. Since OU series databases are the only ones to contain the age information of subjects, there is no available dataset for generalization capability evaluation. In the future, if a new gait database with age information is released, we'll conduct such experiments and update our results.

We will also conduct subjective tests to evaluate whether the age group and identity characteristics have been preserved in the generated gait silhouette videos. Instead of checking by myself, we'll hire more annotators for reliable evaluation from human perspective.

References

[1] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7366–7375. Computer Vision Foundation / IEEE Computer Society, 2018.

[2] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 568–576, 2014.

[3] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4489–4497. IEEE Computer Society, 2015.

[4] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1510–1517, 2018.

[5] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan C. Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *2017 IEEE Conference on Computer Vision and Pattern*

Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 3165–3174. IEEE Computer Society, 2017.

[6] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society, 2017.

[7] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Wang, William Yang Wang, Tamara L. Berg, Mohit Bansal, Jingjing Liu, Lijuan Wang, and Zicheng Liu. VALUE: A multi-task benchmark for video-and-language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

[8] Omar Elharrouss, Noor Almaadeed, and Somaya Al-Maadeed. A review of video surveillance systems. *Journal of Visual Communication and Image Representation*, 77:103116, 2021.

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[10] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

[11] Anju Jose Tom and Sudhish N. George. Simultaneous reconstruction and moving object detection from compressive sampled surveillance videos. *IEEE Trans. Image Process.*, 29:7590–7602, 2020.

[12] Wei Liu, Shengcai Liao, and Weidong Hu. Perceiving motion from dynamic memory for vehicle detection in surveillance videos. *IEEE Trans. Circuits Syst. Video Technol.*, 29(12):3558–3567, 2019.

[13] Weiwei Xing, Yuxiang Yang, Shunli Zhang, Qi Yu, and Liqiang Wang. Noisyotnet: A robust real-time vehicle tracking model for traffic surveillance. *IEEE Trans. Circuits Syst. Video Technol.*, 32(4):2107–2119, 2022.

[14] Dalia Coppi, Simone Calderara, and Rita Cucchiara. Transductive people tracking in unconstrained surveillance. *IEEE Trans. Circuits Syst. Video Technol.*, 26(4):762–775, 2016.

[15] Sijia Zhang, Maoguo Gong, Yu Xie, A. Kai Qin, Hao Li, Yuan Gao, and Yew-Soon Ong. Influence-aware attention networks for anomaly detection in surveillance videos. *IEEE Trans. Circuits Syst. Video Technol.*, 32(8):5427–5437, 2022.

[16] Guodong Guo and Na Zhang. A survey on deep learning based face recognition. *Comput. Vis. Image Underst.*, 189, 2019.

[17] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

[18] Daigo Muramatsu, Akira Shiraishi, Yasushi Makihara, Md. Zasim Uddin, and Yasushi Yagi. Gait-based person recognition using arbitrary view transformation model. *IEEE Trans. Image Process.*, 24(1):140–154, 2015. DOI:10.1109/TIP.2014.2371335.

[19] Yasushi Makihara, Takuya Tanoue, Daigo Muramatsu, Yasushi Yagi, Syunsuke Mori, Yuzuko Utsumi, Masakazu Iwamura, and Koichi Kise. Individuality-preserving silhouette extraction for gait recognition. *IPSJ Trans. Comput. Vis. Appl.*, 7:74–78, 2015. DOI:10.2197/ipsjtcva.7.74.

[20] Joann M Montepare and Leslie Zebrowitz-McArthur. Impressions of people created by age-related qualities of their gaits. *Journal of personality and social psychology*, 55(4):547, 1988.

[21] Asymmetric visual representation of sex from human body shape. *Cognition*, 205:104436, 2020. DOI:10.1016/j.cognition.2020.104436.

[22] Lynn T Kozlowski and James E Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception & psychophysics*, 21(6):575–580, 1977.

[23] Yuhan Zhou, Robbin Romijnders, Clint Hansen, Jos van Campen, Walter Maetzler, Tibor Hortobágyi, and Claudine JC Lamoth. The detection of age groups by dynamic gait outcomes using machine learning approaches. *Scientific reports*, 10(1):1–12, 2020.

[24] Edward R. Morrison, Hannah Bain, Louise Pattison, and Hannah Whyte-Smith. Something in the way she moves: biological motion, body shape, and attractiveness in women. *Visual Cognition*, 26(6):405–411, 2018. DOI:10.1080/13506285.2018.1471560.

[25] Kai Han, Jianyuan Guo, Chao Zhang, and Mingjian Zhu. Attribute-aware attention model for fine-grained representation learning. In Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei, editors, *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 2040–2048. ACM, 2018.

[26] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognit.*, 95:151–161, 2019.

[27] Thomas Wolf, Mohammadreza Babaee, and Gerhard Rigoll. Multi-view gait recognition using 3d convolutional neural networks. In *2016 IEEE International Conference on Image Processing, ICIP 2016*, pages 4165–4169. IEEE, 2016. DOI:10.1109/ICIP.2016.7533144.

[28] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(2):209–226, 2017. DOI:10.1109/TPAMI.2016.2545669.

[29] Gunawan Ariyanto and Mark S. Nixon. Model-based 3d gait biometrics. In *2011 IEEE International Joint Conference on Biometrics, IJCB 2011*, pages 1–7. IEEE Computer Society, 2011. DOI:10.1109/IJCB.2011.6117582.

[30] Gunawan Ariyanto and Mark S. Nixon. Marionette mass-spring model for 3d gait biometrics. In Anil K. Jain, Arun Ross, Salil Prabhakar, and Jaihie Kim, editors, *5th IAPR International Conference on Biometrics, ICB 2012*, pages 354–359. IEEE, 2012. DOI:10.1109/ICB.2012.6199832.

[31] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015*, pages 4489–4497. IEEE Computer Society, 2015. DOI:10.1109/ICCV.2015.510.

[32] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 4710–4719. Computer Vision Foundation / IEEE, 2019. DOI:10.1109/CVPR.2019.00484.

[33] Rijun Liao, Chunshui Cao, Edel B. García Reyes, Shiqi Yu, and Yongzhen Huang. Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations. In *Biometric Recognition CCBR 2017*, volume 10568 of *Lecture Notes in Computer Science*, pages 474–483. Springer, 2017. DOI:10.1007/978-3-319-69923-3_51.

[34] Yuqi Zhang, Yongzhen Huang, Shiqi Yu, and Liang Wang. Cross-view gait recognition by discriminative feature learning. *IEEE Trans. Image Process.*, 29:1001–1015, 2020. DOI:10.1109/TIP.2019.2926208.

[35] Xinhui Wu, Weizhi An, Shiqi Yu, Weiyu Guo, and Edel B. García Reyes. Spatial-temporal graph attention network for video-based gait recognition. In *Pattern Recognition - 5th Asian Conference, ACPR 2019*, volume 12047 of *Lecture Notes in Computer Science*, pages 274–286. Springer, 2019. DOI:10.1007/978-3-030-41299-9_22.

[36] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 8126–8133. AAAI Press, 2019. DOI:10.1609/aaai.v33i01.33018126.

[37] James W. Davis. Visual categorization of children and adult walking styles. In Josef Bigün and Fabrizio Smeraldi, editors, *Audio- and Video-Based Biometric Person Authentication, Third International Conference, AVBPA 2001 Halmstad, Sweden, June 6-8, 2001, Proceedings*, volume 2091 of *Lecture Notes in Computer Science*, pages 295–300. Springer, 2001.

[38] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu. A study on gait-based gender classification. *IEEE Transactions on Image Processing*, 18(8):1905–1910, Aug. 2009.

[39] Yasushi Makihara, Mayu Okumura, Haruyuki Iwama, and Yasushi Yagi. Gait-based age estimation using a whole-generation gait database. In *2011 IEEE International Joint Conference on Biometrics, IJCB 2011, Washington, DC, USA, October 11-13, 2011*, pages 1–6. IEEE Computer Society, 2011.

[40] Yun Fu, Guodong Guo, and Thomas S. Huang. Age synthesis and estimation via faces: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, 2010.

[41] Jin-Li Suo, Xilin Chen, Shiguang Shan, Wen Gao, and Qionghai Dai. A concatenational graph evolution aging model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2083–2096, 2012.

[42] Narayanan Ramanathan and Rama Chellappa. Modeling shape and textural variations in aging faces. In *8th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2008), Amsterdam, The Netherlands, 17-19 September 2008*, pages 1–8. IEEE Computer Society, 2008.

[43] Narayanan Ramanathan and Rama Chellappa. Modeling age progression in young faces. In *2006 IEEE Computer Society Conference on Computer*

Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA, pages 387–394. IEEE Computer Society, 2006.

[44] Xiangbo Shu, Jinhui Tang, Hanjiang Lai, Luoqi Liu, and Shuicheng Yan. Personalized age progression with aging dictionary. In *2015 IEEE International Conference on Computer Vision, ICCV 2015*, pages 3970–3978. IEEE Computer Society, 2015. DOI:10.1109/ICCV.2015.452.

[45] Xiangbo Shu, Jinhui Tang, Zechao Li, Hanjiang Lai, Liyan Zhang, and Shuicheng Yan. Personalized age progression with bi-level aging dictionary learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):905–917, 2018.

[46] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 4352–4360. IEEE Computer Society, 2017. DOI:10.1109/CVPR.2017.463.

[47] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. Face aging with identity-preserved conditional generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pages 7939–7947. Computer Vision Foundation / IEEE Computer Society, 2018. DOI:10.1109/CVPR.2018.00828.

[48] Zhenliang He, Meina Kan, Shiguang Shan, and Xilin Chen. S2GAN: share aging factors across ages and share aging trends among individuals. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9439–9448. IEEE, 2019. DOI:10.1109/ICCV.2019.00953.

[49] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K. Jain. Learning face age progression: A pyramid architecture of gans. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pages 31–39. Computer Vision Foundation / IEEE Computer Society, 2018. DOI:10.1109/CVPR.2018.00011.

[50] Zeqi Li, Ruowei Jiang, and Parham Aarabi. Continuous face aging via self-estimated residual age embedding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15008–15017. Computer Vision Foundation / IEEE, 2021.

[51] Zhizhong Huang, Junping Zhang, and Hongming Shan. When age-invariant face recognition meets face age synthesis: A multi-task learning framework. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7282–7291. Computer Vision Foundation / IEEE, 2021.

[52] Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J. Maybank. General tensor discriminant analysis and gabor features for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1700–1715, 2007.

[53] Toby H. W. Lam, King Hong Cheung, and James N. K. Liu. Gait flow image: A silhouette-based gait representation for human identification. *Pattern Recognition*, 44(4):973–987, 2011.

[54] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006.

[55] Khalid Bashir, Tao Xiang, and Shaogang Gong. Gait recognition using gait entropy image. In *3rd International Conference on Imaging for Crime Detection and Prevention, ICDP 2009, London, UK, December 3, 2009*, pages 1–6. IET / IEEE, 2009.

[56] Chen Wang, Junping Zhang, Liang Wang, Jian Pu, and Xiaoru Yuan. Human identification using temporal information preserving gait template. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(11):2164–2176, nov. 2012.

[57] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Gait recognition using a view transformation model in the frequency domain. In

European Conference on Computer Vision, pages 151–163, Graz, Austria, May 2006.

[58] Chi Xu, Yasushi Makihara, Yasushi Yagi, and Jianfeng Lu. Gait-based age progression/regression: a baseline and performance evaluation by age group classification and cross-age gait identification. *Mach. Vis. Appl.*, 30(4):629–644, 2019. DOI:10.1007/s00138-019-01015-x.

[59] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[60] Yang Feng, Yuncheng Li, and Jiebo Luo. Learning effective gait features using LSTM. In *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, pages 325–330. IEEE, 2016.

[61] Rijun Liao, Chunshui Cao, Edel B. García Reyes, Shiqi Yu, and Yongzhen Huang. Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations. In Jie Zhou, Yunhong Wang, Zhenan Sun, Yong Xu, Linlin Shen, Jianjiang Feng, Shiguang Shan, Yu Qiao, Zhenhua Guo, and Shiqi Yu, editors, *Biometric Recognition - 12th Chinese Conference, CCBR 2017, Shenzhen, China, October 28-29, 2017, Proceedings*, volume 10568 of *Lecture Notes in Computer Science*, pages 474–483. Springer, 2017.

[62] Francesco Battistone and Alfredo Petrosino. TGLSTM: A time based graph deep learning approach to gait recognition. *Pattern Recognit. Lett.*, 126:132–138, 2019.

[63] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4710–4719. Computer Vision Foundation / IEEE, 2019.

[64] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds. *ACM Trans. Graph.*, 38(5):146:1–146:12, 2019.

[65] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4490–4499. Computer Vision Foundation / IEEE Computer Society, 2018.

[66] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5099–5108, 2017.

[67] William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1024–1034, 2017.

[68] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3337–3345. IEEE Computer Society, 2017.

[69] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proc. of the 33rd AAAI Conf. on Artificial Intelligence (AAAI 2019)*, 2019.

[70] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *International Conference on Biometrics, ICB*

2016, Halmstad, Sweden, June 13-16, 2016, pages 1–8. IEEE, 2016. DOI:10.1109/ICB.2016.7550060.

[71] Francisco M. Castro, Manuel J. Marín-Jiménez, Nicolás Guil, Santiago Lopez Tapia, and Nicolas Pérez de la Blanca. Evaluation of cnn architectures for gait recognition based on optical flow maps. In *International Conference of the Biometrics Special Interest Group, BIOSIG 2017, Darmstadt, Germany, September 20-22, 2017*, volume P-270 of *LNI*, pages 251–258. GI / IEEE, 2017.

[72] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[73] Yasushi Makihara and Yasushi Yagi. Silhouette extraction based on iterative spatio-temporal local color transformation and graph-cut segmentation. In *19th International Conference on Pattern Recognition (ICPR 2008), December 8-11, 2008, Tampa, Florida, USA*, pages 1–4. IEEE Computer Society, 2008.

[74] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[75] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[76] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[77] Jiwen Lu and Yap-Peng Tan. Ordinary preserving manifold analysis for human age and head pose estimation. *IEEE Trans. Hum. Mach. Syst.*, 43(2):249–258, 2013.

[78] J.W. Davis. Visual categorization of children and adult walking styles. In *International Conference on Audio- and Video-based Biometric Person Authentication*, pages 295–300, Jun. 2001.

[79] R.K. Begg. Support vector machines for automated gait classification. *IEEE Transactions on Biomedical Engineering*, 52(5):828–838, May 2005.

[80] S. Zhang, Y. Wang, and A. Li. Gait-based age estimation with deep convolutional neural network. In *International Conference on Biometrics*, pages 1–8, 2019.

[81] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren. Make the bag disappear: Carrying status-invariant gait-based human age estimation using parallel generative adversarial networks. In *Proc. of the IEEE 10th Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS 2019)*, pages 1–9, Sep. 2019.

[82] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(2):316–322, 2006. DOI:10.1109/TPAMI.2006.38.

[83] Z. Liu and S. Sarkar. Simplest representation yet for gait recognition: Averaged silhouette. In *International Conference on Pattern Recognition*, volume 1, pages 211–214, Aug. 2004.

[84] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Jian Wan, Nanxin Wang, and Xiaoming Liu. Gait recognition via disentangled representation learning. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Long Beach, CA, June 2019.

[85] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[86] Yasushi Makihara, Hidetoshi Mannami, and Yasushi Yagi. Gait analysis of gender and age using a large-scale multi-view gait database. In *Asian Conference on Computer Vision*, volume 6493, pages 440–451. Springer, 2010.

[87] Benz Kek Yeo Chuen, Connie Tee, Thian Song Ong, and Michael Goh Kah Ong. A preliminary study of gait-based age estimation techniques. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2015, Hong Kong, December 16-19, 2015*, pages 800–806. IEEE, 2015.

[88] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Nikolai Yakovenko, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Neural Information Processing Systems*, pages 1152–1164, 2018.

[89] Xingxing Wei, Jun Zhu, Sitong Feng, and Hang Su. Video-to-video translation with global temporal consistency. In *ACM Multimedia Conference on Multimedia*, pages 18–25. ACM, 2018.

[90] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *European Conference on Computer Vision*, volume 11209, pages 122–138. Springer, 2018.

[91] Yang Chen, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. Mocycle-gan: Unpaired video-to-video translation. In *ACM Multimedia Conference on Multimedia*, pages 647–655. ACM, 2019.

[92] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, pages 2242–2251. IEEE, 2017.

[93] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Computer Vision and Pattern Recognition*, pages 8789–8797. IEEE, 2018.

[94] Changhong Chen, Jimin Liang, Heng Zhao, Haihong Hu, and Jie Tian. Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, 30(11):977–984, 2009.

[95] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *International Conference on Computer Vision*, pages 6201–6210. IEEE, 2019.

[96] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. Spatiotemporal residual networks for video action recognition. In *Neural Information Processing Systems*, pages 3468–3476, 2016.

[97] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Computer Vision and Pattern Recognition*, pages 1933–1941. IEEE, 2016.

[98] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1857–1865. PMLR, 2017.

[99] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *Neural Information Processing Systems*, pages 5767–5777, 2017.

[100] Chi Xu, Yasushi Makihara, Gakuto Ogi, Xiang Li, Yasushi Yagi, and Jianfeng Lu. The OU-ISIR gait database comprising the large population dataset with age and performance evaluation of age estimation. *IPSJ Trans. Comput. Vis. Appl.*, 9:24, 2017.

[101] Barry Bogin. *The human pattern of growth and development in paleontological perspective*. 01 2003.

[102] Barry Bogin, Jared Bragg, and Christopher Kuzawa. Humans are not cooperative breeders but practice biocultural reproduction. *Annals of human biology*, 41:368–380, 07 2014.

[103] Zhou Chengju, Ikuhisa Mitsugami, Fumio Okura, Kota Aoki, and Yasushi Yagi. Growth assessment of school-age children using dual-task observation. *ITE Transactions on Media Technology and Applications*, 6:286–296, 10 2018.

[104] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Gait-based human age estimation using age group-dependent manifold learning and regression. *Multimedia Tools Applications*, 77(21):28333–28354, 2018.

[105] D P Kingma and J Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015,Conference Track Proceedings*, 2015. http://arxiv.org/abs/1412.6980.

[106] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016.

[107] Anil Jain, Ruud Bolle, and Sharath Pankanti. Introduction to biometrics. In *Biometrics*, pages 1–41. Springer, 1996.

[108] Anthony J Mansfield and James L Wayman. Best practices in testing and reporting performance of biometric devices. 2002.

[109] Jian Sun, Jiaya Jia, Chi-Keung Tang, and Heung-Yeung Shum. Poisson matting. *ACM Trans. Graph.*, 23(3):315–321, 2004. DOI:10.1145/1015706.1015721.

[110] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):228–242, 2008. DOI:10.1109/TPAMI.2007.1177.

[111] Yung-Yu Chuang, Brian Curless, David Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 264–271, 2001. DOI:10.1109/CVPR.2001.990970.

[112] Yagiz Aksoy, Tunç Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In *2017 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2017. DOI:10.1109/CVPR.2017.32.

[113] Jue Wang and Michael F. Cohen. Optimized color sampling for robust matting. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007. DOI:10.1109/CVPR.2007.383006.

[114] Ehsan Shahrian Varnousfaderani and Deepu Rajan. Weighted color and texture sample selection for image matting. *IEEE Trans. Image Process.*, 22(11):4260–4270, 2013. DOI:10.1109/TIP.2013.2271549.

[115] Xiao Chen, Fazhi He, and Haiping Yu. A matting method based on full feature coverage. *Multim. Tools Appl.*, 78(9):11173–11201, 2019. DOI:10.1007/s11042-018-6690-1.

[116] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2049–2056, 2011. DOI:10.1109/CVPR.2011.5995495.

[117] Eduardo Simoes Lopes Gastal and Manuel M. Oliveira. Shared sampling for real-time alpha matting. *Comput. Graph. Forum*, 29(2):575–584, 2010. DOI:10.1111/j.1467-8659.2009.01627.x.

[118] Ehsan Shahrian, Deepu Rajan, Brian L. Price, and Scott Cohen. Improving image matting using comprehensive sampling sets. In *2013 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 636–643, 2013. DOI:10.1109/CVPR.2013.88.

[119] Donghyeon Cho, Yu-Wing Tai, and In-So Kweon. Natural image matting using deep convolutional neural networks. In *Computer Vision - ECCV 2016 - 14th European Conference*, volume 9906, pages 626–643. Springer, 2016. DOI:10.1007/978-3-319-46475-6_39.

[120] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. Alphagan: Generative adversarial networks for natural image matting. In *British Machine Vision Conference 2018*, page 259. BMVA Press, 2018. http://bmvc2018.org/contents/papers/0915.pdf.

[121] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018*, pages 618–626. ACM, 2018. DOI:10.1145/3240508.3240610.

[122] Guanying Chen, Kai Han, and Kwan-Yee K. Wong. Tom-net: Learning transparent object matting from a single image. In *2018 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9233–9241, 2018. DOI:10.1109/CVPR.2018.00962.

[123] Ning Xu, Brian L. Price, Scott Cohen, and Thomas S. Huang. Deep image matting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 311–320, 2017. DOI:10.1109/CVPR.2017.41.

[124] Yu Wang, Yi Niu, Peiyong Duan, Jianwei Lin, and Yuanjie Zheng. Deep propagation based image matting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 999–1006, 2018. DOI:10.24963/ijcai.2018/139.

[125] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion CNN for digital matting. In *2019 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7469–7478, 2019. DOI:10.1109/CVPR.2019.00765.

[126] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 8818–8827, 2019. DOI:10.1109/ICCV.2019.00891.

[127] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *2019 IEEE/CVF International Con-*

ference on Computer Vision, ICCV 2019, pages 4129–4138. IEEE, 2019. DOI:10.1109/ICCV.2019.00423.

[128] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *2019 IEEE/CVF International Conference on Computer Vision*, pages 3265–3274. IEEE, 2019. DOI:10.1109/ICCV.2019.00336.

[129] Jingwei Tang, Yagiz Aksoy, Cengiz Öztireli, Markus H. Gross, and Tunç Ozan Aydin. Learning-based sampling for natural image matting. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 3055–3063, 2019. DOI:10.1109/CVPR.2019.00317.

[130] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 11450–11457, 2020. https://aaai.org/ojs/index.php/AAAI/article/view/6809.

[131] J Yu, Z Lin, J Yang, X Shen, X Lu, and T S Huang. Generative image inpainting with contextual attention. In *2018 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018. DOI:10.1109/CVPR.2018.00577.

[132] P Lee and Y Wu. Nonlocal matting. In *2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2193–2200, 2011. DOI:10.1109/CVPR.2011.5995665.

[133] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. KNN matting. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 869–876, 2012. DOI:10.1109/CVPR.2012.6247760.

[134] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *Computer Vision - ECCV 2016 - 14th European Conference*, volume 9905, pages 92–107. Springer, 2016. DOI:10.1007/978-3-319-46448-0_6.

[135] L Chen, G Papandreou, F Schroff, and H Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. http://arxiv.org/abs/1706.05587.

[136] X Wang, R B Girshick, A Gupta, and K He. Non-local neural networks. In *2018 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7794–7803. IEEE, 2018. DOI:10.1109/CVPR.2018.00813.

[137] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard S. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 4898–4906, 2016. proceedings.neurips.cc/c8067ad1937f728f51288b3eb986afaa.html.

[138] A Recasens, P Kellnhofer, S Stent, W Matusik, and A Torralba. Learning to zoom: A saliency-based sampling layer for neural networks. In *Computer Vision - ECCV 2018 - 15th European Conference*, volume 11213, pages 52–67. Springer. DOI:10.1007/978-3-030-01240-3_4.

[139] Haichao Yu, Ning Xu, Zilong Huang, Yuqian Zhou, and Humphrey Shi. High-resolution deep image matting. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 3217–3224, 2021. https://ojs.aaai.org/index.php/AAAI/article/view/16432.

[140] C Rhemann, C Rother, J Wang, M Gelautz, P Kohli, and P Rott. A perceptually motivated online benchmark for image matting. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009. DOI:10.1109/CVPR.2009.5206503.

[141] Yuanjie Zheng and Chandra Kambhamettu. Learning based digital matting. In *IEEE 12th International Conference on Computer Vision, ICCV 2009*, pages 889–896. IEEE, 2009. DOI:10.1109/ICCV.2009.5459326.

[142] Fenfen Zhou, Yingjie Tian, and Zhiquan Qi. Attention transfer network for nature image matting. *IEEE Trans. Circuits Syst. Video Technol.*, 31(6):2192–2205, 2021. DOI:10.1109/TCSVT.2020.3024213.

[143] Yutong Dai, Hao Lu, and Chunhua Shen. Learning affinity-aware upsampling for deep image matting. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, pages 6841–6850, 2021. CVPR2021/html/Learning_Affinity-Aware_Upsampling_for_Deep_Image_Matting.html.

[144] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, pages 11120–11129, 2021. CVPR2021/html/Sun_Semantic_Image_Matting.html.

[145] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *CoRR*, abs/1901.00212, 2019.