INCLG: Inpainting for Non-Cleft Lip Generation with a Multi-Task Image Processing Network

Shuang Chen (Durham University, shuang.chen@durham.ac.uk),
Amir Atapour-Abarghouei (Durham University, amir.atapour-abarghouei@durham.ac.uk),
Edmond S. L. Ho (University of Glasgow, shu-lim.ho@glasgow.ac.uk),
Hubert P. H. Shum (Durham University, hubert.shum@durham.ac.uk, corresponding author)

Abstract
We present a software that predicts non-cleft facial images for patients with cleft lip, thereby facilitating the understanding, awareness and discussion of cleft lip surgeries. To protect patients' privacy, we design a software framework using image inpainting, which does not require cleft lip images for training, thereby mitigating the risk of model leakage. We implement a novel multi-task architecture that predicts both the non-cleft facial image and facial landmarks, resulting in better performance as evaluated by surgeons. The software is implemented with PyTorch and is usable with consumer-level color images with a fast prediction speed, enabling effective deployment.

Keywords
Cleft Lip; Image Inpainting; Deep Neural Network; Multi-task Learning; Face Modelling

Code metadata

Nr.	Code metadata description	Please fill in this column
C1	Current code version	v1
C2	Permanent link to code/repository used for this code version	https://github.com/ChrisChen1023/INCLG
C3	Permanent link to Reproducible Capsule	https://codeocean.com/capsule/4388343/tree/v1
C4	Legal Code License	MIT License
C5	Code versioning system used	git
C6	Software code languages, tools, and services used	Python
C7	Compilation requirements, operating environments & dependencies	Python 3.8, PyTorch 1.10.2, torchvision 0.11.3
C8	If available Link to developer documentation/manual	https://codeocean.com/capsule/4388343/tree/v1
C9	Support email for questions	shuang.chen@durham.ac.uk

1. Introduction

A cleft lip/palate is a medical condition where the lip/palate of a patient does not join completely before birth, which usually occurs in the early stages of pregnancy. In the UK, cleft lips are the most common facial birth defect, with one out of every 700 children suffering from cleft lip and palate every year [1]. This explains the importance of cleft lip and palate surgeries, which are usually performed on orofacial cleft patients at an average age of three months [2]. Although the surgical treatment for cleft lip and palate varies, their common objective is to achieve symmetry and enhance a nasolabial look [3].

As cleft lips are pre-born defects, many parents of the patients would find it hard to imagine what the non-cleft faces of their children would be like. To facilitate the understanding, awareness and discussion of cleft lip surgeries, we have worked with the UK's Royal Victoria Infirmary (RVI) to collect a dataset of cleft lips patients, and designed a system that allows the prediction of non-cleft faces from the cleft lip counterparts.

A core challenge of our software design is to protect the privacy of cleft lip patients. Research has shown that due to the high memory capacity of deep learning models, it is possible to reconstruct original training samples from the network parameters, a scenario known as model leakage [4]. While it may be straightforward to formulate the non-cleft facial-prediction task using a style transfer [5] framework, training such systems requires both cleft and non-cleft facial images, resulting in a risk of model leakage. We present a novel software engineering design by tackling non-cleft facial-prediction with an image in-painting framework. This allows us to train the system using open facial datasets [6] with a tailor-made algorithm to mask out the mouth area, and test the system with cleft lip images. As a result, the model parameters do not store any cleft lip information.

In particular, we develop a PyTorch-based software that utilizes a state-of-the-art image inpainting network [7] as the backbone, and develop a multi-task system that predicts both images and the facial landmarks, thereby generating non-cleft faces. The facial landmark task provides geometric information that facilitates the image generation task. Compared to existing work that utilizes a multi-stage framework to first predict landmarks and then predict images [8, 7], ours is superior as both tasks are performed at the same time, avoiding any error propagation from the first stage to the second stage.

The quality of images produced by our software has been evaluated by NHS surgeons, showcasing its superior performance to alternative designs. It has a fast inference speed and works with color images captured by consumer-level cameras, allowing an effective deployment process. It is open-source, facilitating research and development in this area.

The source code presented in this paper has been originally developed to implement the theory proposed in [9], which is accepted in a biomedicine-focused conference. In this paper, we explain the implementation details

of this software and its impact in the real world. In particular, we focus on the design concepts of the software architecture and the details of the engineering considerations. This is further supported by a validated version of the source code in the CodeOcean environment.

2. System description

To protect the privacy of patients' data, we decide to implement the non-cleft facial image prediction system as an image inpainting framework. One key software engineering decision in this research is the framework we use to implement the solution. Existing style transfer-based frameworks [5] allow effective facial image generation with different features. However, they require training data from both the source (i.e. cleft lip images in our case) and target (i.e. non-cleft lip images) domains, which may lead to model leakage where the trained model memorizes the training images. Conditional image translation frameworks using GAN [5] or VAEs [10] may resolve the issue, but those methods mainly focus on the synthesis of new color patterns instead of geometric structures. Our investigation led us to the image inpainting framework [7] as a suitable solution, as it does not necessitate using cleft facial data for training. Additionally, the binary mask effectively defines the lip area for synthesis with the rest of the face, serving as conditions, making it well-suited to our requirements.

In particular, to implement an image inpainting framework, we utilize the image generation network in [7] as the backbone, which is ameliorated from [8], given its good performance in image inpainting. We also re-implemented the gated convolution algorithm proposed in [11] to dynamically select features for each channel and location, resulting in better inpainting quality.

On top of the backbone, we implement a multi-task system that predicts both the non-cleft facial image and facial landmarks. Facial landmark has shown to be effective in assisting facial image inpainting [8, 7], and is used extensively for cleft lip analysis [12]. Our work differs from existing approaches in that we employ a multi-task model, where two tasks share a part of a common network and facilitate each other.

To prepare the training data, we employ an open facial dataset and a tailor-made masking algorithm. In particular, we use the CelebA dataset [6], which consists of 202,599 face images of over 10,000 celebrities. To prepare the data for training our inpainting network, we apply an irregular mask algorithm following [13], such that our network can learn to inpaint any masked regions of the face.

To test the system, we work with the NHS to collect a dataset of cleft lip images. Due to the sensitive nature of the data, ethical approvals are obtained from the Research Ethics Committee (REC), the Health Research Authority (HRA), and Health and Care Research Wales (HCRW), under Approval Nos. 19/LO/1690 and under IRAS Project ID: 240451. Given a cleft lip image, we manually draw a mask that covers the mouth area. The masked image is fed into our multi-task network to create the non-cleft facial counterpart, with the facial landmark as a side-product. Since cleft lip images are only used in testing, we mitigate any risk of model leakage [4].

2.1 Network Design and Implementation

Here, we provide details for the design and implementation of our deep neural network, as shown in Figure 1.

The encoder is used to encode an image into a feature representation. We develop a gated convolution block that includes a gated convolution layer [11], a normalisation layer and an activation layer (ReLU). A masked image is fed into three gated convolution blocks with decreasing feature sizes from 256×256 to 64×64. Subsequently, the encoded feature is passed into multiple dilated residual blocks to extend the receptive field of the encoder. At the end of the encoder, we follow [14] to implement an attention mechanism to match the masked and unmasked regions. After a skip connection, the encoder outputs the shared feature map f_{share}, which is practically a concise representation of the image. The feature map is passed to both the image generator and predictor.

The image generator is used to predict the non-cleft facial image. Given the shared feature f_{share}, we employ a gated convolution block to implement upsampling. This is followed by two 1×1 convolution layers, F_1 and F_2, for feature fusion. The first one is utilised to fuse the encoder feature with the skip connection, while the second one is responsible for parameter sharing to fuse the landmark indicator. This is followed by another

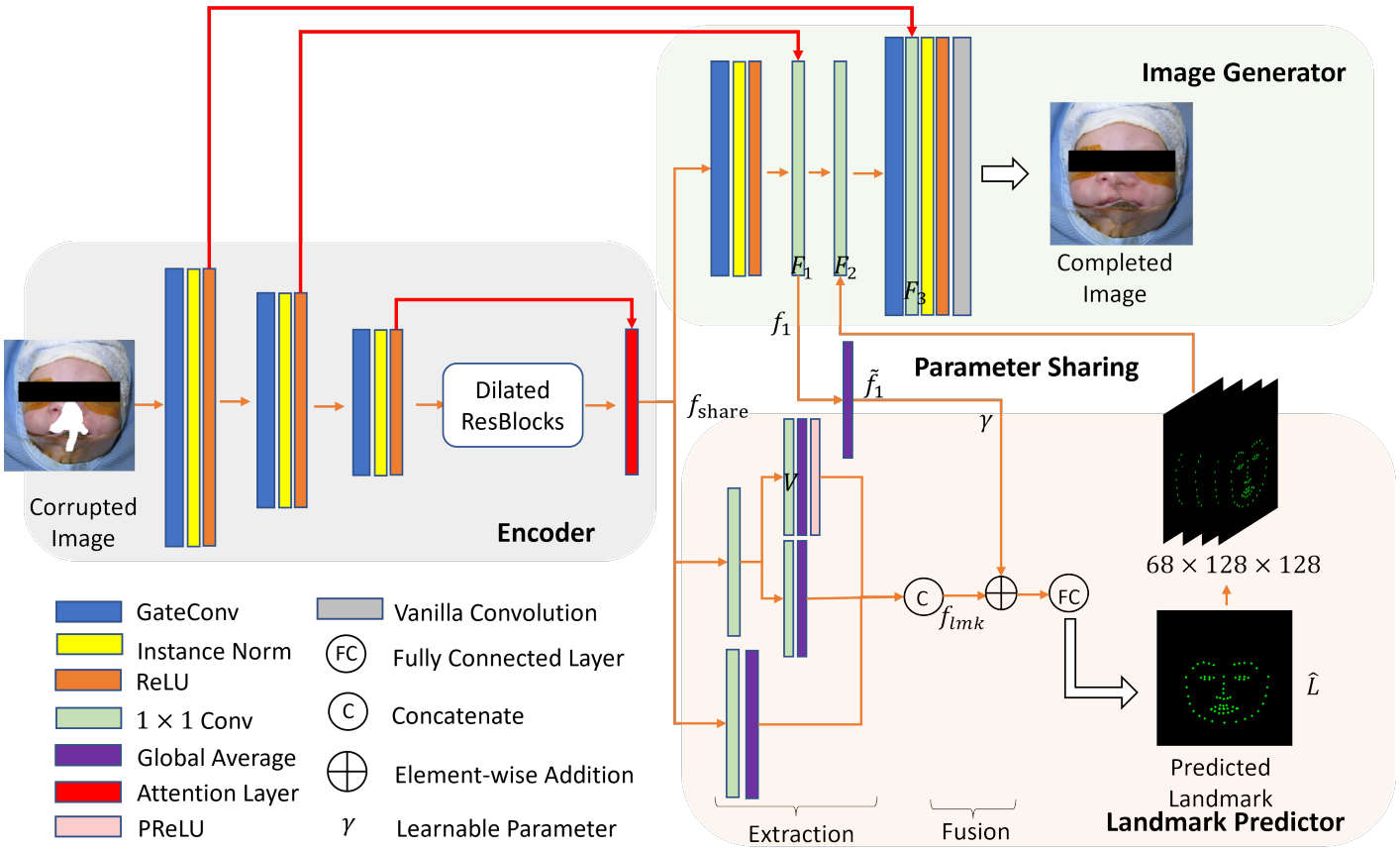

Figure 1: The overview of proposed multi-task architecture.

upsampling gated convolution block with a fusion layer F_3 and a convolutional layer to synthesise the non-cleft lip image.

The landmark predictor is used to predict the landmark of the non-cleft facial image. Following [7], the shared feature $f_{\mathbf{share}}$ is passed into different 1×1 convolution layers followed by global average pooling to extract features of different numbers of channels. The feature with the largest number of channels (i.e. V) is further passed into a PReLU [15] activation layer. These features are concatenated and fused with the image features to predict the landmark.

We develop a parameter-sharing mechanism to share information between the image generator and the landmark predictor. We implement an adaptive feature fusion algorithm, in which the image feature f_1 from the layer F_1 is passed from the image generator to the landmark predictor. This is followed by a fully connected layer to generate the 68 landmark points, $\hat{L}$:

$$\hat{L} = FC(\gamma * \tilde{f}_1 \oplus f_{lmk}), \tag{1}$$

where γ is a trainable parameter with zero initialization, $\oplus$ is element-wise addition, $\tilde{f}_1$ is obtained by passing f_1 through a global average pooling layer and f_{lmk} is the extracted landmark feature map. The predicted landmark points $\hat{L}$ are mapped into a 128×128 image corresponding to the landmark position. The image is stacked channel-wise to increase its influence (68 times in our setup), and passed from the landmark predictor back to F_2 in the image generator.

2.2 Implementation details

This multi-tasking image inpainting system is programmed in PyTorch. The main packages include numpy 1.15.4, torch 1.10.2 and torchvision 0.11.3. All of our experiments are implemented with 256×256 images and masks. We employ the imageio 2.15.0 to load and Pillow 8.4.0 to resize both images and masks. For each facial

Algorithm 1 GAN-based training for proposed Multi-task model MT_θ This is the new pseudo code

Inputs: Generated image x, image ground truth X, predicted landmark k, landmark ground truth K, irregular mask M, max number of iterations T, batch size = 4.

Output: Multi-task model MT_θ

Build dataloader.
Initialize the network.
if t < T **then**:
    Sample 4 images and corresponding landmark from dataloader.
    Sample 4 irregular mask from dataloader.
    $x, k, L_{\text{pixel}}, L_{\text{landmark}}, L_{\text{tv}}, L_{\text{style}}, L_{\text{perceptual}}, L_g, L_d \leftarrow MT_\theta(X, M, L)$.
    $L_G \leftarrow L_{\text{pixel}} + L_{\text{landmark}} + L_{\text{tv}} + L_{\text{style}} + L_{\text{perceptual}} + L_g$
    $L_D \leftarrow L_d$
    Freeze the θ_G in multi-task model, update discriminator with adversarial loss L_D.
    Freeze the θ_D in discriminator, update multi-task model with adversarial loss L_G.
    $t \leftarrow t + 1$
Save the trained Multi-task model MT_θ.

image, Face Align Network (FAN) generates 136 values to denote the x- and y- positions of the 68 landmark points.

To train the network, a GAN-based [16] training flow is employed as shown in Algorithm 1. We first apply FAN [17], a Face Align Network, to generate the ground truth landmark points from CelebA [6], following the default training and testing split of the dataset. We then apply Optuna [18] for hyperparameter tuning. Specifically, we fix all hyperparameters except the weight of landmark loss, and train our model with one epoch using Optuna to sample such weights. With the same method, we also tune other hyperparameters, such as the learning rate, the decay weight of the learning rate and batch size.

The collected dataset is used for inference. From both quantitative and qualitative results, our system generates semantically plausible non-cleft facial images [9]. The results are further evaluated by cleft lip surgeons, showcasing that our proposed network generates better images than state-of-the-arts [19, 8, 7].

The run-time cost of the proposed system is very low. Using our real-world cleft lip data, the inference step is implemented using an NVIDIA GeForce GTX 970 on a laptop, with an inference time of 200ms for a single image. This means that our trained system does not require a particularly powerful computing system to perform the inference, and a standard workstation or laptop computer can use our system. For training the network, one NVIDIA TITAN Xp is used for four days, which is typical in deep learning applications of a similar scale.

2.3 How to use

To retrieve the training dataset for this image inpainting application, users are required to download the CelebA Dataset [6] and the irregular mask dataset [13] from the respective official websites. The CelebA dataset should then be divided into a standard training set and a validation set, according to the official instruction. Additionally, the corresponding landmark points should be generated with FAN [17]. Furthermore, the irregular mask dataset should be divided into three groups according to the mask ratios (0-20%, 20-40%, 40-60%). 3,300 masks are randomly selected from each group, resulting in a total of 9,900 mask images for training. Another 200 masks are selected from each group, resulting in a total of 600 mask images for verification. For the inference step, all cleft facial images and their corresponding masks serve as the image test set and the mask test set, respectively. The user should then run the provided "./scripts/filst.py" script to generate training, test and validation set file lists, and update the information in the "config.yml" file accordingly to set the model configuration. Once the python environment has been set up using the released "requirements.txt" file, the user may proceed to run the "train.py" script for training and the "test.py" script for testing. For the inference process, although we recommend using our system with GPUs for better speed, the system is fully runnable with only CPUs. Due to the sensitivity of patient privacy, we are not allowed to upload the cleft lip data for

an online demonstration. Therefore, we show the reproducibility of our system with the images from CelebA and the irregular masks.

3. Impact overview

While our method primarily focuses on cleft lips, the uses of the implemented source code can be extended to other applications. The key idea of this software is to mask out a particular region of a face, and to employ inpainting techniques for predicting the masked area. The versatility of our system allows for the implementation of extended facial applications, such as makeup and plastic surgery prediction. To utilize these capabilities, a customized dataset is required for training, such as the Facial Beauty Database [20] or a plastic surgery facial dataset [21]. The users then need to retrain our model according to section 2.3. The resulting model can then be tested using a corresponding mask that covers specific facial components, such as nose or eyebrows, to generate the image of the subject after makeup or plastic surgery. Therefore, it can also be used for supporting plastic surgeries and makeup prediction [22] on specific facial components. This would facilitate the understanding and discussion of those operations and applications among stakeholders.

We put a particular effort in selecting a software framework that is robust against model leakage and attack [23, 4]. In particular, we propose the idea of excluding patient data in training deep learning models if possible, mitigating any privacy concerns and risk of data loss. The high-level concept of training with open data and testing with sensitive data can be employed in other machine learning applications to protect data privacy, particularly those in the healthcare domain or involving people of vulnerable groups.

In theory, our system is also capable of synthesising cleft facial images from non-clelf lip ones. In practice, due to the wide variety of cleft lip conditions, training such a system would require a large dataset of cleft images, which is currently not available. Should there be enough data (and we only need the lip area to protect patients' privacy), this system can be used to generate synthetic cleft lip facial images, which enable the training of machine learning algorithms. As the data is artificially created, there is no privacy or model leakage concern, and an unlimited amount of samples can be created. This aligns with the recent trend of using computer graphics techniques to mock up real-world data [24], facilitating the training of machine learning systems for patient-related applications [25]. Since the beginning of this research, there is raising awareness from both UK universities and hospitals in collecting cleft lip data for research purposes. We believe our vision will be made possible in the future.

4. Conclusion

This work implements a multi-task image inpainting model to predict non-cleft lip facial images from cleft lip ones. We make an important software engineering decision to implement the system under an inpainting framework, which does not require patient data for training and mitigates model leakage risks. We design and develop a multi-task neural network that co-predicts a facial image and the corresponding facial landmarks, and we find that the two tasks support each other. Apart from detailing the design and implementation details of our software, we also discuss its impact within and beyond cleft lip applications. The source code is now publicly released on CodeOcean and Github.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] "Cleft lip and palate," 2019. [Online]. Available: https://www.nhs.uk/conditions/cleft-lip-and-palate/

[2] W. Wellens and V. Poorten, "Keys to a successful cleft lip and palate team," *B ENT*, p. 3, 2006.

[3] D. G. Mosmuller, L. M. Mennes, C. Prahl, G. J. Kramer, M. A. Disse, G. M. Van Couwelaar, B. N. Frank, and J. Don Griot, "The development of the cleft aesthetic rating scale: a new rating scale for the assessment

of nasolabial appearance in complete unilateral cleft lip and palate patients," *The Cleft Palate-Craniofacial Journal*, vol. 54, no. 5, pp. 555–561, 2017.

[4] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[6] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[7] Y. Yang, X. Guo, J. Ma, L. Ma, and H. Ling, "Lafin: Generative landmark guided face inpainting," *arXiv preprint arXiv:1911.11394*, 2019.

[8] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[9] S. Chen, A. Atapour-Abarghouei, J. Kerby, E. S. L. Ho, D. C. G. Sainsbury, S. Butterworth, and H. P. H. Shum, "A feasibility study on image inpainting for non-cleft lip generation from patients with cleft lip," in *Proceedings of the 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics*. IEEE, 2022.

[10] N. Nozawa, H. P. H. Shum, Q. Feng, E. S. L. Ho, and S. Morishima, "3d car shape reconstruction from a contour sketch using gan and lazy learning," *Visual Computer*, vol. 38, no. 4, pp. 1317–1330, 2022.

[11] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4471–4480.

[12] Y. Li, J. Cheng, H. Mei, H. Ma, Z. Chen, and Y. Li, "Clpnet: cleft lip and palate surgery support with deep learning," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 3666–3672.

[13] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 85–100.

[14] C. Zheng, T.-J. Cham, and J. Cai, "Pluralistic image completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1438–1447.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[16] J. Goodfellow Ian, M. Pouget-Abadie, B. Mirza, D. Xu, S. Warde-Farley, A. Ozair, Y. Courville, and Bengio, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of NeurIPS*, 2014, p. 2672–2680.

[17] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.

[18] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[19] X. Guo, H. Yang, and D. Huang, "Image inpainting via conditional texture and structure dual generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 134–14 143.

[20] L. Zhang, H. P. Shum, L. Liu, G. Guo, and L. Shao, "Multiview discriminative marginal metric learning for makeup face verification," *Neurocomputing*, vol. 333, pp. 339–350, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231218314577

[21] C. Rathgeb, D. Dogan, F. Stockhardt, M. De Marsico, and C. Busch, "Plastic surgery: An obstacle for deep face recognition?" in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 3510–3517.

[22] D. Organisciak, E. S. L. Ho, and H. P. H. Shum, "Makeup style transfer on low-quality images with weighted multi-scale attention," in *Proceedings of the 2020 International Conference on Pattern Recognition*, ser. ICPR '20, Jan 2020, pp. 6011–6018.

[23] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction {APIs}," in *25th USENIX security symposium (USENIX Security 16)*, 2016, pp. 601–618.

[24] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger, W. Muller-Felber, and A. S. Schroeder, "Learning an infant body model from RGB-D data for accurate full body motion analysis," in *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Sep. 2018.

[25] H. Zhang, H. P. H. Shum, and E. S. L. Ho, "Cerebral palsy prediction with frequency attention informed graph convolutional networks," in *Proceedings of the 2022 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, ser. EMBC '22. IEEE, 2022.