A Review Study on Investing Deep Learning in The Field of Virtual Dressing

دراسة مرجعية عن استثمار التعلم العميق في مجال الإلباس الافتراضي

Sarah SHIKH ALSHABAB : Syrian Virtual University_ Damascus_ Syria.
Chadi ALBITAR : Higher Institute for Applied Sciences and Technology\ Syrian Virtual University_ Damascus_ Syria.
Corresponding: sarah_151923@svuonline.org

ABSTRACT

This study provides a comprehensive review and detailed analysis of existing works based on deep learning for 3D pose estimation, transforming 2D clothing items into a pe son's image using artificial intelligence (AI) techniques to develop virtual fitting room. The current results are presented and discussed, and we summarize the advantages and disadvantages of these existing methods and provide an in-depth understanding of the field. Besides, we present the commonly used standard datasets upon which comprehensive study is conducted for comparison and analysis.
Keywords: Deep Learning, CNN, Virtual Try On, AI, CAPE, SMPL.

الملخص

تقدم هذه الدراسة مراجعة شاملة وتحليلاً مفصلاً للأعمال الحالية القائمة على التعلم العميق لتقدير الوضع ثلاثي الأبعاد، وتحويل عناصر الملابس ثنائية الأبعاد إلى صورة الشخص باستخدام تقنيات الذكاء الاصطناعي (AI) لتطوير غرفة القياس الافتراضية. يتم عرض النتائج الحالية ومناقشتها، وتلخيص مزايا وعيوب هذه الأساليب وتقديم فهم معمق لهذا المجال. إضافة إلى ذلك، نعرض مجموعات البيانات المعيارية الشائعة الاستخدام والتي تجري عليها دراسة شاملة للمقارنة والتحليل.
الكلمات المفتاحية: التعلم العميق، الشبكة العصبية التلافيفية، التجريب الافتراضي للملابس، الذكاء الاصطناعي، الترميز التلقائي لشخص مرتدٍ ملابس، النموذج الخطي متعدد الأشخاص.

INTRODUCTION

For quite some time now, the world has been witnessing a radical shift towards the virtual world in various fields, where financial and commercial transactions and many other fields are now carried out entirely through electronic applications, even official and government transactions, most of which are also took place via Internet. The recent circumstances that the world has witnessed (Corona, wars...) have greatly demonstrated the importance of virtual reality, especially artificial intelligence and deep learning, as these circumstances imposed the importance of many human activities not being linked to the real presence of people. One of the most important activities affected by this new trend is commercial activities, as the localization of e-commerce technology has become of great importance and necessity. On the other hand, the new information fields of artificial intelligence and deep learning provide many advanced tools to localize this technology. There is no doubt that the development in the field of digital transformation has made great strides in many fields, and many activities have now been carried out entirely over Internet using appropriate applications. It is natural that the expansion of digital transformation's control over commercial activity is related to the development of tools and algorithms that enable us creating applications suitable for each field. Another important justification for this study is the lack of programs that support remote shopping in an interactive way that simulates the process of traditional presence in the store. Based on a review of previous studies in the field of virtual clothing [1], it has become clear that most of the proposed models have limitations related to the data on the target clothing and the people wearing those clothing. In this review, we conduct a comparative study on investing deep

learning and deep fake algorithms to choose the best three-dimensional digital models of colored clothing and reveal their attributes [2][3][4]. This topic needs to address several axes, the most important of which are studying the digital models used in deep learning algorithms, studying algorithms for human body detection and determining reference points that correspond to digital models [5] [6], and studying deep fake algorithms to give a real impression about clothing these models [7].

General Methodology

According to the studies, virtual fitting rooms adopt AI techniques, deep learning algorithms, and neural networks to develop virtual try-on clothes. One of the most widely used learning algorithms is convolutional neural networks (CNN) (Figure 1), which is used to prepare 3D networks of the human body and can be generalized to different body shapes and poses.

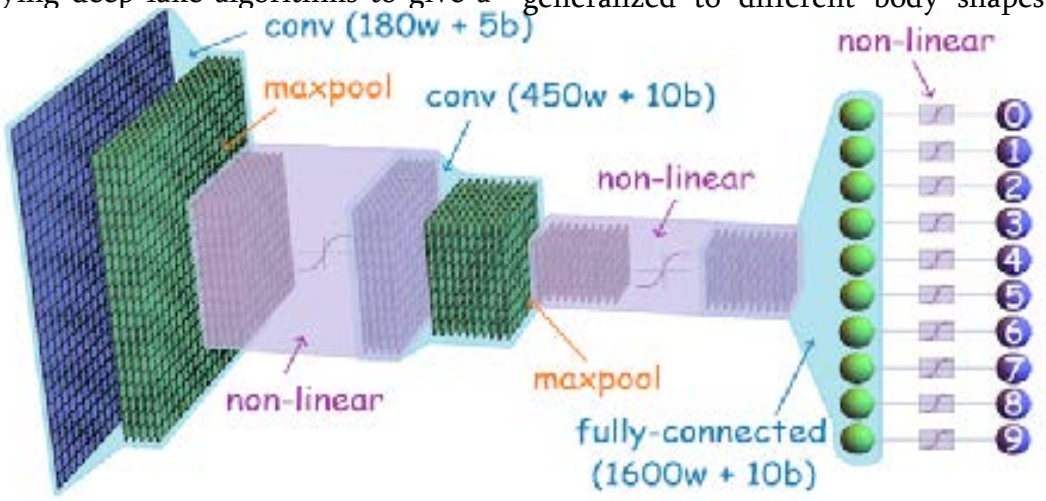

Figure 1: An example of the main layers of a CNN

The general architecture of generative models based on CNN consists of:

- Dataset including multiple meshes of clothed human scans, different types of outfits, and different poses.
- Encoder
- Decoder
- Discriminator

CNN is compatible with the famous Skinned Multi-Person Linear Model (SMPL) body model, which is defined as a realistic 3D model of the human body based on extracting and blending shapes through thousands of 3D body scans [8]. SMPL is more accurate than other models and is compatible with existing graphic pipelines. It is viewed as skinned vertex-based model that accurately represents a wide variety of body shapes in natural human poses. One of the latest network models is the Clothed Auto Person Encoding (CAPE) that provides SMPL mesh-registered 4D scans of people wearing cloths, along with recorded scans of real body shapes under clothing [9]. CAPE adds two stages to the general architecture as follows (Figure 2):

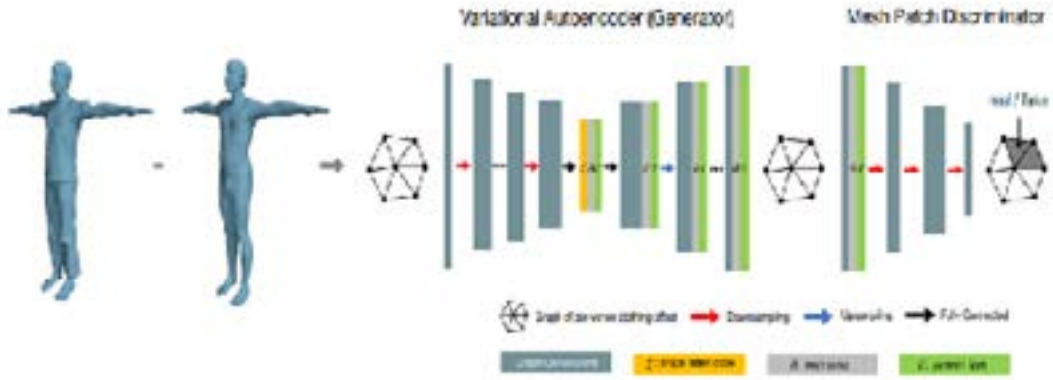

Figure 2: CAPE network architecture [9]

Condition Module

For pose θ, CAPE removes non-clothing parameters, e.g. head, hands, fingers, feet and toes, resulting in 14 valid joints from the body. The pose parameters from each joint are represented by the flattened rotational matrix. This results in the overall pose parameter R, that feeds into a small fully-connected network θ. The clothing type C refers to the type of "outfit", i.e. a combination of upper body clothing and lower body clothing. As the types of clothing are discrete by nature, CAPE represents them using a one-hot vector, C, and feeds it into a linear layer.

Conditional Residual Block (CResBlock):

CAPE adopts the residual block (ResBlock) from Kolotouros et al. [11] which includes ensemble normalization [13], nonlinearity, a graph convolutional layer and a graph linear layer. After input to the residual block, CAPE appends a state vector to each input node along the feature channel. ResBlock is the graph residual block from outputs on each node.

RELATED WORK

The related works can be classified into three axes:

Reconstructing 3D Humans

Reconstruction of 3D humans' bodies from 2D images and videos is a classical computer vision problem. Most approaches [32 ,18] output 3D body meshes from images, but not clothing. This ignores image evidence that may be useful. To reconstruct clothed bodies, methods use volumetric [30 ,20] or bi-planar depth representations [12] to model the body and garments as a whole. While these methods deal with arbitrary clothing topology and preserve a high level of details, the reconstructed clothed body is not parametric, which means that the pose, shape, and clothing of the reconstruction cannot be controlled or animated. Another group of methods is based on SMPL [25, 31]. They represent clothing as an offset layer from the underlying body as proposed in Cloth Cap [28] as shown in (Figure 3). These methods can change the pose and shape of the reconstruction using the deformation model of SMPL. This assumes that clothing deforms like an undressed human body; i.e. clothing shape and wrinkles do not change as a function of pose.

Figure 3: ClothCap approach [28]

Parametric Models For 3D Bodies And Clothes

Statistical 3D human body models learned from 3D body scans [32 ,23] capture body shape and pose and they are an important building block for multiple applications. Most of the time, people are dressed and these models do not represent clothing. In addition, clothes deform as we move, producing changing wrinkles at multiple spatial scales. While clothing models learned from real data exist, few can be generalized to new poses. For example, Neophytou and Hilton [34] proposed to learn a layered garment model from dynamic sequences, but generalization to novel poses is not demonstrated. Yang et al. [27] trained a neural network to regress a PCA-based representation of clothing, but they proved the generalization on the same sequence or on the same subject. Lahner et al. [29] proposed to learn a garment-specific pose-deformation model by regressing low-frequency Principal Components Analysis (PCA) components and high frequency normal maps. While the visual quality was good, the model is garment-specific and does not provide a solution for full-body clothing. Similarly, Alldieck et al. [25] as shown in (Figure 4) used displacement maps with a UV parametrization to represent surface geometry, but the result was only static. Wang et al. [24] allowed manipulation of clothing with sketches in a static pose. The Adam model proposed in [23] can be considered clothed but the shape is very smooth and not pose dependent.

Clothing models have been learned from physics simulation of clothing [33 ,19], but the visual reliability was limited by the quality of the simulations.

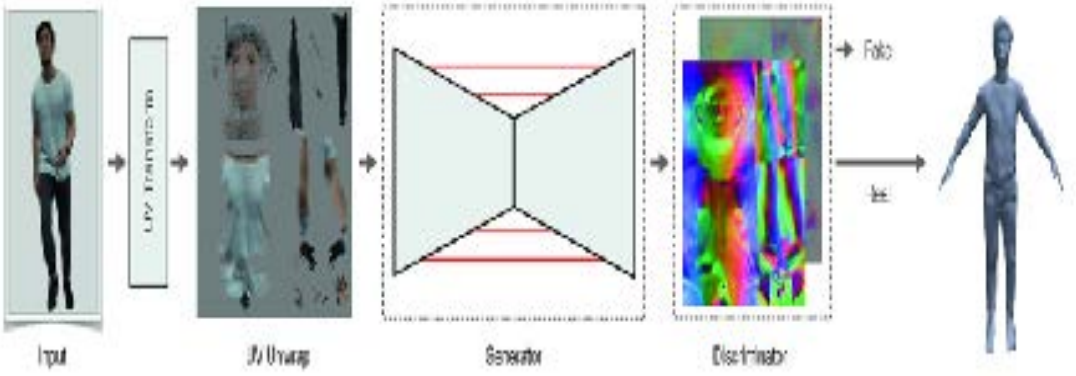

Figure 4: Displacement maps with a UV parametrization to represent surface geometry [25]

Generative Models on 3D Meshes

CAPE model predicts clothing displacements on the graph defined by the SMPL mesh using graph convolutions [10]. However, there is an extensive recent literature on methods and applications of graph convolutions such as [26 ,21]. Most relevant here, Ranjan et al. [26] proposed to learn a convolutional auto encoder using graph convolutions with mesh down- and up-sampling layers [13]. Although it worked well for faces, the mesh sampling layer made it difficult to capture the local details, which are key in clothing, while CAPE captures local details by extending the PatchGAN [22] architecture to 3D meshes (Figure 2).

Comparison and Discussion:

comparing between public 3D clothed human datasets according to six points (Captured, Available Body Shape, Registered, Large Pose Variation, Motion Sequence, High Quality Geometry) leads to the following results:

- Inria Dataset presented an approach to automatically estimate the human body shape under motion based on a 3D input sequence showing a dressed person in possibly loose clothing.
- It has no registered 3D meshes of clothed human scans. It has limited variation in pose and low quality geometry [14] (Figure 5).
- BUFF Dataset introduced a method to estimate a detailed body shape under clothing from a sequence of 3D scans. This method exploits the information in a sequence by merging all clothed registrations into a single frame as shown in (Figure 6).
- BUFF Dataset is like Inria but has high quality geometry [15].

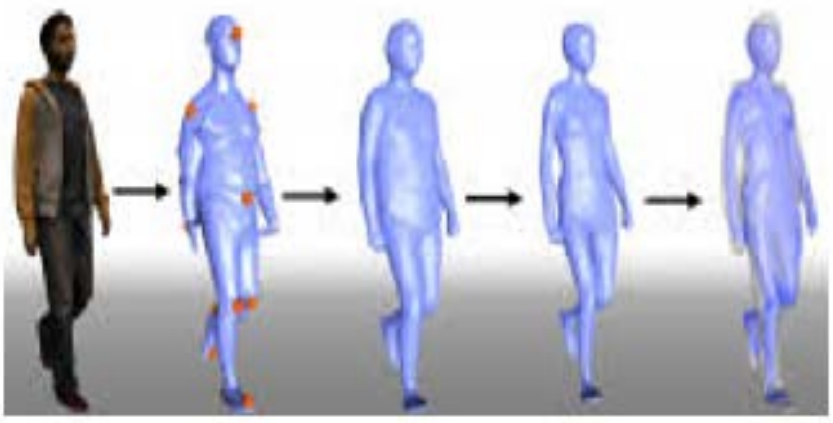

Figure 5: Inria Dataset approach [14]

Figure 6: Qualitative pose estimation results on BUFF dataset [15] Left to right: scan, Yang et al. [27], BUFF result

Adobe Dataset key insight is to use skeletal pose estimation for gross deformation followed by iterative nonrigid shape matching to fit the image data. ●Adobe Dataset does not have human body shapes. It has limited variation in pose and low quality geometry [12]. 3D people dataset proposed a new algorithm to perform spherical parameterizations of elongated body parts, and introduced an end-to-end network to estimate human body and clothing shape from single images, without relying on parametric models, (Figure 7). ●3D People Dataset contains all the points but lacks the ability to capture and

Figure 7: Annotations of the 3D People Dataset [10]

●CAPE Dataset contains all the points, as shown in (Figure 8). Given a SMPL body shape and pose (a), CAPE adds clothing by randomly sampling from a learned model (b, c), and can generate different clothing types — shorts in (b, c) vs. long-pants in (d). The generated clothed humans can generalize to diverse body shapes (e) and body poses (f).

Figure 8: CAPE model for clothed humans [9]

Characterized by accurate alignment, consistent mesh topology, ground truth body shape scans, and a large variation of poses, CAPE features make it suitable not only for studies on human body and clothing, but also for the evaluation of various Graph CNNs. However, CAPE differs from the other methods in learning a parametric model of how clothing deforms with pose. Furthermore, all the methods of Parametric models are regressors that produce single point estimates. In contrast, CAPE is generative, which allows to sample clothing. A conceptually different approach infers the parameters of a physical clothing model from 3D scan sequences was proposed in [17]. This can be generalized to novel poses, but the inference problem is difficult and, unlike CAPE, the resulting physics simulator is not differentiable with respect to the parameters. Since the presented results confirmed the investment of deep fake in the field of virtual dressing, we are on the way towards investing in artificial intelligence and neural networks in this field, in parallel with the very rapid development in electronic clothing marketing, and in response to the requirements of the local and global market in this field.

CONCLUSION

Although the results of the previous studies are significant, the research remains open due to the limitations of the approved methods, which can be summarized as follows:

- The limitation of the offset representation for clothing such as skirts and open jackets differ from the body topology and cannot be represented by offsets as shown in (Figure 9). Mittens and shoes can technically be modelled by the offsets, but their geometry is sufficiently different from that of fingers and toes, making this impractical.
- Dynamics issues: the approved models take a long time to train the algorithms, because the generated clothing depends on pose, and does not depend on dynamics. This does not cause a severe problem for most slow motions but cannot be generalized to faster motions. Future work will address models of clothing, but instead of scanning the entire body, we propose to consider sufficient features on the body to estimate the shape of the body which may be sufficient for virtual dressing. Therefore, one will not need to photograph the body completely naked or with a minimum amount of clothing, as we propose to conduct an investigation of certain points of the body. Another restriction can be added to make the work easier is to deal with the human body as two parts, upper and lower, based on the International Standard Organization (ISO) standards to assign points to the human body [16].

Figure 9: Qualitative results on fashion images [9]
SMPL [8] results are shown in green, CAPE results are in blue

REFERENCES

1. Tatariants ,Maksym., «Deep Learning for Virtual Try On Clothes – Challenges and Opportunities», 2020.
2. Jeon Youngseung., Jin Seungwan., Han Kyungsik., (Apr 2021 ,19),»FANCY: Human-centered, Deep Learning-based Framework for Fashion Style Analysis “, The Web Conference 2021.
3. Liu Jingmiao., Ren Yu., Qin Xiaotong., (September ,8 2021),» Study on 3D Clothing Color Application Based on Deep Learning-Enabled Macro-Micro Adversarial Network and Human Body Modeling «, Computational Intelligence and Neuroscience, Volume 2021.
4. https://medium.com/analytics-vidhya/deep-learning-infashion-industry-dcb897ac3c,33 2021. «Deep Learning in Fashion Industry».
5. Bianco Valentina., “BODY DETECTION USING COMPUTER VISION”, (Oct 2021 ,19). https://blog.xmartlabs.com/blog/computer-vision-techniques-for-body-detection/
6. Wang Jinbao., Tan Shuji., Zhen Xiantong., Xu Shuo., Zheng Feng., He Zhenyu., Shao Ling.”Deep 3D human pose estimation: A review”, Computer Vision and Image Understanding (2021).
7. Whittakera Lucas., Letherenb Kate., Mulcahyc Rory,» The Rise of Deepfakes: A conceptual framework and research agenda for marketing», Australasian marketing journal. 2021.
8. https://star.is.tue.mpg.de/
9. https://cape.is.tue.mpg.de/
10. Albert Pumarola, Jordi Sanchez, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3DPeople: Modeling the Geometry of Dressed Humans. In The IEEE International Conference on Computer Vision (ICCV), 2019.
11. Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for singleimage human shape reconstruction. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
12. Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popovic. Articulated mesh animation from multi- ´view silhouettes. In ACM Transactions on Graphics (TOG), volume 27, page 97. ACM, 2008.
13. Yuxin Wu and Kaiming He. Group normalization. In The European Conference on Computer Vision (ECCV). Springer, 2018.
14. Jinlong Yang, Jean-Sebastien Franco, Franck H ´ etroy-´Wheeler, and Stefanie Wuhrer. Estimation of human body shape in motion with wide clothing. In European Conference on Computer Vision. Springer, 2016.
15. Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
16. https://www.iso.org/obp/ui/en/#iso:std:iso:1-:8559:ed-1:v1:en
17. Carste Stoll, Jurgen Gall, Edilson de Aguiar, Sebastian Thrun, and ¨ Christian Theobalt. Video-based reconstruction of animatable human characters. In ACM SIGGRAPH ASIA, 2010.
18. David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. FACSIMILE: Fast and Accurate Scans From an Image in Less Than a Second. In The IEEE International Conference on Computer Vision (ICCV), 2019.
19. Igor Santesteban, Miguel A. Otaduy, and Dan Casas. Learning-Based Animation of Clothing for Virtual Try-On. Computer Graphics Forum (Proc. Eurographics), 2019.
20. Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In The IEEE International Conference on Computer Vision (ICCV), 2019.
21. Nitika Verma, Edmond Boyer, and Jakob Verbeek. Feastnet: Featuresteered graph convolutions for 3D shape analysis. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
22. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
23. Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total Capture: A 3D deformation model for tracking faces, hands, and bodies. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
24. Tuanfeng Y Wang, Duygu Ceylan, Jovan Popovic, and Niloy J Mitra. ´Learning a shared shape space for multimodal garment design. In ACM SIGGRAPH ASIA, 2018.
25. Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2Shape: Detailed full human body geometry from a single image. In The IEEE International Conference on Computer Vision (ICCV), 2019.
26. Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3D faces using convolutional mesh autoencoders. In The European Conference on Computer Vision (ECCV), 2018.
27. Jinlong Yang, Jean-Sebastien Franco, Franck H ´ etroy-Wheeler, and Stefanie Wuhrer. Analyzing clothing layer deformation statistics of 3D human motions. In The European Conference on Computer Vision (ECCV). Springer, 2018.
28. Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4D clothing capture and retargeting. ACM Transactions on Graphics (TOG), 2017 ,73:(4)36.
29. Zorah Lahner, Daniel Cremers, and Tony Tung. DeepWrinkles: Ac-curate and realistic clothing modeling. In European Conference on Computer Vision. Springer, 2018.
30. Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3D human reconstruction from a single image. In The IEEE International Conference on Computer Vision (ICCV), 2019.
31. Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
32. Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
33. Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. The Virtual Tailor: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
34. Alexandros Neophytou and Adrian Hilton. A layered

AUTHOR CONTRIBUTIONS:

Conceptualization: Eng. Sarah Shikh Alshabab, Dr. Chadi Albitar
Methodology: Eng. Sarah Shikh Alshabab, Dr. Chadi Albitar
Investigation: Eng. Sarah Shikh Alshabab, Dr. Chadi Albitar
Project administration: Dr. Chadi Albitar
Supervision: Dr. Chadi Albitar
Writing – original draft: Eng. Sarah Shikh Alshabab
Writing – review & editing: Dr. Chadi Albitar
Competing interests: "Authors declare that they have no competing interests."
Data and materials availability: "All data are available in the main text or the supplementary materials."