Enhancing Automatic Placenta Analysis Through Distributional Feature Recomposition in Vision-Language Contrastive Learning

Yimu Pan[1](✉), Tongan Cai[1], Manas Mehta[1], Alison D. Gernand[1],
Jeffery A. Goldstein[2], Leena Mithal[3], Delia Mwinyelle[4], Kelly Gallagher[1],
and James Z. Wang[1]

[1] The Pennsylvania State University, University Park, PA, USA
ymp5078@psu.edu
[2] Northwestern University, Chicago, IL, USA
[3] Lurie Children's Hospital, Chicago, IL, USA
[4] The University of Chicago, Chicago, IL, USA

Abstract. The placenta is a valuable organ that can aid in understanding adverse events during pregnancy and predicting issues post-birth. Manual pathological examination and report generation, however, are laborious and resource-intensive. Limitations in diagnostic accuracy and model efficiency have impeded previous attempts to automate placenta analysis. This study presents a novel framework for the automatic analysis of placenta images that aims to improve accuracy and efficiency. Building on previous vision-language contrastive learning (VLC) methods, we propose two enhancements, namely Pathology Report Feature Recomposition and Distributional Feature Recomposition, which increase representation robustness and mitigate feature suppression. In addition, we employ efficient neural networks as image encoders to achieve model compression and inference acceleration. Experiments validate that the proposed approach outperforms prior work in both performance and efficiency by significant margins. The benefits of our method,

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (NIH) under award number R01EB030130 and the College of Information Sciences and Technology of The Pennsylvania State University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work used computing resources at the Pittsburgh Supercomputer Center through allocation IRI180002 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants Nos. 2138259, 2138286, 2138307, 2137603, and 2138296.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_12.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 116–126, 2023.
https://doi.org/10.1007/978-3-031-43987-2_12

including enhanced efficacy and deployability, may have significant implications for reproductive healthcare, particularly in rural areas or low- and middle-income countries.

Keywords: Placenta Analysis · Representation · Vision-Language

1 Introduction

World Bank data from 2020 suggests that while the infant mortality rate in high-income countries is as low as 0.4%, the number is over ten times higher in low-income countries (approximately 4.7%). This stark contrast underlines the necessity for accessible healthcare. The placenta, as a vital organ connecting the fetus to the mother, has discernable features such as meconium staining, infections, and inflammation. These can serve as indicators of adverse pregnancy outcomes, including preterm delivery, growth restriction, respiratory or neurodevelopmental conditions, and even neonatal deaths [9].

In a clinical context, these adverse outcomes are often signaled by morphological changes in the placenta, identifiable through pathological analysis [19]. Timely conducted placental pathology can reduce the risks of serious consequences of pregnancy-related infections and distress, ultimately improving the well-being of newborns and their families. Unfortunately, traditional placenta pathology examination is resource-intensive, requiring specialized equipment and expertise. It is also a time-consuming task, where a full exam can easily take several days, limiting its widespread applications even in developed countries. To overcome these challenges, researchers have been exploring the use of automatic placenta analysis tools that rely on photographic images. By enabling broader and more timely placental analysis, these tools could help reduce infant fatalities and improve the quality of life for families with newborns.

Related Work. Considerable progress has been made in segmenting [17,20,23] and classifying [1,8,10,13,15,21,26] placenta images using histopathological, ultrasound, or MRI data. However, these methods are dependent on expensive and bulky equipment, restricting the accessibility of reproductive healthcare. Only limited research has been conducted on the gross analysis of post-birth placenta photographs, which have a lower equipment barrier. AI-PLAX [4] combines handcrafted features and deep learning, and a more recent study [29] relies on deep learning and domain adaptation. Unfortunately, both are constrained by issues such as data scarcity and single modality, which hinder their robustness and generalizability. To address these, Pan et al. [16] incorporated vision-and-language contrastive learning (VLC) using pathology reports. However, their method struggles with variable-length reports and is computationally demanding, making it impractical for low-resource communities.

With growing research in vision-and-language and contrastive learning [18,28], recent research has focused on improving the performance and efficiency of VLC approaches. They propose new model architectures [2,24], better visual representation [7,27], loss function design [14,16], or sampling strategies [5,12].

However, these methods are still not suitable for variable-length reports and are inefficient in low-resource settings.

Our Contributions. We propose a novel framework for more accurate and efficient computer-aided placenta analysis. Our framework introduces two key enhancements: Pathology Report Feature Recomposition, a first in the medical VLC domain that captures features from pathology reports of variable lengths, and Distributional Feature Recomposition, which provides a more robust, distribution-aware representation. We demonstrate that our approach improves representational power and surpasses previous methods by a significant performance margin, without additional data. Furthermore, we boost training and testing efficiency by eliminating the large language model (LLM) from the training process and incorporating more efficient encoders. To the best of our knowledge, this is the first study to improve both the efficiency and performance of VLC training techniques for placenta analysis.

2 Dataset

We use the exact dataset from Pan et al. [16] collected using a professional photography instrument in the pathology department of the Northwestern Memorial Hospital (Chicago) from 2014 to 2018 and an iPad in 2021. There are three parts of the dataset: 1) the pre-training dataset, containing 10,193 image-and-text pairs; 2) the primary fine-tuning dataset, comprising 2,811 images labeled for five tasks: *meconium*, *fetal inflammatory response* (FIR), *maternal inflammatory response* (MIR), and *histologic chorioamnionitis*, and *neonatal sepsis*; and 3) the iPad evaluation dataset, consisting of 52 images from an iPad labeled for MIR and *clinical chorioamnionitis.* As with the original study, we assess the effectiveness of our method on the primary dataset, while utilizing iPad images to evaluate the robustness against distribution shifts. All images contain the fetal side of a placenta, the cord, and a ruler for scale. The pre-training data is also accompanied by a corresponding text sequence for the image containing a part of the corresponding pathology report as shown in Fig. 1. A detailed breakdown of the images is provided in the supplementary materials.

3 Method

This section aims to provide an introduction to the background, intuition, and specifics of the proposed methods. An overview is given in Fig. 1.

3.1 Problem Formulation

Our tasks are to train an encoder to produce placenta features and a classifier to classify them. Formally, we aim to learn a function f^v using a learned function f^u, such that for any pair of input $(\mathbf{x}_i, \mathbf{t}_i)$ and a similarity function `sim`, we have

$$\texttt{sim}(\mathbf{u}_i, \mathbf{v}_i) > \texttt{sim}(\mathbf{u}_i, \mathbf{v}_j),\ i \neq j\ , \tag{1}$$

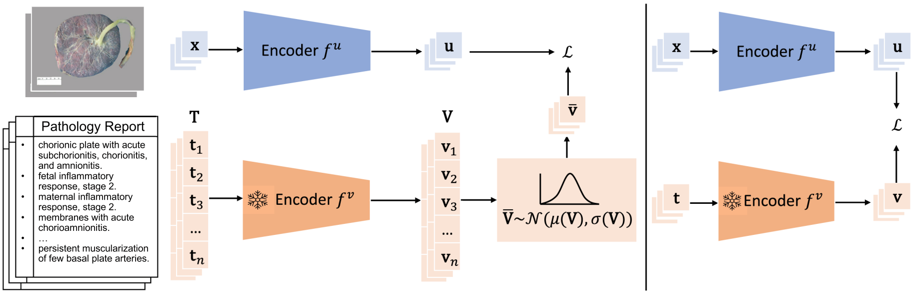
Pathology Report
chorionic plate with acute subchorionitis, chorionitis, and amnionitis.
fetal inflammatory response, stage 2.
maternal inflammatory response, stage 2.
membranes with acute chorioamnionitis.
...
persistent muscularization of few basal plate arteries.
x
Encoder f^u
u
$\mathcal{L}$
$\bar{\mathbf{v}}$
T
$\mathbf{t}_1$
$\mathbf{t}_2$
$\mathbf{t}_3$
...
$\mathbf{t}_n$
Encoder f^v
V
$\mathbf{v}_1$
$\mathbf{v}_2$
$\mathbf{v}_3$
...
$\mathbf{v}_n$
$\bar{\mathbf{V}} \sim \mathcal{N}(\mu(\mathbf{V}), \sigma(\mathbf{V}))$
x
Encoder f^u
u
$\mathcal{L}$
t
Encoder f^v
v

Fig. 1. A diagram illustrating the difference between the proposed approach (left) and the traditional VLC approach (right). $\mathbf{x}$ and $\mathbf{t}$ are images and text inputs, respectively. One sample input image and text are shown on the left. The loss function is defined as $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda\tilde{\ell}_i^{(u\to v)} + (1-\lambda)\tilde{\ell}_i^{(v\to u)}\right)$, following the notations in Sec. 3.

where $\mathtt{sim}(\mathbf{u}, \mathbf{v})$ represents the cosine similarity between the two feature vectors $\mathbf{u} = f^u(\mathbf{x})$, $\mathbf{v} = f^v(\mathbf{t})$. The objective function for achieving inequality (1) is:

$$\ell_i^{(v\to u)} = -\log \frac{\exp(\mathtt{sim}(\mathbf{u}_i, \mathbf{v}_i)/\tau)}{\sum_{k=1}^{N}\exp(\mathtt{sim}(\mathbf{u}_i, \mathbf{v}_k)/\tau)}\ , \tag{2}$$

where τ is the temperature hyper-parameter and N is the mini-batch size.

To train a classifier, we aim to learn a function f_t^c using the learned function f^v for each task $t \in [1:T]$, such that for a pair of input $(\mathbf{x}_i, l_i^t)$, $f_t^c(f^v(\mathbf{x}_i)) = l_i^t$.

3.2 Pathology Report Feature Recomposition

Traditional VLC approaches for medical image and text analysis, such as ConVIRT [28], encode the entire natural language medical report or electronic health record (EHR) associated with each patient into a single vector representation using a language model. However, solely relying on a pre-trained language model presents two significant challenges. First, the encoding process can result in suppression of important features in the report as the encoder is allowed to ignore certain placental features to minimize loss, leading to a single dominant feature influencing the objective (1), rather than the consideration of all relevant features in the report. Second, the length of the pathology report may exceed the capacity of the text encoder, causing truncation (e.g., a BERT [6] usually allows 512 sub-word tokens during training). Moreover, recent LLMs may handle text length but not feature suppression. Our method seeks to address both challenges simultaneously.

Our approach addresses the limitations of traditional VLC methods in the medical domain by first decomposing the placenta pathology report into set $\mathbf{T}$ of arbitrary size, where each $\mathbf{t}_i \in \mathbf{T}$ represents a distinct placental feature; the individual items depicted in the pathology report in Fig. 1 correspond to distinct placental features. Since the order of items in a pathology report does not impact its integrity, we obtain the set of vector representations of the features $\mathbf{V}$ using an expert language model f^v, where $\mathbf{v}_i = f^v(\mathbf{t}_i)$ for $\mathbf{v}_i \in \mathbf{V}$. These resulting vectors are weighted equally to recompose the global representation (see Fig. 1), $\bar{\mathbf{v}} = \sum_{\mathbf{v}\in\mathbf{V}} \mathbf{v}$, which is subsequently used to calculate the cosine similarity $\texttt{sim}(\mathbf{u}, \bar{\mathbf{v}})$ with the image representation $\mathbf{u}$. The recomposition of feature vectors from full medical text enables the use of pathology reports or EHRs of any length and ensures that all placental features are captured and equally weighted, thereby improving feature representation. Additionally, our approach reduces computational resources by precomputing text features, eliminating the need for an LLM in training. Moreover, it is adaptable to any language model.

3.3 Distributional Feature Recomposition

Since our pathology reports are decomposed and encoded as a set of feature vectors, to ensure an accurate representation, it is necessary to consider potential limitations associated with vector operations. In the context of vector summation, we anticipate similar representations when two sets differ only slightly. However, even minor changes in individual features within the set can significantly alter the overall representation. This is evident in the substantial difference between $\bar{\mathbf{v}}_1$ and $\bar{\mathbf{v}}_2$ in Fig. 2, despite $\mathbf{V}_1$ and $\mathbf{V}_2$ differing by only one vector magnitude. On the other hand, two distinct sets may result in the same representation, as shown by $\bar{\mathbf{v}}_1$ and $\bar{\mathbf{v}}_3$ in Fig. 2, even when the individual feature vectors have drastically different meanings. Consequently, it is crucial to develop a method that ensures $\texttt{sim}(\mathbf{V}_1, \mathbf{V}_2) > \texttt{sim}(\mathbf{V}_1, \mathbf{V}_3)$.

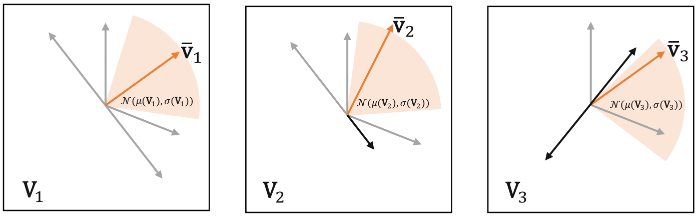
$\bar{\mathbf{v}}_1$
$\mathcal{N}(\mu(\mathbf{V}_1), \sigma(\mathbf{V}_1))$
$\mathbf{V}_1$
$\bar{\mathbf{v}}_2$
$\mathcal{N}(\mu(\mathbf{V}_2), \sigma(\mathbf{V}_2))$
$\mathbf{V}_2$
$\bar{\mathbf{v}}_3$
$\mathcal{N}(\mu(\mathbf{V}_3), \sigma(\mathbf{V}_3))$
$\mathbf{V}_3$

Fig. 2. A diagram illustrating the idea of the proposed distributional feature recomposition. $\bar{\mathbf{v}}_i$ denotes the point estimate sum of the placenta pathological text vectors set $\mathbf{V}_i$. $\mathcal{N}(\mu(\mathbf{V}_i), \sigma(\mathbf{V}_i))$ represents the distribution of the mean placental feature estimated from each $\mathbf{V}_i$. The dark vectors represent the changing vectors from $\mathbf{V}_1$.

To address these limitations, we extend the feature recomposition in Sect. 3.2 to *Distributional Feature Recomposition* that estimates a stable high-dimensional vector space defined by each set of features. We suggest utilizing the distribution $\mathcal{N}(\mu(\mathbf{V}), \sigma(\mathbf{V}))$ of the feature vectors $\mathbf{V}$, instead of point estimates (single vector sum) as a more comprehensive representation, where $\mu(\mathbf{V})$ and $\sigma(\mathbf{V})$ denote the mean and standard deviation, respectively. As shown by the shaded area in Fig. 2, the proposed distributional feature recomposition is more stable and representative than the point estimate sum of vector: $\mathcal{N}(\mu(\mathbf{V}_1), \sigma(\mathbf{V}_1))$ is similar to $\mathcal{N}(\mu(\mathbf{V}_2), \sigma(\mathbf{V}_2))$, but significantly different from $\mathcal{N}(\mu(\mathbf{V}_3), \sigma(\mathbf{V}_3))$.

Implementation-wise, we employ bootstrapping to estimate the distribution of the mean vector. We assume that the vectors adhere to a normal distribution with zero covariance between dimensions. During each training iteration, we randomly generate a new bootstrapped sample set $\tilde{\mathbf{V}}$ from the estimated normal distribution $\mathcal{N}(\mu(\mathbf{V}), \sigma(\mathbf{V}))$. Note that a slightly different sample set is generated in each training epoch to cover the variations in the feature distribution. We can therefore represent this distribution by the vector $\tilde{\mathbf{v}} = \sum_{\mathbf{v} \in \tilde{\mathbf{V}}} \mathbf{v}$, the sum of the sampled vectors, which captures the mean feature distribution in its values and carries the feature variation through epochs. By leveraging a sufficient amount of training data and running multiple epochs, we anticipate achieving a reliable estimation. The distributional feature recomposition not only inherits the scalability and efficiency of the traditional sum of vector approach but also provides a more robust estimate of the distribution of the mean vector, resulting in improved representational power and better generalizability.

3.4 Efficient Neural Networks

Efficient models, which are smaller and faster neural networks, facilitate easy deployment across a variety of devices, making them beneficial for low-resource communities. EfficientNet [22] and MobileNetV3 [11] are two notable examples of such networks. These models achieve comparable or better performance than state-of-the-art ResNet on ImageNet. However, efficient models generally have shallower network layers and can underperform when the features are more difficult to learn, particularly in medical applications [25]. To further demonstrate the representation power of our proposed method and expedite the diagnosis process, we experimentally substitute our image backbone with two efficient models, EfficientNet-B0 and MobileNetV3-Large-1.0, both of which exhibit highly competitive performance on ImageNet when compared to the original ResNet50. This evaluation serves two purposes: First, to test the applicability of our proposed method across different models, and second, to provide a more efficient and accessible placenta analysis model.

4 Experiments

4.1 Implementation

We implemented the proposed methods and baselines using the Python/PyTorch framework and deployed the system on a computing server. For input images, we used PlacentaNet [3] for segmentation and applied random augmentations such as random rotation and color jittering. We used a pre-trained BERT[1] [6] as our text encoder. EfficientNet-B0 and MobileNetV3-Large-1.0 followed official PyTorch implementations. All models and baselines were trained for 400 epochs.

[1] https://tfhub.dev/google/experts/bert/pubmed/2.

The encoder in the last epoch was saved and evaluated on their task-specific performance on the test set, measured by the AUC-ROC scores (area under the ROC curve). To ensure the reliability of the results, each evaluation experiment was repeated five times using different fine-tuning dataset random splits. The same testing procedure was adopted for all our methods. We masked all iPad images using the provided manual segmentation masks. For more information, please refer to the supplementary material.

4.2 Results

We compare our proposed methods (**Ours**) with three strong baselines: a ResNet-50 classification network, the ConVIRT [28] Medical VLC framework, and Pan et al. The mean results and confidence intervals (CIs) reported for each of the experiments on the two datasets are shown in Table 1. Some qualitative examples are in the supplementary material.

Table 1. AUC-ROC scores (in %) for placenta analysis tasks. The mean and 95% CI of five random splits. The highest means are in bold and the second-highest means are underlined. Primary stands for the main placenta dataset, and iPad stands for the iPad dataset. (*Mecon.*: meconium; *H.Chorio.*: histologic chorioamnionitis; *C.Chorio.*: clinical chorioamnionitis)

Method	Primary Task					iPad Task	
	Mecon.	FIR	MIR	H.Chorio.	Sepsis	MIR	C.Chorio
Supervised (ResNet-50)	77.0±2.9	74.2±3.3	68.5±3.4	67.4±2.7	88.4±2.0	50.8±21.6	47.0±16.7
ConVIRT (ResNet-50)	77.5±2.7	76.5±2.6	69.2±2.8	68.0±2.5	89.2±3.6	52.5±25.7	50.7±6.6
Pan et al. (ResNet-50)	79.4±1.3	77.4±3.4	70.3±4.0	68.9±5.0	89.8±2.8	61.9±14.4	53.6±4.2
Ours (ResNet-50)	81.3±2.3	**81.3**±3.0	**75.0**±1.6	**72.3**±2.6	**92.0**±0.9	**74.9**±5.0	59.9±4.5
Ours (EfficientNet)	79.7±1.5	78.5±3.9	71.5±2.6	67.8±2.8	87.7±4.1	58.7±13.3	**61.2**±4.6
Ours (MobileNet)	**81.4**±1.6	80.5±4.0	73.3±1.1	70.9±3.6	88.4±3.6	58.3±10.1	52.3±11.2

Our performance-optimized method with the ResNet backbone consistently outperforms all other methods in all placental analysis tasks. These results confirm the effectiveness of our approach in reducing feature suppression and enhancing representational power. Moreover, compared to Pan et al., our method generally has lower variation across different random splits, indicating that our training method can improve the stability of learned representations. Furthermore, the qualitative examples provided in the supplementary material show that incorrect predictions are often associated with incorrect salient locations.

Table 2 shows the speed improvements of our method. Since the efficiency of Pan et al. and ConVIRT is the same, we only present one of them for brevity. By removing the LLM during training, our method reduces the training time by a factor of 2.0. Moreover, the efficient version (e.g., MobileNet encoder) of our method has 2.4 to 4.1 times the throughput of the original model while still outperforming the traditional baseline approaches in most of the tasks, as shown

Table 2. Training and inference efficiency metrics. All these measurements are performed on a Tesla V100 GPU with a batch size of 32 at full precision (fp32). ResNet-50 s have the same inference efficiency and the number of parameters. (*#params*: number of parameters; *Time*: total training time in hours; *throughput*: examples/second; *TFLOPS*: Tera FLoating-point Operations/second). Improvements are in green.

Method	#params↓	Training	Inference	
		Time↓	Throughput↑	TFLOPS↓
Pan et al. (ResNet-50)	27.7M	38 hrs	–	–
Ours (ResNet-50)	27.7M	20 hrs ÷1.9	334	4.12
Ours (EfficientNet)	6.9M÷4.01	19 hrs÷2.0	822×2.46	0.40÷10.3
Ours (MobileNet)	7.1M÷3.90	18 hrs÷2.1	1368×4.10	0.22÷18.7

in Table 1. These results further support the superiority of the proposed representation and training method in terms of both training and testing efficiency.

4.3 Ablation

To better understand the improvements, we conduct a component-wise ablation study. We use the ConVIRT method (instead of Pan et al.) as the starting point to keep the loss function the same. We report the mean AUC-ROC across all tasks to minimize the effects of randomness.

Table 3. Mean AUC-ROC scores over placenta analysis tasks on the primary dataset. The mean and 95% CI of five random splits. *+Recomposition* means the use of Pathology Report Feature Recomposition over the baseline, ∼*+Distributional* stands for the further adoption of the Distributional Feature Recomposition. Improvements are in green. The abbreviations follow Table 1.

	Mecon.	FIR	MIR	H. Chorio.	Sepsis	Mean
Baseline (ConVIRT)	77.5±2.7	76.5±2.6	69.2±2.8	68.0±2.5	89.2±3.6	76.1
+ Recomposition	80.8±1.9	80.2±3.1	74.6±1.8	71.8±3.2	92.0±1.4	79.9+3.8
∼ + Distributional	81.3±2.3	81.3±3.0	75.0±1.6	72.3±2.6	92.0±0.9	80.4+4.3

As shown in Table 3, the text feature recomposition resulted in a significant improvement in performance since it treats all placental features equally to reduce the feature suppression problem. Moreover, applying distributional feature recomposition further improved performance, indicating that using a distribution to represent a set produces a more robust representation than a simple sum. Additionally, even the efficient version of our approach outperformed the performance version that was trained using the traditional VLC method. These improvements demonstrate the effectiveness of the proposed methods across different model architectures. However, we observed that the additional improvement from the distributional method was relatively small compared to that from the recomposition method. This may be due to the fact that the feature suppression problem is more prevalent than the misleading representation problem, or that the improvements may not be linearly proportional to the effectiveness–it may be more challenging to improve a better-performing model.

5 Conclusions and Future Work

We presented a novel automatic placenta analysis framework that achieves improved performance and efficiency. Additionally, our framework can accommodate architectures of different sizes, resulting in better-performing models that are faster and smaller, thereby enabling a wider range of applications. The framework demonstrated clear performance advantages over previous work without requiring additional data, while significantly reducing the model size and computational cost. These improvements have the potential to promote the clinical deployment of automated placenta analysis, which is particularly beneficial for resource-constrained communities.

Nonetheless, we acknowledge the large variance and performance drop when evaluating the iPad images. Hence, further research is required to enhance the model's robustness, and a larger external validation dataset is essential. Moreover, the performance of the image encoder is heavily reliant on the pre-trained language model, and our framework does not support online training of the language model. We aim to address these limitations in our future work.

References

1. Asadpour, V., Puttock, E.J., Getahun, D., Fassett, M.J., Xie, F.: Automated placental abruption identification using semantic segmentation, quantitative features, SVM, ensemble and multi-path CNN. Heliyon **9**(2), e13577:1–13 (2023)
2. Bakkali, S., Ming, Z., Coustaty, M., Rusiñol, M., Terrades, O.R.: VLCDoC: vision-language contrastive pre-training model for cross-modal document classification. Pattern Recogn. **139**(109419), 1–11 (2023)
3. Chen, Y., Wu, C., Zhang, Z., Goldstein, J.A., Gernand, A.D., Wang, J.Z.: PlacentaNet: automatic morphological characterization of placenta photos with deep learning. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 487–495. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_54
4. Chen, Y., et al.: AI-PLAX: AI-based placental assessment and examination using photos. Comput. Med. Imaging Graph. **84**(101744), 1–15 (2020)
5. Cui, Q., et al.: Contrastive vision-language pre-training with limited resources. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision - ECCV 2022. ECCV 2022. LNCS, vol. 13696, pp. 236–253. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20059-5_14
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

7. Dong, X., et al.: MaskCLIP: masked self-distillation advances contrastive language-image pretraining. arXiv preprint arXiv:2208.12262 (2022)
8. Dormer, J.D., et al.: CascadeNet for hysterectomy prediction in pregnant women due to placenta accreta spectrum. In: Proceedings of SPIE-the International Society for Optical Engineering, vol. 12032, pp. 156–164. SPIE (2022)
9. Goldstein, J.A., Gallagher, K., Beck, C., Kumar, R., Gernand, A.D.: Maternal-fetal inflammation in the placenta and the developmental origins of health and disease. Front. Immunol. **11**(531543), 1–14 (2020)
10. Gupta, K., Balyan, K., Lamba, B., Puri, M., Sengupta, D., Kumar, M.: Ultrasound placental image texture analysis using artificial intelligence to predict hypertension in pregnancy. J. Matern.-Fetal Neonatal. Med. **35**(25), 5587–5594 (2022)
11. Howard, A., et al.: Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)
12. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the International Conference on Machine Learning, pp. 4904–4916. PMLR (2021)
13. Khodaee, A., Grynspan, D., Bainbridge, S., Ukwatta, E., Chan, A.D.: Automatic placental distal villous hypoplasia scoring using a deep convolutional neural network regression model. In: Proceedings of the IEEE International Instrumentation and Measurement Technology Conference (I2MTC), pp. 1–5. IEEE (2022)
14. Li, T., et al.: Addressing feature suppression in unsupervised visual representations. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1411–1420 (2023)
15. Mobadersany, P., Cooper, L.A., Goldstein, J.A.: GestAltNet: aggregation and attention to improve deep learning of gestational age from placental whole-slide images. Lab. Invest. **101**(7), 942–951 (2021)
16. Pan, Y., Gernand, A.D., Goldstein, J.A., Mithal, L., Mwinyelle, D., Wang, J.Z.: Vision-language contrastive learning approach to robust automatic placenta analysis using photographic images. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science, vol. 13433, pp 707–716. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_68
17. Pietsch, M., et al.: APPLAUSE: automatic prediction of PLAcental health via U-net segmentation and statistical evaluation. Med. Image Anal. **72**(102145), 1–11 (2021)
18. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
19. Roberts, D.J.: Placental pathology, a survival guide. Arch. Pathol. Labor. Med. **132**(4), 641–651 (2008)
20. Specktor-Fadida, B., et al.: A bootstrap self-training method for sequence transfer: state-of-the-art placenta segmentation in fetal MRI. In: Sudre, C.H., et al. (eds.) UNSURE/PIPPI -2021. LNCS, vol. 12959, pp. 189–199. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87735-4_18
21. Sun, H., Jiao, J., Ren, Y., Guo, Y., Wang, Y.: Multimodal fusion model for classifying placenta ultrasound imaging in pregnancies with hypertension disorders. Pregnancy Hypertension **31**, 46–53 (2023)
22. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)

23. Wang, Y., Li, Y.Z., Lai, Q.Q., Li, S.T., Huang, J.: RU-net: an improved U-Net placenta segmentation network based on ResNet. Comput. Methods Program. Biomed. **227**(107206), 1–7 (2022)
24. Wen, K., Xia, J., Huang, Y., Li, L., Xu, J., Shao, J.: COOKIE: contrastive cross-modal knowledge sharing pre-training for vision-language representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2208–2217 (2021)
25. Yang, Y., et al.: A comparative analysis of eleven neural networks architectures for small datasets of lung images of COVID-19 patients toward improved clinical decisions. Comput. Biol. Med. **139**(104887), 1–26 (2021)
26. Ye, Z., Xuan, R., Ouyang, M., Wang, Y., Xu, J., Jin, W.: Prediction of placenta accreta spectrum by combining deep learning and radiomics using T2WI: A multicenter study. Abdom. Radiol. **47**(12), 4205–4218 (2022)
27. Zhang, P., et al.: Vinvl: revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5579–5588 (2021)
28. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Proceedings of the Machine Learning for Healthcare Conference, pp. 2–25. PMLR (2022)
29. Zhang, Z., Davaasuren, D., Wu, C., Goldstein, J.A., Gernand, A.D., Wang, J.Z.: Multi-region saliency-aware learning for cross-domain placenta image segmentation. Pattern Recogn. Lett. **140**, 165–171 (2020)