Vision-Language Contrastive Learning Approach to Robust Automatic Placenta Analysis Using Photographic Images*

Yimu Pan[1], Alison D. Gernand[1], Jeffery A. Goldstein[2], Leena Mithal[3], Delia Mwinyelle[4], and James Z. Wang[1]

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

Abstract. The standard placental examination helps identify adverse pregnancy outcomes but is not scalable since it requires hospital-level equipment and expert knowledge. Although the current supervised learning approaches in automatic placenta analysis improved the scalability, those approaches fall short on robustness and generalizability due to the scarcity of labeled training images. In this paper, we propose to use the vision-language contrastive learning (VLC) approach to address the data scarcity problem by incorporating the abundant pathology reports into the training data. Moreover, we address the feature suppression problem in the current VLC approaches to improve generalizability and robustness. The improvements enable us to use a shared image encoder across tasks to boost efficiency. Overall, our approach outperforms the strong baselines for fetal/maternal inflammatory response (FIR/MIR), chorioamnionitis, and sepsis risk classification tasks using the images from a professional photography instrument at the Northwestern Memorial Hospital; it also achieves the highest inference robustness to iPad images for MIR and chorioamnionitis risk classification tasks. It is the first approach to show robustness to placenta images from a mobile platform that is accessible to low-resource communities.

Keywords: Placenta analysis · mHealth · Vision-language pre-training.

1 Introduction

The placenta is a temporary organ that forms during pregnancy and acts as fetal life support prior to delivery. Adverse pregnancy outcomes, including chorioamnionitis and sepsis (infection) and meconium staining (fetal distress), produce reproducible morphologic changes in the placenta that can be identified by

* This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

pathologic examination. The current standard placental examination consists of macroscopic examination, production of microscopic slides, manual examination of the slide by a pathologist, and production of a report. This process requires hospital-level equipment and human input at each step, introducing variation and limiting opportunities for scaling. Automatic placenta analysis using a photographic image is more scalable and can benefit low-resource communities with no access to a pathologist.

Related Work. A recent automatic placenta photo analysis approach, the AI-PLAX [4], used a combination of handcrafted features/rules and deep learning methods; A later approach [19] used only deep learning methods. Although these approaches have achieved promising results, their models suffered from data scarcity; a large portion of the collected images was discarded to balance the positive and negative sample ratio and meet certain quality standards; all pathology reports were ignored since their models were not designed to use text data. Recent advances in self-supervised learning [8, 1] and vision-language contrastive learning (VLC) [18, 14] have shown promising results in pre-training tasks and can potentially benefit the model performance by including the discarded data as part of the pre-training dataset. However, current contrastive loss used in both the self-supervised methods and VLC methods suffered from the feature suppression problem [2]. Although recent work has addressed such a problem in a self-supervised setting [16, 15, 12, 5], to our knowledge, no work has been done in a VLC setting.

Our Contributions. We tackle the data scarcity problem by using an improved VLC technique to train a shared image encoder using the placenta image and the corresponding pathology report. Our technique is designed to learn generalizable placental features that can be applied to many downstream placental analysis tasks without training a separate image encoder for every task. This approach requires less data for the downstream tasks thus alleviating the data scarcity problem. To our knowledge, this is the first work to address the feature suppression problem in VLC. This is also the first automatic placenta analysis approach tested on iPad images. Our work improves both efficiency and robustness over the existing work in automatic placenta analysis.

2 Method

The proposed method is illustrated in Fig. 1. It consists of a pre-training stage and a fine-tuning stage. The pre-trained text encoder is frozen using a stop gradient operation in the pre-training stage. The trained image encoder is frozen using a stop gradient operation and shared for all tasks in the fine-tuning stage.

2.1 Problem Formulation

We have two tasks, the pre-training and the downstream classification. Formally, for the former, we want to learn a function f_v using a learned function f_u such

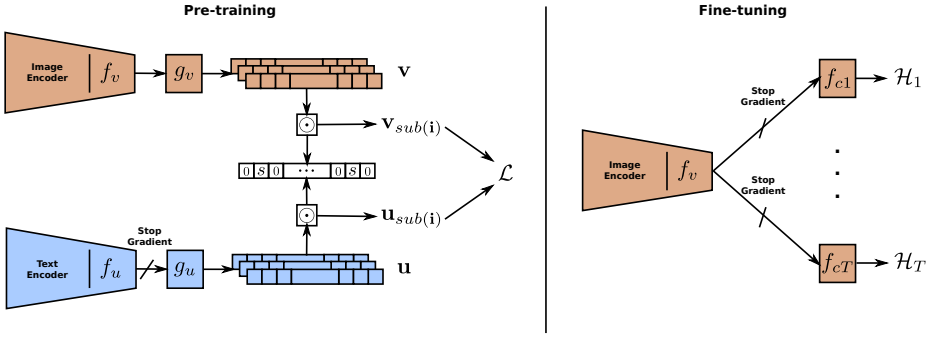

Fig. 1: A diagram illustrating our approach where all the notations correspond to the descriptions in Section 2. The inputs are omitted for simplicity.

that, for any pair of input $(\mathbf{x}_i, \mathbf{t}_i)$ and an similarity function $\texttt{sim}$, we have

$$\texttt{sim}(g_v(f_v(\mathbf{x}_i)), g_u(f_u(\mathbf{t}_i))) > \texttt{sim}(g_v(f_v(\mathbf{x}_i)), g_u(f_u(\mathbf{t}_j))) \ , \tag{1}$$

where g is a linear projection function to map the output vector to the same shape and $i \neq j$. The details on the objective function for achieving inequality (1) are discussed later.

For the latter task, we want to learn a function f_{ct} using the learned function f_v for each task $t \in [1:T]$ such that, for a pair of input $(\mathbf{x}_i, l_{ti})$,

$$f_{ct}(f_v(\mathbf{x}_i)) = l_{ti} \ , \tag{2}$$

which can be achieved by using the cross-entropy loss $\mathcal{H}$ as the objective function.

2.2 Hypothetical Cause of Feature Suppression Problem

Current VLC methods minimize the following contrastive loss:

$$\ell_i^{(v \to u)} = -\log \frac{\exp(\langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau)} \ , \tag{3}$$

where $\langle \mathbf{v}, \mathbf{u} \rangle$ represents the cosine similarity between the two feature vectors $\mathbf{v}$, $\mathbf{u}$ from $g_v(f_v(x))$ and $g_u(f_u(t))$, respectively, τ is the temperature hyper-parameter, and N is the total number of sample is a mini-batch. CLIP [14] and ConVIRT [18] have shown that models trained using this method are more robust. However, as demonstrated in [2], such a contrastive loss suffers from the feature suppression problem where the model only learns the most important feature. This effect is especially problematic in our application since we have multiple tasks for the same image and different tasks may require different features. Current contrastive loss uses cosine similarity $\langle \mathbf{v}, \mathbf{u} \rangle$ which is defined as

$$\langle \mathbf{v}, \mathbf{u} \rangle = \frac{\sum_{i=1}^{n} v_i u_i}{\sqrt{\sum_{i=1}^{n} v_i^2} \sqrt{\sum_{i=1}^{n} u_i^2}} \ , \tag{4}$$

where n is the length of vector $\mathbf{v}$ or $\mathbf{u}$ and v_i and u_i are the ith element in $\mathbf{v}$ and $\mathbf{u}$, respectively. Although previous work [16, 15, 12, 5] has studied the cause of the feature suppression problem by analyzing the entire loss function in a self-supervised setting, we believe the similarity metric alone can play an important role. The objection function (3) tries to achieve $\langle \mathbf{v}, \mathbf{u} \rangle > \langle \mathbf{v}, \mathbf{w} \rangle$ when $\mathbf{v}$ and $\mathbf{u}$ are the corresponding pair. Given $\sqrt{\sum_{i=1}^{n} u_i^2} = \sqrt{\sum_{i=1}^{n} w_i^2}$ (*i.e.*, two text features have the same L^2 norm) and $v_j = 0$ (*i.e.*, some elements of image features are not important), $u_i > w_i$ is enough for $\langle \mathbf{v}, \mathbf{u} \rangle > \langle \mathbf{v}, \mathbf{w} \rangle$. In this case, both u_j and w_j are ignored although they are not necessarily zero. This effect takes place when $v_j \approx 0$. In other words, the elements of features with small value contribute very little to the loss function (4) but the true importance of a feature is unknown before the downstream tasks. We hypothesize such an effect in the similarity metric is one cause of the feature suppression problem in VLC methods and we can address it by simply replacing the similarity metric.

2.3 Negative Logarithmic Hyperbolic Cosine Similarity

To minimize the feature suppression problem, we propose to use the Negative Logarithmic Hyperbolic Cosine (`NegLogCosh`) as the similarity metric:

$$\texttt{NegLogCosh}(\mathbf{v}, \mathbf{u}) = -\frac{1}{n} \sum_{i=1}^{n} \log(\cosh(s(v_i - u_i))) \ , \tag{5}$$

where s is a scaling factor, The advantage of $\texttt{NegLogCosh}(\mathbf{v}, \mathbf{u})$ over $\langle \mathbf{v}, \mathbf{u} \rangle$ is that the value change of any v_i or u_i is reflected in the result, thus the trained model tends to focus on more features. Although L1 and L2 loss functions have the same property, $\texttt{NegLogCosh}(\mathbf{v}, \mathbf{u})$ has more advantages. First, `NegLogCosh` has less emphasis than L2 loss when v_i and u_i are very different thus reducing the effect of the dominant feature from either the text side or the image side. Second, `NegLogCosh` is more stable than L1 loss when $v_i - u_i \approx 0$. The proposed objective function is the following:

$$\tilde{\ell}_i^{(v \to u)} = -\log \frac{\exp(\texttt{NegLogCosh}(\mathbf{v}_i, \mathbf{u}_i)/\tau)}{\sum_{k=1}^{N} \exp(\texttt{NegLogCosh}(\mathbf{v}_i, \mathbf{u}_k)/\tau)} \ . \tag{6}$$

Same as ConVIRT [18], the final loss function given $\lambda \in [0, 1]$ is

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda \tilde{\ell}_i^{(u \to v)} + (1 - \lambda) \tilde{\ell}_i^{(v \to u)} \right) \ . \tag{7}$$

2.4 Sub-feature Comparison

Because the similarity metric (5) compares two feature vectors element-wise instead of the angle between the two vectors, we can compare a random subset of the elements in the two features vector to reduce the feature suppression

problem further. Inspired by Dropout [17], we can randomly set some element of a feature vector to zero so that we have a sub-feature vector. For a feature vector $\mathbf{v}$ and an index vector $\mathbf{l} = (l_1, l_2, ..., l_k)$ where k is the size of $\mathbf{v}$ and i_j is a sample from a Bernoulli distribution with probability p, a sub-feature $\mathbf{v}_{\text{sub}}$ given $\mathbf{l}$ is defined as:

$$\mathbf{v}_{\text{sub}(\mathbf{l})} = \mathbf{v} \odot \mathbf{l} \,, \tag{8}$$

where $\odot$ is the element-wise multiplication. Sub-features contain many zero entries determined by $\mathbf{l}$. Replacing $\mathbf{v}$ with $\mathbf{v}_{\text{sub}(\mathbf{l})}$ in the metric (5) produces

$$\texttt{NegLogCosh}(\mathbf{v}_{\text{sub}(\mathbf{l})}, \mathbf{u}_{\text{sub}(\mathbf{l})}) = -\frac{1}{n}\sum_{i=1}^{n} \log(\cosh(sl_i(v_i - u_i))) \,, \tag{9}$$

where the index vector $\mathbf{l}$ is shared within the same mini-batch. Sharing the index vector is the main difference between this sub-feature approach and Dropout. Based on inequality (1), once the loss function (7) is minimized, we have:

$$\texttt{NegLogCosh}(\mathbf{v}_{\text{sub}(\mathbf{l})}, \mathbf{u}_{\text{sub}(\mathbf{l})}) > \texttt{NegLogCosh}(\mathbf{v}_{\text{sub}(\mathbf{l})}, \mathbf{w}_{\text{sub}(\mathbf{l})}) \,, \tag{10}$$

for any $\mathbf{v}, \mathbf{u}$ pair and $\mathbf{w}$ that comes from other pairs and for any $\mathbf{l}$. If we construct the corresponding index vector, $\mathbf{1} - \mathbf{l} = (1 - l_1, l - l_2, ..., 1 - l_k)$, which satisfies (10), we obtain

$$\begin{aligned} &\texttt{NegLogCosh}(\mathbf{v}, \mathbf{u}) \\ &= \texttt{NegLogCosh}(\mathbf{v}_{\text{sub}(\mathbf{l})}, \mathbf{u}_{\text{sub}(\mathbf{l})}) + \texttt{NegLogCosh}(\mathbf{v}_{\text{sub}(\mathbf{1}-\mathbf{i})}, \mathbf{u}_{\text{sub}(\mathbf{1}-\mathbf{l})}) \\ &> \texttt{NegLogCosh}(\mathbf{v}_{\text{sub}(\mathbf{l})}, \mathbf{w}_{\text{sub}(\mathbf{l})}) + \texttt{NegLogCosh}(\mathbf{v}_{\text{sub}(\mathbf{1}-\mathbf{i})}, \mathbf{w}_{\text{sub}(\mathbf{1}-\mathbf{l})}) \\ &= \texttt{NegLogCosh}(\mathbf{v}, \mathbf{w}) \end{aligned} \tag{11}$$

from equation (9). Thus, achieving inequality (10) implies achieving inequality (11). However, achieving inequality (11) does not imply inequality (10) because we can select a sub-feature to flip the inequality, and the loss function does not rule out such a possibility. This one-way implication shows that sub-feature comparison enables a VLC model to learn more image and text relationships in the feature space than a traditional approach. In other words, instead of just learning the features presented in the text, we have a chance to learn a more general feature representation. This advantage should both help alleviate the feature suppression problem and reduce over-fitting.

3 Dataset

The primary dataset was collected using a professional photography instrument in the pathology department at the Northwestern Memorial Hospital (Chicago) between 2014 and 2018. After filtering out blurry images and images with sliced placenta, we were left with 13,004 fetal side placenta images and pathology report pairs. We selected 2,811 images from 2017 for fine-tuning and the rest for pre-training.

The final pre-training data set consists of 10,193 image-and-text pairs. Each image contains the fetal side of a placenta, the cord, and a ruler. Each text sequence for the image contains a part of the corresponding pathology report.

The fine-tuning dataset consists of 2,811 images; we first manually checked the images to ensure the placenta is complete and free from obscures. We labeled each image based on the pathology report on four tasks presented in [4, 19]: *meconium*, *fetal inflammatory response* (FIR), *maternal inflammatory response* (MIR), and *chorioamnionitis*. There are different levels or stages for each symptom in the pathology report. We labeled the images as positive for meconium and chorioamnionitis regardless of the level. For FIR and MIR, we labeled the image as negative if the report does not contain any related information or identified the placenta as negative; we labeled the image as positive if the report identifies the placenta as stage 2 or higher; we dropped the image if the stage is higher than 0 but lower than 2 to improve the model's ability to distinguish significant cases. To assess the generalizability, we also labeled 166 images with neonatal sepsis based on the results that are diagnosed by treating physicians using clinical criteria on the infant charts. We then used all the positive examples for each task and uniformly sampled a similar number of negative samples. We then randomly split the data into training, validation, and testing sets with the ratio of 0.25:0.25:0.5 since we do not have the exact test set as in [4, 19]. We selected more images for testing to reduce the randomness in the testing result.

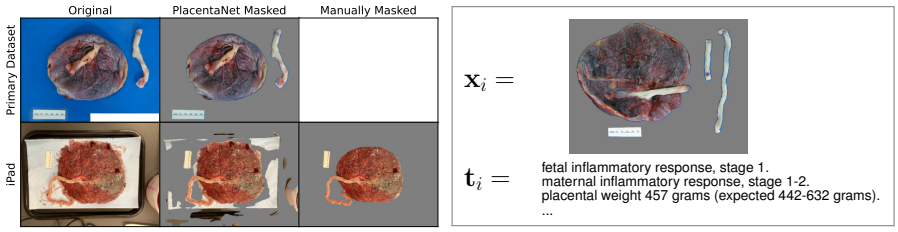

Fig. 2: **Left:** Example images from the two datasets. There is no manually generated background mask for the primary dataset. The image from the primary dataset and the iPad images have different white balance (see ruler color) and different backgrounds. **Right:** An example input image-and-text pair used in the pre-training. **Best viewed in color.**

To understand the robustness of the proposed method, we collected 52 placenta images at the same hospital in the summer of 2021 using an iPad (2021 model). The placentae were placed on a surgical towel and wiped clean of excess blood, the lighting was adjusted to minimize glare, and the iPad was held near parallel to the bench surface. As shown in Fig. 2, the performance in semantic segmentation from PlacentaNet is sub-optimal, and the white balance is different for the iPad image. We obtained labels for MIR and FIR from microscopical diagnoses [10] by an expert perinatal pathologist and clinical chorioamnionitis

from the infant charts. Note that the clinical chorioamnionitis is different from the histologic chorioamnionitis in the primary dataset. Since all data have FIR lower than stage 2, we discarded this label. For the rest of the tasks, we labeled images using the same criteria as the primary dataset. We acknowledge that this dataset is too small to serve as a benchmark, thus we considered a method outperforming others when the difference is significant (*e.g.*, by a few percentage points or higher). We used the iPad images to test the robustness; the main evaluation dataset is the primary dataset. A table containing the detailed breakdown of the data is in the supplementary material.

4 Experiments

4.1 Training and Testing

We used the ResNet50 [9] as our image encoder architecture and a pre-trained BERT [6] as our text encoder. The projection layers for both encoders were one-layer fully-connected neural networks (FC) with no activation. The classifiers were all two-layer FC with ReLU activation in the first layer but no activation in the output layer. For image preprocessing, we masked out the background from each image using PlacentaNet [3] and applied random augmentations. We randomly sampled topics in the text strings with replacements for the text preprocessing. We applied the Adam optimizer [11] and cosine decay learning rate scheduler with warm-up [13]. We selected the hyper-parameters on the baselines and applied them to our method. The details are in the supplementary material.

One independent image encoder was jointly trained with a classifier for each task for the baseline ResNet50. We trained the model on the training set for 100 epochs and saved the model with the highest validation accuracy.

For VLC models, we used the ConVIRT method as the baseline. We changed the projection layer from two-layer FC with ReLU activation to one-layer FC with no activation but kept the essential methodology the same. We trained the model for 400 epochs and saved the encoder in the last epoch. The training procedure for each downstream task was the same as for the baseline ResNet50, but the pre-trained encoder was frozen.

Our proposed method used the `NegLogCosh` similarity with the sub-feature comparison technique instead of cosine similarity. The training procedure followed the baseline ConVIRT.

We used the same testing procedure for all methods; we used the same prepossessing steps for all images in the primary dataset but two methods to mask out the background on the iPad images. The first method uses the segmentation map from PlacentaNet, which is sub-optimal (see Fig. 2) due to the difference in image quality. We included manually labeled segmentation maps as the second method to address this issue. In practice, this issue can be minimized by [19].

4.2 Results and Discussion

The mean results and confidence intervals (CIs) for the five experiments on the primary dataset are shown in Table 1. Since we do not have the exact model ar-

Table 1: AUC-ROC scores (in %) for placenta analysis tasks. **Top:** The mean and 95% CI of five random splits. The highest means are in bold. **Bottom:** The estimated mean improvements over the baseline ResNet50 and 95% CI using 100 bootstrap samples on the five random splits. The statistically significant improvements (CIs above 0) are underlined.

	Primary					iPad			
	Mecon.	FIR	MIR	H. Chorio.	Sepsis	PlacentaNet		Manual	
						MIR	C. Chorio.	MIR	C. Chorio.
RN50	77.0±2.9	74.2±3.3	68.5±3.4	67.4±2.7	88.4±2.0	46.7±20.9	42.8±14.0	50.8±21.6	47.0±16.7
ConVIRT	77.5±2.7	76.5±2.6	69.2±2.8	68.0±2.5	89.2±3.6	53.0±8.0	42.4±4.8	52.5±25.7	50.7±6.6
Ours	**79.4**±1.3	**77.4**±3.4	**70.3**±4.0	**68.9**±5.0	**89.8**±2.8	**58.4**±7.2	**45.4**±2.7	**61.9**±14.4	**53.6**±4.2
ConVIRT	0.6±1.8	2.2±1.9	0.8±1.3	0.6±2.3	0.8±1.4	6.4±6.7	-0.7±8.0	1.8±7.5	3.4±10.0
Ours	2.5±1.3	3.1±1.9	1.8±1.3	1.5±2.3	1.4±1.3	11.8±6.5	2.4±7.2	11.3±7.8	6.4±7.7

chitectures for all the experiments in [4, 19], we are reporting the widely adopted ResNet50 as the baseline. Our method achieved the highest area under ROC [7] (AUC-ROC) for all tasks in the experiment. Many of the improvements are statistically significant as the 95% CIs estimated using bootstrap samples do not contain 0. Although ConVIRT also outperformed the baseline on all the tasks in the pre-training data, only the improvement on FIR is significant. Moreover, the text features directly correspond to sepsis, which is not in the pre-training tasks, could be suppressed by other features. ConVIRT did not outperform the baseline without addressing the feature suppression problem, even with additional pre-training data. In contrast, our method showed significant improvement.

Additionally, our method also achieved the best when testing on the iPad images regardless of the segmentation map generation method, as shown in Table 1. The performance on clinical chorioamnionitis was much lower for the iPad images because we trained the model using histologic chorioamnionitis. As expected, the manually labeled segmentation maps resulted in higher AUC-ROC scores. Moreover, the proposed method always has smaller CIs, which also confirms the improvement in robustness. Although a larger iPad dataset would be necessary for confirming the improved performance of our approach on the mobile platform, the better robustness of our approach is apparent.

The qualitative examples are in the supplementary materials. Those examples show that all the experimented methods are sensitive to placenta color. We need more control over the lighting when collecting placenta images or better prepossessing to balance the placenta color for better performance.

Moreover, the shared encoder makes our method more efficient than the previous approach as the number of tasks grows.

5 Conclusions and Future Work

We proposed a robust, generalizable, and efficient framework for automatic placenta analysis. We showed that our pre-training method outperformed the popular approaches in almost all the placenta analysis tasks in the experiments. Our approach's robustness on photos taken with an iPad has high clinical value to low-resource communities. We expect our approach to perform better if we have a better image encoder, more data, or a domain-specific text encoder.

In the future, it would be interesting to extend our approach to a zero-shot setting [14] to further reduce the computation cost. More qualitative analysis can be performed to understand the improvement better. Lastly, we can collect a larger clinical dataset to improve the accuracy and robustness.

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning. pp. 1597–1607. PMLR (2020)
2. Chen, T., Luo, C., Li, L.: Intriguing properties of contrastive losses. Advances in Neural Information Processing Systems **34** (2021)
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: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 487–495. Springer (2019)
4. Chen, Y., Zhang, Z., Wu, C., Davaasuren, D., Goldstein, J.A., Gernand, A.D., Wang, J.Z.: AI-PLAX: Ai-based placental assessment and examination using photos. Computerized Medical Imaging and Graphics **84**, 101744:1–15 (2020)
5. Denize, J., Rabarisoa, J., Orcesi, A., Hérault, R., Canu, S.: Similarity contrastive estimation for self-supervised soft contrastive learning. arXiv preprint arXiv:2111.14585 (2021)
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. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters **27**(8), 861–874 (2006)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. Khong, T.Y., Mooney, E.E., Ariel, I., Balmus, N.C., Boyd, T.K., Brundler, M.A., Derricott, H., Evans, M.J., Faye-Petersen, O.M., Gillan, J.E., et al.: Sampling and definitions of placental lesions: Amsterdam placental workshop group consensus statement. Archives of Pathology & Laboratory Medicine **140**(7), 698–713 (2016)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

12. Li, T., Fan, L., Yuan, Y., He, H., Tian, Y., Feris, R., Indyk, P., Katabi, D.: Addressing feature suppression in unsupervised visual representations. arXiv preprint arXiv:2012.09962 (2020)
13. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
15. Rezaei, M., Soleymani, F., Bischl, B., Azizi, S.: Deep bregman divergence for contrastive learning of visual representations. arXiv preprint arXiv:2109.07455 (2021)
16. Robinson, J., Sun, L., Yu, K., Batmanghelich, K., Jegelka, S., Sra, S.: Can contrastive learning avoid shortcut solutions? arXiv preprint arXiv:2106.11230 (2021)
17. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(56), 1929–1958 (2014)
18. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:2010.00747 (2020)
19. 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 Recognition Letters **140**, 165–171 (2020)

Supplementary Materials for "Vision-Language Contrastive Learning Approach to Robust Automatic Placenta Analysis Using Photographic Images"

Yimu Pan[1], Alison D. Gernand[1], Jeffery A. Goldstein[2], Leena Mithal[3], Delia Mwinyelle[4], and James Z. Wang[1]

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

Table 1: Hyper-parameters for the pre-training and the fine-tuning models. The hyper-parameters are selected to make the baselines converge; no other tunning is made on our method. The average runtime is in the end.

	Pre-train	Supervised/Fine-tune
Optmizer	Adam	Adam
$\beta_1/\beta_2/\epsilon$	$0.9/0.999/10^{-7}$	$0.9/0.999/10^{-7}$
Learning Rate Schedule	WarmUp&CosineDecay	WarmUp&CosineDecay
Initial Learning Rate	0.00025	0.00025
Final Learning Rate	0	0
Warm-up Epochs	10	0
Weight Decay	10^{-6}	10^{-6}
Class Weight	N/A	$(\log(1.03 + \frac{\#\text{samples}}{\#\text{all data}}))^{-1}$
Batch Size	32	32
Epochs	400	100
Random Left Right Flip	Yes	Yes
Random Up Down Flip	Yes	Yes
Random Brightness	0.05	0.01
Random Hue	0.05	0.01
Input Size	512×384	512×384
Projection Output Size	768	N/A
Sub-feature Drop Ratio	0.2	N/A
τ/c (Variables in the paper)	0.1/1.25	N/A
Classifier FC1 Units	N/A	256
Classifier FC2 Units	N/A	1
Classifier Dropout Ratio	N/A	0.2
ResNet50 Average Runtime	N/A	(1.6/0.7) sec/batch
ConVIRT Average Runtime	0.9 sec/batch	(1.1/0.7) sec/batch
Ours Average Runtime	0.9 sec/batch	(1.1/0.7) sec/batch

Table 2: An example random split of the fine-tuning dataset (negative/positive).

	Mecon.	FIR	MIR	H. Chorio.	Sepsis	MIR (iPad)	C. Chorio. (iPad)
Train	177/173	88/79	166/188	119/102	41/44	-	-
Val.	174/176	79/88	198/157	100/122	43/42	-	-
Test	370/330	190/145	332/378	228/215	90/80	10/14	22/23
Total	721/679	357/312	696/723	435/439	174/166	10/14	22/23

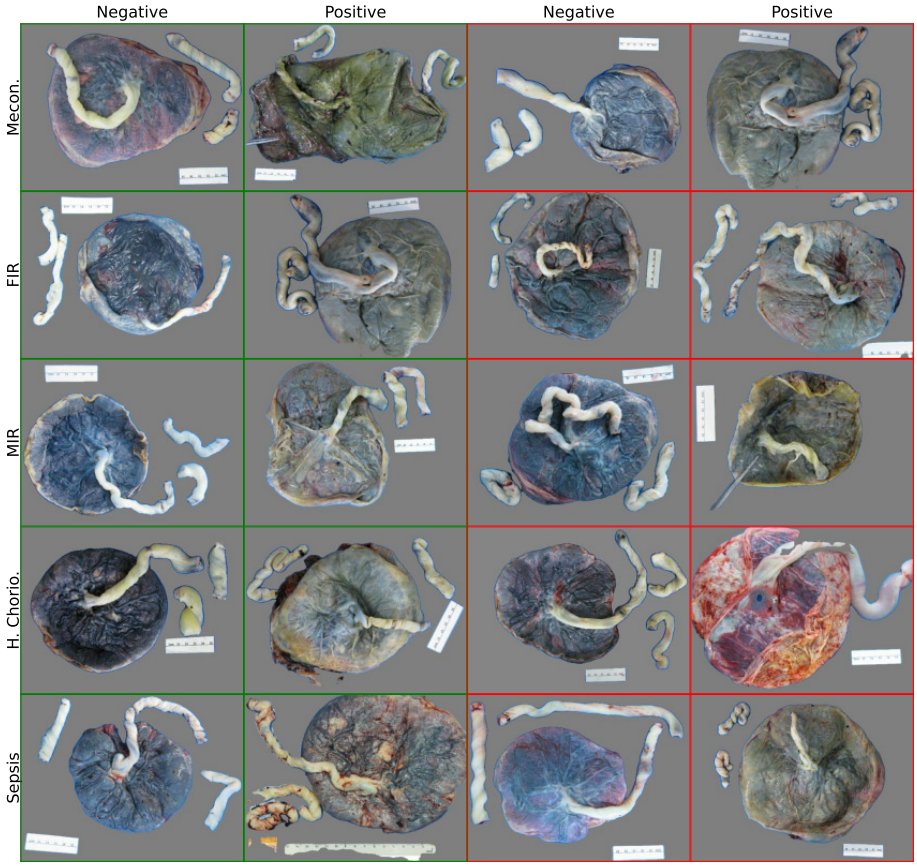

Fig. 1: Qualitative examples of the high confidence predictions produced by all methods on the primary dataset. The predictions are labeled above the images. The correct (left two columns) and incorrect (right two columns) predictions are boxed with the corresponding color. Placentas with warmer colors tend to receive positive predictions. **Best viewed in color.**