Pattern Recognition Letters
journal homepage: www.elsevier.com

Multi-region Saliency-aware Learning for Cross-domain Placenta Image Segmentation

Zhuomin Zhang[a,**], Dolzodmaa Davaasuren[a], Chenyan Wu[a], Jeffery A. Goldstein[b], Alison D. Gernand[a], James Z. Wang[a]

[a]*The Pennsylvania State University, University Park, Pennsylvania, USA*
[b]*Northwestern University, Chicago, Illinois, USA*

ABSTRACT

We propose a multi-region saliency-aware learning (MSL) method for cross-domain placenta image segmentation. Unlike most existing image-level transfer learning methods that fail to preserve the semantics of paired regions, our MSL incorporates the attention mechanism and a saliency constraint into the adversarial translation process, which can realize multi-region mappings in the semantic level. Specifically, the built-in attention module serves to detect the most discriminative semantic regions that the generator should focus on. Then we use the attention consistency as another guidance for retaining semantics after translation. Furthermore, we exploit the specially designed saliency-consistent constraint to enforce the semantic consistency by requiring the saliency regions unchanged. We conduct experiments using two real-world placenta datasets we have collected. We examine the efficacy of this approach in 1) segmentation and 2) prediction of the placental diagnoses of fetal and maternal inflammatory response (FIR, MIR). Experimental results show the superiority of the proposed approach over the state of the art.

Keywords: transfer learning; placenta; photo image analysis; pathology

1. Introduction

The placenta is the essential connection between mother and fetus, sensing nutrient availability and needs, producing hormones to drive physiologic changes, and protecting the fetus from pathogens (Roberts, 2008). Because the main function of the placenta is to support important metabolic activities, pathological analysis of the placenta should be an integral part of the health examination during pregnancy and after delivery. Yet, as a result of costly examination charges and limited expertise and facilities in developing countries, we estimate that only a small proportion of placentas ever examined by a pathologist worldwide. Automating pathological analysis by advanced image processing techniques is an inevitable trend because it can substantially augment the productivity of pathologists, shorten examination time and enhance examination accuracy. Particularly, accurate image segmentation is essential for integrated computerized placenta photo analysis (Chen et al., 2019a, 2020). As shown in Fig. 1, the placenta disc, umbilical cord, ruler (for measuring the scale), and background are four common categories that must be segmented in a photo for further analysis.

Although convolutional neural networks (CNNs) have triumphed over conventional object segmentation in many clinical image segmentation and analysis applications (Chen et al., 2019a, 2020; Kamnitsas et al., 2017; Milletari et al., 2016), the generalizability of a trained CNN is often inadequate when applied to a new dataset (*e.g.*, photos from a different hospital) because the two datasets often have vastly different data distributions (Yan et al., 2019; Yu et al., 2019).

This limits the applicability of AI models trained on data from high-resource settings, such as academic medical centers, to lower resource settings including community hospitals and low income countries. As illustrated in Fig. 1(a), although the appearances of the disc and cord are reasonably consistent across different data sources, the ruler and background can be vastly distinctive in terms of the color, texture, and amount of distraction. These largely different visual appearances can cause a well-trained CNN model, such as UNet (Ronneberger et al., 2015), to be vulnerable when generating segmentation

**Corresponding author
e-mail: zxz78@psu.edu (Zhuomin Zhang)

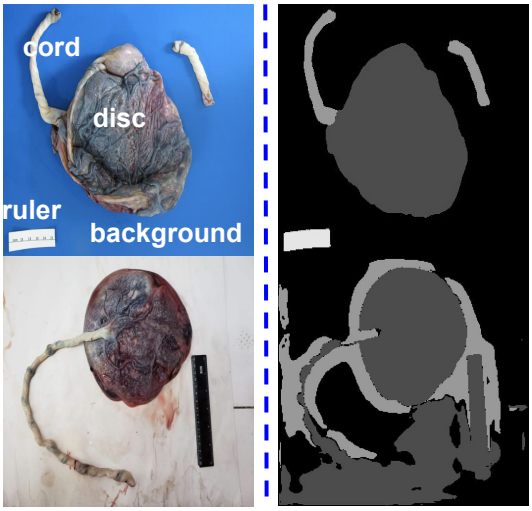

Fig. 1. Challenges for cross-domain placenta image segmentation. (a) Cross-domain images: Placenta images from two different hospitals. (b) Their corresponding segmentation results using the trained model on the first dataset.

results in an unseen domain, as shown in Fig. 1(b). Because of the difficulties lying in data collection and annotation, it is practically infeasible to retrain a label-dependent CNN model to adapt to different datasets every time.

Therefore, it is highly desirable to close the gap across domains by an effective domain adaptation method. This paper focuses on the domain adaptation problem for placenta image segmentation when data in the target domain are insufficient and the corresponding labels are unavailable.

Most existing unsupervised domain adaptation methods can be categorized into: (1) feature distribution alignment, (2) task-specific output space relation preservation, and (3) image-to-image translation. Feature-level adaptation methods aim at aligning the two domains in a latent feature space by minimizing the distribution distance, such as the maximum mean discrepancy (Long et al., 2015), or leveraging adversarial learning strategies (Ganin and Lempitsky, 2014). However, the aligned feature space is not guaranteed to be semantically consistent and some low-level visual cues that are crucial for the end segmentation task may be lost. The output space adaptation method aims to make the predicted maps close to each other (Tsai et al., 2018; Chen et al., 2019b). It is not suitable for placenta image segmentation because the spatial relations among objects are not consistent across domains. Image-to-image translation tries to alleviate the domain shift from headstream by forcing the cross-domain images to look like those from the original domain. Some regularization terms such as the cycle strategy (Zhu et al., 2017) are exploited to make the training process free of paired data. Attention mechanism is also introduced to pair the regions of interests across domains for more realistic image generation (Chen et al., 2018; Mejjati et al., 2018). Nevertheless, these attention methods only focus on one specific type of object, so the translation can still be mismatched when multiple objects are attended to simultaneously. How to enforce semantic consistency for multi-region translation is the research problem our MSL model aims to solve.

In this paper, we propose a multi-region saliency-aware learning method to realize cross-domain placenta image translation by enforcing both the attention and saliency consistency. An attention module serving as the semantic guidance is firstly coupled with the classic generator-discriminator game to find the most discriminative regions (*i.e.*, ruler and background). Out of the motivation for semantic-consistent transfer of multi regions, an attention-consistent loss is added as an extra constraint to enforce the translation to preserve the attention-related information. Notably, we devise a new saliency-consistent constraint as another semantic guidance by restraining the saliency relation unchanged after translation. Finally, we feed the translated target domain images to a well-trained CNN model from the annotation-sufficient source domain for the ultimate segmentation task.

2. Approach

The overall framework of our MSL is presented in Fig. 2, where two attention networks, A_S and A_T, work together with corresponding generators $G_{S\to T}$ and $G_{T\to S}$, along with two discriminators D_S and D_T, to form our MSL model. Due to the high demand of multi-region semantic consistency before and after image translation for our end segmentation task, we add an attention-consistent loss L_{att} to alleviate the influence of unattended regions and a simple but effective saliency-consistent constraint L_{sal} to guarantee the salient regions to stay unaltered during the generation. Meanwhile, the classic adversarial loss L_{gan} and cycle loss L_{cyc} are exploited together to accomplish semantic mappings, which will be described in detail as follows.

We formulate this task as an image-to-image translation for segmentation map prediction. We assume that the source image set X_S together with the source label set Y_S are accessed, while only the image set X_T in the target domain is available. Our final goal is to adapt the pre-trained segmentation model F_S to the translated source-like images in the target domain.

2.1. Cycle Generative Adversarial Network

The goal of image translation is to learn a mapping between the source domain and the target domain. A generative model annotated as $G_{S\to T}$ is exploited to learn this kind of data mapping to generate target-like images $x_{s\to t} = G_{S\to T}(x_s)$, which can deceive the discriminator. On the contrary, the discriminator D_T aims to distinguish the genuine image x_t from the translated images $x_{s\to t}$, to constitute a dynamic min-max two-player game. We adopt the adversarial loss function in LSGAN (Mao et al., 2017) into our model, and L_{gan} is denoted as:

$$\begin{aligned} L_{\text{gan}}(G_{S\to T}, D_T, X_S, X_T) = {} & E_{x_t\sim P_{X_T}(x_t)}[\log(D_T(x_t))] \\ & + E_{x_s\sim P_{X_S}(x_s)}[\log(1 - D_T(x_{s\to t}))] \,. \end{aligned} \tag{1}$$

Similarly, the corresponding loss function for the target-to-source translation $L_{\text{gan}}(G_{T\to S}, D_S, X_T, X_S)$ is defined in the same way. That is,

$$\begin{aligned} L_{\text{gan}}(G_{T\to S}, D_S, X_T, X_S) = {} & E_{x_s\sim P_{X_S}(x_s)}[\log(D_S(x_s))] \\ & + E_{x_t\sim P_{X_T}(x_t)}[\log(1 - D_S(x_{s\to t}))] \,. \end{aligned} \tag{2}$$

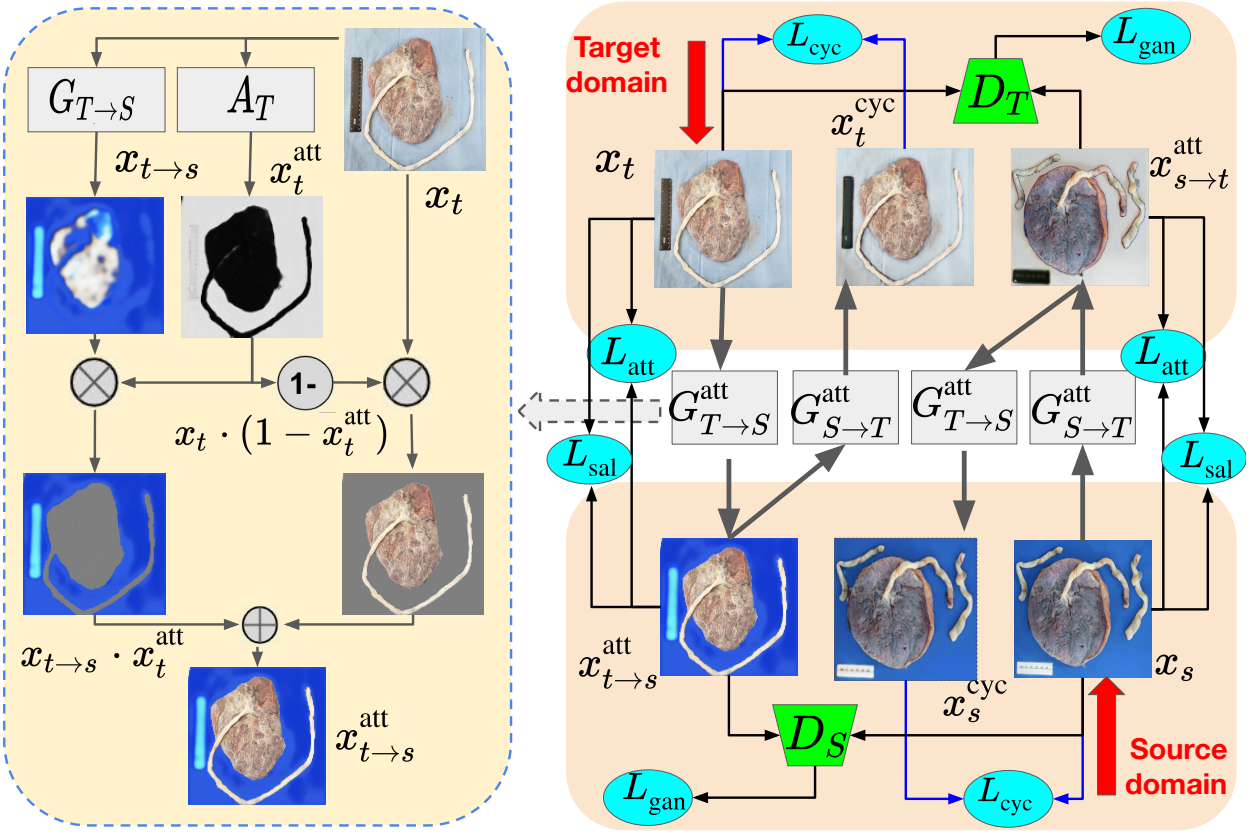

Fig. 2. The pipeline of the proposed approach. As shown in the left part, the introduced attention network A_T can divide the placenta image x_t into attended regions such as ruler and background, and unattended regions that include disc and cord. The translated image $x_{t\to s}^{\text{att}}$ is a combination of translated attended parts and original unattended parts. The attention-consistent loss L_{att} and a saliency-consistent L_{sal} loss are added to preserve the semantics in together with the image-level adaptation as composed of the pixel GAN loss L_{gan} and the cycle loss L_{cyc}.

The discriminators D_S and D_T attempt to maximize the loss, while the generators $G_{S\to T}$ and $G_{T\to S}$ strive to minimize the loss. To make the translated image preserve the structure and local content of the original image, a cycle consistency loss (Isola et al., 2017) is also designed as follows by measuring the pixel-wise difference between the reconstructed image and the original image.

$$\begin{aligned} L_{\text{cyc}}(G_{T\to S}, G_{S\to T}) =& E_{x_s\sim P_{X_S}(x_s)}[\|G_{T\to S}(x_{s\to t}) - x_s\|_1] \\ &+ E_{x_t\sim P_{X_T}(x_t)}[\|G_{S\to T}(x_{t\to s}) - x_t\|_1] \,. \end{aligned} \tag{3}$$

2.2. Multi-region Saliency-aware Constraints

To realize multi-region translation simultaneously by a single generator, we made the assumption that fixing the foreground regions (placenta disc and cord) would improve the translation quality and subsequent cross-domain pathological diagnosis adaptation. The rationales are: 1) The characteristics of placenta datasets: Although the background material and the ruler of placenta photos vary a lot across different clinical datasets, visual differences of the foreground disc and cord are relatively negligible because placentas share the same basic structure and appearance. We have collected over 1000 placentas in the source dataset covering most variations in placenta morphology in the target domain, so the feature extracted from the placenta is insignificant for fake image discrimination. That also explains why the foreground is learned as unattended regions. Hence, transferring the foreground placenta itself helps little to the quality of the final generated images. 2) The requirement of real clinical application: The pathological indicator prediction relies on the extracted visual features from the placenta, so keeping the original foreground features unchanged is essential for adapting the pretrained source-domain diagnosis models to the translated images. Otherwise, the diagnosis results could be unreliable if the foreground is also translated. 3) The functionality of the generator: Incorporating the separation and conversion of foreground objects with the existing background translation in a single network would confuse the aims of the generator, leading to unmatched contents in generated images. This viewpoint has been justified in [1,13], where fixing the unattended regions can help generate more realistic images than transferring the entire image.

Inspired by AGGAN (Mejjati et al., 2018), we decompose the generative module into two separated parts: (1) the attention networks A_S and A_T attempting to attend to the regions that the discriminator considers to be the most descriptive within its domain (*i.e.*, the ruler and background in our application), and (2) the classic generative networks focusing on transforming the whole image from one domain to another. The ultimate generated image is therefore a combination of the attended regions from the transformed image and unattended areas in the

original image by using the attention map as mask. With the attention mechanism, the discriminator can force the attention networks to find the most domain-descriptive regions of interest and, therefore, make the generators pay more attention to the attended objects.

We denote the attention maps induced from X_S and X_T as $x_s^{\text{att}} = A_S(x_s)$ and $x_t^{\text{att}} = A_T(x_t)$, respectively. Attention-guided generated map can then be computed as:

$$x_{s\to t}^{\text{att}} = x_s^{\text{att}} \cdot G_{S\to T}(x_s) + (1 - x_s^{\text{att}}) \cdot x_s \ , \tag{4}$$

where $\cdot$ denotes the element-wise product. This attention network is jointly adversarially trained with the generators and discriminators. We replace $x_{s\to t}$ and $x_{t\to s}$ in the (1) and (3) with $x_{s\to t}^{\text{att}}$ and $x_{t\to s}^{\text{att}}$, respectively.

Attention-consistent loss: One regulation of generators is that the transformed image should have the same semantics as the original image. If the semantics are aligned, in other words, the attended regions should maintain the same before and after translation: $A_S(x_s) \approx A_T(x_{s\to t})$. Instead of using the segmentation label to supervisedly preserve the semantics (Li et al., 2018), we treat the learned attention map as an important form of semantics. Besides, because we have the segmentation maps in the source domain, we can add extra supervision to attention map generation in the source domain. To that end, the attention-consistent losses are formulated as:

$$\begin{aligned} L_{\text{att}}(A_S) &= E_{x_s\sim P_{X_S}(x_s)}[\|A_S(x_s) - y_s\|_1] \\ &\quad + E_{x_s\sim P_{X_S}(x_s)}[\|A_T(x_{s\to t}) - y_s\|_1] \ , \\ L_{\text{att}}(A_T) &= E_{x_t\sim P_{X_T}(x_t)}[\|A_T(x_t) - A_S(x_{t\to s})\|_1] \ . \end{aligned} \tag{5}$$

Saliency-consistent loss: In most attention-guided image translation cases, the most salient object in the foreground is likely to be learned as the region of interest. To add the additional saliency-consistent loss is hence meaningless. However, this attention map is learned by the discriminator, which is not always consistent with the visual attention (*i.e.*, saliency). For instance, what if both the background and foreground objects are included into the attended region? In our case, it is observed that the cord and disc in the foreground stay changeless across domains, while the ruler and background suffer a lot of variations, leading them to be classified as attended areas by the discriminator. Even if we force the generator to focus on this region, the generated ruler and background can still be mismatched. Therefore, adding the saliency-consistent loss as a constraint is indispensable to help maintain the semantic consistency before and after translation to prevent label flipping.

Because the attended areas can be obtained at the early training stage, we only compute the saliency value for pixels in the attended regions. We employ the simple but effective FT (Qin et al., 2019) method for saliency detection:

$$S(i,j) = \|I(i,j) - I_\mu\|_2 \ , \tag{6}$$

where $I(i,j)$ represents the pixel color vector value after Gaussian blurring (Gedraite and Hadad, 2011), and I_μ is the mean image color vector. $\|\cdot\|_2$ is the Euclidean distance between color vectors. We denote the saliency maps for both the images in the original domain and the translated images as S_S, S_T and $S_{S\to T}$, $S_{T\to S}$, respectively. We binarize these saliency maps S_S, S_T as the saliency ground truth to formulate the loss function. To overcome the problem that the number of pixels in different categories are highly unbalanced, the saliency consistency loss is defined using the dice coefficient (Sudre et al., 2017):

$$L_{\text{sal}}(G_{S\to T}) = 1 - \frac{\sum_{i,j} S_S(i,j) \cdot S_{S\to T}(i,j)}{\sum_{i,j} S_S(i,j) + \sum_{i,j} S_{S\to T}(i,j)} \ . \tag{7}$$

The $L_{\text{sal}}(G_{T\to S})$ is defined in a similar way.

We obtain the final loss function by combining the adversarial, cycle consistency, attention consistency, and saliency consistency losses for both the source and target domains, defined as:

$$\begin{aligned} L_{\text{total}} &= L_{\text{gan}}(G_{S\to T}, D_T, X_S, X_T) + L_{\text{gan}}(G_{T\to S}, D_S, X_T, X_S) \\ &\quad + \lambda_{\text{cyc}} L_{\text{cyc}}(G_{T\to S}, G_{S\to T}) + \lambda_{\text{att,S}} L_{\text{att}}(A_S) + \lambda_{\text{att,T}} L_{\text{att}}(A_T) \\ &\quad + \lambda_{\text{sal,S}} L_{\text{sal}}(G_{S\to T}) + \lambda_{\text{sal,T}} L_{\text{sal}}(G_{T\to S}) \ . \end{aligned} \tag{8}$$

This ultimate model parameters can be obtained by solving the mini-max optimization problem:

$$G_{S\to T}^*, G_{T\to S}^*, D_S^*, D_T^*, A_S^*, A_T^* = \arg\min_{G_{S\to T}, G_{T\to S}, A_S, A_T} \arg\max_{D_S, D_T} L_{\text{total}} \ . \tag{9}$$

Segmentation loss: We adopt the same structure of the segmentation module from PlacentaNet (Chen et al., 2019a) to train a segmentation model in the source domain. We use $p(i,j,k)$ to denote the prediction probability of the pixel (i,j) belonging to class k and $g(i,j,k)$ to represent the corresponding ground truth. Sharing the same spirit of saliency consistency loss to balance labels, the dice loss for 4-class segmentation is defined as:

$$L_{seg} = 1 - \frac{\sum_{i,j}\sum_{k=0}^{3} p(i,j,k) \cdot g(i,j,k)}{\sum_{i,j}\sum_{k=0}^{3} (p(i,j,k) + g(i,j,k))} \ , \tag{10}$$

We apply this pretrained model to the translated target-domain images to obtain the final segmentation results.

3. Experiments

3.1. Datasets and Experimental Settings

We curated real-world post-delivery datasets, including a relatively clean image set together with comprehensive pathology reports (de-identified) from a large urban academic hospital,the Northwestern Memorial Hospital, as the source domain, and images of non-professional quality taken from a hospital in Mongolia (only images, no accompanying pathology reports) as the target domains. A web-based annotation tool was developed to: (1) discard images that don't meet our image quality standard (disc and cord should be fresh and not occluded by irrelevant objects); and (2) get pixel-wise segmentation maps for the disc, cord, ruler and background annotated by trained labelers. The dataset collected as the source domain contained 1,003 placenta images together with segmentation maps and extracted diagnoses from the pathology reports, while the target dataset has 76 images and corresponding annotated segmentation maps for evaluation purpose.

We divided the dataset in the source domain into training and testing sets with the ratio of 0.8 : 0.2 for training the segmentation model. For the image translation task, 200 images from the source domain and 60 images from the target domain were used for training. We used cross-validation to demonstrate the translation performance and the segmentation result in the target domain.

PlacentaNet. We adopted the same encoder-decoder structure from PlacentaNet (Chen et al., 2019a, 2020) to train a segmentation model in the source domain. We used the Adam optimizer (Kingma and Ba, 2014) with a mini-batch size of 5 and a learning rate of 0.001 for training. The pixel-wise accuracy and mean IoU are 0.9693 and 0.9131 respectively for testing in the same domain.

Translation network. Our training process can be separated into three steps: (1) We first trained the discriminators on full images for 20 epochs to help the attention module well trained with the guidance of the attention consistency loss; (2) Then, we make the discriminator to focus on the salient region (*i.e.*, the ruler) within the next 5 epochs by multiplying the saliency map (threshold= 0.7) to the image, which can alleviate the unbalanced label distribution problem. (3) Finally, we multiply the binarized attention map (threshold= 0.2) to the generated images to make the discriminator only consider attended regions. The saliency loss is then added to guide the overall translation performance.

For all steps, the training images were rescaled to 512×512 pixels, following by random flipping for data augmentation. We used Adam with a batch size of 1 and a linearly decaying learning rate from 0.0002 for the training of all the three networks. As for the network structure, we used the residual attention module introduced in (Wang et al., 2017) as attention network, Resnet-9blocks (Johnson et al., 2016) as the generator and PatchGAN (Isola et al., 2017) as the discriminator. We set hyper-parameters $\lambda_{\text{cyc}} = 5$, $\lambda_{\text{att,S}} = 2$, $\lambda_{\text{att,T}} = 4$, $\lambda_{\text{sal,S}} = 1$, and $\lambda_{\text{sal,T}} = 1$, respectively.

3.2. Results

To show the improvement on segmentation brought by the cross-domain adaptation, we first compare our model with the baseline scenario (*i.e.*, segmentation without adaptation). Then two state-of-the-art image translation models, CycleGAN (Zhu et al., 2017) and AGGAN (Mejjati et al., 2018), are compared to demonstrate the the superiority of our MSL model. These two methods are pioneering in the GAN-based translation methods and most relevant to our placenta segmentation problem. The segmentation performance is evaluated using standard segmentation metrics, including pixel accuracy, mean accuracy, and mean IoU. The definition of those metrics are as follows: we use $P_{i,j}$denote the number of pixels that are annotated as class i but predicted as class j. The total number of pixels that belong to class i in the ground truth are denoted as G_i. Because there are 4 classes (ruler, disc, cord and background) in our case, $i, j \in \{0, 1, 2, 3\}$. The pixel accuracy, mean class accuracy, and mean IoU are then defined as follows:

- Pixel accuracy: $\frac{\sum_{i=0}^{3} P_{i,i}}{\sum_{i=0}^{3} G_i}$.
- Mean class accuracy: $\frac{1}{4}\sum_{i=0}^{3} \frac{p_{i,i}}{G_i}$.
- Mean IoU: $\frac{1}{4}\sum_{i=0}^{3} \frac{P_{i,i}}{G_i + \sum_{j \neq i} P_{i,j}}$.

Table 1. Segmentation evaluation accuracy.

Method	Pixel Accu.	Mean Accu.	Mean IoU
No adaptation	0.5256	0.4164	0.2049
CycleGAN	0.7022	0.5850	0.3508
AGGAN	0.7807	0.6497	0.4672
MSL(w/o L_{att})	0.8291	0.6502	0.4812
MSL(w/o L_{sal})	0.7852	0.6591	0.4722
MSL(w/o L_{cyc})	0.7593	0.6203	0.3874
MSL	**0.8743**	**0.8691**	**0.7380**

The quantitative results are shown in Table. 1. We observe that the segmentation performance has been greatly improved from 0.2049 to 0.7380 in mean IoU by adding proper adaptation. The success of the background translation leads to a big leap on the pixel accuracy. The improvement of Mean IoU should be credited to our accurate ruler translation with the saliency-aware guidance. We also show a few segmentation examples in Fig. 3 for qualitative comparison. Some special cases, including non-uniform background, poor-quality cloth, different color, and messy surrounding, are shown to demonstrate the robustness of our proposed translation method. Besides, to illustrate detailed segmentation performance in different categories,we also compare our approach with the baseline, AGGAN and CycleGAN respectively using pixel-wise prediction confusion matrices as shown in Fig. 4. According to the confusion matrices, it is noticeable that our method greatly improve the segmentation accuracy of cord and ruler with the added losses.

3.3. Ablation Study

To show the indispensability of each loss term, we conducted an ablation study as follows.

Effectiveness of attention consistency: As shown in the Fig. 5(a), the attention-consistent loss enforces the cord region to be unattended by keeping the consistency of attention regions. Without the attention-consistent loss, the cord can sometimes be learned as the attended background due to the similar light colors and bloody background in the target domain.

Effectiveness of saliency consistency: Fig. 5(b) shows the saliency loss constrains the saliency relation between ruler and background to be consistent. Without the saliency constraint, the translation often suffers from random label flipping because the generator fails to produce matched semantics.

Effectiveness of cycle consistency: From Fig. 5(c), we observe that the translation back to the original domain sometimes fails because there is no reconstruction guarantee without the cycle loss, which may cause the structure of the generated image altered. The quantitative results of removing each loss term can be found in the Table. 1.

3.4. Enhancing Diagnosis of Chorioamnionitis

Fetal and Maternal Inflammatory Responses (FIR, MIR) are components of ascending infection and chorioamnionitis in

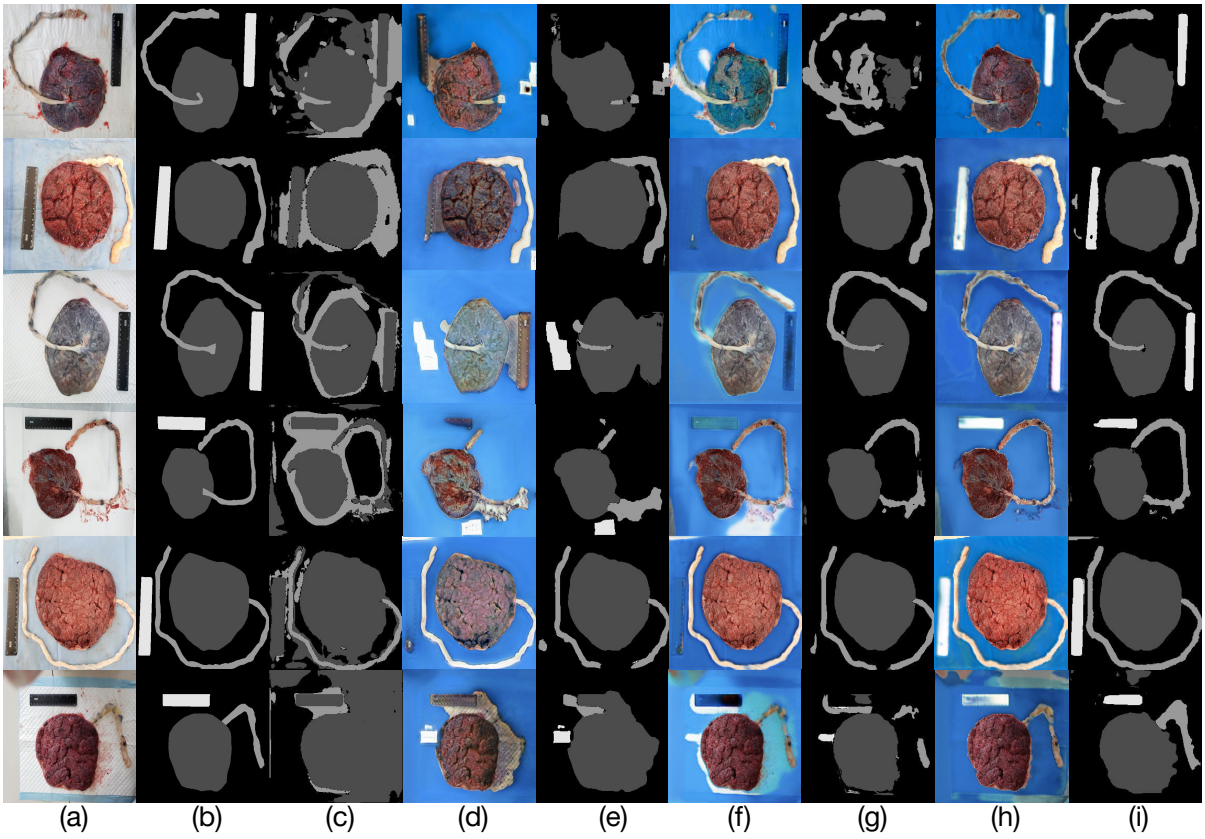

Fig. 3. Segmentation result comparisons. (a) Original images. (b) Ground truth. (c) Segmentation results without adaptation. (d)(f)(h) Translation results using CycleGAN, AGGAN, and our model, respectively. (e)(g)(i) Segmentation results using the translated images to the left of it.

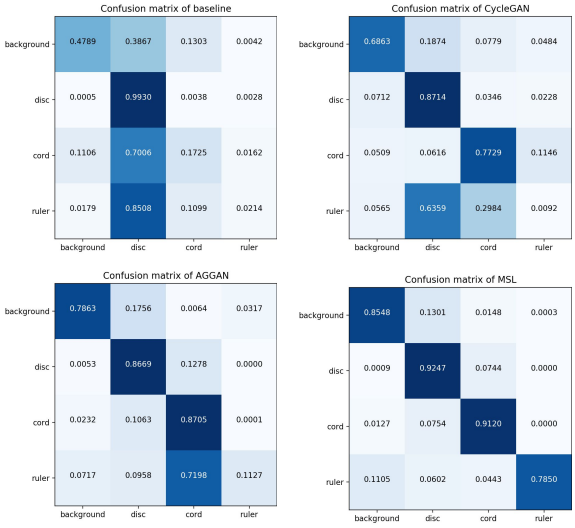

Fig. 4. Confusion matrices of baseline, CycleGAN, AGGAN and MSL.

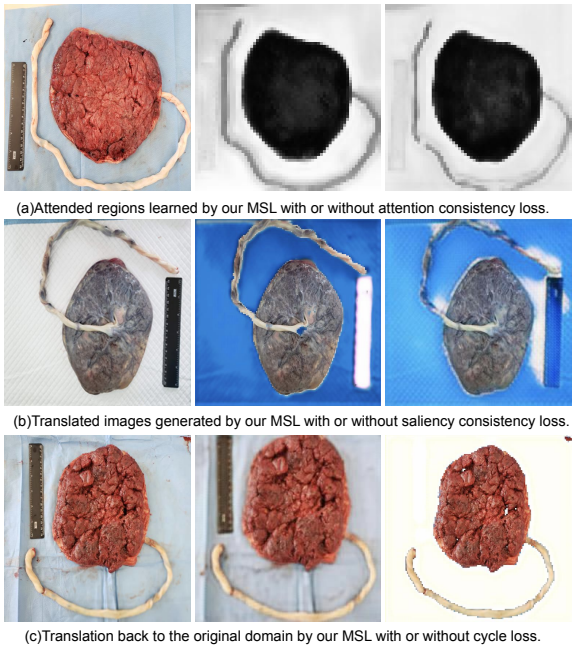

Fig. 5. Ablation Study. Left: original images. Middle: results of MSL. Right: results without the corresponding loss.

pregnancy and predictive of infant sepsis (Romero et al., 2007). To show the application potential of the MSL, we applied a pre-trained source-domain classification model to the target domain to predict if FIR and/or MIR of Stage 2 or 3 are observed within an image. The labels for training are from the collected pathological records (prepared via traditional histological exams), while the images in the target domain were annotated by a perinatal pathologist (from the images alone). Instead of using full translated images as the input, we fed the network with segmented images only occupied by cord and disk to remove

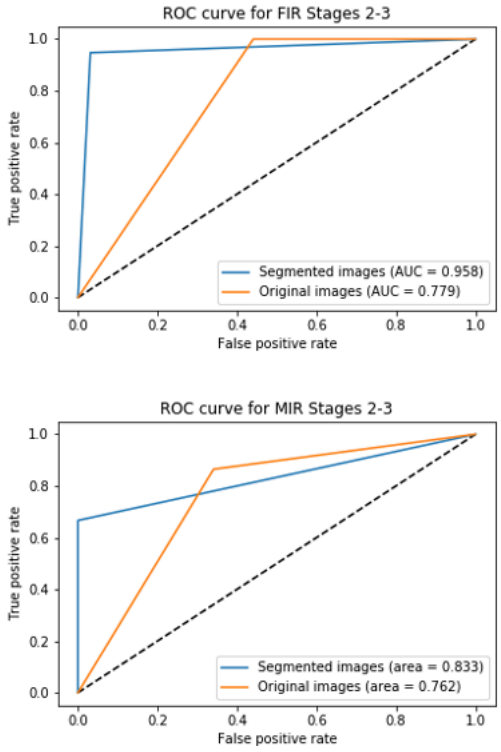

Fig. 6. Prediction results comparison using original or segmented images for FIR (top) and MIR (bottom).

distractions. We compare the ROC curves (Fawcett, 2006) of the FIR and MIR prediction either employing full images or segmented images in Fig. 6, where the AUC has been substantially enhanced when the segmented images are utilized for both cases. Specifically, the Area Under Curve(AUC) for FIR and MIR prediction changed from 0.78 and 0,76 to 0.96 and 0.83 correspondingly.

4. Conclusions

To enable the use of machine learning based pathology image analysis models in very different hospital environments, we have proposed a new unified pipeline adopting multi-region saliency-aware learning for cross-domain placenta image segmentation. Our approach guides the translation between domains by enforcing both the image-level and semantic-level consistency. By introducing the attention and saliency consistency constraints, the translation performance is substantially improved and the segmentation accuracy is enhanced. To our knowledge, this is the first approach for cross-domain placenta image segmentation, with real clinical datasets involving hundreds of patients, to demonstrate clinical relevance/viability in clinical practice. We showed successful use of of our proposed model in instantly detecting significant pathological/abnormal indicators, MIR and FIR, which traditionally need histology to diagnose. This application lays the foundation for further pathological indicator analysis in real clinical situations.

Acknowledgments

This work was supported primarily bythe Bill& Melinda Gates Foundation, Seattle, WA (Grant No. OPP1195074).

References

Chen, X., Xu, C., Yang, X., Tao, D., 2018. Attention-GAN for object transfiguration in wild images, in: Proceedings of the European Conference on Computer Vision, pp. 164–180.

Chen, Y., Wu, C., Zhang, Z., Goldstein, J.A., Gernand, A.D., Wang, J.Z., 2019a. Placentanet: Automatic morphological characterization of placenta photos with deep learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 487–495.

Chen, Y., Zhang, Z., Wu, C., Davaasuren, D., Goldstein, J.A., Gernand, A.D., Wang, J.Z., 2020. Ai-plax: Ai-based placental assessment and examination using photos. Computerized Medical Imaging and Graphics , 101744.

Chen, Y.C., Lin, Y.Y., Yang, M.H., Huang, J.B., 2019b. Crdoco: Pixel-level domain transfer with cross-domain consistency, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1791–1800.

Fawcett, T., 2006. An introduction to roc analysis. Pattern recognition letters 27, 861–874.

Ganin, Y., Lempitsky, V., 2014. Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495 .

Gedraite, E.S., Hadad, M., 2011. Investigation on the effect of a gaussian blur in image filtering and segmentation, in: Proceedings of the ELMAR, IEEE. pp. 393–396.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: Proceedings of the European Conference on Computer Vision, Springer. pp. 694–711.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected crf for accurate brain lesion segmentation. Medical Image Analysis 36, 61–78.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Li, P., Liang, X., Jia, D., Xing, E.P., 2018. Semantic-aware Grad-GAN for virtual-to-real urban scene adaption. arXiv preprint arXiv:1801.01726 .

Long, M., Cao, Y., Wang, J., Jordan, M.I., 2015. Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791 .

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802.

Mejjati, Y.A., Richardt, C., Tompkin, J., Cosker, D., Kim, K.I., 2018. Unsupervised attention-guided image-to-image translation, in: Advances in Neural Information Processing Systems, pp. 3693–3703.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: Proceedings of the IEEE International Conference on 3D Vision, pp. 565–571.

Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M., 2019. Basnet: Boundary-aware salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7479–7489.

Roberts, D.J., 2008. Placental pathology, a survival guide. Archives of pathology & laboratory medicine 132, 641–651.

Romero, R., Gotsch, F., Pineles, B., Kusanovic, J.P., 2007. Inflammation in pregnancy: its roles in reproductive physiology, obstetrical complications, and fetal injury. Nutrition Reviews 65, S194–S202.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 240–248.

Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M., 2018. Learning to adapt structured output space for semantic segmentation,

in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7472–7481.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X., 2017. Residual attention network for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164.

Yan, W., Wang, Y., Gu, S., Huang, L., Yan, F., Xia, L., Tao, Q., 2019. The domain shift problem of medical image segmentation and vendor-adaptation by unet-gan, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 623–631.

Yu, F., Zhao, J., Gong, Y., Wang, Z., Li, Y., Yang, F., Dong, B., Li, Q., Zhang, L., 2019. Annotation-free cardiac vessel segmentation via knowledge transfer from retinal images, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 714–722.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232.