Using the triplet loss for domain adaptation in WCE

Pablo Laiz
University of Barcelona
laizpablo@ub.edu

Jordi Vitrià
University of Barcelona
jordi.vitria@ub.edu

Santi Seguí
University of Barcelona
santi.segui@ub.edu

Abstract

Wireless Capsule Endoscopy (WCE) is a minimally-invasive procedure that, based on a vitamin-size camera that is swallowed by the patient, allows the visualization of the entire gastrointestinal tract. This technology was developed 20 years ago to perform useful and safe studies of different bowel disorders. However, especially the number of captured images and their difficult interpretation has hindered its deployment in some clinical scenarios.

Deep learning methods have the necessary capacity to deal with WCE image interpretation, but training good models is still an open problem for some bowel disorders due to the fact that obtaining a sufficiently large set of positive cases, for the creation and validation of the model, is an arduous task. Moreover, technological advances are rapidly moving forward proposing new hardware able to obtain images with a substantially improved quality. Given these two facts, it is obvious that highly accurate models can only be built by considering heterogeneous datasets composed of images captured by different cameras, and if training methods are able to find invariances with respect to the image acquisition systems.

In this paper, we study the use of deep metric learning, based on the triplet loss function, to improve the generalization of a model over different datasets from different versions of WCE hardware. The obtained results show evidence that with just a few labeled images from a newer camera set, a model that has been trained with images from older systems can be easily adapted to the new environment.

1. Introduction

Wireless Capsule Endoscopy (WCE) is a medical procedure that enables the visualization of the entire gastrointestinal tract. WCE is based on a vitamin-size capsule, equipped with a light source, camera, an optical lens, radio transmitter, and a battery, that is swallowed by the patient and propelled by the peristalsis along all GastroIntestinal (GI) tract, allowing the full visualization of it, from inside, without pain or sedation.

The use of a WCE capsule produces a long video that contains thousands of images that must be individually reviewed by a medical specialist, making the interpretation and analysis of WCE data a complex and time-consuming activity. To overcome this drawback we can use a Computer-Aided Decision system (CAD) to support human interpretation.

The first difficulty that researchers must tackle when developing CADs for WCE is the need to build representative databases for some specific disease or condition. The creation of these databases is time-consuming and economically expensive because of technical questions and also because of the scarcity of positive cases. For this reason, most of the methods we can find in the literature are built and validated with very small datasets.

Another important point that should not be overlooked is that, in the medical field, technological advances are rapidly moving forward. Since the presentation of the first WCE device in 2001, new devices have been periodically presented with better image resolution, illumination or larger field of view. Today, we can find different WCE devices, coming from different manufacturers, that present different technical specifications. Table 1 illustrates some of the most known WCE devices with their main specs, and Figure 1 shows images captured by two different capsules from Medtronic: PillCam SB2 and PillCam SB3. As it can be appreciated, images from PillCam SB3 are better. It is clear that if a model is trained with data from an older capsule, it might not give the expected results when it is evaluated on a newer one since the same data distribution is not guaranteed. However, when the cost of creating a database is that high, it is not acceptable to lose previous databases and build a new one from scratch each time a new device is released.

To overcome this problem, we propose a domain adaptation method based on deep metric learning using the triplet loss. The proposed method aims to adapt the embedding space trained with a large training data set to a new domain where only comparatively few labeled images are available. The embedding space is adapted by generating triplets of images from both domains, with the goal that two images

Capsule	Size (mm)	Weight (g)	Battery life (h)	Resolution (pixels)	Frames per second	Field of view
PillCam SB2 - Given Imaging	26.0 × 11.0	3.40	8	256×256	2 fps	156°
PillCam SB3 - Given Imaging	26.2 × 11.4	3.00	>8	340×340	2-6 fps	156°
EndoCapsule - Olympus America	26.0 × 11.0	3.50	>8	512×512	2 fps	145°
MiroCam - IntroMedic Company	24.5 × 10.8	3.25 - 4.70	>11	320×320	3 fps	170°
OMOM Jinshan - Science and Technology	27.9 × 13.0	6.00	>6 - 8	640×480	2 fps	140°

Table 1: Capsule endoscopy devices used to perform endoscopy operations. The table contain a summary of the main features of each one.

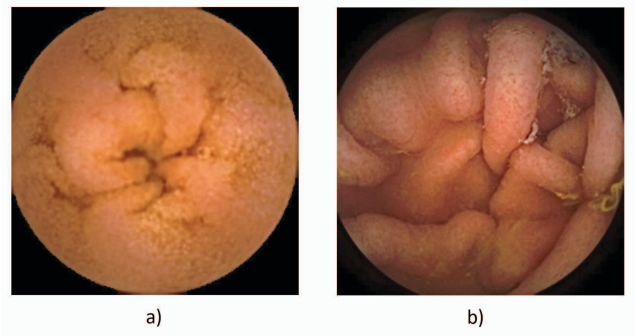

Figure 1: Frames from different capsules present different technical quality. (a) PillCam SB2 capsule image (b) PillCam SB3 capsule image. Image in (b) is clearly better that image in (a).

in the same category are closer than images belonging to different domains. The obtained results show that by using a small labeled dataset from the new domain, the embedding space can be adapted to work **in both domains** with high performance.

The rest of the paper is organized as follows: first, we give an overview of the related work in the field. This is followed by a description of our methodology, presenting the system architecture, followed by the experimental setup and results. Finally, we conclude the paper and give directions for future work.

2. Related Work

Deep Learning for WCE analysis. Several deep learning methods have been proposed for WCE image analysis, dealing with different pathologies such as bleeding, hemorrhages, angiectasia, polyps/cancer, ulcers and hookworms. For example, Zou et al. [24] proposed a CNN-based method to classify the different organs of the digestive system such as stomach, small intestine, and colon; Segui et al. proposed in [20] a classification method of motility events such as turbid, bubbles, clear blob, wrinkle, and wall; finally, Yuan et al. [23] proposed a stacked sparse autoencoder-based approach for detecting polyps [23].

Metric Learning. Metric learning has been extensively used in many machine learning and computer vision applications [15]. Inspired by the success of deep neural networks, deep metric learning has become popular in the last years. These methods aim to learn a discriminative feature embedding, using deep neural networks, such that similar samples are represented by similar embedding vectors and different samples are represented by dissimilar ones. In order to learn these features, embedding deep neural networks are trained using special loss functions such as the Contrastive loss [10], the Triplet loss [13] or the Quadruplet loss [4]. Triplet loss has shown very good results on several image retrieval tasks [2, 9] and in many image classification problems such as facial recognition [19], person re-identification [5, 12] or action recognition [16].

The selection of the triplets is one of the key factors when implementing the triplet loss. In the literature, we can find several methodologies, such as the *Batch All* or *Batch Hard* [7], that face the problem of triplet sampling for each batch.

Domain Adaptation. Domain adaptation methods are designed to deal with the problem of distribution shift across domains. Many domain adaptation (or transfer learning) approaches have been proposed for computer vision applications [18, 8, 3]. To our knowledge, the use of triplet loss in the domain adaptation problem has been limited. Huang et al. [14] defined a triplet visual similarity constraint for learning to rank across two sub-networks using online and offline images. Yu et al. [22] used the triplet loss to correct the selection bias in the triplet selection. Deng et al. [6] used the triplet loss and pseudo-labels for unsupervised domain alignment.

3. Method

The architecture of our system is illustrated in Figure 2. As it can bee seen, the system architecture consists of a classical neural network architecture followed by a normalization layer L_2 and an embedding layer which is optimized with the triplet loss.

In this section, we firstly introduce the triplet loss func-

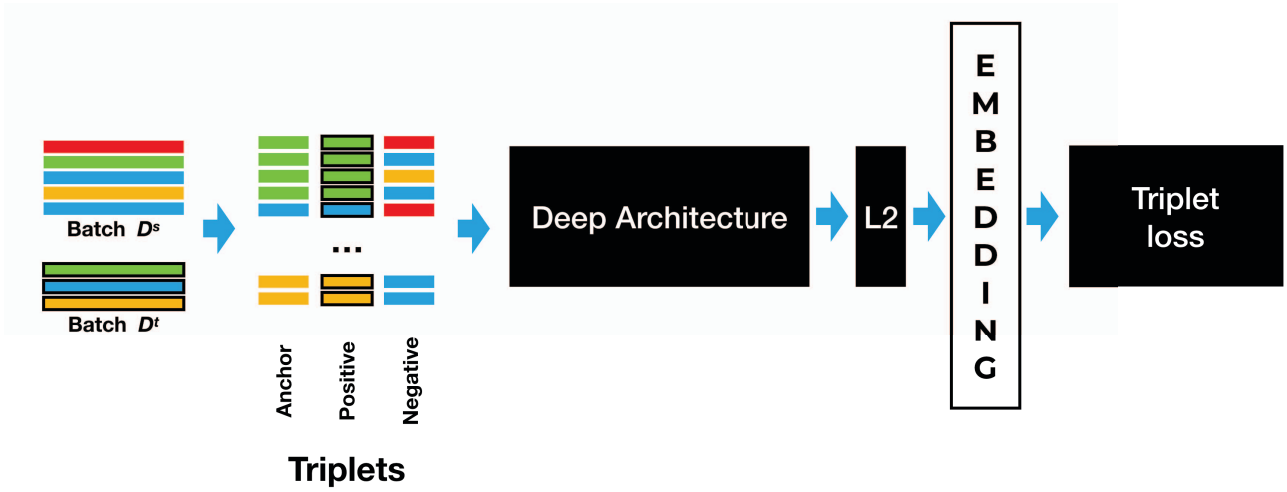

Figure 2: Overview of the proposed CNN structure. The input of the network consists of a batch of images from both domains, source D^s and target D^t. A set of triplets is generated, where anchor images and negative images are from the same domain while positive images are from a different domain but belonging to the same class as the anchor image. The architecture is defined as a standard CNN Architecture followed by L_2 normalization and an embedding layer. It is optimized by using the triplet loss over the generated triplets.

tion for deep metric learning and then we consider its role in the problem of domain adaptation in our scenario.

3.1. Triplet loss for Deep Metric Learning

Let X, Y denote two random variables, which indicate data and label, respectively. Let D be the set of data sampled from $P(X, Y)$. The goal of metric learning is to learn a distance function that assigns small (or large) distance values to a pair of similar (or dissimilar) images. Deep metric learning uses a deep neural network to learn a feature embedding $x' = \Phi(x)$ with the goal of learning a non-linear distance function as follow:

$$d^2(x^i, x^j) = ||\Phi(x^i) - \Phi(x^j)||_2^2$$

In order to learn this embedding representation $\Phi(x_i)$, the triplet loss function is defined as follows:

$$L_{triplet} = \sum_{(x^a, x^p, x^n) \in D} \Big[d^2(x^a, x^p) - d^2(x^a, x^n) + \alpha \Big]_+$$

where $\Big[\cdot\Big]_+ = max(\cdot, 0)$, $\alpha > 0$ and x^a, x^p and x^n refers to anchor, positive and negative examples respectively. Hence, the set of triplets used to train the network is defined as:

$$\tau = \{(x^a, x^p, x^n) | y^a = y^p \text{ and } y^a \neq y^n\}$$

This loss function has shown excellent results learning feature embedding mappings, requiring that the distance between $\Phi(x_a)$ and $\Phi(x_p)$ is smaller than the distance between $\Phi(x_a)$ and $\Phi(x_n)$.

The selection of triplets during training is one of the key factors in order to optimize the network using the triplet loss. As it was said before, there exist two main methodologies to face the sampling problem of triplets: *Batch All* and *Batch Hard* [7]. In *Batch All* strategy a batch of images from the training set is selected and then all possible triplets are generated to optimize the loss. On the other hand, in *Batch Hard* strategy, triplets are generated by seeking, for each sample x_a in the batch, the hardest positive sample, or farthest positive sample $argmax_{x^p}(\|\Phi(x^a) - \Phi(x^p)\|_2^2)$, and the hardest negative sample, or closest negative sample $argmin_{x^n}(\|\Phi(x^a) - \Phi(x^n)\|_2^2)$. Depending on the data set, *Batch All* can lead to a sub-optimal solution while *Batch Hard* can have some convergence problems as a consequence of only considering the hardest samples. For our problem, we will consider the *Batch All* strategy due to the visual heterogeneity of our classes.

3.2. Domain adaption using triplet loss

In our problem, it is assumed that two datasets from different domains are available, the source domain dataset, D^s, and the target domain dataset, D^t, obtained by different capsules. Both datasets are fully labeled but D^s is expected to contain a larger amount of images while D^t is expected to be smaller. To goal is to adapt the model trained with images from D^s to the new environment D^t with minimal efforts.

We assume that there is a covariate shift on the marginal distribution $P(X)$ across domains while the conditional distribution $P(Y|X)$ remains equal. To correct the distribu-

tion shift across domains, we first learn a model that defines the embedding function using the large labeled training set D^s. This model is trained using the standard *Batch All* strategy using all the images from the training set D^s. Then, in order to align the data distributions from both domains and then reduce the whole distribution discrepancy between the source and target datasets, new triplets are generated using both domains, D^s and D^t. Triplets are generated from a batch of N images, where K images are selected from D^t while $N-K$ from D^s. Each triplet consists of an anchor sample x^a that can be from D^s or D^t indifferently, a positive sample x^p that is from a different domain than x^a but with the same label and a negative sample x^n which is from the same domain than the anchor image x^a. Formalizing, the set of triplets used to train the system is defined as follows:

$$\tau = \{(x_i^a, x_j^p, x_i^n) | y_i^a = y_j^p \text{ and } y_j^a \neq y_j^n \text{ and } i \neq j\}$$

where i and j represent any of the classes of the dataset.

4. Experimental Results

4.1. Dataset

In order to validate the proposed system, two different datasets have been used, named **SB2D** and **SB3D**. These datasets have been created using two different versions of the capsules. $SB2D$ has been created using the PillCam SB2 version while the $SB3D$ dataset with the PillCam SB3 version. The most remarkable difference between these two capsules is a 30% improvement in resolution quality (see Table 1 and Figure 1) but also the improvement in illumination, color and the overall image quality.

Both datasets were labeled by expert physicians into 6 different classes: *bubbles, turbid, clear blob, wrinkles, wall*, and *undefined*. All images are resized to 256×256 pixels. $SB2D$ contains a total of 120K labeled images, 20K images per class from a total of 50 different procedures. $SB3D$ contains a total of 6K images, 1K images per class obtained from a total of 10 different procedures. Figure 3 shows six exemplary images for each class.

4.2. Implementation Details and Evaluation Methodology

We implemented the methods using Tensorflow [1]. The system architecture is based on the ResNet-50 [11] with an additional normalization L_2 layer and embedding layer of size 2048. ResNet parameters are preloaded from a trained network using Imagnet dataset. The network is trained for a total of 50 epochs using the stochastic gradient descent (SGD) algorithm with a cyclic learning [21] rate that moves between 0.01 and 1e-5 with stepsize 4000. The batch size is set to 64. All experiments are executed using the standard 2-fold cross-validation methodology where images of the same procedure strictly belong to only one partition.

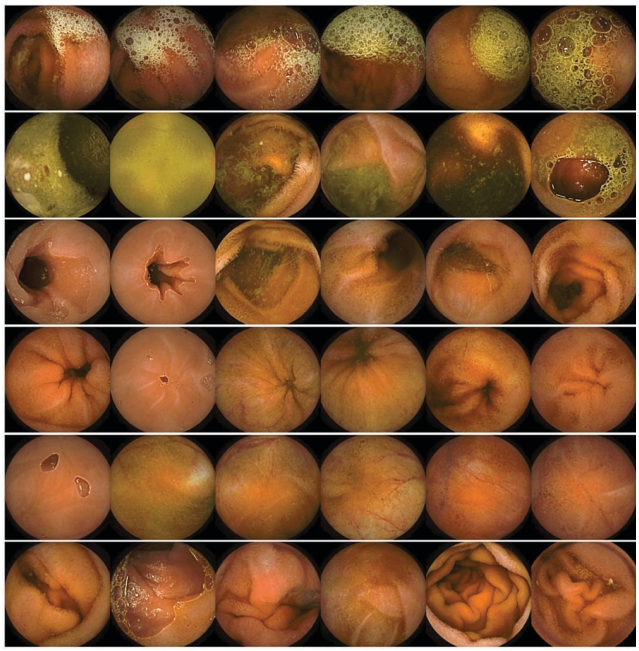

Figure 3: Each row shows six exemplary images for each category in the database: *bubbles, turbid, clear blob, wrinkles, wall*, and *undefined*, respectively

4.3. Results

A first experiment is done to compare the proposed methodology, **TL_SB2-3**, against 3 classical training alternatives **CE_SB2**, **CE_SB2-FT-SB3** and **TL_SB2**. CE_SB2 refers to ResNet-50 trained on $SB2D$ with the classical cross-entropy loss function. CE_SB2-FT-SB3 consists of the CE_SB2 model where the classification layer is fine-tunned using the standard methodology with the dataset $SB3D$. TL_SB2 refers to the proposed architecture presented in Figure 2, based on the ResNet-50 and optimized with the triplet loss function with the dataset $SB2D$. Finally, the proposed method TL_SB2-3 which is optimized with the triplet loss function using data from both domains, $SB2D$ and $SB3D$. In order to avoid overfitting, the parameters of the network are initialized using the TL_SB2 model which is trained using $SB2D$.

As it can be seen in Table 2, CE_SB2 and TL_SB2 obtain good results on $SB2D$ but very poor results when using $SB3D$. On the other hand, CE_SB2-FT-SB3, that uses the classical fine-tunning procedure, obtains satisfactory results on $SB3D$ but its accuracy on $SB2D$ drops. The proposed methods, TL_SB2-3 is able to obtain good results on $SB3D$ without deteriorating its accuracy on the source domain $SB2D$.

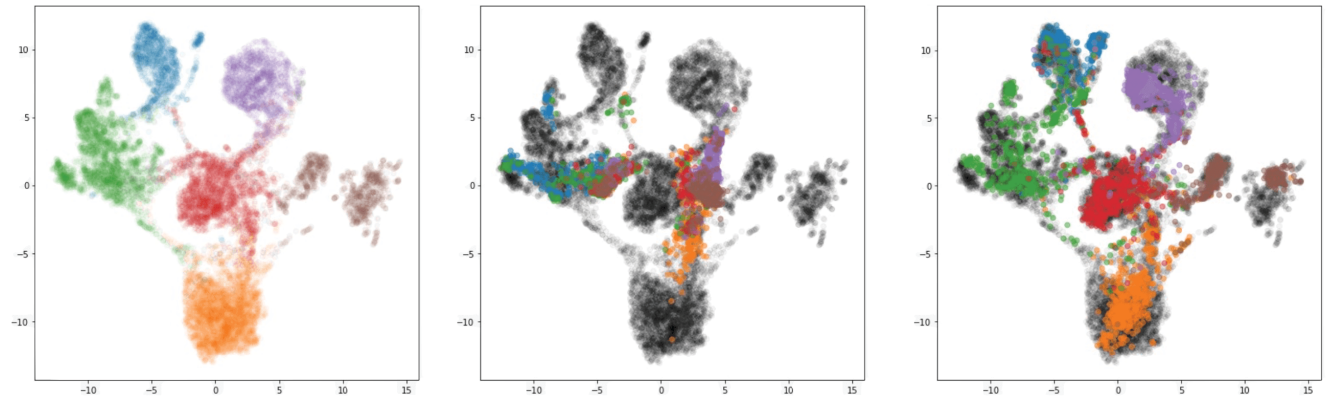

Figure 4: UMAP plots of the learned embedding spaces. Each color represent a different class while (a) illustrates the embedding space obtained with $SB2D$; (b) colored points represents $SB3D$ data projected into the $SB2D$ embedding space (gray); and (c) illustrates the adapted embedding space with $SB2D$ (gray) and $SB3D$ (colored).

Methods	**Accuracy (%)**	
	SB2	**SB3**
CE_SB2	92.5	51.7
CE_SB2-FT-SB3	62.7	87.0
TL_SB2	93.3	41.2
TL_SB2-3	93.1	89.3

Table 2: Comparison of the different proposed methods evaluated in target and source domains respectively, $SB2D$ and $SB3D$.

Method	**SB3 Images**	**Accuracy** (%)
TL_SB2-3	0	41.2
	150	84.6
	300	86.1
	750	87.3
	1500	88.6
	3000	89.3

Table 3: Accuracy of the proposed system evaluated on $SB3D$ with different size of training samples from the target domain. Data is obtained uniformly per class (k = 6) and procedure (n = 5).

Figure 4 shows the UMAP [17] plots of the learned embedding spaces. Each color represent a different class. Plot (a) illustrates the embedding space obtained with $SB2D$; in plot (b) colored points represent $SB3D$ data projected into the $SB2D$ embedding space (gray); and plot (c) illustrates the adapted embedding space with $SB2D$ (grey) and $SB3D$ (colored). As it can observed, there exists a clear shift between the distribution from different domains which is adapted after training with both domains.

In the second experiment (see Table 3), we evaluated the accuracy of the system using different amount of images per procedure. A total of 10 procedures were selected using the standard 2-fold cross-validation strategy. As it can be seen, with just 30 images per procedure (5 images per class), i.e. a total of 150 images since 5 videos are used for creating the training data, the accuracy of the system is increased from 41.28% to 84.64%. As more images per procedure are used, the accuracy increases, obtaining an accuracy of 89.32% when all images of all procedures are used.

Finally, Table 4 shows the behavior of the system when more diversity of the target domain is introduced. To perform this experiment, the accuracy of the system is evaluated when a different amount of videos are used but setting the same amount of labeled data, 600 images. As it can be seen, the accuracy of the system is enhanced as the number of different used videos is increased. Hence, it is more important to use a diverse set of data, for example using more procedures, than using a large amount of images from the same procedure.

Finally, Figure 5 shows a set of anchor images acquired with PillCam SB3, the target domain, and its top more similar images from the PillCam SB2 dataset, the source domain. Central images in each row represent anchor images while the three images at left are the top most similar images before adapting the domain, and the three images at the right are the top most similar images when the embedding has been adapted. As it can bee seen, similar images when using the adapted embedding are really similar in shape and

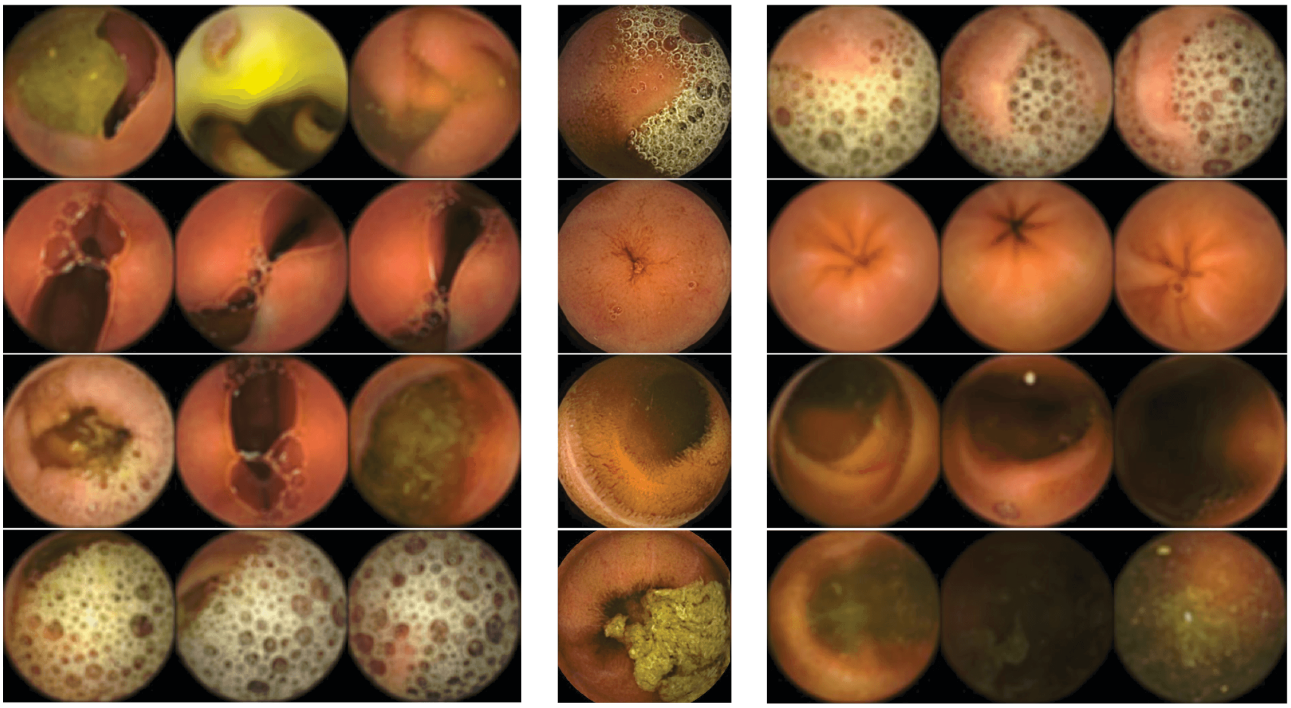

Figure 5: Each row shows a query where (b) is the anchor images from the target domain $SB3D$ capsule, (a) the three most similar images to the anchor image without adapting the model to the target domain and (c) The thee most similar images adapting the model to the target domain.

Method	SB3 Videos	Accuracy (%)
TL_SB2-3	1	80.2
	2	85.7
	3	86.8
	4	86.8
	5	86.9

Table 4: Accuracy of the proposed system evaluated on $SB3D$ trained with 600 from $SB3D$ using different number of procedures.

color to the anchor images, although their look and feel is blurrier.

5. Conclusions

In this work, we have explored the use of deep metric learning, based on the triplet loss function, to improve the generalization of a model over different datasets from different versions of WCE capsules. The proposed method is trained using a larger dataset from a source domain, using an old WCE device, and is adapted to work on a target domain that represents images obtained by a new WCE device, with minimal labeling efforts. Results show evidence that with just a few labeled images from a newer camera, a model that has been trained with images from older systems can be readily used in the new environment.

We also explored, evaluated and compared several different transfer learning solutions when dealing with small target domain datasets. We have shown that the triplet loss function may be well suited for dealing with the problem of data distribution shift over different domains. Particularly, we study the effects of using different amounts of images and procedures, concluding that diversity is more important than the amount.

Acknowledgements

We want to thank Carolina Malagelada and Fernando Azpiroz from Hospital General de la Vall d'Hebron for their clinical insights, the team from CorporateHealth International ApS for their feedback and economic support and NVIDIA for their GPU donations. This work has been also supported by MINECO Grant RTI2018-095232-B-C21 and SGR 1742.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 4

[2] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(Mar):1109–1135, 2010. 2

[3] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5315–5324, 2015. 2

[4] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2017. 2

[5] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the iEEE conference on computer vision and pattern recognition*, pages 1335–1344, 2016. 2

[6] W. Deng, L. Zheng, and J. Jiao. Domain alignment with triplets. *arXiv preprint arXiv:1812.00893*, 2018. 2

[7] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48, 04 2015. 2, 3

[8] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014. 2

[9] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016. 2

[10] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[12] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2

[13] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015. 2

[14] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 2

[15] B. Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013. 2

[16] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016. 2

[17] L. McInnes, J. Healy, N. Saul, and L. Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018. 5

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014. 2

[19] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[20] S. Seguí, M. Drozdzal, G. Pascual, P. Radeva, C. Malagelada, F. Azpiroz, and J. Vitrià. Generic feature learning for wireless capsule endoscopy analysis. *Computers in biology and medicine*, 79:163–172, 2016. 2

[21] L. N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472, March 2017. 4

[22] B. Yu, T. Liu, M. Gong, C. Ding, and D. Tao. Correcting the triplet selection bias for triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–87, 2018. 2

[23] Y. Yuan and M. Q.-H. Meng. Deep learning for polyp recognition in wireless capsule endoscopy images. *Medical physics*, 44(4):1379–1389, 2017. 2

[24] Y. Zou, L. Li, Y. Wang, J. Yu, Y. Li, and W. Deng. Classifying digestive organs in wireless capsule endoscopy images based on deep convolutional neural network. In *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pages 1274–1278. IEEE, 2015. 2