Learning with Synthesized Data for Generalizable Lesion Detection in Real PET Images

Xinyi Yang[1], Bennett Chin[1], Michael Silosky[1], Daniel Litwiller[2],
Debashis Ghosh[1], and Fuyong Xing[1](✉)

[1] University of Colorado Anschutz Medical Campus, Aurora, USA
fuyong.xing@cuanschutz.edu
[2] GE Healthcare, Denver, USA

Abstract. Deep neural networks have recently achieved impressive performance of automated tumor/lesion quantification with positron emission tomography (PET) imaging. However, deep learning usually requires a large amount of diverse training data, which is difficult for some applications such as neuroendocrine tumor (NET) image quantification, because of low incidence of the disease and expensive annotation of PET data. In addition, current deep lesion detection models often suffer from performance degradation when applied to PET images acquired with different scanners or protocols. In this paper, we propose a novel single-source domain generalization method, which learns with human annotation-free, list mode-synthesized PET images, for hepatic lesion identification in real-world clinical PET data. We first design a specific data augmentation module to generate out-of-domain images from the synthesized data, and incorporate it into a deep neural network for cross domain-consistent feature encoding. Then, we introduce a novel patch-based gradient reversal mechanism and explicitly encourage the network to learn domain-invariant features. We evaluate the proposed method on multiple cross-scanner ^{68}Ga-DOTATATE PET liver NET image datasets. The experiments show that our method significantly improves lesion detection performance compared with the baseline and outperforms recent state-of-the-art domain generalization approaches.

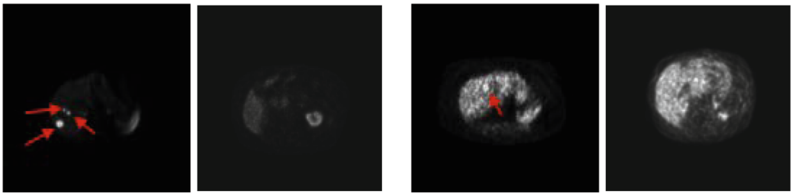
Keywords: Lesion detection · PET images · domain generalization

1 Introduction

Deep neural networks have recently shown impressive performance on lesion quantification in positron emission tomography (PET) images [6]; however, they usually rely on a large amount of well-annotated, diverse data for model training.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43904-9_12.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 116–126, 2023.
https://doi.org/10.1007/978-3-031-43904-9_12



(a) Synthesized (b) Synthesized (c) Real (d) Real

Fig. 1. Example PET images. Diseased subjects with (a) simulated and (c) real lesions (red arrows), and normal (b) synthesized and (d) real subjects without lesions. (Color figure online)

This is difficult or even infeasible for some applications such as lesion identification in neuroendocrine tumor (NET) images, because NETs are rare tumors and lesion annotation in low-resolution, noisy PET images is expensive. To address the data shortage issue, we propose to train a deep model for lesion detection with synthesized PET images generated from list mode PET data, which is low-cost and does not require human effort for manual data annotation.

Synthesized PET images may exhibit a different data distribution from real clinical images (see Fig. 1), i.e., a domain shift, which can pose significant challenges to model generalization. To address domain shifts, domain adaptation requires access to target data for model training [5,29], while domain generalization (DG) trains a model with only source data [39] and has recently attracted increasing attention in medical imaging [1,13,15,18]. Most of current DG methods rely on multiple sources of data to learn a generalizable model, i.e., multi-source DG (MDG); however, multi-source data collection is often difficult in real practice due to privacy concerns or budget deficits. Although single-source DG (SDG) using only one source dataset has been applied to medical images [12,14,32], very few studies focus on SDG with PET imaging and the current SDG methods may not be suitable for lesion identification on PET data. For instance, many existing methods use a complicated, multi-stage model design pipeline [10,23,30], which introduces an additional layer of algorithm variability. This situation will become worse for PET images, which typically have a poor signal-to-noise ratio and low spatial resolution. Several other SDG approaches [26,31,34] leverage unique characteristics of the imaging modalities, e.g., color spectrum of histological stained images, which are not applicable to PET data.

In this paper, we propose a novel single-stage SDG framework, which learns with human annotation-free, list mode-synthesized PET images for generalizable lesion detection in real clinical data. Compared with domain adaptation and MDG, the proposed method, while more challenging, is quite practical for real applications due to the relatively cheaper NET data collection and annotation. Specifically, we design a new data augmentation module, which generates out-of-domain samples from single-source data with multi-scale random convolutions. We integrate this module into a deep lesion detection neural network and introduce a cross-domain consistency constraint for feature encoding between

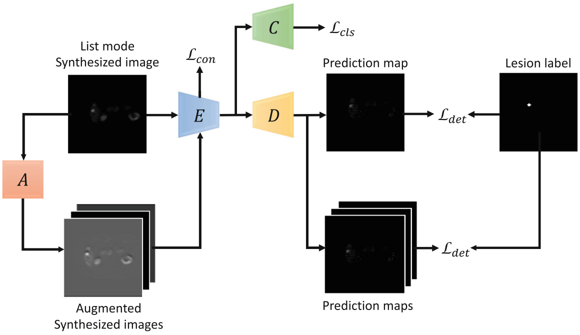


Fig. 2. The proposed SDG framework for generalizable lesion detection. The A, E, D and C represents the data augmentation module, feature encoder, decoder and domain classifier, respectively. The $\mathcal{L}_{det}, \mathcal{L}_{cls}$ and $\mathcal{L}_{con}$ denote the losses for lesion detection, domain classification and cross-domain consistency, respectively.

original synthesized and augmented images. Furthermore, we incorporate a novel patch-based gradient reversal mechanism into the network and accomplish a pretext task of domain classification, which explicitly promotes domain-invariant, generalizable representation learning. Trained with a single-source synthesized dataset, the proposed method provides superior performance of hepatic lesion detection in multiple cross-scanner real clinical PET image datasets, compared with the reference baseline and recent state-of-the-art SDG methods.

2 Methodology

Figure 2 presents the proposed SDG framework. Given a source-domain dataset of list mode-synthesized 3D PET images and corresponding lesion labels ($\boldsymbol{X}_S$, $\boldsymbol{Y}_S$), the goal of the framework is to learn a lesion detection model H, composed of E and D, which generalizes to real clinical PET image data. The framework first feeds synthesized images $\boldsymbol{X}_S$ into a random-convolution data augmentation module A and generates out-of-domain samples $\boldsymbol{X}_A = A(\boldsymbol{X}_S)$. Then, it provides both original and augmented images, $\boldsymbol{X}_S$ and $\boldsymbol{X}_A$, to a feature encoder E, which is followed by a decoder D for lesion detection. The framework imposes a cross-domain consistency constraint on the encoder E to promote consistent feature learning between $\boldsymbol{X}_S$ and $\boldsymbol{X}_A$. Meanwhile, it uses a patch gradient reversal-based domain classifier C to differentiate $\boldsymbol{X}_A$ from $\boldsymbol{X}_S$ and further encourages the encoder E to learn domain-agnostic representations for H.

2.1 Synthesized Data Augmentation

In the synthesized PET image dataset, each subject have multiple simulated lesions of varying size with known boundaries [11], and thus no human annotation is required. However, this synthesized dataset presents a significant domain shift from real clinical data, as they have markedly different image textures and voxel intensity values (see Fig. 1). Inspired by previous domain generalization work [39], we introduce a specific data augmentation module to generate out-of-domain samples from this single-source synthesized dataset for generalizable model learning (see Fig. 2). Specifically, we tailor a random convolution technique [33] for synthesized PET image augmentation with the following substantial improvement: 1) extend it from single value-prediction image classification to a more challenging dense prediction task of lesion detection; 2) refine it to produce realistic augmented images where the organ regions are brighter than image background, instead of randomly switching the foreground and background intensity values; 3) place a cross-domain consistency constraint on the encoding features, rather than output predictions, of original synthesized and augmented images, so as to directly encourage consistent representation learning between the source and other domains. This module can preserve global shapes or the structure of objects (e.g., lesions and livers) in images but distorts local textures, so that the lesion detection model learned with these augmented images can generalize to unseen real-world PET image data, which typically have high lesion heterogeneity and divergent texture styles.

Given a synthesized input image $\boldsymbol{x}_S \in \boldsymbol{X}_S$, our data augmentation module A first performs a *random* convolution operation $R(\boldsymbol{x}_S)$ with a $k \times k$ kernel R, where the kernel size k and the convolutional weights are randomly sampled from a multi-scale set $\mathcal{K} = \{1, 3, 5, 7\}$ and a normal distribution $\mathcal{N}(0, 1/k^2)$, respectively. Then, inspired by [7,35,36], we mix $R(\boldsymbol{x}_S)$ and $\boldsymbol{x}_S$ to generate a new mixed image $\boldsymbol{x}_M$ via a convex combination, $\boldsymbol{x}_M = \alpha \boldsymbol{x}_S + (1 - \alpha) R(\boldsymbol{x}_S)$, where $\alpha \in [0, 1]$ is randomly sampled from a uniform distribution $\mathcal{U}(0, 1)$. This data mixing strategy allows continuous interpolation between the source domain and a randomly generated out-of-distribution domain to improve model generalizability. Finally, if the foreground (i.e., lesion region) of $\boldsymbol{x}_M$ has a higher mean intensity value than the background (non-lesion region), we use $\boldsymbol{x}_M$ as the final augmented image, $\boldsymbol{x}_A = \boldsymbol{x}_M$. Otherwise, we invert the image intensity of $\boldsymbol{x}_M$ to obtain $\boldsymbol{x}_A = x_M^{max}\mathbf{1} + x_M^{min}\mathbf{1} - \boldsymbol{x}_M$, where x_M^{max}/x_M^{min} is the maximum/minimum intensity of $\boldsymbol{x}_M$ and $\mathbf{1}$ is a matrix with all elements being one and the same dimension as $\boldsymbol{x}_M$. This intensity inversion operation is to ensure the lesion region has higher intensity values than other regions, mimicking the image characteristics of real-world PET data in our experiments. Here we calculate the mean intensity value of the background from the regions that have a distance greater than half of the image width from the closest lesion center.

In our modeling, for each synthesized training image $\boldsymbol{x}_S$, we generate multiple augmented images (i.e., 3), $\{\boldsymbol{x}_A^i\}_{i=1}^3$, and feed them into the encoder E for feature learning. Due to the distance preservation property of random convolutions [33], the module A changes local textures but preserves object shapes at different scales, and thus $\boldsymbol{x}_S$ and $\{\boldsymbol{x}_A^i\}_{i=1}^3$ should have identical semantic content, such as lesion presence, quantity and positions. Therefore, they should have consistent representations in the feature space, i.e., $E(\boldsymbol{x}_S) \approx E(\boldsymbol{x}_A^i)$, $i = 1, 2, 3$. To this end, we place a cross-domain consistency loss $\mathcal{L}_{con}$ on top of the encoder E as

$$\mathcal{L}_{con} = \frac{1}{3}\sum_{i=1}^{3}\mathcal{L}_{con}^{i}, \tag{1}$$

$$\mathcal{L}_{con}^{i} = \mathbb{E}_{\boldsymbol{x}_S \sim \boldsymbol{X}_S}[\frac{1}{|E(\boldsymbol{x}_S)|}||E(\boldsymbol{x}_S) - E(\boldsymbol{x}_A^i)||_F^2],\ \forall i \in \{1, 2, 3\}, \tag{2}$$

where $\mathbb{E}$ is an expectation operator, $|E(\boldsymbol{x}_S)|$ is the number of elements in $E(\boldsymbol{x}_S)$, and $||\cdot||_F$ denotes the Frobenius norm. Unlike the previously reported work [33] promotes consistent output-layer predictions, the loss $\mathcal{L}_{con}$ in Eq. (1) directly encourages the encoder E to extract cross-domain consistent representations, which improves model generalization more effectively for dense prediction tasks [8], such as lesion detection. We hypothesize that forcing similar feature encoding between $\boldsymbol{x}_S$ and $\boldsymbol{x}_A^i$ can facilitate image content preservation for lesion detection. In addition, we adopt a mean squared error (MSE) to measure the consistency, different from [33] using the Kullback-Leibler divergence for image classification, which is not suitable for our application. Note that the convolution weights in module A are randomly sampled within each iteration and are not updated during model training.

2.2 Patch Gradient Reversal

Because of random convolution weights, the original synthesized $\boldsymbol{X}_S$ and augmented $\boldsymbol{X}_A$ data can have substantially different image appearances. Consequently, the use of the loss $\mathcal{L}_{con}$ in Eq. (1) may not be sufficient to enforce consistent feature encoding. To address this issue, we propose to use a pretext task as an additional information resource for the encoder E and to further promote domain-agnostic representation learning. Specifically, we incorporate a domain classifier C on top of the encoder E to perform a pretext task of domain discrimination, i.e., predict whether each input image is from the original synthesized data $\boldsymbol{X}_S$ or augmented data $\boldsymbol{X}_A$. This domain classification accompanies the main task of lesion detection (see Fig. 2) to assist with feature learning. In this way, the encoder E improves feature invariance to domain changes by penalizing domain classification accuracy, while retaining feature discriminativeness to lesion prediction via the decoder D. This is different from other methods [2,25] that use intrinsic supervision signals within a single image to perform an auxiliary task, e.g., solving jigsaw puzzles, for model generalization enhancement.

In general, the classifier C will encourage the encoder E to learn discriminative features for accurate domain classification. In order to make features invariant to different domains, we reverse the gradient propagated from the domain classifier C with a multiplication of -1 [3] and send this reversed gradient to the encoder E, while keeping all the other gradient flows unchanged during the backpropagation for model training. Note that the computation in forward propagation of our network is the same as that in a standard feed-forward neural network. Compared with [3], we make the following significant improvements: 1) Instead of back propagating the reversed gradient from a single-valued prediction of the domain label of the entire input image, we introduce a patch-based gradient reversal to enhance feature representation invariance to local texture or style changes. Inspired by [9], we design the domain classifier C with a fully convolutional network and produce a prediction map, where each element corresponds to a local patch of input image, i.e., conducting small patch categorization. We then apply the reversal operation to the gradient propagated from the prediction map and feed it into the encoder E for feature learning. 2) Motivated by [17], we remove the sigmoid layer in [3] and replace the cross-entropy loss by an MSE loss, which can facilitate the adversarial training caused by the gradient reversal. With the MSE loss, the patch-based gradient reversal penalizes image structures and enhances feature robustness and invariance to style shifts at the local-patch level, so that the lesion detection model H (i.e., E followed by D) learned with source data annotations is directly applicable to unseen domains [4,24], based on the covariate shift assumption [20].

Formally, let $\boldsymbol{X} = \{\boldsymbol{X}_S, \boldsymbol{X}_A\}$ denote the input data for the encoder E and $\boldsymbol{Z} = \{\boldsymbol{Z}_S, \boldsymbol{Z}_A\}$ represent the corresponding domain category labels, with $\boldsymbol{Z}_S$ and $\boldsymbol{Z}_A$ for the original source images $\boldsymbol{X}_S$ and corresponding random convolution-augmented image $\boldsymbol{X}_A$, respectively. Each label $\boldsymbol{z} \in \boldsymbol{Z}$ is a 3D image with all voxel intensity being 0's for $\boldsymbol{z} \in \boldsymbol{Z}_S$ or 1's for $\boldsymbol{z} \in \boldsymbol{Z}_A$. We define the domain classification objective $\mathcal{L}_{cls}$ as follows

$$\mathcal{L}_{cls} = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{z})\sim(\boldsymbol{X},\boldsymbol{Z})}[\frac{1}{|\boldsymbol{z}|}||\boldsymbol{z} - \hat{\boldsymbol{z}}||_F^2], \tag{3}$$

where $\hat{\boldsymbol{z}} = C(E(\boldsymbol{x}))$ is the prediction of $\boldsymbol{x}$.

For source-domain data $(\boldsymbol{X}_S,\ \boldsymbol{Y}_S)$, the augmented images $\boldsymbol{X}_A$ have the same gold-standard lesion labels $\boldsymbol{Y}_A = \boldsymbol{Y}_S$, each of which is a 3D binary image with $1's$ for lesion voxels and $0's$ for non-lesion regions. Let $\boldsymbol{Y} = \{\boldsymbol{Y}_S,\ \boldsymbol{Y}_A\}$. We formulate the lesion detection objective $\mathcal{L}_{det}$ as

$$\begin{aligned}\mathcal{L}_{det} = \beta\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim(\boldsymbol{X},\boldsymbol{Y})}[\frac{-1}{|\boldsymbol{y}|}\sum_{j=1}^{|\boldsymbol{y}|}(\gamma y^j \log \hat{y}^j + (1-y^j)\log(1-\hat{y}^j))] \\ +\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim(\boldsymbol{X},\boldsymbol{Y})}[1 - \frac{2\sum_{j=1}^{|\boldsymbol{y}|} y^j \hat{y}^j + \epsilon}{\sum_{j=1}^{|\boldsymbol{y}|} y^j + \sum_{j=1}^{|\hat{\boldsymbol{y}}|} \hat{y}^j + \epsilon}],\end{aligned} \tag{4}$$

where the first and second terms in Eq. (4) are a weighted binary cross-entropy loss and a Dice loss, respectively. We add a smooth term, $\epsilon = 10^{-6}$, to the Dice loss to avoid division by zero. The y^j and $\hat{y}^j$ are the j-th values of $\boldsymbol{y}$ and corresponding prediction $\hat{\boldsymbol{y}}$, respectively. The β controls the relative importance between the two losses, and γ emphasizes the lesions in each image. The combo loss $\mathcal{L}_{det}$ can further help address the data imbalance issue [22], i.e., lesions have significantly fewer voxels than the non-lesion regions including the background.

With the losses in Eqs. (1)–(4), we define the following full objective as

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_{con}\mathcal{L}_{con} + \lambda_{cls}\mathcal{L}_{cls}, \tag{5}$$

where λ_{con} and λ_{cls} are weighting parameters. Note that while we minimize $\mathcal{L}$ for model training, we reverse the gradient propagated from the domain classifier C before sending it to the encoder E during the backpropagation.

3 Experiments

Datasets. We evaluate the proposed method with multiple ^{68}Ga-DOTATATE PET liver NET image datasets that are acquired using different PET/CT scanners and/or imaging protocols. The synthesized source-domain dataset contains 103 simulated subjects, with an average of 5 lesions and 153 transverse slices per subject. This dataset is synthesized using list mode data from a single real, healthy subject acquired on a GE Discovery MI PET/CT scanner with list mode reconstruction [11,37]. We collect two additional real ^{68}Ga-DOTATATE PET liver NET image datasets that serve as unseen domains. The first dataset ($Real1$) has 123 real subjects with about 230 hepatic lesions in total and is acquired using clinical reconstructions with a photomultiplier tube-based PET/CT scanner (GE Discovery STE). The second real-world dataset ($Real2$) consists of 65 cases with around 113 lesions and is acquired from clinical reconstructions using a digital PET/CT scanner (GE Discovery MI). Following [28,38], we randomly split the synthesized dataset and the $Real1$ dataset into 60%, 20% and 20% for training, validation and testing, respectively. Due to the relatively small size of $Real2$, we use a two-fold cross-validation for model evaluation on this dataset. Here we split the real datasets to learn fully supervised models for a comparison with the proposed method.

Table 1. Domain generalization evaluation on different datasets. Each method is run 5 times, and the mean and standard deviation (std) of each metric (%) are reported: *mean* (*std*). The $*$ means a statistically significant difference (p-value < 0.05) between our method and others. The highest F$_1$ score is highlighted with **bold** font.

	Real1			Real2		
	F$_1$	Precision	Recall	F$_1$	Precision	Recall
CMSDG [18]	57.7* (3.2)	64.2 (13.0)	54.6 (7.9)	49.8 (6.5)	45.9 (8.3)	56.5 (9.8)
RandConv [33]	58.1* (1.3)	72.8 (10.4)	49.4 (5.9)	48.1* (2.3)	81.2 (4.4)	34.8 (2.1)
L2D [27]	46.4* (1.6)	41.1 (4.8)	54.7 (7.6)	41.9* (4.4)	30.1 (4.5)	71.0 (3.6)
Baseline	39.4* (1.9)	30.9 (2.4)	54.6 (2.2)	45.8* (5.0)	66.7 (10.0)	37.4 (4.2)
Aug.	51.5* (5.0)	45.4 (7.4)	60.3 (1.1)	44.3* (3.7)	53.5 (12.6)	41.6 (8.3)
Aug.+$\mathcal{L}_{con}$	55.1* (3.3)	60.0 (14.5)	52.3 (7.5)	44.5* (2.2)	67.0 (12.3)	35.0 (5.5)
Aug.+$\mathcal{L}_{con}$+*gGR*	58.7* (1.8)	62.5 (4.7)	55.5 (3.2)	48.2* (1.7)	69.6 (6.5)	37.8 (3.5)
Ours	**63.1** (0.5)	74.2 (9.2)	55.6 (5.4)	**53.8** (2.1)	58.0 (4.3)	50.9 (6.1)
Upper-bound	75.5 (4.3)	81.7 (3.6)	70.6 (7.0)	63.5 (7.6)	57.5 (7.3)	75.7 (5.1)

Implementation Details and Evaluation Metrics. We implement the encoder E and the decoder D with a U-Net architecture [19], with four downsampling and upsampling layers in the encoder and decoder, respectively. We build the domain classifier C using three stacked stride-1 convolutional layers of kernel size of 4, and each convolution is followed by a batch normalization and a leaky ReLU activation [16]. We set $\beta = 6, \gamma = 5$ in Eq. (4) and $\lambda_{con} = 1, \lambda_{cls} = 1$ in Eq. (5). We train the model using stochastic gradient descent with Nesterov momentum with learning rate $= 5 \times 10^{-4}$, momentum $= 0.99$ and batch size $= 1$. We perform standard image augmentation including random scaling, noise adding and image contrast adjustment before applying random convolutions in the module A. In the testing stage, we adopt the model H to produce a prediction map for each input image, and identify lesions with a threshold (i.e., 0.1) to binarize the map followed by a connected component analysis, which helps detect individual lesions by identifying connected regions from the binarized map. We use precision, recall and F_1 score as model evaluation metrics [21,28,38].

Comparison with State of the Art. We compare our method with several recent state-of-the-art SDG approaches, including causality-inspired SDG (*CISDG*) [18], *RandConv* [33], and learning to diversify (*L2D*) [27]. We run each model 5 times with different random seeds and report the mean and standard deviation. Table 1 presents the comparison results on the two unseen-domain datasets. Our method significantly outperforms the state-of-the-art approaches in terms of F_1 score, with p-value < 0.05 in Student's t-test for almost all cases on both datasets. In addition, our method gives lower standard deviation of F_1 than others. This indicates that compared with the competitor approaches, our method is relatively more effective and stable in learning generalizable representations for lesion detection in a very challenging situation, i.e., learning with a single-source synthesized PET image dataset to generalize to real clinical data. The qualitative results are provided in the Supplementary Material.

Ablation Study. In Table 1, the *Baseline* represents a lesion detection model trained with the source data but without the data augmentation module A, $\mathcal{L}_{con}$ or $\mathcal{L}_{cls}$. We then evaluate different variants of our method by sequentially adding one component to the *Baseline* model: 1) using only the module A for model training (*Aug.*), 2) using module A and $\mathcal{L}_{con}$ (*Aug.*+$\mathcal{L}_{con}$), and 3) using module A, $\mathcal{L}_{con}$ and $\mathcal{L}_{cls}$ (*Ours*). We also report the performance of the model, *Aug.*+$\mathcal{L}_{con}$+*gGR*, which does not use the proposed patch-based gradient reversal but outputs a single-value prediction for the entire input image, i.e., global gradient reversal (gGR). The *Upper-bound* means training with real-world images and gold-standard labels from the testing datasets. We note that using the data augmentation module A can significantly improve the lesion detection performance compared with the *Baseline* on the *Real*1 dataset, and combining data augmentation and patch gradient reversal can further close the gap to the *Upper-bound* model. Our method also outperforms the *Baseline* model by a large margin on the *Real*2 dataset, suggesting the effectiveness of our method.

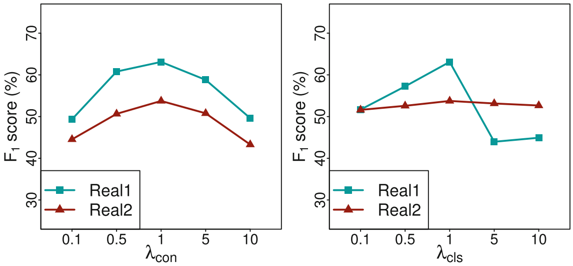


Fig. 3. The F_1 score of our method with different values of λ_{con} (left) and λ_{cls} (right).

Effects of Parameters. We evaluate the effects of λ_{con} and λ_{cls} of our method on lesion detection in Fig. 3. The lesion detection performance improves when increasing λ_{con} from 0.1 to 1. However, a further emphasis on consistent feature encoding, e.g., $\lambda_{con} \geq 5$, decreases the F_1 score. This suggests the importance of an appropriate λ_{con} value. In addition, we observe a similar trend of the F_1 curve for the λ_{cls}, especially for the *Real*1 dataset, and this indicates the necessity of the domain classification pretext task.

4 Conclusion

We propose a novel SDG framework that uses only a single dataset for hepatic lesion detection in real clinical PET images, without any human data annotations. With a specific data augmentation module and a new patch-based gradient reversal, the framework can learn domain-invariant representations and generalize to unseen domains. The experiments show that our method outperforms the reference baseline and recent state-of-the-art SDG approaches on cross-scanner or -protocol real PET image datasets. A potential limitation may be the need of a proper selection of weights for different tasks during model training.

References

1. Cai, J., et al.: Generalizing nucleus recognition model in multi-source Ki67 immunohistochemistry stained images via domain-specific pruning. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 277–287. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_27
2. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: CVPR, pp. 2229–2238 (2019)
3. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML, pp. 1180–1189 (2015)
4. Geirhos, R., et al.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: ICLR, pp. 1–12 (2019)
5. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE TBME **69**(3), 1173–1185 (2021)

6. Hatt, M., Laurent, B., Ouahabi, A., Fayad, H., Tan, S.: The first MICCAI challenge on pet tumor segmentation. MedIA **44**, 177–195 (2018)
7. Hendrycks, D., et al.: AugMix: a simple data processing method to improve robustness and uncertainty. In: ICLR, pp. 1–11 (2020)
8. Hong, W., Wang, Z., Yang, M., Yuan, J.: Conditional generative adversarial network for structured domain adaptation. In: CVPR, pp. 1335–1344 (2018)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR, pp. 1125–1134 (2017)
10. Kamraoui, R.A., et al.: DeepLesionBrain: towards a broader deep-learning generalization for multiple sclerosis lesion segmentation. MedIA **76**, 102312 (2022)
11. Leung, K.H., et al.: A physics-guided modular deep-learning based automated framework for tumor segmentation in pet. Phys. Med. Biol. **65**(24), 245032 (2020)
12. Li, H., et al.: Domain generalization for medical imaging classification with linear-dependency regularization. In: NeurIPS, pp. 3118–3129 (2020)
13. Li, Z., et al.: Domain generalization for mammography detection via multi-style and multi-view contrastive learning. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12907, pp. 98–108. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87234-2_10
14. Liu, Q., Chen, C., Dou, Q., Heng, P.A.: Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In: AAAI, pp. 1756–1764 (2022)
15. Liu, Q., Dou, Q., Heng, P.-A.: Shape-aware meta-learning for generalizing prostate MRI segmentation to unseen domains. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12262, pp. 475–485. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_46
16. Maas, A., Hannun, A., Ng, A.: Rectifier nonlinearities improve neural network acoustic models. In: ICML, pp. 1–6 (2013)
17. Mao, X., et al.: Least squares generative adversarial networks. In: ICCV, pp. 2813–2821 (2017)
18. Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C.: Causality-inspired single-source domain generalization for medical image segmentation. IEEE TMI, 1–12 (2022)
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. J. Stat. Plan. Inference **90**(2), 227–244 (2000)
21. Song, Y., et al.: Lesion detection and characterization with context driven approximation in thoracic FDG PET-CT images of NSCLC studies. IEEE TMI **33**(2), 408–421 (2014)
22. Taghanaki, S.A., et al.: Combo loss: handling input and output imbalance in multi-organ segmentation. CMIG **75**, 24–33 (2019)
23. Vesal, S., et al.: Domain generalization for prostate segmentation in transrectal ultrasound images: a multi-center study. MedIA **82**, 102620 (2022)
24. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: NeurIPS, pp. 10506–10518 (2019)
25. Wang, S., Yu, L., Li, C., Fu, C.-W., Heng, P.-A.: Learning from extrinsic and intrinsic supervisions for domain generalization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 159–176. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_10

26. Wang, X., et al.: A generalizable and robust deep learning algorithm for mitosis detection in multicenter breast histopathological images. MedIA **84**, 102703 (2023)
27. Wang, Z., Luo, Y., Qiu, R., Huang, Z., Baktashmotlagh, M.: Learning to diversify for single domain generalization. In: ICCV, pp. 834–843 (2021)
28. Wehrend, J., et al.: Automated liver lesion detection in 68Ga DOTATATE PET/CT using a deep fully convolutional neural network. EJNMMI Res. **11**(1), 1–11 (2021)
29. Wilson, G., Cook, D.J.: A survey of unsupervised deep domain adaptation. ACM TIST **11**(5), 1–46 (2020)
30. Xie, L., Wisse, L.E., Wang, J., Ravikumar, S., Khandelwal, P.: Deep label fusion: a generalizable hybrid multi-atlas and deep convolutional neural network for medical image segmentation. MedIA **83**, 102683 (2023)
31. Xu, C., Wen, Z., Liu, Z., Ye, C.: Improved domain generalization for cell detection in histopathology images via test-time stain augmentation. In: Wang, L., et al. (eds.) MICCAI 2022, pp. 150–159. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_15
32. Xu, Y., et al.: Adversarial consistency for single domain generalization in medical image segmentation. In: Wang, L., et al. (eds.) MICCAI 2022, pp. 671–681. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_64
33. Xu, Z., Liu, D., Yang, J., Raffel, C., Niethammer, M.: Robust and generalizable visual representation learning via random convolutions. In: ICLR, pp. 1–12 (2021)
34. Yamashita, R., Long, J., Banda, S., Shen, J., Rubin, D.L.: Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. IEEE TMI **40**(12), 3945–3954 (2021)
35. Yun, S., et al.: CutMix: regularization strategy to train strong classifiers with localizable features. In: ICCV, pp. 6022–6031 (2019)
36. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. In: ICLR, pp. 1–13 (2018)
37. Zhang, Z., et al.: Optimization-based image reconstruction from low-count, list-mode TOF-pet data. IEEE TBME **65**(4), 936–946 (2018)
38. Zhao, Y., et al.: Deep neural network for automatic characterization of lesions on 68Ga-PSMA-11 PET/CT. EJNMMI **47**, 603–613 (2020)
39. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: a survey. IEEE TPAMI **45**(4), 4396–4415 (2022)