sciendo

JAISCR, 2022, Vol. 12, No. 1, pp. 29– 39
10.2478/jaiscr-2022-0003

MIXUP (SAMPLE PAIRING) CAN IMPROVE THE PERFORMANCE OF DEEP SEGMENTATION NETWORKS

Lars J. Isaksson[1,*], Paul Summers[2], Sara Raimondi[3], Sara Gandini[3], Abhir Bhalerao[4], Giulia Marvaso[1,5], Giuseppe Petralia[2,5], Matteo Pepa[1], Barbara A. Jereczek-Fossa[1,5]

[1]*Division of Radiotherapy, European Institute of Oncology IRCCS, via Ripamonti 435, Milan, Italy*

[2]*Division of Radiology, European Institute of Oncology IRCCS, via Ripamonti 435, Milan, Italy*

[3]*Department of Experimental Oncology, European Institute of Oncology IRCCS, via Ripamonti 435, Milan, Italy*

[4]*Department of Computer Science, University of Warwick, Coventry CV4 7AL, Warwick, United Kingdom*

[5]*Department of Oncology and Hemato-oncology, University of Milan, via Festa del Perdono 7, Milan, Italy*

[*]*E-mail: larsjohannes.isaksson@ieo.it*

Submitted: 11th March 2021; Accepted: 2nd July 2021

Abstract

Researchers address the generalization problem of deep image processing networks mainly through extensive use of data augmentation techniques such as random flips, rotations, and deformations. A data augmentation technique called mixup, which constructs virtual training samples from convex combinations of inputs, was recently proposed for deep classification networks. The algorithm contributed to increased performance on classification in a variety of datasets, but so far has not been evaluated for image segmentation tasks. In this paper, we tested whether the mixup algorithm can improve the generalization performance of deep segmentation networks for medical image data. We trained a standard U-net architecture to segment the prostate in 100 T2-weighted 3D magnetic resonance images from prostate cancer patients, and compared the results with and without mixup in terms of Dice similarity coefficient and mean surface distance from a reference segmentation made by an experienced radiologist. Our results suggest that mixup offers a statistically significant boost in performance compared to non-mixup training, leading to up to 1.9% increase in Dice and a 10.9% decrease in surface distance. The mixup algorithm may thus offer an important aid for medical image segmentation applications, which are typically limited by severe data scarcity.

Keywords: magnetic resonance imaging, segmentation, prostate, data augmentation, mixup

1 Introduction

Much of the success of deep learning (DL) models depends on their ability to generalize well to instances outside the training set. However, the ability to generalize well is not guaranteed, and researchers typically employ several training techniques to bolster this capacity. This is particularly important in medical image analysis due to the limited availability of relevant data (typically on the order of 10^2 cases for medical imaging studies, compared to for instance the 10^6 training images in the ImageNet-2012 data set for generic image classification). One widely used technique in the field of image analysis is data augmentation in which the dataset is enlarged by constructing virtual training samples via transformations such as random flips and rotations to the original images.

A common task for medical image analysis is that of organ or region-of-interest segmentation, where the goal is to delineate a particular zone or organ of interest. The current state of the art for this task is represented by variants of the U-net DL architecture [1]. In recent medical image segmentation challenges, these DL architectures have consistently achieved the top ranking positions, with new top contenders appearing regularly [2–4]. Likewise, a number of research groups have also published DL architecture variants with similar high-end performance (see e.g. [5–12]). An important factor uniting these high-end models is the extensive use of data augmentation.

Three new data augmentation techniques were introduced at the ICLR conference in 2018: mixup [13], Between-Class (BC) learning [14, 15], and Sample-Pairing [16]; all of which work by constructing augmented samples ($\hat{\boldsymbol{x}}$) by linearly combining training data ($\boldsymbol{x}$), e.g. as $\hat{\boldsymbol{x}} = \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_j$. Mixup and BC learning further combine the corresponding data labels (for classification) in the same manner as the images. The mixup technique differs from BC learning in the way that λ is chosen: in mixup, λ is drawn from a beta distribution with shape parameter α ($\lambda \sim \beta(\alpha, \alpha)$) rather than fixing it explicitly. As such, one can choose from a range of distributions for λ that passes from a uniform distribution (for $\alpha = 1$), to a delta function centered at $\lambda = 0.5$ (for $\alpha \to \infty$). The authors of mixup reported that the optimal choice for α was problem dependent, with α falling within the range of 0.2 to 0.4 for normal image classification.

Linearly combining training data is completely data-agnostic and orthogonal to most other common augmentation practices, and introduces little computational overhead. Intuitively, mixing images in this manner may not appear sensible (see Figure 1), especially for instances from different classes, where it implies simultaneous membership of two (or more) classes, e.g. a cat and a dog at the same time (note that this is different from an image containing both a cat and a dog). The authors argue, however, that the algorithms work by promoting linear behavior between training samples [13], and reducing the ratio of the inter-class distance to the intra-class variance [14]. Such behavior may be beneficial for machine learning while not being directly comprehensible by humans.

Most of the focus in the aforementioned publications are on image classification tasks, with some discussion about sound recognition [15] and sample synthesising with generative adversarial networks (GANs) [13, 17–19]. The question remains whether these techniques also can improve the performance in image segmentation tasks – an issue that was briefly mentioned in [13] but never explicitly tested.

In this paper, we tested the utility of the mixup data augmentation technique for a medical image segmentation task. In a data set of 100 magnetic resonance imaging (MRI) scans of prostate cancer patients with delineated prostates, we examined the impact of different parameterizations of mixup for image segmentation with a simple variant of the standard U-net. If successful, the technique could help improve the performance of medical segmentation models (and potentially even other segmentation models), which are often constrained by their limited training data.

2 Methods

2.1 The mixup algorithm

The mixup algorithm involves constructing new samples $\hat{\boldsymbol{x}}$ from training samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ according

[2] Images designed by Freepik.com

(a) *Cat* (b) *Dog* (c) $0.5 \cdot Cat + 0.5 \cdot Dog$

Figure 1. Example of a simple linear combination of two images. The combination of the two may not be very helpful for teaching humans how to distinguish between cats and dogs, but it may be beneficial for teaching machines to generalize well.[2]

to $\hat{\boldsymbol{x}} = \lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_j$. The lambda is drawn from a symmetric beta distribution, i.e. $\lambda \sim \beta(\alpha, \alpha)$, where α is a hyperparameter of the transformation.

2.2 Data set

We conducted this study on a set of 100 T2-weighted MRI scans of the prostate from patients at IEO European Institute of Oncology IRCCS, Milan, Italy. All patients gave their consent for use of their data for research and educational purposes, and this study was performed under ethical committee approval. All images had accompanying ground-truth prostate labels generated by a radiologist with more than 5 years experience and checked by one or more other radiologists who revised the contours if they felt it necessary. The images were acquired using a 1.5 T scanner (slice thickness 3.0-3.6 mm, slice gap 0.3 mm, pixel spacing 0.59×0.59 mm, echo time 118 ms, and repetition time 3780 ms).

Prior to training, we resampled the MRI volumes to a common size of $320 \times 320 \times 32$ using bi-linear interpolation (in the three cases where the image matrix was larger than 320×320) and zero padding (in the cases where there were fewer than 32 slices). The image intensities were then normalized to zero mean and unit variance.

2.3 Experiments

2.3.1 Hyperparameter search

A hyperparameter exploration was first conducted in order to select the best value for the mixup parameter α. For this experiment, we randomly split the data into two equally sized (N=50) training and hold-out test sets. After splitting, the clinical variable distributions of the two sets were tested to be similar (Wilcoxon signed rank test), which allowed us to control for adverse effects stemming from different clinical traits between the two groups. For each value of α (including $\alpha = 0$ for no mixup), the network was trained until convergence eight times with different initializations (see sections below for training and implementation details). The best value of α was selected for the remainder of our experiments.

2.3.2 Quantifying the impact of mixup

We performed a random five-fold cross validation procedure comparing the best value of the mixup parameter α with no mixup. In addition to mixup, the samples were augmented with a standard scheme of random horizontal flips, random zoom and translation ($scalefactor \in [0.5, 1.5]$), and random rotations in the $[-\frac{\pi}{4}, \frac{\pi}{4}]$ range. Each session was trained for 350 epochs, which was sufficient for convergence, and the best weights were restored prior to evaluation.

Segmentation performance was quantified by the Sørensen–Dice similarity coefficient (Dice) and mean surface distance (MSD) between the network outputs and the ground truths. In addition, we included the following associated metrics sometimes seen in related research: absolute relative volume difference, 95th percentile Hausdorff distance, and sensitivity.

The Dice coefficient is defined by

$$\text{Dice} = \frac{2|X \cap Y|}{|X| + |Y|} \tag{1}$$

where X and Y are the sets of pixels of the structures being compared. As such, the Dice coefficient ranges from 0 to 1, with 1 corresponding to a perfect overlap. In Boolean algebra, it's equivalent to the F_1 score.

The MSD (measured in pixels) is defined as

$$\mathrm{MSD} = \frac{\sum_{x \in X_s} d(x, Y_s) + \sum_{y \in Y_s} d(y, X_s)}{|X_s| + |Y_s|} \quad (2)$$

for the surfaces X_s and Y_s of X and Y, using the Euclidean distance from a pixel $x \in X_s$ to surface Y_s given by $d(x, Y_s) = min_{y \in Y_s}||x - y||$. In evaluating (2), only 2D (within-slice) distances were considered, and surface voxel connectivity was defined on an eight-neighbour basis.

2.4 Network architecture

The architecture used for the segmentation network was a variation of the classic U-net [1] as illustrated in Figure 2 and inspired by high performing submissions to the PROMISE12 online prostate segmentation challenge [20]. In contrast to other models, the main features of our network are that:

1. Fewer filters were used, and the number of filters at deeper levels of the network was increased by a constant amount (+14) instead of a multiplying factor (typically $\times 2$).
2. Each level performs a single convolutional operation instead of two or more in series.
3. The first operation is strided instead of a normal convolution in order to more efficiently manage computational resources.
4. A PReLU [21] activation function was used instead of the more common ReLU function.

All design choices above were motivated by heuristic comparisons, favoring simplicity and speed over minor performance gains. In addition, we use heavy dropout with a fraction of 0.5 for the encoding part of the network.

2.5 Training and implementation

The models were trained with the Adam optimizer [22] with default learning parameters ($\beta_1 = 0.9, \beta_2 = 0.999$), a learning rate of 0.0005, and extended with the lookahead mechanism [23] (sync period = 6, step size = 0.5) to reduce its variance. As a loss function, we used the top-k cross entropy [24], which calculates the pixel-wise cross entropy, but only considers the top k pixels as contributing to the final loss. This allows the network to focus training on hard to classify pixels (likely from the edges of the prostate) and also speeds up training [24]. The parameter k was set to 5% of the number of pixels of the median image size (in this case $k = 143\,360$) for each sample in the minibatch, and the batch size was set to 8. The implementation was done in Python 3.7 with TensorFlow 2.3 running on an Nvidia Tesla K80 GPU (16 GB).

3 Results

A visualization of the mixup technique for two samples in our data set is presented in Figure 3. In this case (as in most), the mixup image does not depict a realistic appearance for a prostate MRI, and the target mask of the augmented sample is no longer binary.

Our mixup implementation did not introduce any significant computational overhead; less than 0.04 s per batch of 8 images (roughly 0.1% increase).

The parameter search concluded that the best choice for α was $\alpha = 0.5$ in terms of both Dice and MSD (see Figure 4). Both $\alpha = 0.5$ and $\alpha = 0.7$ were significant improvements (Mann-Whitney U-test) compared to non-mixup training for both evaluation metrics. Other α values also outperformed non-mixup training, albeit without statistical significance. The mean performance of $\alpha = 0.5$ resulted in a 1.46% increase in Dice and a 10.9% decrease in MSD over non-mixup training.

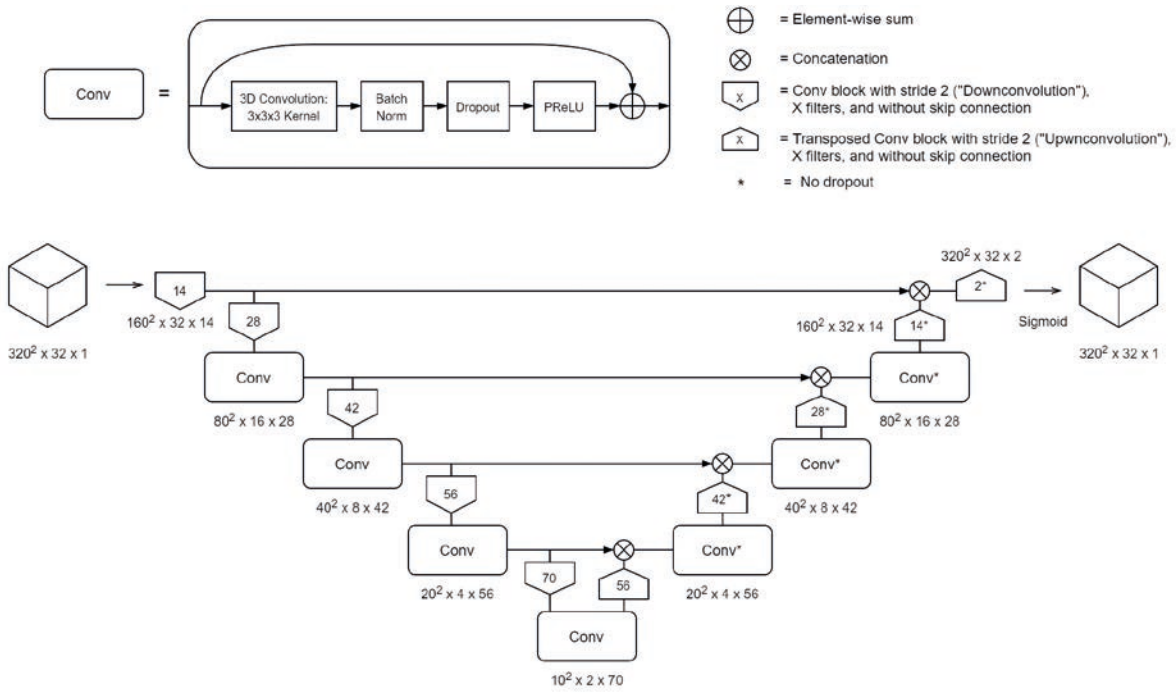


Figure 2. Network architecture.

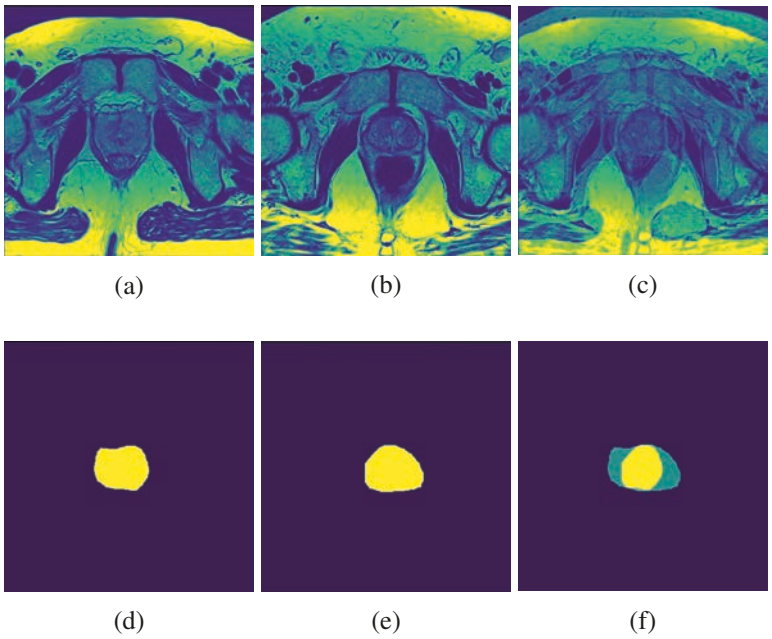


Figure 3. Examples of prostate MRI scans (a, b) from two patients with accompanying prostate masks (d, e). The right images (c, f) show the results of the mixup data augmentation technique ($\lambda = 0.5$) of the two patients.

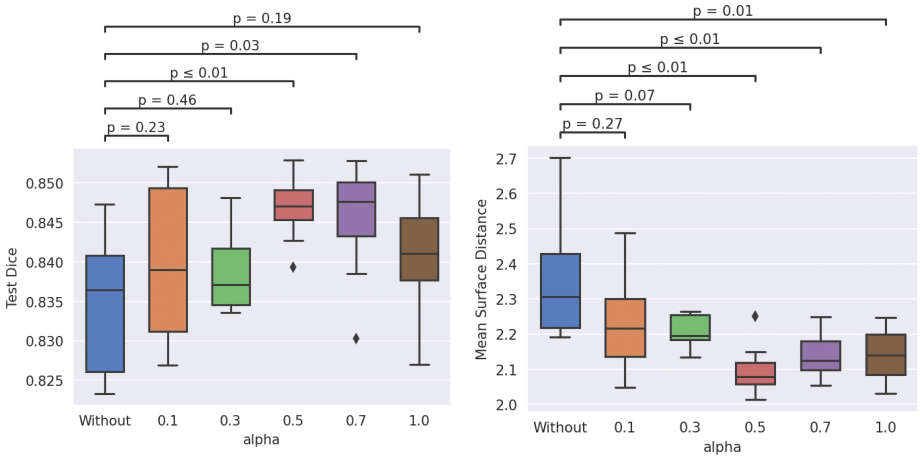


Figure 4. Test results of the α parameter search on the holdout test set. Results are aggregated from eight different initializations. A good performance is characterized by a high Dice and low surface distance. *p*-values indicate the significance of the Mann–Whitney U-test.

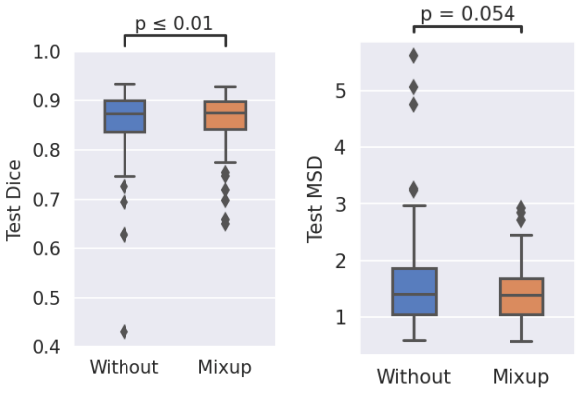


Figure 5. Cross validation results from training without mixup and training with mixup with $\alpha = 0.5$. *p*-values indicate the significance of the pair-wise Wilcoxon signed-rank test. The mean pairwise differences of Dice and MSD were 0.016 (3.15%) and -0.207 (-13.9%), respectively.

In our cross-validation experiment with $\alpha = 0.5$, mixup training increased mean Dice by 1.91% ($p = 0.005$) and decreased mean MSD by 10.7% ($p = 0.054$) over non-mixup training (see Figure 5). An additional analysis of other common segmentation metrics (absolute relative volume difference, 95th percentile Hausdorff distance, and sensitivity) reveals that mixup training is superior in every case, although not significantly so for MSD and absolute relative volume difference (see Table 1).

Training curves from the *k*-fold cross validation are displayed in Figure 6. The time to convergence seems to be largely unaffected by the use of mixup, and mixup training seems to generalize better and with lower variance compared to non-mixup training, even though the train losses are very similar.

Table 1. Mean values of different commonly used performance metrics for the cross validated test performance. Parentheses indicate the 95% CI and p-values show the significance of the Wilcoxon signed rank test. aRVD: absolute relative volume difference, HD95: 95th percentile Hausdorff distance, Sen: sensitivity.

	No mixup	mixup	*p*-value
Dice	0.839 (±0.03)	0.855 (±0.02)	0.005
MSD	1.93 (±0.57)	1.73 (±0.49)	0.054
aRVD	0.088 (±0.04)	0.066 (±0.03)	0.155
HD95	6.46 (±1.41)	5.78 (±1.45)	0.011
Sen	0.812 (±0.03)	0.832 (±0.03)	0.036

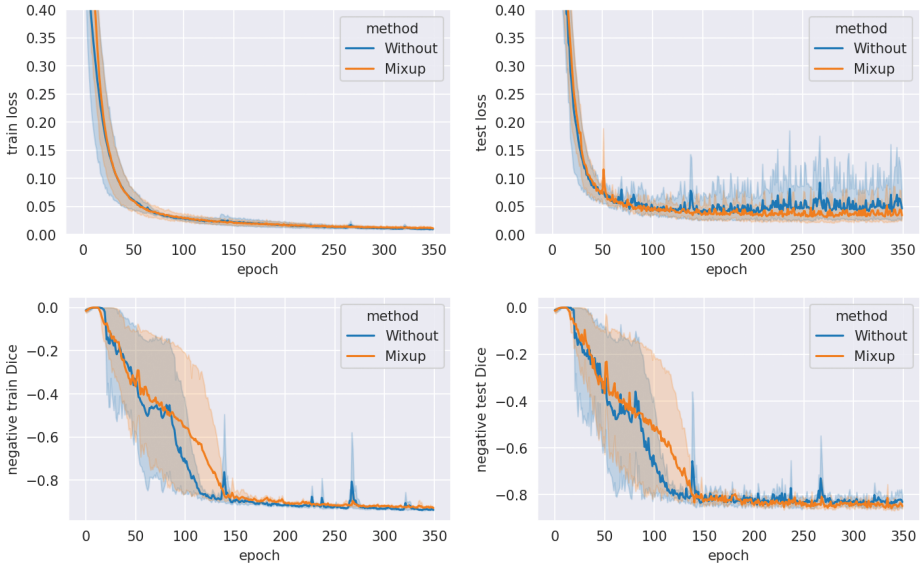


Figure 6. Training curves from the five-fold cross validation without mixup and mixup with with $\alpha = 0.5$. The large Dice variance in the early stages of training is a feature of the loss function: the time it takes for top-k pixel loss to generalize globally varies strongly between runs.

4 Discussion

Our results suggest that mixup data augmentation offers a clear and prominent generalization improvement for segmentation of prostates in T2-weighted MRI images with convolutional neural networks.

It is possible that the generalization improvement from mixup is dependent on the network architecture, but we expect networks of the U-net family to behave similarly since they all follow the same principle of systematic feature extraction with convolutions. Our network was designed to be as simple and general as possible without sacrificing performance, and therefore it did not incorporate attention blocks [25] or residual refinement blocks [26] which are utilized in many state-of-the-art architectures today. Since generalization improvement attributed to mixup for classification has been demonstrated to be greater for larger networks (i.e. networks with more parameters) [13], it is also plausible that the segmentation performance gain from mixup could be greater for more complex anatomical structures that require larger networks, such as bone or blood vessels, compared to the relatively simple prostate geometry.

Because the mixup procedure renders the target mask non-binary, caution needs to be taken when designing the desired loss function. For example, the standard definition of the Dice coefficient in (1) may be extended to continuous cases by simply defining the intersection as the product between X and Y, but such an implementation is not always maximized when $X = Y$ because $\sum_i x_i y_i \leq \sum_i x_i$ for $x_i, y_i \in [0,1]$. Another continuous extension to the Dice coefficient has been proposed in [27]. However, in this paper we used top-k cross entropy because of the previously demonstrated success of standard cross entropy in combination with mixup for classification [13], and because it outperformed Dice in preliminary testing. Further exploration in the interest of optimizing the loss function for mixup segmentation is warranted, but is outside the scope of this article. We do not expect the generalization improvement from mixup to be exclusive to the cross entropy loss.

Additional theoretical explorations as to why and how the mixup procedure leads to better generalization performance is also warranted since this is still not well understood [13, 16, 18]. Two principally distinct intuitive explanations have been proposed: first, it can be viewed as regularizing the network by simply increasing the sample size, and sec-

ondly, as simultaneously learning multiple samples such that the network learns to better distinguish between their corresponding classes. It has also been suggested that mixing promotes more robust detection of low level features such as lines or edges. If this is indeed the case, the strategy ought to be less effective for transfer-learned networks, where the low level feature weights are typically frozen.

It has also been suggested that mixup training significantly improves the calibration issue and uncertainty of deep convolutional networks [28]. This could be an incentive to use mixup for medical image segmentation even if the performance itself is not improved by mixup, since uncertainty in medicine is much more detrimental and may impact patients well-being.

It should also be noted that, although our experiments focused exclusively on T2-weighted MRI images, it is likely that the benefits transfer to other modalities or domains as well such as CT and PET images, or even non-medical image segmentation. Since medical image data tend to be severely more scarce, this is where we believe the benefit to be greatest.

5 Conclusion

We have tested data augmentation by linearly combining training samples (referred to as mixup, sample-pairing, or between-class learning) for the task of medical image segmentation with deep convolutional neural networks. This procedure, which has been successful in improving the generalization performance of image classification, has not previously been studied for image segmentation tasks. Our results show a clear and statistically significant generalization improvement of up to 1.9% increased Dice score and 10.9% decreased mean surface distance compared to non-mixup training. Our best performance was achieved with sample mixing weights drawn from a beta distribution, $\beta(\alpha, \alpha)$, with $\alpha = 0.5$

Acknowledgements

This study was partially supported by the Italian Ministry of Health with Ricerca Corrente and 5x1000 funds. The study was supported by the University of Milan with APC funds.

Declaration of Interest

None.

Patient Characteristics and Dataset Split

A summary of the clinical characteristics within the entire cohort of 100 prostate cancer patients is presented in Table 2. For the parameter search, the full data set was split into two equally sized (N=50) training and hold-out test sets on which the models were trained and evaluated on, respectively. The split was conducted randomly and the clinical characteristics of the respective sets were tested to be similar with the Wilcoxon rank-sum test. The clinical characteristics on which the test was performed were the volume of the prostate ($p = 0.339$) and the ECE score ($p = 0.345$). According to the radiologists, these were the clinical characteristics that were deemed to have the most potential impact on the segmentation results.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[2] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang et al., Evaluation of prostate segmentation algorithms for mri: the promise12 challenge, Medical image analysis, vol. 18, no. 2, pp. 359–373, 2014.

[3] MICCAI challenges, http://www.miccai.org/events/challenges/, 2020, accessed: 2020-08-03.

[4] grand-challenge.org challenges, https://grand-challenge.org/challenges/, 2020, accessed: 2020-08-03.

[5] R. Cuocolo, A. Comelli, A. Stefano, V. Benfante, N. Dahiya, A. Stanzione, A. Castaldo, D. R. De Lucia, A. Yezzi, and M. Imbriaco, Deep learning whole-gland and zonal prostate segmentation on a public mri dataset, Journal of Magnetic Resonance Imaging, 2021.

Table 2. Summary of clinical prostate cancer characteristics within the study cohort and the training and test sets.

Characteristic		Number of patients		
		Total	Train set	Test set
ECE score	1	5	3	2
	2	22	12	10
	3	27	12	15
	4	30	15	15
	5	14	6	8
T-stage	cT1	18	14	4
	cT2	71	32	29
	cT3	9	3	6
Gleason Score	6	48	25	23
	7	51	24	27
	8	1	1	0
PIRADS score	2	2	2	0
	3	20	15	5
	4	35	13	22
	5	42	20	22
Volume		42.4 (22.0)*	43.5 (26.9)*	41.3 (15.3)*
PSA		11.2 (33.4)*	13.9 (46.7)*	8.53 (5.31)*

*Mean (standard deviation)

[6] A. Comelli, N. Dahiya, A. Stefano, F. Vernuccio, M. Portoghese, G. Cutaia, A. Bruno, G. Salvaggio, and A. Yezzi, Deep learning-based methods for prostate segmentation in magnetic resonance imaging, Applied Sciences, vol. 11, no. 2, p. 782, 2021.

[7] M. Penso, S. Moccia, S. Scafuri, G. Muscogiuri, G. Pontone, M. Pepi, and E. G. Caiani, Automated left and right ventricular chamber segmentation in cardiac magnetic resonance images using dense fully convolutional neural network, Computer Methods and Programs in Biomedicine, vol. 204, p. 106059, 2021.

[8] Y. Xie, J. Zhang, C. Shen, and Y. Xia, Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation, arXiv preprint arXiv:2103.03024, 2021.

[9] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, arXiv preprint arXiv:2102.04306, 2021.

[10] Y. Shu, J. Zhang, B. Xiao, and W. Li, Medical image segmentation based on active fusion-transduction of multi-stream features, Knowledge-Based Systems, vol. 220, p. 106950, 2021.

[11] H. H. Bo Wang, Shuang Qiu, Dual encoding u-net for retinal vessel segmentation, Medical Image Computing and Computer Assisted Intervention, vol. 11764, pp. 84–92, 2019.

[12] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, Bi-directional convlstm u-net with densley connected convolutions. institute of electrical and electronics engineers (ieee); 2019; 406–415, 2020.

[13] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412, 2017.

[14] Y. Tokozume, Y. Ushiku, and T. Harada, Between-class learning for image classification, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5486–5494.

[15] Y. Tokozume, Y. Ushiki, and T. Harada, Learning from between-class examples for deep sound recognition, arXiv preprint arXiv:1711.10282, 2017.

[16] H. Inoue, Data augmentation by pairing samples for images classification, arXiv preprint arXiv:1801.02929, 2018.

[17] L. Perez and J. Wang, The effectiveness of data augmentation in image classification using deep learning, arXiv preprint arXiv:1712.04621, 2017.

[18] D. Liang, F. Yang, T. Zhang, and P. Yang, Understanding mixup training methods, IEEE Access, vol. 6, pp. 58 774–58 783, 2018.

[19] C. Summers and M. J. Dinneen, Improved mixed-example data augmentation, in 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019, pp. 1262–1270.

[20] Promise12 online challenge leaderboard, https://promise12.grand-challenge.org/evaluation/leaderboard/, 2020, accessed: 2020-08-04.

[21] K. He, X. Zhang, S. Ren, and J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

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

[23] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton, Lookahead optimizer: k steps forward, 1 step back, in Advances in Neural Information Processing Systems, 2019, pp. 9597–9608.

[24] Z. Wu, C. Shen, and A. v. d. Hengel, Bridging category-level and instance-level semantic image segmentation, arXiv preprint arXiv:1605.06885, 2016.

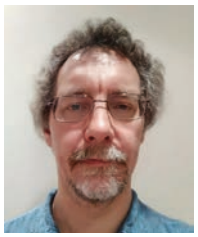
[25] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz et al., Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999, 2018.

[26] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation, arXiv preprint arXiv:1802.06955, 2018.

[27] R. R. Shamir, Y. Duchin, J. Kim, G. Sapiro, and N. Harel, Continuous dice coefficient: a method for evaluating probabilistic segmentations, arXiv preprint arXiv:1906.11031, 2019.

[28] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak, On mixup training: Improved calibration and predictive uncertainty for deep neural networks, in Advances in Neural Information Processing Systems, 2019, pp. 13 888–13 899.

Lars Johannes Isaksson is a Ph.D. student of computational biology at the European School of Molecular Medicine. While trained as a theoretical physicist with a B.Sc. and M.Sc. from Lund University, his recent research interests mostly include machine learning and its application to medical images and radiomics.



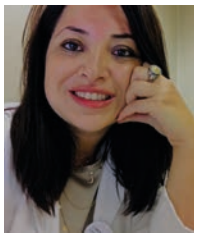
Dr. **Paul Summers** has received B.Sc. degrees in Physics (1987) and Mathematics (1989) from the University of Alberta, Edmonton, Canada, and in Tecniche di radiologia medica, per immagini e radioterapia (2018) from the University of Milan, Milan, Italy. He received a Ph.D. in Medical Physics (1999) from the University of London, London, England. He is currently a researcher at IEO - European Institute of Oncology in Milan, Italy. His research interests center around clinical applications of magnetic resonance imaging in oncology and neurovascular disorders.

Sara Raimondi is a staff biostatistician and epidemiologist at the Department of Experimental Oncology at the European Institute of Oncology in Milan, Italy. She received her M.Sc. Degree in Biostatistics and Experimental Statistics in 2005 and a Ph.D. in Medical Statistics in 2013. Her main research interest are molecular epidemiology and big data analysis, with special focus on radiomic studies. She is the referent statistician for radiomic studies at the IEO. She is co-author of 115 publications and her H-index of 37 (Google Scholar) - 32 (Scopus) places her on the Top Italian Scientists list.

Sara Gandini is a biostatistician and epidemiologist (Ph.D.) Group Leader at the Department of Experimental Oncology at the IEO. She is adjunct professor in medical statistic at University "Statale di Milano" (National Academic Qualification as Associate Professor in medical statistics from 2017) and faculty member of System Medicine PhD (SEMM) University ("Statale di Milano"). Chair Epidemiology subgroup EORTC melanoma. She published more than 250 publications in peer journals. H-index=68, Google scholar (July 2021).

Abhir Bhalerao is Reader in Computer Science. He received his B.Sc. and Ph.D. degrees in 1986 and 1992 respectively. He joined as faculty at Warwick in 1998 having been a research scientist with the NHS and Kings Medical School, London and a Research Fellow at Harvard Medical School. His current research interests are in computer assisted medical diagnosis, computer vision for intelligent vehicles, biometrics and security. He has published around 80 refereed articles in image analysis, medical imaging, computer vision and machine learning. In 2007 he

was the co-chair of the British Machine Vision Conference, and Medical Image Understanding and Analysis, 2010. He was a co-founder of Warwick Warp Ltd., a university spin out company specializing in biometric technologies worked a research officer for SMEs.



Giulia Marvaso is a Radiation Oncologist working, since September 2015, at the Istituto Europeo di Oncologia (IEO) IRCCS as medical assistant. In March 2020 she also obtained the position as Researcher in the Department of Oncology and Onco-Hematology at the University of Milan. Her major clinical interests include the management of patients with urological and head and neck cancer, oligometastatic status, palliative setting and stereotactic body radiotherapy (SBRT). Dr. Marvaso's research focuses on clinical research related to the use of radiotherapy in the management of cancer patients, moreover the technological and biological issues of high precision radiotherapy with particular attention to hypofractionation (i.e. delivery of high radiotherapy doses to small volumes called also ablative radiotherapy if doses > 5 Gy/fraction).

Giuseppe Petralia is Professor in Radiology at the University of Milan and Director of the Precision Imaging and Research Unit at the European Institute of Oncology, Milan (Italy). The mission of this Unit is to deliver the promise of Precision Medicine to cancer patients, by the use of the most advanced imaging techniques in conjunction with clinical parameters and other biomarkers. His main research areas include prostate cancer (multiparametric MRI, in-bore MRI–targeted biopsy) and whole-body MRI (prostate and breast cancer, multiple myeloma, and cancer screening in high-risk and general populations).

Matteo Pepa is a Biomedical Engineer at the Division of Radiation Oncology and at the Radiomic Group of the IEO European Institute of Oncology IRCCS in Milan, Italy. He is also adjunct professor at the University of Milan and his areas of expertise include medical imaging, artificial intelligence, radiomics and radiation oncology.

Barbara Jereczek-Fossa MD Ph.D. is an Associate Professor of Radiation Oncology at the University of Milan, Italy and a Head of the Department of Radiation Oncology at the European Institute of Oncology in Milan, Italy. Her other commitments include among others:

- Director of the School of Specialization in Radiation Oncology.
- Chair of the RTT BSc Degree of the University of Milan, Italy.
- Chair of the Scientific Committee of the Italian Association of Radiotherapy and Clinical Oncology (AIRO).
- Chair of the National Societies Committee of the European Society for Radiotherapy and Oncology (ESTRO).

Prof. Jereczek-Fossa is an active member of numerous national and international societies and scientific committees and coordinates many research projects. Her main research and clinical interests include: urological malignancies, breast cancer, combined modality approach, high precision radiotherapy, oligometastatic cancer, and new prognostic and predictive factors. Prof. Jereczek-Fossa is an author of over 330 peer-reviewed scientific papers and 7 book chapters and her H-index of 42 places her on the Top Italian Scientists list.