Investigating a Feature Unlearning Bias Mitigation Technique for Cancer-type Bias in AutoPet Dataset

Duc Thang HOANG[1], Quentin FERRE[1], Elsa SCHALCK[1], Olivier HUMBERT[2], Rosana EL JURDI[1]
[1]Euranova , Marseille, France; [2]Université Côte d'Azur, CNRS, Inserm, iBV, Nice, France.

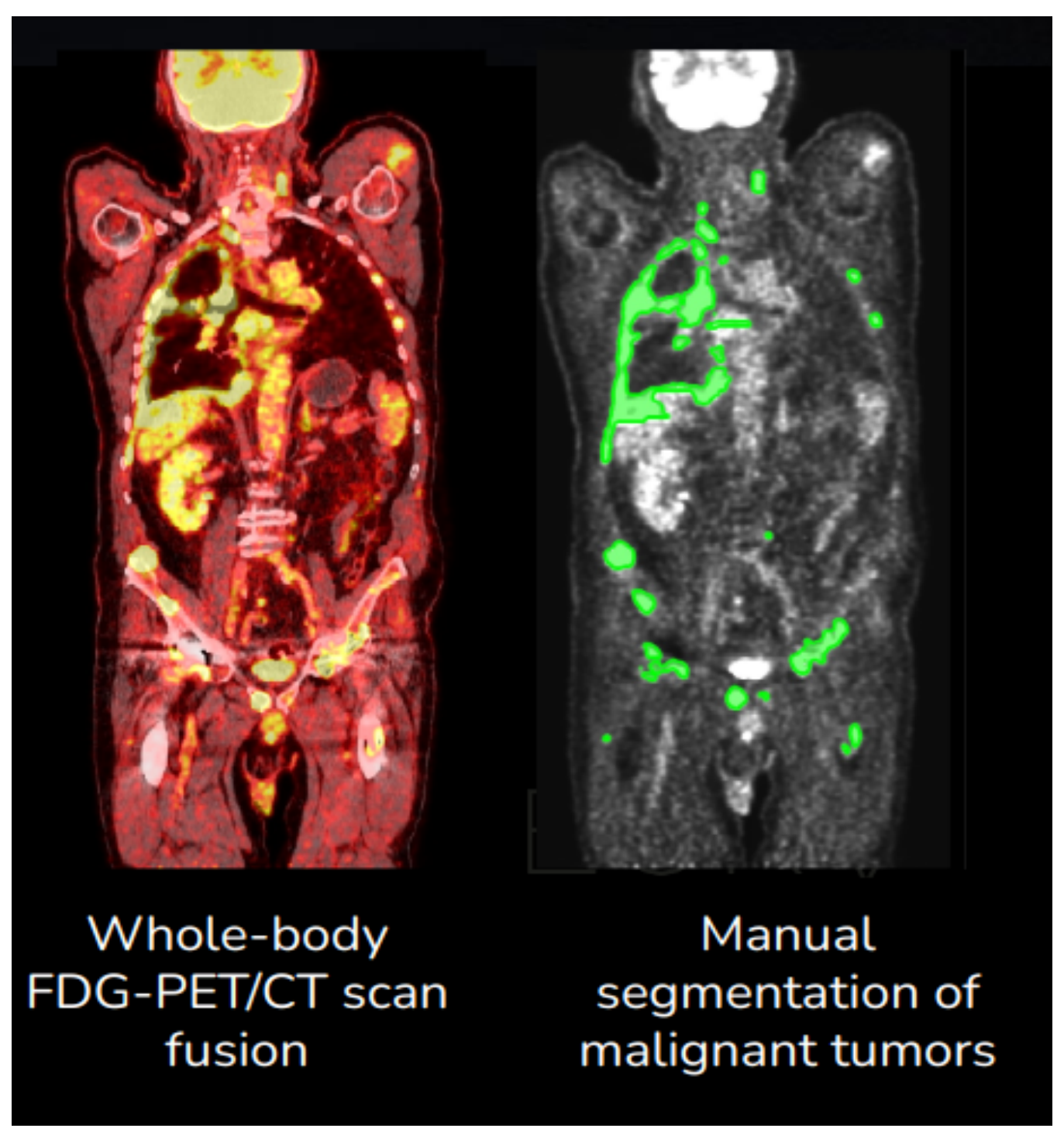
Introduction

Integrating datasets from different cancer types can improve diagnostic accuracy, as deep learning models tend to generalize better with more data. However, this benefit is often limited by performance variance caused by biases, such as under- or over-representation of certain diseases.

Contribution : In this work, we propose a **cancer-type-invariant model** capable of segmenting tumors from both lymphoma and lung cancer, irrespective of their frequency or representation bias. We frame the problem as a **transfer learning** task; we introduce a **discriminator** dedicated to learning bias-group specific features and a **confusion loss** that **preserve generic features** while **unlearning the domain-specific ones**.

Results : The model is trained on 154 lung-cancer and 132 lymphoma FDG-PET/CT scans from AutoPET [4]. The proposed method achieves state-of-the-art results, improves fairness, and remains effective under subgroup imbalance.

AutoPet Dataset

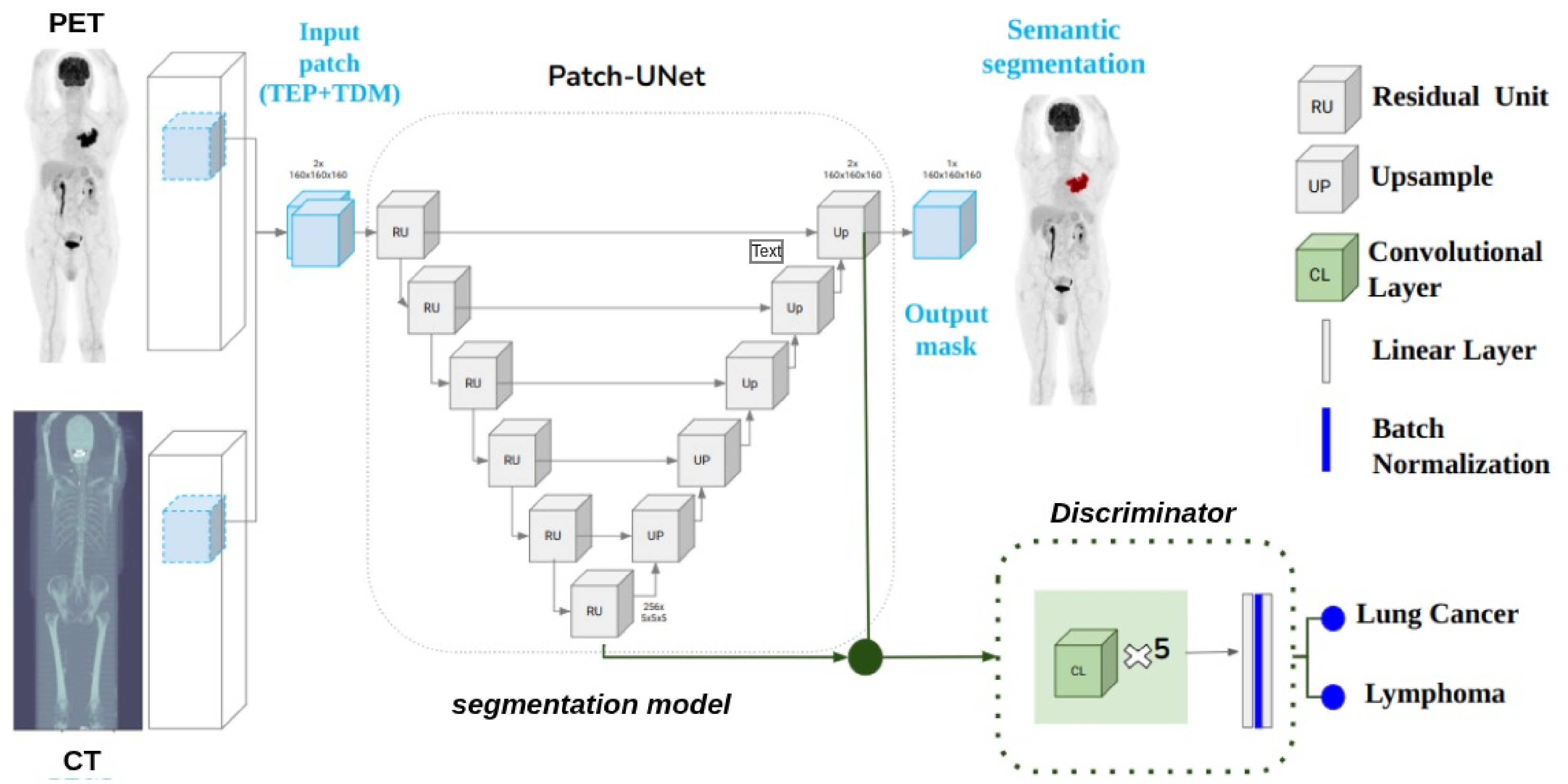

Whole-body FDG-PET/CT scan fusion
Manual segmentation of malignant tumors

Bias Mitigation Pipeline

Pipeline Components: The debiasing pipeline is constituted of 3 main components (Proposed method : Mixed Train-Unlearn-BatchRes):

- **Segmentation Model [8]**: Patch-based 3D U-Net (stride-2 conv + transpose conv) on CT + PET patches 160^3 → binary tumor map.
- **Discriminator**: 3D CNN (5 conv layers, 32→512 ch., instance norm) with U-Net bottleneck input → lung vs. lymphoma.
- **Batch Resampling**: Balances subgroup counts per batch, mitigating sampling bias.
- **Evaluation Metrics**: Dice accuracy; subgroup standard deviation SD [7]; Skewed Error Rate (SER, ratio of max to min subgroup error) [7]; Equity-Scaled Segmentation Performance (ESSP) [10] penalizing accuracy disparities.

Architecture Diagram



Loss Functions & Optimization Strategy

- **A 3-step iterative optimization process is adopted with 3 adversarial losses via 3 backward and forward passes [2]**:
  - Step 1: Optimize a segmentation loss $\mathcal{L}_p$: $\mathcal{L} = (1-\lambda)\mathcal{L}_{\text{Dice}} + \lambda\mathcal{L}_{\text{CE}}$
  - Step 2: Optimize a domain loss $L_d = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$.
  - Step 3: Optimize a confusion loss : $L_{\text{conf}} = -\frac{1}{N}\sum_{n=1}^{N}\log(p_n)$ to confuse the discriminator via the feature extractor.

Experimental Setup: Data were split into training, validation, and test sets, with 15 % reserved for testing. Test set is fixed across all experiments: 23 lung cancer samples and 20 lymphoma samples.

	Training		Validation	
	Lung Cancer	Lymphoma	Lung Cancer	Lymphoma
LungCancer-Only	108	-	23	-
Lymphoma-Only	-	92	-	20
Balanced Scheme (100 %)	108	92	23	20
Moderate Exclusion (50 %)	108	50	23	10
Limited Access (25 %)	108	23	23	5

- method encourages **domain-invariant** features via adversarial training while ensuring a **trade-off** between **accuracy** and **fairness**.

Results & Analysis

Results on the test sets			**Lung Cancer (23)**	**Lymphoma (20)**	**Lung Cancer + Lymphoma**			
Method / metrics	**Batch Res.**	**Unlearn**	Mean dice ± std	Mean dice ± std	**Av.**	**SD**	**SER**	**ESSP**
			Subgroup-only					
(1) LungCancer-Only			0.72 ± 0.18	0.45 ± 0.28	0.59	0.13	1.92	0.47
(2) Lymphoma-Only			0.57 ± 0.19	0.64 ± 0.25	0.59	0.03	1.19	0.55
			Balanced Scheme					
(3) Mixed Train-RandomRes.			0.74 ± 0.15	0.64 ± 0.23	0.69	0.04	1.37	0.63
(5) Mixed Train-Unlearn		✓	0.76 ± 0.15	0.67 ± 0.22	0.71	0.04	1.35	0.66
(6) Mixed Train-Unlearn-BatchRes	✓	✓	**0.74 ± 0.13**	**0.68 ± 0.24**	**0.71**	**0.03**	**1.25**	**0.67**
			Moderate Exclusion					
(7) Mixed Train-RandomRes.			0.76 ± 0.14	0.64 ± 0.24	0.70	0.05	1.48	0.62
(9) Mixed Train-Unlearn		✓	0.71 ± 0.17	0.64 ± 0.25	0.67	0.03	1.23	0.63
(10) Mixed Train-Unlearn-BatchRes	✓	✓	**0.73 ± 0.14**	**0.67 ± 0.21**	**0.70**	**0.02**	**1.19**	**0.67**
			Limited Access					
(11) Mixed Train-RandomRes.			0.75 ± 0.15	0.54 ± 0.27	0.65	0.10	1.85	0.54
(13) Mixed Train-Unlearn		✓	0.72 ± 0.18	0.58 ± 0.26	0.65	0.07	1.52	0.57
(14) Mixed Train-Unlearn-BatchRes	✓	✓	**0.73 ± 0.15**	**0.60 ± 0.25**	**0.67**	**0.06**	**1.48**	**0.59**

- **Subgroup-only:** Models failed to generalize, performing poorly on cancer types absent from their training data.
- **Balanced Training:** The method notably improved fairness (9% STD reduction, 3% ESSP increase) while preserving strong segmentation.
- **Moderate Exclusion:** Unlearning with batch resampling yielded highest fairness (8.06% ESSP increase, 19.59% SER reduction), showing resilience to partial data exclusion.
- **Limited Access:** Despite severe data limits, unlearning and resampling boosted fairness (9.26% ESSP increase) and closed performance gaps.

Conclusion

- **Summary:** We proposed a feature unlearning technique to reduce cancer-type bias, which improved segmentation accuracy while promoting fairness across subgroups, even with limited data.
- **Future work** will involve generalizing our approach to other datasets and bias types, as well as exploring normalizing flows for generating common unbiased feature embeddings.

Acknowledgements

The authors acknowledge the ANR – FRANCE (French National Research Agency) for its financial support under reference ANR-23-CE23-0019 (FAMOUS).

SCAN ME

References