PET-guided Attention Network for Segmentation of Lung Tumors from PET/CT images (Supplementary Information)

Varaha Karthik Pattisapu\textsuperscript{1}, Imant Daunhawer\textsuperscript{1}, Thomas Weikert\textsuperscript{2}, Alexander Sauter\textsuperscript{2}, Bram Stieltjes\textsuperscript{2}, and Julia E. Vogt\textsuperscript{1}

\textsuperscript{1} Department of Computer Science, ETH Zurich, Switzerland
\textsuperscript{2} Clinic of Radiology & Nuclear Medicine, University Hospital Basel, Switzerland

1 Data

The dataset consists of whole-body PET/CT scans of 397 lung cancer patients with ICD-10 diagnosis code C34 with histologically proven NSCLC [4]. The scans were obtained between Jan 2008 to Dec 2016 from University Hospital Basel, Switzerland. The patient population consisted of 28\% females and 72\% males with ages between 38-97 years (71 ± 10.5 years). PET/CT examinations were performed on an integrated PET/CT system (Discovery STE, GE Healthcare, Chalfont St Giles, UK) with 16-slice CT from Jan 2008 to Dec 2016 and on a PET/CT with 128-slice CT (Biograph mCT-X RT Pro Edition, Siemens Healthineers, Erlangen, Germany) from Dec 2015 to Dec 2016. Annotation and volumetric image segmentation of each lesion on transversal slices was performed in random order by a dual-board-certified radiologist and nuclear medicine physician with 9 years of professional experience in PET/CT reading as well as a supervised radiology resident with 2 years of professional experience.

The dataset consists of approximately seven labeled tumors per image. Each of the tumors is labeled as either a primary tumor, a nodular tumor, or a metastasis tumor. Primary tumors are the sites of origin of cancer. While there is usually only a single site of origin of cancer, it is however, not uncommon to have more than one. When the cancer is more advanced, it will have started to invade neighboring nodular regions. Such invasions are annotated as nodular tumors. In the most advanced stage, the cancer will have spread to the neighboring and distant tissues. Such malignant regions are labeled as metastasis tumors. Figure 1 gives a visual example of each of the tumor types. Primary tumors are bound by red bounding boxes, nodular tumors by green bounding boxes, and metastasis tumors by blue bounding boxes. It can be observed that the primary tumors and nodular tumors often appear in the vicinity of each other. While there is usually a distinction from the primary & nodular tumors from metastasis tumors, it might so happen that they also appear in the vicinity of primary tumors. It thus makes the task of distinguishing each of the tumors apart quite challenging, and in the most extreme cases, they can appear so close to one another that they practically start merging. This makes the predictions very hard as a model can get tricked into predicting the combined region of primary and...


Fig. 1. PET/CT image with malignant tumorous regions bounded by bounding boxes. Primary tumors are bound by red bounding boxes, nodular tumors by green bounding boxes, and metastasis tumors are bound by blue bounding boxes. Note that nodular tumors can appear in close proximity to primary tumors. They often merge with one another making the task of distinguishing them apart difficult. On a similar note, nodular tumors and metastases can appear in close proximity with one another.

nodular tumors as one single malignant region. Figure 2 gives the distribution for each of the tumor types for the current dataset. While the number of nodular and metastasis tumors is on a comparable scale to that of primary tumors, the primary tumors occupy a large fraction of the total tumor volume, which is reasonable because primary tumors are regions of active and unrestricted growth. In contrast, nodular and metastasis tumors typically appear in more advanced stages of cancer.

Figures 3-4 give a more in-depth view of the kind of variation that exists in the current dataset. Figure 3 gives the distribution of the stages of the tumors for each of the different types of tumors. Similarly, Figure 4 gives the distribution of the locations of the tumors for each of the different types of tumors. The stages and the locations are labelled according to TNM staging protocol. (See Tables 5 - 6 for a detailed description). The stages of the primary tumors are more or less uniformly distributed. In contrast, N2 stage is predominant for the nodular tumors and M1b stage stage is predominant for the metastases tumors. Typically, tumors of more advanced stages start to metastasise all across the lungs and distant parts of the body such as the chest, liver, colon etc. (see Figure 4 for the extent of distribution across the lungs and beyond). The current dataset encompasses tumors ranging from early to the most advanced/mature state not confined to one single location but spread across the regions of the lungs and distant organs such as the nearby lymph nodes, peripheral tissues, bones etc,
with varying shapes and morphologies. This variation makes the detection of tumors particularly challenging.

**Fig. 2.** (Left) Distribution of tumor types. There exist three different types of tumor: primary, nodular and metastasis. Primary tumors are the sites of origin of cancer. When the tumor has metastasized to nearby lymph nodes, it is a nodular tumor and when it has metastasized to distant tissues and organs, it is a metastasis tumor. (Right) While the number of each of the tumor types is on a comparable scale, primary tumors account for a large fraction of total tumor volume.

**Fig. 3.** The Figure shows the distribution of the tumor stages for each of the different types of tumors. The stages are defined according to TNM staging protocol. (See 5-6 for TNM staging protocol). The dataset consists of tumors ranging from early to most advanced state.
The Figure shows the distribution of the locations of the tumors for each of the different types of tumors. The locations are labeled according to TNM staging protocol. (See 5-6 for TNM staging protocol). The tumors are not localised to one region of the lungs but spread across the lungs.
1.1 Data pre-processing

The raw dataset consists of high-resolution whole-body PET/CT scans (512 × 512 × z) with varying axial dimension z. As our primary focus is lung tumor segmentation, we decided to focus only on the lungs and their vicinity. From the very start, we decided to use 3D convolutions instead of 2D convolutions because 3D convolutions can take into account spatial information across all the three dimensions necessary for tumor segmentation. As input we use a single cropped image that encloses the lungs and their vicinity.

CT image data consists of a large value range. In order to design an effective CAD solution, it is helpful to restrict the value range for training/testing. For instance, CT image data which is obtained in terms of Hounsfield units has a range of -3000 to +3000. Since air occupies most of the parts of the lungs, the lungs have an HU value of ∼ -1000 which is the HU value of air. Most of the soft tissues found around the lung region have HU values in the range of 0 to +300.

In order to account for the problem of large value range, we used contrast adjustment by clipping out the extreme values to adjust the CT value range to be within -1000 and +1000 HU. These threshold values are chosen based on reported values in relevant literature [1]. Each pixel is clipped based on the lower and upper thresholds. Accordingly, if a pixel value is greater than the upper threshold, then it is clipped to the upper threshold. Likewise, if a pixel value is less than the lower threshold, then it is clipped to the lower threshold. Concretely let the raw pixel value of a CT image be $I$ and let $I_{\text{min}} = -1000$ and $I_{\text{max}} = +1000$ be the lower and upper thresholds respectively. Then each pixel value is transformed given by the Equation 1 where $\tilde{I}$ is the result of clipping operation.

$$\tilde{I} = \min(\max(I, I_{\text{min}}), I_{\text{max}})$$ (1)

After adjusting for the large value ranges, each pixel is linearly adjusted to a range of [0,1] according to Equation 2. Similarly PET images are also linearly scaled to a range of [0,1]. Further, the resolution of the PET/CT images is reduced to 152 × 232 × 96 for computational reasons (so that an entire PET/CT with 3D convolutions image would fit on GPU memory). The reduction is done using bilinear interpolation on each of the 96 axial slices. (Note that the third dimension is the axial dimension).

$$I_{\text{final}} = (\tilde{I} - I_{\text{min}})/(I_{\text{max}} - I_{\text{min}})$$ (2)

Specifically, we used the following preprocessing steps:

- Clip the whole body PET/CT scan so as to include only the regions around the lungs.
- Clip out the extreme values for CT images so that they lie within a range of -1000 to +1000 HU.
– Linearly scale the PET/CT images to a range of [0,1].
– Use bi-linear interpolation to reduce the image resolution by almost a factor of two to $152 \times 232 \times 96$.

2 Network Architecture

2.1 Network architecture: Encoder

The encoder consists of a series of ResNet blocks. Each ResNet block consists of two consecutive sub-blocks, each consisting of Group Normalization layer [5], ReLU layer, and a Convolutional Layer stacked one after the other followed by an identity skip connection as shown in Figure 5. The convolutional layers, along with non-linear activations, extract the image features while the Group Normalization layers, along with identity skip connections, allow for faster convergence. We chose Group Normalization against the standard Batch Normalization technique because the former has been shown to work well with small batch sizes, which is the problem at hand (batch size of 2) [5]. The convolutional layers of the ResNet block retain the spatial resolution of the input. Convolutional layers of stride two follow these ResNet blocks at selected locations, which downsample the spatial resolution of the input by a factor of two along with increasing the number of filter channels by a factor of two. Three such downsampling layers together reduce the spatial resolution of the input image by a factor of eight along all the three dimensions of the input image. At its bottleneck layer, the encoder consists of a series of ResNet blocks in order to extract rich high dimensional features from the input image. All the convolutions are $3 \times 3 \times 3$, with the initial number of filters equal to 32. A summary of the layers of the encoder block is given in Table 1. It is noteworthy that the input to the encoder branch is only the 3D CT image and not the PET image, even if it is present.

2.2 Network architecture: Decoder

The decoder structure is similar to the encoder structure but with a single ResNet block per spatial level. The encoder is intended to be larger than the decoder in order to extract rich high dimensional image features from the input CT images. Each level of the decoder begins with a convolutional layer (using $1 \times 1 \times 1$ convolutions) that reduces the number of filter channels by a factor of two and 3D bilinear interpolation that increases the spatial resolution by a factor of two, followed by addition of the filtered skip connection from the encoder at the equivalent spatial resolution. The skip connection is filtered by passing it through an attention gate, as proposed earlier (shown in Figure 6). The last layer is a convolutional layer to reduce the filter channels to one, followed by a sigmoid activation function to squash the outputs to a range of $[0,1]$. The output of the decoder block has the same spatial resolution as that of the original input.

The model architecture is an encoder-decoder architecture similar to a U-Net architecture with three skip connections, as shown in Figure 5. The three
skip connections are filtered through their respective attention gates, with each attention gate having its own set of parameters. The attention gate, part of the attention mechanism has been discussed in the main section.

Fig. 5. Schematic of the model architecture. The input (colored in blue) is a single channel 3D CT image, followed by an initial $3 \times 3 \times 3$ convolution with 32 filters. Each of the green blocks is a ResNet block that consists of two consecutive sub-blocks. Each sub-block consists of Group Normalization, ReLU, and Convolutional layers stacked one after another. The attention gates filter the skip connections before being added to the downstream layers. The output of the model (colored in blue) is a single channeled probability map of the tumorous region with the same spatial resolution as the input CT image. The details of the attention gate are shown in Figure 6.

2.3 Training strategy: Exclude PET images randomly

One of the key features of the proposed PAG model is its flexibility to accommodate PET images as and when available. However, this robustness of the model is not achieved by training the network on CT images and then on PET/CT images in a separate step. Instead, the network should be able to adapt to the missing PET images, which is not possible with such a two-step training strategy. We devise a training strategy of random exclusion (or inclusion) of PET images at every training step and argue why this is better than the former training strategy.
Fig. 6. The Figure shows a schematic representation of the proposed PET Guided Attention Gate. The input feature representation $x^l$ is scaled by attention mask $\alpha$, which is computed by the attention gate. Spatial and contextual information are captured by two gating signals: the encoded feature representation $g$ and the composite function $h(x)$. The composite function is zero when the PET images are missing and the output of the function $f(x^{PET})$ when PET images are available. PET image representation and not the PET image itself is fed to the attention gate.
| Layer     | Details                        | Output size   |
|-----------|-------------------------------|---------------|
| input     | CT Image                      | $1 \times 152 \times 232 \times 96$ |
| Conv-1    | $32 \times 3 \times 3 \times 3$; stride=1; ReLU | $32 \times 152 \times 232 \times 96$ |
| ResNet-1  | ResNet block                  | $32 \times 152 \times 232 \times 96$ |
| Downsample-1 | $64 \times 1 \times 1 \times 1$; stride=2; ReLU | $64 \times 76 \times 116 \times 48$ |
| ResNet-2  | ResNet block                  | $64 \times 76 \times 116 \times 48$ |
| ResNet-3  | ResNet block                  | $64 \times 76 \times 116 \times 48$ |
| Downsample-2 | $128 \times 1 \times 1 \times 1$; stride=2; ReLU | $128 \times 38 \times 58 \times 24$ |
| ResNet-4  | ResNet block                  | $128 \times 38 \times 58 \times 24$ |
| ResNet-5  | ResNet block                  | $128 \times 38 \times 58 \times 24$ |
| Downsample-3 | $256 \times 1 \times 1 \times 1$; stride=2; ReLU | $256 \times 19 \times 29 \times 12$ |
| ResNet-6  | ResNet block                  | $256 \times 19 \times 29 \times 12$ |
| ResNet-7  | ResNet block                  | $256 \times 19 \times 29 \times 12$ |
| ResNet-8  | ResNet block                  | $256 \times 19 \times 29 \times 12$ |
| ResNet-9  | ResNet block                  | $256 \times 19 \times 29 \times 12$ |

*Table 1.* The Table shows the layers of the encoder. The ResNet block is a series of two sub-blocks stacked one after another. Each of the sub-blocks consists of GroupNorm layer, ReLU layer and a Convolutional layer. A schematic representation of the ResNet block is shown in Figure 5.
### Table 2

The Table shows the layers of the decoder. UpLinear represents 3D bilinear interpolation and AddId represents an addition of filtered skip connection. The definition of ResNet block is analogous to the one defined in Table 1.

| Layer       | Details          | Output size           |
|-------------|------------------|-----------------------|
| Encoder output | Encoder output  | $256 \times 19 \times 29 \times 12$ |
| Upsample-1 | $128 \times 1 \times 1 \times 1$; stride=$1$; LeakyReLU; UpLinear; AddId | $128 \times 38 \times 58 \times 24$ |
| ResNet-10 | ResNet block     | $128 \times 38 \times 58 \times 24$ |
| Upsample-2 | $64 \times 1 \times 1 \times 1$; stride=$1$; LeakyReLU; UpLinear; AddId | $64 \times 76 \times 116 \times 48$ |
| ResNet-11 | ResNet block     | $64 \times 76 \times 116 \times 48$ |
| Upsample-3 | $32 \times 1 \times 1 \times 1$; stride=$1$; LeakyReLU; UpLinear; AddId | $32 \times 152 \times 232 \times 96$ |
| ResNet-12 | ResNet block     | $32 \times 152 \times 232 \times 96$ |
| Conv-2    | $1 \times 3 \times 3 \times 3$; stride=$1$; Sigmoid | $1 \times 152 \times 232 \times 96$ |
PET images are excluded as such with a probability $p$ (or included with a probability $1 - p$) while training the PAG model. When the PET images are included, the PET image feature extractor $f(x^{\text{PET}})$ along with the encoder, decoder, and the attention gates are trained. However, with the absence of PET images, only the encoder, decoder, and the attention gates are trained to leave out $f(x^{\text{PET}})$. This random exclusion/inclusion of PET images at every training step ensures that the model does not over-fit to one of the two scenarios i.e., when PET images are available or not. In contrast, allowing the model to be trained first on CT only images and then on PET/CT images or any other pre-determined strategy as such would run the risk of over-fitting to one of the particular scenarios. Accordingly, the robustness of the model to be able to take the PET images whenever they are available without over-fitting to either of the scenarios arises from this custom training strategy.

### 2.4 Loss functions

The model is trained by a composite loss function that is a weighted combination of two-loss functions: soft dice loss [3] and binary cross-entropy loss. The role of binary cross-entropy and the dice loss becomes evident when we examine their respective gradients. Let $p$ and $t$ denote the output of the sigmoid activation function towards the end of the network applied to the value $\alpha$ ($\alpha$ is also known as logits) and its respective target. Then the gradient of binary cross-entropy loss with respect to the logits ($\alpha$) is given by $p - t$. On the other hand, the gradient of the dice loss with respect to the output $p$ is given by $\frac{2(t^2-p^2)}{(p^2 + t^2)^2}$. It can be noticed that for very small values of $p$ or $t$, the gradient of the dice loss can easily explode to a huge value due to which the training can easily become unstable. This problem is however, not present with the gradient of the binary cross-entropy loss. The problem of the dice loss becomes even more acute when the target $t$ is zero. Zero target $t$ results in zero gradients, which means the model is not being trained at all. However, we want our network to maximize
the dice coefficient, which can be achieved by minimizing the dice loss directly. In addition, dice loss has been shown to perform extremely well for datasets with severe class imbalance [3]. Therefore, in order to accommodate the best of both the worlds, we consider a composite loss function that is a weighted combination of the two-loss functions.

\[ L_{\text{composite}} = L_{\text{dice}} + \lambda \cdot L_{\text{bce}} \]  

(3)

where \( \lambda \) is a hyper-parameter that controls the relative contribution of each of the two loss functions to the composite loss function and hence to the composite gradient step.

We observe the contribution of each of the two-loss functions while training our model. There was no apparent training when the model was trained with binary cross-entropy loss alone, probably because of severe class imbalance. A similar problem occurred while training the model with dice loss alone due to the issue of zero gradients raised above. However, the composite loss function performs quite well, suggesting that the two-loss functions contribute well to the training of the model in conjunction with one another.

### 2.5 Training algorithm

Let \( X^{ct} \) and \( X^{pet} \) represent the domain of CT and PET images respectively. Likewise, let \( Y \) represent the domain of segmented tumorous regions. Given is a dataset consisting of \( N \) PET/CT images \( \{x^{ct}_i, x^{pet}_i\}_{i=1}^N \) and \( M \) CT images \( \{x^{ct}_i \in X^{ct} \text{ and } x^{pet}_i \in X^{pet}\}_{i=1}^M \) for which corresponding PET images are missing. A typical scenario would be when \( N \geq 0 \) and \( M > 0 \). For every CT image we have a segmentation mask of tumorous regions i.e. \( \{x^{ct}_j, y_j\}_{j=1}^{N+M} \) where \( x^{ct}_j \in X^{ct} \text{ and } y_j \in Y \).

- A set of PET/CT and labeled images \( A: = \{\langle x^{ct}_i, x^{pet}_i, y_i\rangle\}_{i=1}^N \).
- A set of PET/CT and labeled images derived from samples of \( A \)
  \( A_{\text{aug}}: = \{\langle \phi(x^{ct}_i), \phi(x^{pet}_i), \phi(y_i)\rangle\}_{i=1}^N \).
- A set of CT and labeled images \( B: = \{\langle x^{ct}_j, y_j\rangle\}_{j=1}^{N+M} \).
- A set of CT and labeled images derived from samples of \( B \)
  \( B_{\text{aug}}: = \{\langle \phi(x^{ct}_j), \phi(y_j)\rangle\}_{j=1}^{N+M} \).

where \( \phi(x) \) is the result of applying a data pre-processing pipeline to input image \( x \) as described earlier.

---

**Algorithm 1** Training algorithm for PAG model

# Assume \( \Theta \) to be network parameters
# \( m \) := Batch size
# \( N + M \) := Total number of CT images
# \( E \) := Number of epochs
# $p$: probability of exclusion of PET images at every training step
# $p_{\text{data-aug}}$: probability of choosing augmented data samples
# \text{uniform}(0,1): = \text{uniform random sampler in the range } [0,1]
$T := (N + M)/m \# \text{Number of iterations for every epoch}$

Initialize $\Theta$ randomly

\textbf{for} epoch $e \in [1, 2, \ldots E]$

\{ 
  shuffle $A, A_{\text{aug}}, B, B_{\text{aug}}$ to prevent cycles
  \textbf{if} uniform$(0,1) < p_{\text{data-aug}}$
  \{ 
    \# Consider augmented samples for training
    $A_{\text{choose}} = A_{\text{aug}}$
    $B_{\text{choose}} = B_{\text{aug}}$
  \}
  \textbf{else}
  \{ 
    \# Consider true samples for training
    $A_{\text{choose}} = A$
    $B_{\text{choose}} = B$
  \}
  \textbf{for} iteration $i \in [0, 1, 2, 3, \ldots T]$
  \{ 
    \textbf{if} uniform$(0,1) < p$
    \{ 
      \# Exclude PET images
      draw $m$ random samples without replacement \{$(a_j, c_j) \in B_{\text{choose}}$\}$
      compute predictions $\hat{c}_j = H(a_j, 0)$
    \}
    \textbf{else}
    \{ 
      \# Include PET images
      draw $m$ random samples without replacement \{$(a_j, b_j, c_j) \in A_{\text{choose}}$\}$
      compute predictions $\hat{c}_j = H(a_j, b_j)$
    \}
    compute loss $L_{e,i} = \frac{1}{m} \sum_{j=1}^{m} L_{\text{composite}}(c_j, \hat{c}_j)$
    compute gradients $\Delta \Theta = \partial L_{e,i}/\partial \Theta$
    update the parameters $\Theta := \Theta - \eta \Delta \Theta \# \eta$: Decaying learning rate
  \}
\}

The decay in the learning rate is given by Equation 4. In addition, Table 4 gives the list of hyper-parameters used. Tables 5-7 give the TNM staging protocol.

$$\alpha = \alpha (1 - \frac{e}{N_e})^{0.9} \quad (4)$$
| Hyper-parameter | Description                                                                 | Value   |
|----------------|------------------------------------------------------------------------------|---------|
| $\lambda$     | Weight of BCE loss in the combined loss function                            | 10      |
| $\alpha_0$    | Initial learning rate                                                        | 0.0001  |
| $\beta$       | The L2 regularisation parameter                                              | $1e^{-5}$|
| $p_{\text{data-aug}}$ | Probability of inclusion of data augmentation for every training epoch | 0.25    |
| $p$           | Probability of inclusion of PET images at every training step                | 0.5     |

**Table 4.** Hyper-parameters set across all the models. ($p$ parameter appropriate for PAG model only.) The hyper-parameters were set based on the performance on one of the folds at a different seed than that is used for evaluation of models.

**T1:** Tumor 3 cm or less in greatest dimension, surrounded by lung or visceral pleura, without bronchoscopic evidence of invasion more proximal than the lobar bronchus

**T2:** Tumor more than 3 cm but 7 cm or less or tumor with any of the following features: involves main bronchus, 2 cm or more distal to the carina; invades visceral pleura; associated with atelectasis or obstructive pneumonitis that extends to the hilar region but does not involve the entire lung

**T3:** Tumor more than 7 cm or one that directly invades any of the following: parietal pleural (PL3), chest wall (including superior sulcus tumors), diaphragm, phrenic nerve, mediastinal pleura, parietal pericardium; or tumor in the main bronchus less than 2 cm distal to the carina but without involvement of the carina; or associated atelectasis or obstructive pneumonitis of the entire lung or separate tumor nodule(s) in the same lobe.

**T4:** Tumor of any size that invades any of the following: mediastinum, heart, great vessels, trachea, recurrent laryngeal nerve, esophagus, vertebral body, carina, separate tumor nodule(s) in a different ipsilateral lobe.

**M1a:** Separate tumor nodule(s) in a contralateral lobe, tumor with pleural nodules or malignant pleural (or pericardial) effusion

**M1b:** Distant metastasis (in extrathoracic organs)

**N1:** Metastasis in ipsilateral peribronchial and/or ipsilateral hilar lymph nodes and intrapulmonary nodes, including involvement by direct extension

**N2:** Metastasis in ipsilateral mediastinal and/or subcarinal lymph node(s)

**N3:** Metastasis in contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene, or supraclavicular lymph node(s)

**Table 5.** Definitions for the Primary, nodular and metastases tumors stage classifications according to TNM staging protocol. Definitions adapted from [2, p 257-263].
1  Low cervical, supraclavicular, and sternal notch nodes.
2R  Upper Paratracheal (right)
2L  Upper Paratracheal (left)
3a  Pre-vascular
3p  Retrotracheal
4R  Lower Paratracheal (right)
4L  Lower Paratracheal (left)
5   Subaortic
6   Para-aortic (ascendic aorta or phenic)
7   Subcarinal
8   Paraesophageal (below carina)
9   Pulmonary ligament
10  Hilary
11  Interlobar
12  Lobar
13  Segmental
14  Subsegmental

Table 6. Labels for the locations of the nodular tumors defined according to TNM staging protocol. Definitions adapted from [2, p 257-263].

| T-Descriptor          | N-Descriptor          | M-Descriptor          |
|-----------------------|-----------------------|-----------------------|
| T1                    | N1_{10-11i}           | M1a_{contralat}       |
| T2\_main\_bronchus    | N1_{12-15i}           | M1a_{pleura}          |
| T2\_visc\_pleura      | N2_{2i}               | M1b_{adrenal}         |
| T2\_obstr\_lobe       | N2_{3}                | M1b_{liver}           |
| T3\_main\_bronchus    | N2_{4i}               | M1b_{bone}            |
| T3\_obstr\_lung       | N2_{5i}               | M1b_{node}            |
| T3\_nodule\_same\_lobe| N2_{6}                |                       |
| T4\_inv\_mediastinum  | N2_{7}                |                       |
| T4\_nodule\_diff\_lobe| N2_{8i}               |                       |
|                       | N2_{9}                |                       |
|                       | N3_{1}                |                       |
|                       | N3_{2c}               |                       |
|                       | N3_{4c}               |                       |
|                       | N3_{5c}               |                       |
|                       | N3_{8c}               |                       |
|                       | N3_{9c}               |                       |
|                       | N3_{10-11c}           |                       |

Table 7. The T-labels, N-labels ad M-labels that are present in the current dataset. The T-labels are for the primary tumors, the N-labels are for the nodular tumors and the M-Labels are for the Metastasis tumors. They are defined by their respective stages and locations according to TNM staging protocol. The T-label stage is the one that primarily decides the stage of cancer.
References

1. Ben-Cohen, A., Klang, E., Raskin, S.P., Soffer, S., Ben-Haim, S., Konen, E., Ami-
tai, M.M., Greenspan, H.: Cross-modality synthesis from ct to pet using fcn and
gan networks for improved automated lesion detection. Engineering Applications of
Artificial Intelligence 78, 186–194 (2019)
2. Edge B.S., Byrd D.R., C.C.F.A.G.F.T.A.: AJCC Cancer Staging Manual (7th edi-
tion), vol. 2010. Springer (2010)
3. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks
for volumetric medical image segmentation. In: 2016 Fourth International Confer-
ence on 3D Vision (3DV). pp. 565–571. IEEE (2016)
4. Sexauer, R., Weikert, T., Mader, K., Wicki, A., Schädelin, S., Stieltjes, B., Bre-
merich, J., Sommer, G., Sauter, A.W.: Towards more structure: comparing tnm
staging completeness and processing time of text-based reports versus fully seg-
mented and annotated pet/ct data of non-small-cell lung cancer. Contrast media &
molecular imaging 2018 (2018)
5. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference
on Computer Vision (ECCV). pp. 3–19 (2018)