Prediction of FDG uptake in Lung Tumors from CT Images Using Generative Adversarial Networks

Annika Liebgott*, Darius Hinderer†, Karim Armanious*, Alexander Bartler†,
Konstantin Nikolaou*, Sergios Gatidis* and Bin Yang†
* Department of Diagnostic and Interventional Radiology, University Hospital of Tübingen, Germany
†Institute of Signal Processing and System Theory, University of Stuttgart, Germany
Email: annika.liebgott@iss.uni-stuttgart.de

***Abstract*—In modern medicine, combined PET-CT is a commonly-used tool in clinical diagnostics, which is especially important in oncology for staging or treatment planning. Variations in FDG uptake visible in a PET image, which indicate variances in metabolic activity, are not visually recognizable within a CT scan from the same region, making both imaging modalities necessary for diagnosis and exposing the patient to a high amount of radiation. In this study, we investigate the possibility of using generative adversarial networks (GANs) to synthesize a PET image from a CT scan to predict metabolic activity.**

***Index Terms*—PET/CT, lung cancer, FDG uptake, machine learning, GANs**

I. Introduction

The value of using medical imaging techniques like Computed Tomography (CT) in present-day clinical diagnostic is unambiguous. When combined with Positron Emission Tomography (PET) [1], CT becomes an even more powerful diagnostic tool able to visualize both detailed physiological structures and metabolic activities within a patient's body. To have access to both kinds of information is crucial in the diagnosis of some diseases, especially in the field of oncology. The basic principle of PET is injecting a positron-emitting tracer into the patient's body. Upon decay, the tracer releases positrons which then annihilate with electrons, creating gamma radiation that can be measured by the PET detector. The higher the concentration of the detected gamma particles, the higher the metabolic activity is in a particular body region.

The majority of clinical PET scans utilizes the radioisotope fluorine-18 synthesized into fludeoxyglucose (FDG) as tracer. The PET scanner then detects variations in regional glucose uptake, which is measured in Standardized Uptake Values (SUV). That way, ongoing metabolic processes can be localized and quantified. Image resolution, however, is rather low, making the combination with a CT or MRI scan necessary to get sufficient physiological information beyond glucose uptake. In oncology, combined FDG-PET/CT is used to get information from the glucose uptake regarding location and size of a tumor and/or metastases and simultaneously receive detailed information about the patient's physiology [2]–[4]. An example is given in Figure 1, where the CT shows anatomic details of the thorax while the corresponding PET scan can be regarded as a map of the metabolic activity.

While being a powerful tool in clinical diagnostic, one drawback of combined PET/CT is the higher amount of radiation a patient is exposed to. Repeating a PET/CT scan within a short time frame, e.g. with additional, more specialized tracers, is normally not possible.

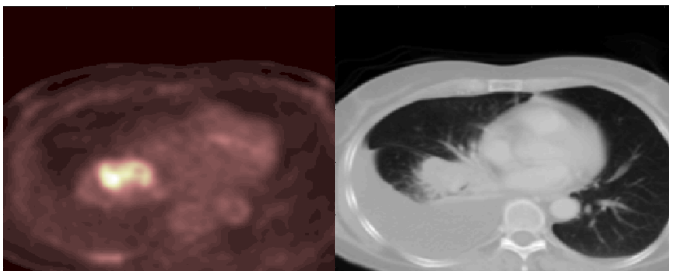


Fig. 1: Example of co-registered images from a PET/CT scan of the thorax. Left: PET image, right: CT scan

Estimating the metabolic activity normally shown on an FDG-PET image from a CT scan only is hence an interesting, yet challenging task. While the physiological processes causing variations in CT attenuation and FDG uptake are different, there are some diagnostic findings for which the same processes are the underlying cause. While tumor necrosis results in decreased perfusion, leading to reduced CT density, it also causes a decrease in FDG uptake, making CT density a possible surrogate for metabolic activity in this case. However, finding visual indicators for varying FDG uptake in CT images is hardly possible.

In a previous study we proposed a feature-based machine learning approach to locate regions of especially high SUVs within CT images of lung tumors [5], which is described in section II of this paper. Since the ultimate goal of our work is to generate detailed PET-like images and our first approach was not suitable for this task, we decided to investigate the ability of generative adversarial networks (GANs) [6] to create a more realistic estimation of variations in FDG uptake within a tumor from CT input data.

Works in related research fields generally focus on other questions, e.g. using features derived from PET and CT to distinguish between benign and malignant structures, not to describe the metabolic activity within a tumor. For instance, Arimura et al. [7] and Bi et al. [8] use SVMs trained on features extracted from PET-CT images to localize regions of high FDG uptake in whole-body scans with the goal of automatically finding or contouring tumors. A study by Kerhet et al. [9] proposes to use a classifier trained on PET image features to predict an optimum SUV threshold for tumor segmentation. Another popular research area, which actually focuses on intratumor heterogeneity, is the estimation of treatment success, i.e. finding links between patient outcome and PET or PET/CT features derived from regional FDG uptake [10]–[12]. We, in contrast,

978-9-0827-9703-9/19/$31.00

investigate possible relationships between both imaging modalities.

Image-to-imge translation of PET and CT images using deep learning have been proposed in various studies in the past years. Most works focus on denoising or dose reduction [13]–[15], but there have also been some efforts investigating PET synthesis employing GANs. Ben-Cohen et al. proposed to generate full-body PET images from CT scans combining a fully convolutional network with a GAN to improve automated lesion detection on CT scans [16]. Another study by Bi et al. proposes to use multi-channel GANs to synthesize FDG-PET images aiming to detect lung tumors [17]. In Contrast, we are not interested in finding unusually high FDG uptake in a full-body setup to find tumors, but in investigating the variations in metabolic activity within regions of high FDG uptake.

II. Previous work

In a previous study, we proposed to use an SVM to classify patched CT images with labels corresponding to regional glucose uptake of the co-registered PET images of a PET/CT scan [5].

Although this approach showed promising results, it had some obvious shortcomings. First of all, we categorized the FDG uptake in only two classes (high and very high SUVs), which resulted in an at best coarse representation of the actual textural distribution within the tumor. Secondly, the labeling strategy only allowed for very small image patches, resulting in a high number of 261032 patches and corresponding computational complexity. Finally, the most critical shortcoming of our previous approach was the quite low overall test accuracy of 76.97% on 29 unseen patients despite the simplified problem. While visual comparison of the predicted FDG uptake classes and the original PET images converted to the binary label representation yielded quite good results for larger tumors, the classifier was not able to correctly predict FDG uptake in smaller structures. Experiments using more than 2 intensity classes resulted in significantly worse performance. To overcome these issues, we decided to abandon the feature-based approaches and small image patches for now and started investigating deep learning techniques.

III. Proposed method

The general idea of the approach is based on image-to-image translation, i.e. generating one type of image from another (e.g. photograph from sketch). We propose to employ a generative adversarial network (GAN) to generate PET-like images from the CT scans of lung tumors to avoid the need for explicitly extracted image features and class labels.

A. Conditional generative adversarial network

Our proposed approach is based on pix2pix [18], a particularly popular GAN architecture for image-to-image translation tasks proposed by Isola et al. in 2016. It is a conditional GAN, i.e. rather than learning the mapping of random noise z to output image y, the model is given an input image x as additional input and trained to map x and z to y. The generator G is hence trained to generate images as similar to the target y as possible while the objective of the discriminator D is to detect the fake images generated by G correctly. The training objective can be formulated as $\min_G \max_D \mathcal{L}_{GAN}$ with the adversarial loss

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{x,y}\left[\log D(x,y)\right] + \mathbb{E}_{x,z}\left[\log\left(1 - D\left(x, G(x,z)\right)\right)\right].$$

TABLE I: $\overline{\text{SSIM}}$, $\overline{\text{MSE}}$ and $\overline{\text{PSNR}}$ of validation images after training the GAN with all CT/PET images and differently preprocessed input images: no preprocessing (NP), CLAHE, histogram matching (HM) and Gaussian blurring (GB).

	NP	CLAHE	CLAHE+HM	CLAHE+HM+GB
$\overline{\text{SSIM}}$	0.3510	0.3815	0.4833	0.4470
$\overline{\text{MSE}}$	1.1410	1.1022	0.7473	0.7107
$\overline{\text{PSNR}}$	14.1778	14.4273	17.6506	17.5422

Since the generator needs to not only produce output the dicriminator is unable to distinguish from real images, but which is also close to the ground truth, an additional L1 loss is introduced:

$$\mathcal{L}_{\text{L1}} = \mathbb{E}_{x,y,z}\left[\|y - \log G(x,z)\|_1\right],$$

leading to overall training objective

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{GAN}} + \lambda \mathcal{L}_{\text{L1}},$$

where λ is a parameter to weight the L1 loss.

B. Image preprocessing

Initial attemtps of training the GAN showed poor results. The structural similarity index (SSIM) between the generated PET images and the ground truth PET images had a mean value of 0.35, the calculated mean peak signal-to-noise ratio was 14.1778. As already discussed, a direct connection between PET and CT depends on the presence of physiological processes affecting both modalities. This means estimating a PET image from a CT may not be possible in each tumor. Hence, we started analyzing our data set for possible visual connections between CT and PET to see which images might be better suited for training the GAN than others. Such connections are, however, not necessarily directly visible in the CT without enhancing textural variations first. To this end, we employed various preprocessing techniques.

We used contrast-limited adaptive histogram equalization (CLAHE) [19] to enhance the contrast of the CT images and additionally matched the histogram of the CT images to that of the corresponding PET (HM) to increase performance. However, the contrast enhancement accentuated tiny textural variations in the CT which are not present in the smoother PET limiting the ability of the GAN to correctly learn a relationship between input and target. This assumption was supported by the finding that smaller tumors, which did not have as many textural variations in the CT, could be estimated better than larger ones. Thus, we limited the detailed variations in the CT images by applying Gaussian blurring (GB).

IV. Experiments and results

A. Data set and experimental setup

For our experiments, we used 3D FDG-PET/CT images of 202 patients with non-small cell lung cancer. All images were acquired on a clinical PET/CT scanner (Siemens Biograph mCT). An experienced radiologist manually segmented the primary lung tumor to define the volumes of interest (VOIs). Segmentation was performed on the CT images. Since the GAN needs images to be perfectly aligned pairwise, we used re-sampling and interpolation to achieve the same resolution in x, y and z direction for PET, CT and

TABLE II: $\overline{\text{PSNR}}$ and $\overline{\text{MSE}}$ of validation images after training the GAN with training sets composed with different thresholds s. For the varying s (column A), training was conducted with an augmented training set (AUG). As preprocessing, a combination of CLAHE + HM + GB (in the following denoted as COMB) was used. TA denotes results of validation on all available validation data, TS on validation data selected with respect to s. Columns B and C show notable exemplary results of experiments without data augmentation (B) and with other preprocessing (C).

	A				B	C
	$s = 0$	$s = 0.5$, AUG	$s = 0.6$, AUG	$s = 0.7$, AUG	$s = 0.7$	$s = 0.7$, AUG, CLAHE+GB
$\overline{\text{PSNR}}$, TS	17.5422	20.4661	21.2475	23.6733	21.7307	18.0713
$\overline{\text{PSNR}}$, TA	–	18.4465	17.8969	18.0448	17.2061	13.1140
$\overline{\text{MSE}}$, TS	0.7107	0.5928	0.6315	0.7928	1.0566	0.5468

the segmentation masks. As input data, we only used the segmented tumor regions, amounting to a total number of 3575 2D images.

We used 173 patients for training, the remaining 29 were saved for validation. We explicitly used the same validation patients as in our SVM study to be able to compare the results. The baseline pix2pix architecture employs ADAM optimizer. For the learning rate we found 0.0002 to give the best results for our task. Additionally, we adapted the baseline model to train the generator and discriminator in a 1:3 ratio. Each model was then trained for 100 epochs on a GPU, since more epochs did not improve results during early analyses.

For each experiment, we calculated the mean structural similarity index ($\overline{\text{SSIM}}$), mean mean square error ($\overline{\text{MSE}}$) and mean peak signal-to-noise ratio ($\overline{\text{PSNR}}$) between the generated images and the target PET images in the validation set. We only evaluated similarity within the tumor region. In addition to these quantitative measures, visual resemblance between the GAN output and the target PET images gives also a qualitative feeling about the quality of the synthesized PET images. Overall, we found $\overline{\text{PSNR}}$ to be the best of the 3 calculated metrics as surrogate for visual resemblance.

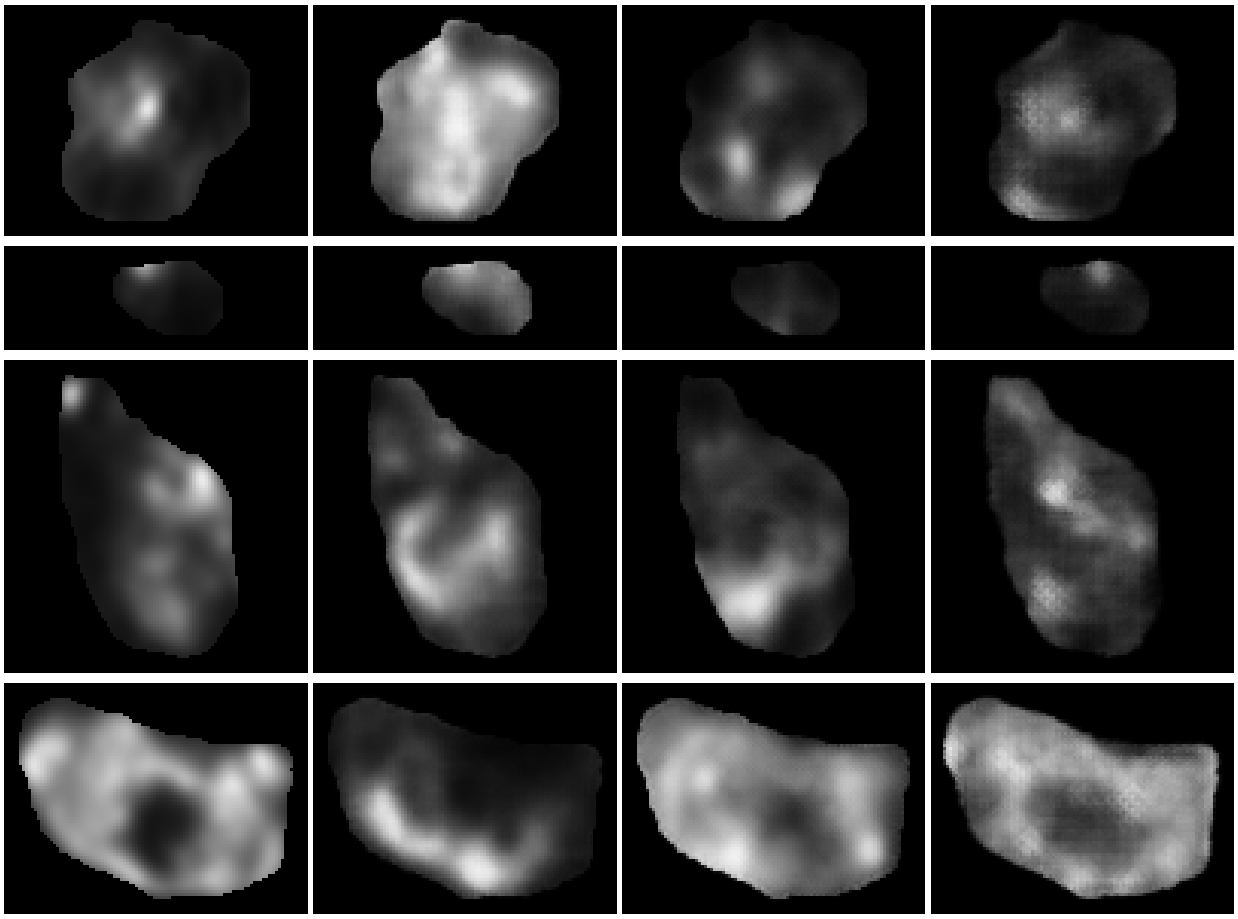
We investigated the ability of the GAN to generate realistic PET images from the input CT scans depending on the various preprocessing steps. As previously mentioned, we assume that a certain relationship between the structure of the input images and the targets needs to exist in order for a GAN to be able to generate realistic images. As an indicator for the presence of such a connection, we calculated the SSIM between the ROIs of the preprocessed CT images and the corresponding PET scans and used only images with an SSIM over a pre-defined threshold s for training. We conducted several analyses with different values for s. Since the training data did not include enough images of sufficiently high visual connection for some values of s to still be able to properly train the GAN, we employed data augmentation in such cases (rotation, stretching, translation). The results shown in the following were achieved when using COMB for image preprocessing. Using other combinations of preprocessing techniques led to lower performance.

B. Results of PET prediction

Table I shows the resulting $\overline{\text{SSIM}}$, $\overline{\text{MSE}}$ and $\overline{\text{PSNR}}$ of the validation set for training with images with and without preprocessing. The numbers clearly indicate the importance of these preprocessing steps for the GAN to be able to correctly learn to generate PET images from CT. We thoroughly tested all possible combinations of the applied preprocessing, in Table I we listed the ones that worked best in our experiments. Although $\overline{\text{SSIM}}$ and $\overline{\text{MSE}}$ are lower for COMB than for CLAHE+HM while training on all available images, we got significantly better results for COMB than for any other combination during our experiments with varying values s.

In Table II, exemplary results of our analyses are shown. As we found it to correlate best with visual perception of the output image quality, we focused on $\overline{\text{PSNR}}$ for quantitative evaluation, $\overline{\text{MSE}}$ is given as additional metric. Although $\overline{\text{SSIM}}$ was behaving similarly to $\overline{\text{PSNR}}$ in most cases, we chose not to rely on this measure for quality assessment in our experiments involving varying s, as an increase in $\overline{\text{SSIM}}$ would be expected when this measure is used for dataset selection. In column A, results for the analyses we conducted with different values for similarity threshold s are given. Due to a lack of enough data with higher SSIM, we only conducted analyses up to $s = 0.7$. For increasing values of s, $\overline{\text{PSNR}}$ improves, indicating that a sufficient given connection between input CT and target PET enhances the performance of the GAN. Column B highlights the importance of data augmentation. For training with images with SSIM > 0.7 without augmenting the training set, $\overline{\text{PSNR}}$ is significantly lower, results for other values of s confirmed this finding. The numbers in column C clarify the importance of the combined preprocessing. We chose CLAHE+GB as example because it additionally shows, why $\overline{\text{MSE}}$ should not be used as a measure for output quality. While we achieved the best values for $\overline{\text{MSE}}$ in all our experiments involving training on images with CLAHE+GB, all other measures and visual resemblance between output and target image deteriorated significantly. In general, we found that $\overline{\text{MSE}}$ does not reflect visual resemblance well in our case.

Figure 2 further encourages the conclusions drawn from the calculated results. The output images of training with preprocessed images in column (c) mostly have a stronger resemblance to the real PET images in column (a) than the images the GAN generated from CT images without any preprocessing (b). However, they are not as similar to the target as the images in column (d), where only training data with $\text{SSIM} > 0.7$ was used. The results in Table II support this perception: for training on data with $\text{SSIM} > 0.7$, $\overline{\text{PSNR}}$ was not only higher when validating on data selected accordingly ($\overline{\text{PSNR}}$, TS), but there was also an improvement when validating the trained model on all available data ($\overline{\text{PSNR}}$, TA). Artifacts present in the generated images in (d) are due to the lack of a sufficient amount of training data. However, augmenting the training data too much



(a) target PET image (b) output NP (c) output COMB (d) output COMB, $s = 0.7$, AUG

Fig. 2: Examples of images the GAN was able to generate well from the CT images. The far left column shows the target PET image (a), the middle columns the output for training without first applying preprocessing (b) and with the most successful combination of all proposed preprocessing steps (c). In the column on the far right, the results for augmented training data with all preprocessing with a similarity over $s = 0.7$ are shown (d).

resulted in decreased GAN perfomance, probably due to overfitting.

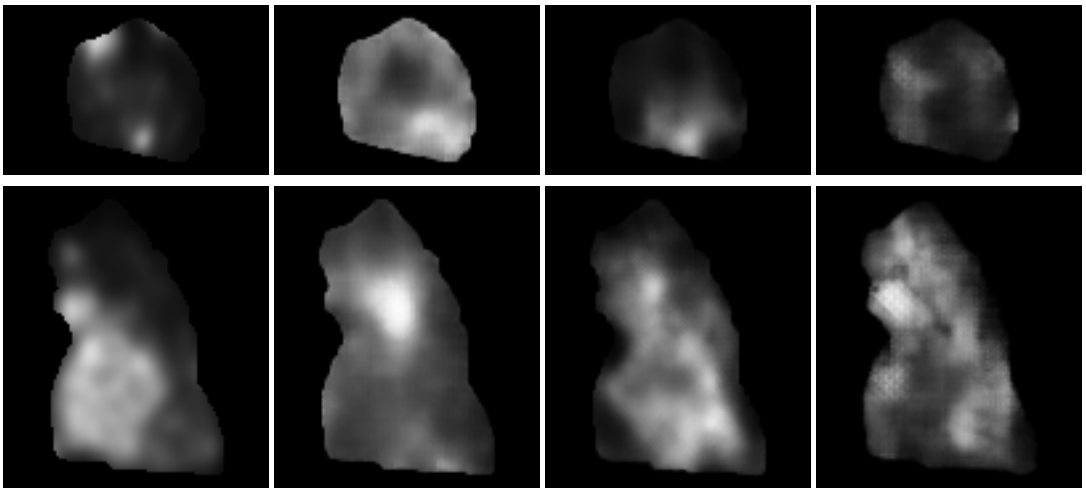
Figure 3 shows examples where the GAN was not able to synthesize a realistic PET, even when training on images with a similarity over $s = 0.7$. Some parts of the predicted images even look reversed compared to the target.

V. Discussion

Our experiments showed that at least under the proper circumstances, i.e. when suitable training data is available, it is possible to predict metabolic activity within a tumor from a CT scan. We achieved good resemblance between output and target when the GAN was trained with images where a visual connection between PET and CT images was already present, e.g. originating from tumor necrosis leading to decreased perfusion in the CT and decreased glucose uptake in the PET. However, our approach has limitations as GAN performance is strongly dependent on the underlying physiological/biological conditions. We used image similarity as a simple approach to test our hypothesis that, in order to be able to solve the task, the GAN needs data with a relationship between input and target images. It hence is crucial to first find a way to determine wether or not a CT image of a tumor is suitable without needing to compare it to a PET image first before this could be used in a real-world application.

Overall, our previous approach and our goal in this study are two completely different tasks, making a fair comparison difficult. The SVM was trained for a binary classification between high and very high SUVs, giving only a coarse representation of the structure within the tumor. The GAN was trained to synthesize realistic images, which is a much more complicated task. However, one major issue we experienced with our SVM approach was the fact that the classifier was not able to correctly identify regions of high FDG uptake in smaller ROIs. A particularly interesting result of this study is that the GAN behaves the opposite way: most of the smaller tumors in our validation set are generated more accurately than the larger ones. The reason for this finding might be that smaller tumors tend to have less chaotic textural variations than some of the bigger tumors, making it harder for the GAN to generate realistic images. The feature-based SVM, in contrast, would benefit from a more diverse textural information.

While the proposed approach using GANs has been able to



(a) target PET image (b) output NP (c) output COMB (d) output COMB, $s = 0.7$, AUG

Fig. 3: Examples of bad results for the same training as for the images in Figure 2.

overcome some of the shortcomings of the previous feature-based method (i.e. problems with small ROIs, only simplified SUV representation), there are still some issues that need to be addressed in future studies. Special focus should be on the analysis of biological and physiological conditions of the data to get a better understanding about which images of tumors are suitable for prediction of FDG uptake from CT images and which are not. With this knowledge, a classifier could be trained to determine whether or not the PET of a tumor can be synthesized from a CT scan. Additionally, we aim to investigate other GAN architectures more specialized on medical image-to-image translation to hopefully increase performance. Simultaneously, we plan experiments with other deep learning techniques (e.g. convolutional neural networks) and also to try combining our feature-based approach with deep learning by investigating hybrid feature learning.

VI. Conclusion

In this study, we presented an approach to estimate metabolic activity within a tumor by training a GAN to generate PET-like images from CT input. We found that, while the GAN performed well for part of the input images, the ability to create realistic PET images strongly depends on the available data. From our results, we conclude that predicting FDG uptake in tumors is indeed possible for certain cases, but only if an underlying physiological process influencing both PET and CT imaging is present in the data.

References

[1] R. E. Coleman, "Clinical PET in oncology," *Clinical Positron Imaging*, vol. 1, no. 1, pp. 15 – 30, 1998.

[2] I. R. Francis, R. K. J. Brown, and A. M. Avram, "The clinical role of CT/PET in oncology: an update," *Cancer Imaging*, vol. 5, pp. S68 – S75, 2005.

[3] A. Gallamini, C. Zwarthoed, and A. Borra, "Positron emission tomography (PET) in oncology," *Cancers*, vol. 6, no. 4, pp. 1821–1889, 2014.

[4] N. C. Nguyen, I. Akduman, and M. M. Osman, "F-18 FDG-PET and PET/CT imaging of cancer patients," *J. of Radiology Nursing*, vol. 27, no. 2, pp. 61 – 69, 2008.

[5] A. Liebgott, S. Gatidis, F. Liebgott, K. Nikolaou, and B. Yang, "Automated detection of high FDG uptake regions in CT images," in *2018 IEEE ICASSP*, April 2018, pp. 1065–1069.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Int. Conf. on Neural Information Processing Systems*. 2014, pp. 2672–2680, MIT Press.

[7] H. Arimura, Z. Jin, Y. Shioyama, K. Nakamura, T. Magome, and M. Sasaki, "Automated method for extraction of lung tumors using a machine learning classifier with knowledge of radiation oncologists on data sets of planning CT and FDG-PET/CT images," in *2013 35th Annual Int. Conf. of the IEEE EMBS*, July 2013, pp. 2988–2991.

[8] L. Bi, J. Kim, D. Dagan Feng, and M. Fulham, "Classification of thresholded regions based on selective use of PET, CT and PET-CT image features," in *2014 36th Annual Int. Conf. of the IEEE EMBS*, Aug 2014, pp. 1913–1916.

[9] A. Kerhet, C. Small, H. Quon, T. Riauka, L. Schrader, R. Greiner, D. Yee, A. McEwan, and W. Roa, "Application of machine learning methodology for PET-based definition of lung cancer," in *Current oncology*, 2010.

[10] T. Win, K. A. Miles, S. M. Janes, B. Ganeshan, M. Shastry, R. Endozo, M. Meagher, R. I. Shortman, S. Wan, I. Kayani, P. J. Ell, and A. M. Groves, "Tumor heterogeneity and permeability as measured on the CT component of PET/CT predict survival in patients with non–small cell lung cancer," *Clinical Cancer Research*, vol. 19, no. 13, pp. 3591–3599, 2013.

[11] S. Chicklore, V. Goh, M. Siddique, A. Roy, P. K. Marsden, and G. J. R. Cook, "Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 40, no. 1, pp. 133–140, Jan 2013.

[12] I. El Naqa, P. W. Grigsby, A. Apte, E. Kidd, E. Donnelly, D. Khullar, S. Chaudhari, D. Yang, M. Schmitt, Richard Laforest, W. L. Thorstad, and J. O. Deasy, "Exploring feature-based approaches in PET images for predicting cancer treatment outcomes," *Pattern Recogn.*, vol. 42, no. 6, pp. 1162–1171, June 2009.

[13] Y. Wang, B. Yu, L. Wang, C. Zu, D. S. Lalush, W. Lin, X. Wu, J. Zhou, D. Shen, and Lu. Zhou, "3D conditional generative adversarial networks for high-quality PET image estimation at low dose," *NeuroImage*, vol. 174, pp. 550 – 562, 2018.

[14] L. Xiang, Y. Qiao, D. Nie, L. An, W. Lin, Q. Wang, and D. Shen, "Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI," *Neurocomputing*, vol. 267, pp. 406–416, 12 2017.

[15] S. Kaplan and Y.-M. Zhu, "Full-dose PET image estimation from low-dose PET image using deep learning: a pilot study," *J. of Digital Imaging*, pp. 1–6, 2018.

[16] A. Ben-Cohen, E. Klang, S. Raskin, S. Soffer, S. Ben-Haim, E. Konen, M. Amitai, and H. Greenspan, "Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection," *Engineering Applications of Artificial Intelligence*, vol. 78, 02 2018.

[17] L. Bi, J. Kim, A. Kumar, D. D. Feng, and M. J. Fulham, "Synthesis of positron emission tomography (PET) images via multi-channel generative adversarial networks (GANs)," in *CMMI/RAMBO/SWITCH@MICCAI*, 2017.

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[19] K. Zuiderveld, "Contrast limited adaptive histogram equalization," *Graphics Gems IV*, pp. 474–485, 12 1994.