Disease-image-specific Learning for Diagnosis-oriented Neuroimage Synthesis with Incomplete Multi-Modality Data – *Supplementary Materials*

Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen, *Fellow, IEEE*

In what follows, we first show the advantages of the proposed method comparing to our previous works in Section 1 and more details about the datasets we used in Section 2. Then, we present more discussion on hyper-parameter settings in Section 3, including the batch sizes, training epochs, and backbones. We also explain more details about the advantages of our DSDL framework (Sections 4 and 5) and more samples of synthetic neuroimages (Section 6). In addition, we discussed the potential applications in other scenarios (Section 7).

1 Technical Novelty

Due to its ability to provide complementary structural and functional information, multi-modal neuroimaging (e.g., MRI and PET) has been commonly used for the diagnosis of neurodegenerative disorders such as the Alzheimer's disease (AD). However, the missing data problem is almost inevitable in clinical practice due to various reasons, e.g., PET scanning may be rejected by some subjects due to high cost or concern of radioactive exposure. Many studies on multi-modal neuroimages simply discard the subjects with missing PET, leading to a significant decrease of the number of training subjects. However, deep learning-based diagnostic models, which have become the de facto standard in medical image analysis, are prone to over-fit the training dataset and hence exhibit unreliable performance, if the training dataset is small.

To solve the missing data problem, we attempted, in our previous works (e.g., [1]), to impute the missing PET data using the available MRI data based on the observation that there probably exists underlying relevance between the images acquired from the same subject but using different modalities. We first resorted 3D Cycle-consistency Generative Adversarial Networks (3D-cGAN) to learn the bi-directional mapping between relevant image domains (i.e., across PET and MRI), where the cycle-consistency loss is used to capture their probable underlying relationship. Then, we developed a landmark-based multi-modal multi-instance learning (LM3IL) model to use the complete MRI and PET data (i.e., real or synthetic PET + real MRI) for AD diagnosis and mild cognitive impairment (MCI) conversion prediction. This previous work achieve good performance because (1) we directly imputed missing PET scans to almost double the number of training subjects, leading to a more reliably-learned diagnostic model; (2) we performed the classification only on the patches around disease-related landmarks, which are pre-defined manually.

However, this previous work is an early attempt to learn data imputation and classification under a two-stage framework in a data-driven manner and has three limitations. First, the image synthesis and disease diagnosis are treated as two standalone tasks, and hence the difference of specificities conveyed by two modalities is ignored. Second, the cycle consistency we used is a weak constraint to preserve the disease information, since it only encourages pixel/voxel consistency after two transformations (i.e., transformed through two synthesis models), not encouraging the consistency of disease-relevant information. Third, the classification performance relies highly on

- *Y. Pan and Y. Xia were partially supported by the National Natural Science Foundation of China under Grant 61771397, the Science and Technology Innovation Committee of Shenzhen Municipality, China, under Grant JCYJ20180306171334997, and the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX201835. M. Liu and D. Shen were partially supported by NIH grant (No. AG041721). Corresponding authors: Yong Xia, Mingxia Liu, and Dinggang Shen.*

- *Y. Pan and Y. Xia are with School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China. Y. Pan is also with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. (E-mail: {yspan@mail.,yxia@}nwpu.edu.cn) M. Liu and D. Shen are with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. D. Shen is also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea. (E-mails: mxliu@med.unc.edu, Dinggang.Shen@gmail.com)*

This work was finished when Y. Pan was visiting the University of North Carolina at Chapel Hill. Part of the data used in this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The investigators within the ADNI contributed to the design and implementation of ADNI and provided data but did not participate in analysis or writing of this article.

the precision of landmarks. However, no landmark set is universally recognized as precise and comprehensive, since the pathological changes can be subtle in the early course of the disease and there can be some overlap with other neurodegenerative types.

To address these issues, we proposed a disease-image-specific deep learning (DSDL) framework for *joint neuroimage synthesis and disease diagnosis* using incomplete multi-modality neuroimages. Specifically, we first designed a Disease-image-Specific Network (DSNet) with a spatial cosine module to implicitly model the disease-image specificity, and then developed a Feature-consistency Generative Adversarial Network (FGAN) to impute missing neuroimages. During the image synthesis, we aim to preserve the disease-image-specific information via using the feature-consistency constraint, which encourages the multi-layer feature maps (generated by DSNet) of a synthetic image and its corresponding real image to be consistent. For instance, at an early stage of the AD process, subtle physiological changes can be detected by PET long before any changes are apparent on MRI. In this case, when imputing the missing PET scan based on the available MRI scan, our previous method cannot generate the physiological changes since they are not apparent on the MRI scan. However, with the help of the proposed the feature-consistency constraint, our FGAN can generate the synthetic PET scan that contains the physiological changes via implicitly learning from same-class PET scans in the training dataset. In this way, our FGAN is correlated with DSNet, leading the missing neuroimages to be imputed in a diagnosis-oriented manner. Hence the synthetic neuroimages are more consistent with real neuroimages from a diagnostic point of view.

2 Dataset Introduction

ADNI: The Alzheimer's Disease Neuroimaging Initiative (ADNI) study [2] is the most widely open-access study for Alzheimer's Disease (AD) and is jointly funded by the National Institutes of Health (NIH) and industry via the Foundation for the NIH. It is a longitudinal multi-site observational study of elderly individuals with normal cognition, mild cognitive impairment (MCI), or AD. Healthy elderly controls are sampled at 0, 6, 12, 24, and 36 months. Subjects with MCI are sampled at 0, 6, 12, 18, 24, and 36 months. AD subjects are sampled at 0, 6, 12, and 24 months. The follow-up study assesses how well the information (alone or in combination) obtained from MRI, 18F- FDG PET, etc., can measure the disease progression in three groups of elderly subjects mentioned above.

- The ADNI-1 phase was launched in October 2004 and has lasted for 5 years, during which more than 800 subjects have been collected. All subjects are from multiple participating sites in North America (United States and Canada). All subjects were scanned with 1.5 T MRI at each time point, and half subjects were scanned with FDG PET.
- The ADNI-2 phase was launched in September 2011 and has lasted for 5 years. Except for the subjects in ADNI-1, more than 800 additional subjects have been collected during this phase. In ADNI-2, all subjects were scanned with 3T MRI using similar T1-weighted imaging parameters to those used in ADNI-1. About half subjects were scanned with FDG PET using similar parameters to those for ADNI-1.

AIBL: The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL) [3] is a study to discover which biomarkers, cognitive characteristics, and health and lifestyle factors determine subsequent development of symptomatic AD. It is a 4.5-year prospective longitudinal study of cognition, launched in November 2006, which is the largest study of its kind in Australia. This study collected more than 1000 subjects with AD, MCI, and healthy volunteers from two sites (i.e., Perth and Melbourne). In AIBL, MRI was performed at 1.5T/3T MRI with similar T1-weighted imaging parameters to those used in ADNI, but PET was performed with the parameters different from those used in ADNI.

For more details, we listed the statistics of the protocols used for acquiring the baseline MRI and PET scans in Table SII and Table SIII, respectively in the Supplementary Materials. The ADNI organization requires a uniform protocol for data acquirement. As for the AIBL database, the MRI scanning protocols are similar to the one used for ADNI (ADNI-1 and ADNI-2), but the PET scanning protocols are slightly different. For example, the slice thickness of most PET scans in AIBL is either 2.0 mm or 3.0 mm, which is different from that of ADNI PET scans, and the radioisotope for most PET scans in ADNI is F-18 but for 43% PET scans in AIBL is C-11. Meanwhile, most ADNI contributors provide data for both ADNI1 and ADNI2 cohorts. Since similar scanning protocols and the same imaging site may lead to less-diverse imaging quality, the model learned on ADNI-1 can adapt to the data in ADNI-2. To cope with the quality diversity of PET scans in AIBL, we directly used the synthetic PET for this study.

3 Hyper-parameter Discussion

3.1 Influence of Batch Size

Due to the limitation of GPU memory, we cannot use a large batch size. To assess the impact of this hyperparameter on the model's performance, we set different batch sizes and performed the AD *vs.* CN classification task again using different batch sizes. It shows that, when setting the batch size to 1, 2, 3, and 4, the proposed model achieves an AUC of 0.9631, 0.9622, 0.9588, and 0.9596, respectively. The results indicate that the performance of our model is not sensitive to the batch size. Hence, considering the performance and complexity, we empirically set the batch size to 1.

3.2 Influence of Maximum Number of Epochs

In Fig. 6, we reported associated performance of image synthesis obtained after training FGAN different epochs. It shows that the performance has a broad dynamic range and has tolerable changes when the number of epochs approaches to 100. In Fig. S1, we further plotted the training loss and test loss of DSNet during the first 100 epochs. It reveals that the test loss of DSNet become relatively stable after 40 epochs. Therefore, we empirically set the maximum number of epochs to 100 and 40 for FGAN and DSNet, respectively.

TABLE S1
Protocols of baseline MRI scans in ADNI-1, ADNI-2, and AIBL.

Protocol	Dataset	Parameter (Number of Subjects)
Acquisition Plane	ADNI-1	SAGITTAL (844)
	ADNI-2	SAGITTAL (846)
	AIBL	SAGITTAL (665)
Slice Thickness	ADNI-1	1.2 (844)
	ADNI-2	1.2 (846)
	AIBL	1.0 (137), 1.2 (528)
Matrix Z	ADNI-1	146.0 (1), 160.0 (332), 162.0 (1), 166.0 (284), 170.0 (90), 180.0 (124), 184.0 (12)
	ADNI-2	170.0 (166), 176.0 (466), 196.0 (214)
	AIBL	160.0 (527), 170.0 (137), 159.0 (1), 153.0 (1)
Acquisition Type	ADNI-1	3D (844)
	ADNI-2	3D (846)
	AIBL	3D (665)
Manufacturer	ADNI-1	GE Medical Systems (410), Philips Medical Systems (103), SIEMENS (331)
	ADNI-2	GE Medical Systems (214), Philips Medical Systems (165), SIEMENS (466), Philips Healthcare (1)
	AIBL	SIEMENS (665)
Mfg Model	ADNI-1	Achieva (13), Intera (69), Intera Achieva (6), Avanto (64), GENESIS_SIGNA (78), Gyroscan Intera (12), Intera (3), SIGNA EXCITE (307), SIGNA HDx (25), Sonata (99), SonataVision (7), Symphony (161)
	ADNI-2	Achieva (108), Discovery MR750 (90), Discovery MR750w (9), GEMINI (14), Intera (28), SIGNA HDx (9), SIGNA HDxt (9), Skyra (52), TrioTim (279), Verio (135)
	AIBL	Avanto (116), TrioTim (426), Verio (123)
Field Strength	ADNI-1	1.5 (844)
	ADNI-2	3.0 (846)
	AIBL	1.5 (116), 3.0 (549)
Weighting	ADNI-1	T1 (844)
	ADNI-2	T1 (846)
	AIBL	T1 (665)

TABLE S2
Protocols of baseline PET scans in ADNI-1, ADNI-2, and AIBL.

Protocol	Dataset	Parameter (Number of Subjects)
Slice Thickness	ADNI-1	1.2 (51), 2.0 (71), 2.4 (110), 3.3 (21), 3.4 (53), 4.3 (93)
	ADNI-2	1.2 (41), 2.0 (254), 2.4 (157), 3.3 (156), 3.4 (20), 4.3 (45)
	AIBL	2.0 (355), 3.0 (142), 3.3 (109)
Manufacturer	ADNI-1	CPS (59), GE MEDICAL SYSTMS (67), GEMS (47), Philips Medical Systems (39), Siemens ECAT (51) Siemens/CTI (136)
	ADNI-2	CPS (53), GE MEDICAL SYSTMS (184), GEMS (17), Philips Medical Systems (91), Siemens (123), Siemens ECAT (41), Siemens/CTI (164)
	AIBL	GE MEDICAL SYSTMS (109), SIEMENS (142), Philips Medical Systems (355)
Mfg Model	ADNI-1	ACCEL (17), Advance (47), Allegro Body (C) (10), Discovery HR (5), Discovery LS (46), Discovery RX (3), Discovery ST (13), EXACT (ACS 1) (3), EXACT (ACS 2) (6), G-PET Brain (C), Gemini TF(C) (4), Guardian Body(C) (17), HR+ (110), HRRT (51), LSO PET/CT (5), LSO PET/CT (Pico electronics) (22), LSO PET/CT HI-REZ (32)
	ADNI-2	1093 (14), 1094 (51), ACCEL (7), Advance (17), Allegro Body (C) (3), Biograph64 (22), Discovery 600 (6), Discovery LS (28), Discovery RX (11), Discovery ST (45), Discovery STE (94), GEMINI TF Big Bore (14) , GEMINI TF TOF 16 (30) , GEMINI TF TOF 64 (20), Guardian Body(C) (12), HR+ (157), HRRT (41), LSO PET/CT (Pico electronics) (13), LSO PET/CT HI-REZ (72), SOMATOM Definition AS_mCT (4)
	AIBL	Allegro Body (C) (353), Biograph128 (123), Biograph128_mCT (19), Discovery 710 (109), GEMINI TF TOF 64 (2)
Radioisotope	ADNI-1	C-11 (8), F-18 (391)
	ADNI-2	F-18 (673)
	AIBL	C-11 (262), F-18 (344)
Radio Pharmaceutical	ADNI-1	11C-PIB (8), 18F-FDG (391)
	ADNI-2	18F-AV45 (284), 18F-FDG (389)
	AIBL	11C-PIB (61), 18F-Flutemetamol (48), Flutemetamol (142), Other (355)
Frames	ADNI-1	1.0 (70), 4.0 (2), 5.0 (1), 6.0 (252), 7.0 (40), 12.0 (1), 15.0 (2), 17.0 (1), 27.0 (2), 28.0 (1), 30.0 (1), 33.0 (19), 38.0 (1), 39.0 (2)
	ADNI-2	1.0 (83), 2.0 (2), 4.0 (247), 6.0 (339), 16.0 (2)
	AIBL	1.0 (498), 4.0 (49), 6.0 (59)

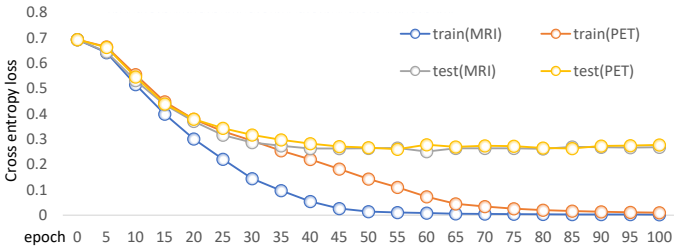

Fig. S1. Training and testing losses for image classification versus different numbers of training epochs.

3.3 Influence of Generative Model Backbone

In general, there are two alternative backbone structures, i.e., Decoder-Encoder (DE) backbone and UNet backbone [4], available for the generative components. The structures of DE and UNet were displayed in Fig. S2 (left), where the major differences are the skip connections and feature concatenation used in UNet. Fig. S2 (right) shows the performance metrics of AD $vs.$ CN classification obtained by using either the MRI scan or PET scan synthesized by either DE backbone or UNet backbone. It reveals that these two structures achieve similar results, e.g., the AUC values of

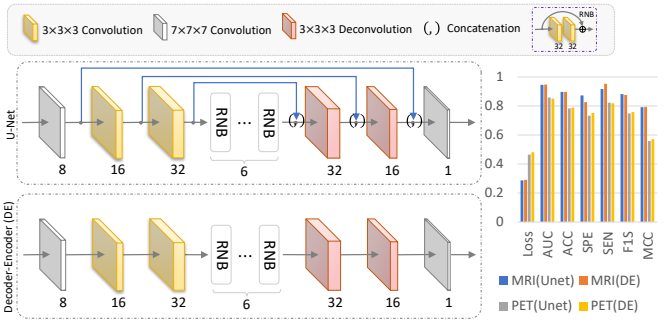

Fig. S2. The structure (left) of Decoder-Encoder (DE) and UNet backbone are displayed in the left while the metric values on image classification of scans synthesized by DE and UNet for MRI and PET modality, respectively.

MRI scans generated by UNet backbone and DE backbone are 0.9459 and0.9475, respectively. The reason may lie in the fact that the spatial consistency between a pair of MRI and PET scans are not strictly held. Following the principle of Occam's razor, we use the DE backbone, which is simpler than the UNet backbone, for this study.

4 Advantages of Synthetic Data

Compared to "unseen-modality classification methods", namely, using different networks for feature extraction and a shared network for classification, our model has the following three major advantages.

- For each incomplete case, the information of the missing modality is transformed not only from the other modality, but also from the complete cases in the training set via training the image synthesize model. For instance, the missing PET image of an incomplete AD subject is synthesized based on both the MRI image of that subject and the PET images of AD subjects in the training set, since FGAN used for image synthesis was trained on both MRI and PET images in the training set. Therefore, even if a disease specificity has not become evidenced on the structural MRI image, our model can still "guess" it according to the PET images of AD cases in the training set, given that the image synthesis and classification are performed in a unified framework in our solution.
- Suppose our model, in the worst case, cannot benefit from the complete cases in the training set at all. From the information point of view, our model uses the fake "multi-modality" information, which is only the information of the available modality. In this case, our model becomes a "real unseen-modality classification method".
- The proposed multi-modality classification model is much simpler than the "real unseen-modality classification" model, in which one or two of the feature extraction branches is active in each training epoch. Alternatively, if training independently two models for feature extraction and another model for classification, the system cannot benefit from the unique advantage of "learning image representation and classification in a unified framework for simultaneous optimization".

5 Advantages of Cascade Model

In our experiments, we followed the cascade strategy to train DSNet and FGAN components rather than training then together in an end-to-end strategy. There are four major reasons for using the cascade model, instead of an end-to-end one.

- *First*, the feature-consistency constraint is defined in the feature extraction part, and hence is varying during training DSNet. If jointly training DSNet and FGAN in an end-to-end model, the less-optimal feature-s obtained in the early training stage will undermine the convergence of FGAN.
- *Second*, we design the feature-consistency constraint to capture the disease-specific information in real data and then use the information to guide FGAN to synthesize real-like scans. Thus, it derives the information only from existing real data and is independent to training FGAN. Therefore, jointly or step-wise iteratively training DSNet and FGAN will not capture more disease-specific information, but requires more GPU memory and computational resources.
- *Third*, we need a feature extraction module to measure the effect of feature-consistency constraint on FGAN. Thus, we utilize the well-trained DSNet with freezing weights while training FGAN.
- *Finally*, as a classification model, DSNet can hardly converge synchronously with FGAN. Hence, if we jointly train DSNet and FGAN in an end-to-end way, the asynchronous convergence of both components will lead one component to under-fitting or the other component to over-fitting.

6 More disease-related Visual Examples

Besides samples in Fig. 4 in the main text, we supplied more views of high-resolution examples in Figs. S3-S5, where four typical subjects (Roster IDs: 4386, 4765, 4997, and 4417) in ADNI-2 are shown. Basically, it can be seen from Fig. S3 (sagittal MRI views) that the sizes of ventricle in the 4^{th}, 5^{th}, 6^{th} columns are more like the ground truth (7^{th} column) than the 1^{st}, 2^{nd}, and 3^{rd} columns. It suggests the feature-consistency constraint can coexist with other consistency constraint, with minimal impact of voxel-wise-consistency and cycle-consistency constraints.

Taking into account the results listed in Table 2 (main text), the conclusion can be supported that the feature-consistency constraint can help preserve more diagnosis information during the transformation between two modalities without dropping the visual quality. However, there may be no metric that can cover all concerned aspects. Therefore, we still suggest considering the suitable choice of constraints for a specific task, e.g., using the adversarial loss to keep the distribution (texture, structure) similarity, using the pixel-wise-consistency constraint to keep intensity consistency, and using the feature-consistency constraint to keep diagnosis consistency.

7 Application in Another Scenario

We applied the proposed DSDL framework to a natural image classification scenario: using grayscale images to

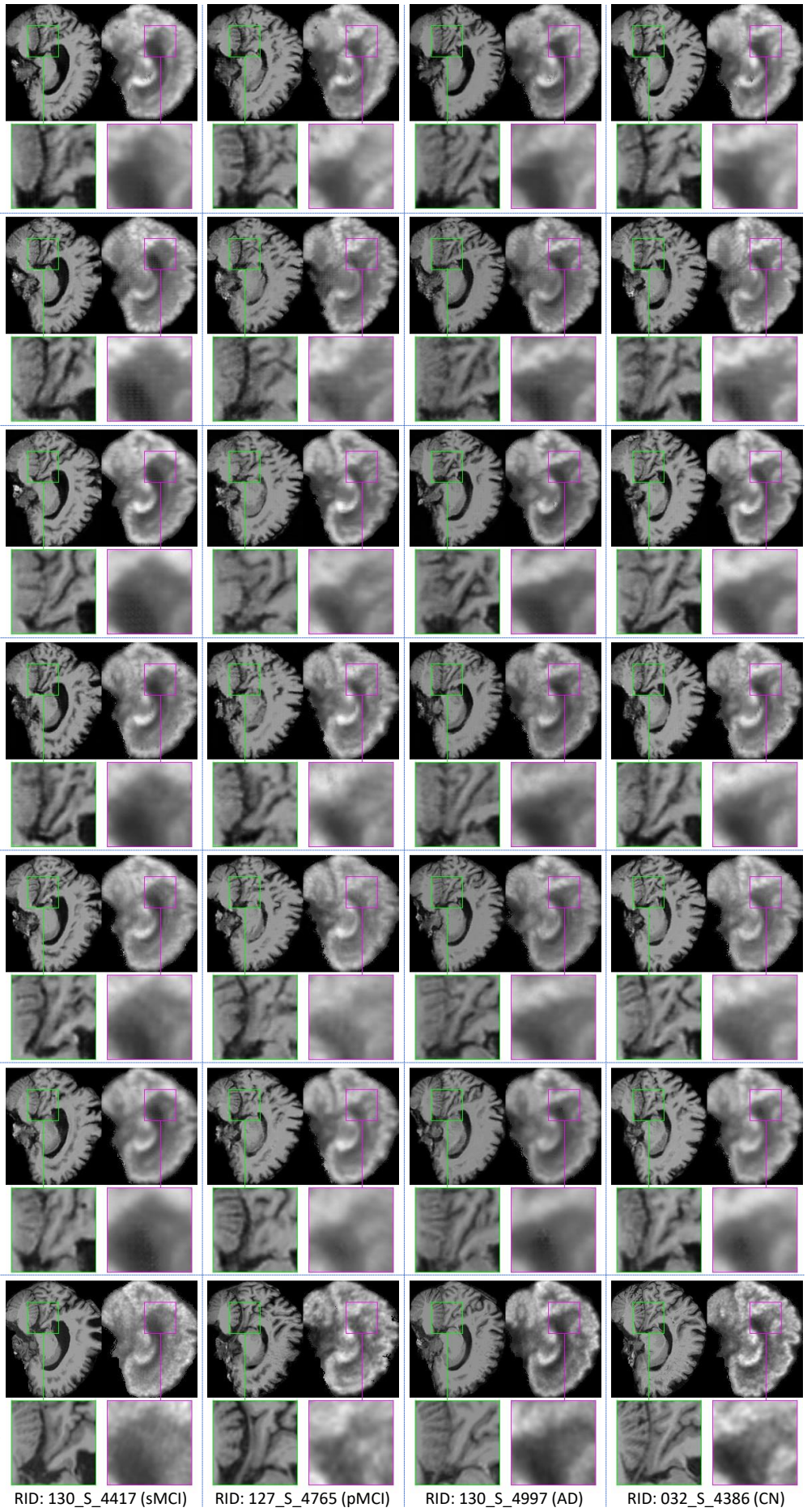

Fig. S3. Sagittal views of PET and MRI scans synthesized by six methods for four typical subjects (Roster ID: 4386, 4765, 4997, and 4417) in ADNI-2, along with their corresponding ground-truth images. All six image synthesis models are trained on ADNI-1.

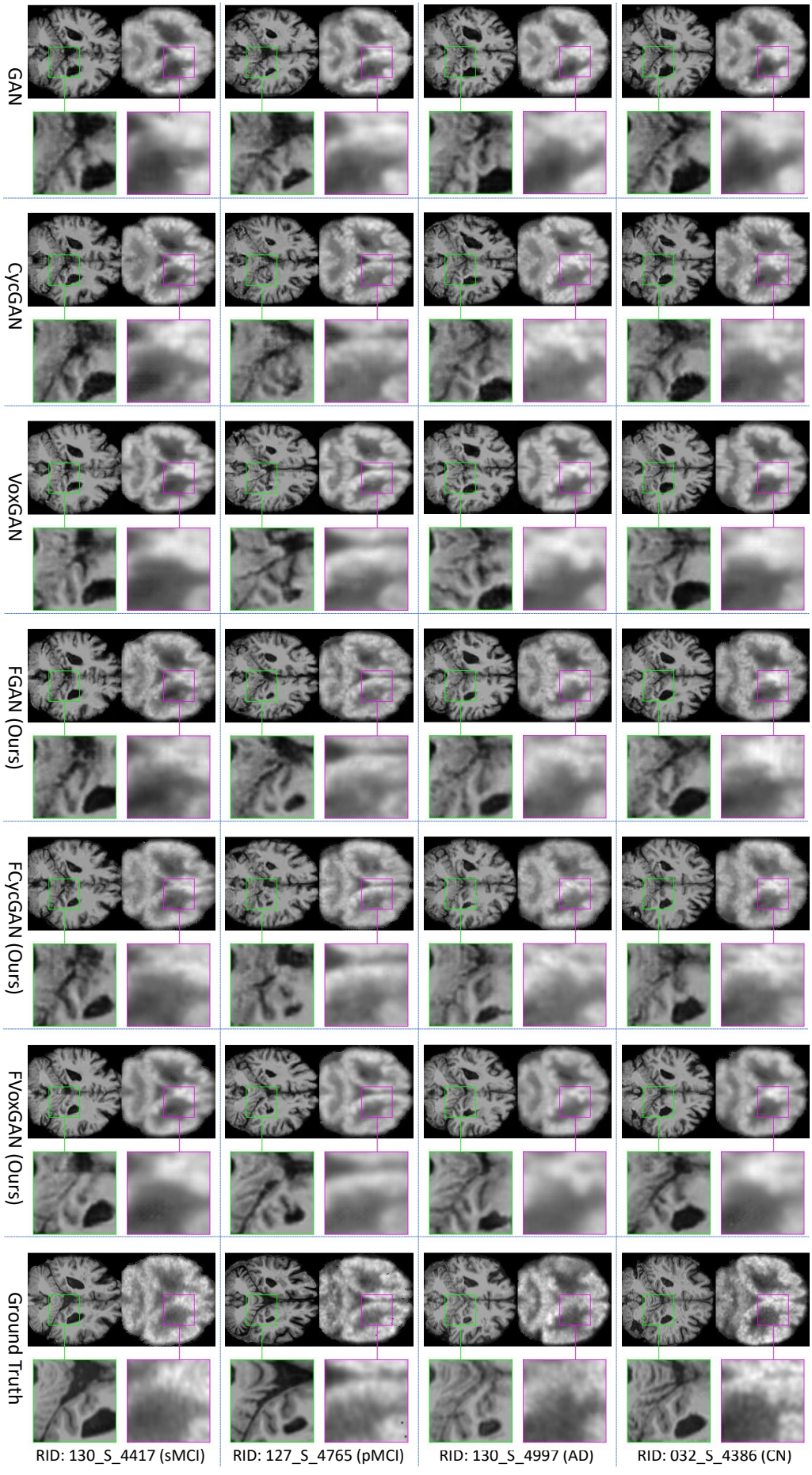

Fig. S4. Coronal views of PET and MRI scans synthesized by six methods for four typical subjects (Roster ID: 4386, 4765, 4997, and 4417) in ADNI-2, along with their corresponding ground-truth images. All six image synthesis models are trained on ADNI-1.

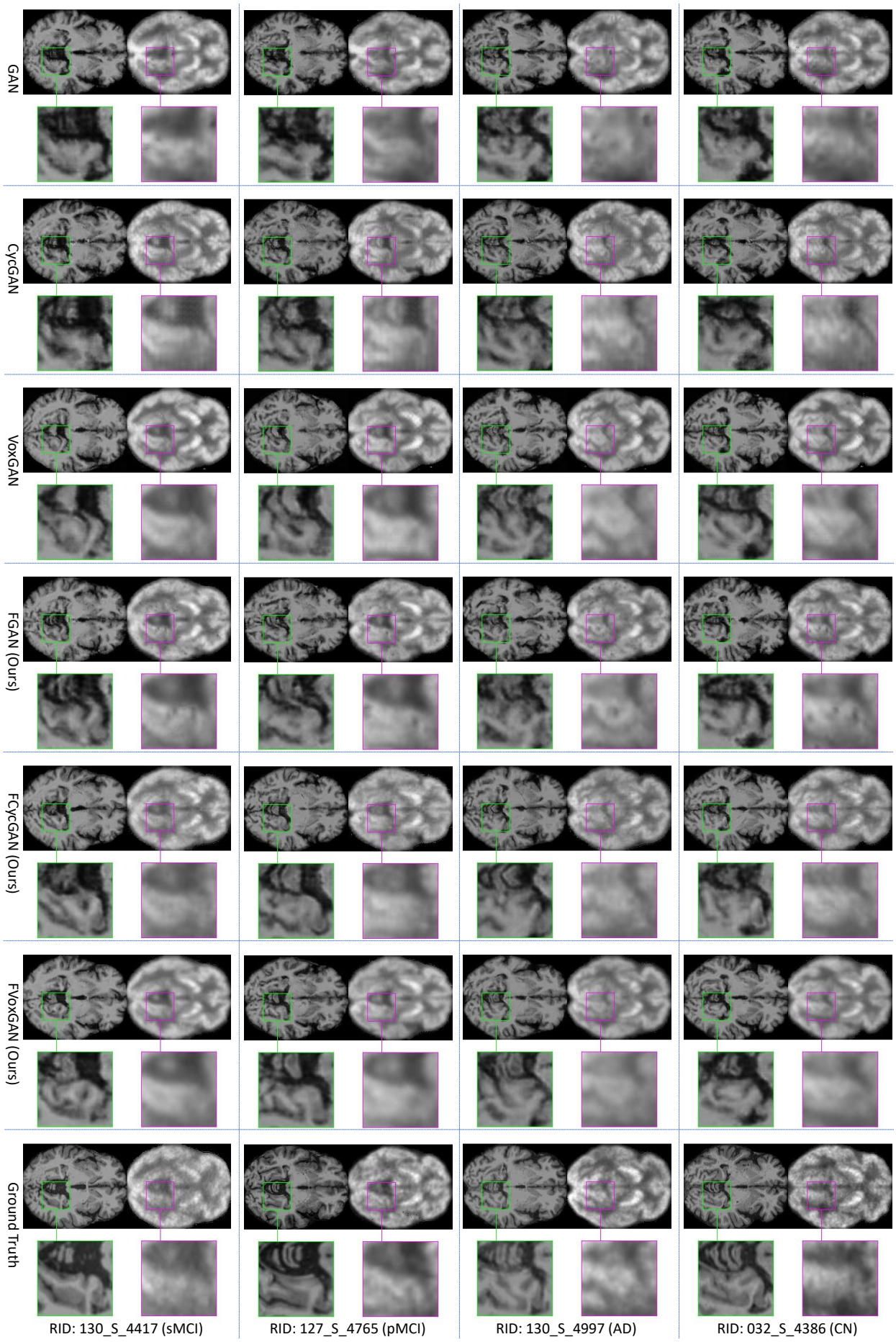

Fig. S5. Axial views of PET and MRI scans synthesized by six methods for four typical subjects (Roster ID: 4386, 4765, 4997, and 4417) in ADNI-2, along with their corresponding ground-truth images. All six image synthesis models are trained on ADNI-1.

generate the unknown color images [4] and then jointly using both grayscale and color images for classification. Two benchmark datasets for fine-grained classification were considered. The Oxford Flower-17 (F17) [5] dataset contains 17 flower species with 80 color images per species. The Oxford Pet-37 (P37) [6] dataset contains $7,349$ color images of 12 kinds of cats and 25 kinds of dogs with roughly 200 images per class. In both datasets, the images suffer from large variations in scale, pose, viewpoint angle and illumination. We randomly selected 75% images per category for training and used the rest for test.

To adapt our method to this problem, we replace 3D (de)convolutional kernels to 2D kernels and set the input size to 224×224. Since these images have no rigid structure consistency, the effect of spatial cosine kernel is suppressed. Noted that, although our DSNet may not be the best choice for 2D image classification, it is useful for verifying the effectiveness of the proposed feature-consistency constraint. The quality of synthetic color images was measured by the mean absolute error (MAE), mean square error (MSE), structural similarity index measure (SSIM), and peak signal-to-noise ratio (PSNR), and the performance of image classification was measure by the area under receiver operating characteristic (AUC), accuracy (ACC), average precision score (APS), and F1-Score (F1S).

The experiments include four stages. In the *first* stage, we trained two DSNets on color images and grayscale images, respectively, and reported the classification performance of each DSNet on the test set in Table S3. In the *second* stage, we trained different image generative models to transfer grayscale images to color images and reported the quality of synthetic color images in Table S4. In the *third* stage, the synthetic color images generated in the second stage were fed to the DSNet trained on real color images in the first stage, and the classification performance was reported in Table S5. In the *fourth* stage, the predicted scores achieved by DSNets on grayscale images and synthetic color images were simply averaged to mimic the multi-modality data. The corresponding classification performance was reported in Table S6. Besides, we show several samples from the F17 dataset and P37 dataset and the corresponding synthetic color images obtained by PixGAN, FGAN, and FPixGAN in Fig. S6 and Fig. S7. The following four conclusions can be drawn from these results.

$First$, the classification performance achieved by using grayscale images is obviously lower than that achieved by using color images on both datasets (see Table S3). Thus, it is possible to boost the performance of grayscale image classification as long as we can use grayscale images to generate the missing color images reasonably.

$Second$, the experimental evidence provided in [7] demonstrates that (1) the distribution matching constraints used in GANs may not be able to preserve discriminative information for either unpaired or paired data translation, leading to mis-diagnosis of medical conditions, and (2) using the l_1 (MAE) loss, equivalent to a pixel-wise-consistency constraint, seems to be helpful when the image quality metric is MAE, which matches the l_1 loss rather than measuring the classification performance. To evaluate how much discriminative information is preserved by each generative model, we gave the performance of using only the discriminative loss (GAN-d), feature-consistency loss (GAN-f), and pixel-wise consistency loss (GAN-p) in the 1^{rd} - 3^{th} rows of Table S4 and Table S5, respectively. It shows that GAN-f achieves the best image classification performance, GAN-p achieves best quality of synthetic images, and GAN-d, which may be good at distribution matching, achieves lower performance than GAN-f and GAN-p in color image generation and classification. This conclusion is consistent with the conclusion that distribution matching can hardly preserve discriminative information [7]. Therefore, we suggest considering a suitable constraint for each specific task, e.g., using the adversarial loss to keep the distribution (texture / structure) similarity, using the pixel-wise-consistency constraint to keep intensity consistency, and using the proposed feature-consistency constraint to keep classification consistency.

$Third$, jointly using different constraints may lead to balanced performance. PixGAN jointly uses the adversarial loss pixel-wise consistency loss, FGAN jointly uses the adversarial loss and feature-consistency loss, and FPixGAN jointly uses all three losses. The performance of PixGAN, FGAN, and FPixGAN was displayed in the 4^{rd} - 6^{th} rows of Table S4 and Table S5, respectively. Some samples from the F17 dataset and P37 dataset and the corresponding synthetic color images obtained by PixGAN, FGAN, and FPixGAN were illustrated in Fig. S6 and Fig. S7. It reveals that FGAN outperforms PixGAN in terms of image classification, but underperforms it in terms of image synthesis. FpixGAN achieves similar quality of synthetic images to PixGAN and similar classification performance to FGAN. It demonstrates again that the feature-consistency constraint is effective to preserve discriminative information. Moreover, multiple constraints can be jointly used to balanced performance in terms of both image generation and classification.

$Fourth$, combining the grayscale images with the synthetic color images to form pseudo multi-modality images and using them to perform the classification task may lead to improved performance (see Table S6). Especially, the performance of classifying pseudo multi-modality data, in which the missing color images were generated by the GANs with the feature-consistency constraint (e.g., FGAN and FPixGAN), is even compatible to that of classifying real color images on the F17 dataset. It suggests that the proposed feature-consistency constraint can be successfully applied to the transform of grayscale images to color images for classification purpose.

REFERENCES

[1] Y. Pan, M. Liu, C. Lian, T. Zhou, Y. Xia, and D. Shen, "Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 455–463.

[2] C. R. J. Jr., M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell *et al.*, "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," *J. Magn. Reson. Imag.*, vol. 27, no. 4, pp. 685–691, 2008.

[3] K. A. Ellis, A. I. Bush, D. Darby, D. De Fazio, J. Foster, P. Hudson, N. T. Lautenschlager, N. Lenzo, R. N. Martins, P. Maruff *et al.*, "The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease," *International Psychogeriatrics*, vol. 21, no. 4, pp. 672–687, 2009.

Fig. S6. Some samples from the Oxford Flower-17 dataset. The top and bottom rows are the paired grayscale and color images, while the 2^{nd}-4^{th} rows are the synthesized images generated by PixGAN, FGAN, and FPixGAN, respectively.

Fig. S7. Some samples from and Oxford Pet-37 dataset. The top and bottom rows are the paired grayscale and color images, while the 2^{nd}-4^{th} rows are the synthesized images generated by PixGAN, FGAN, and FPixGAN, respectively.

TABLE S3
Classification results achieved by DSNet on grayscale images and color images

Data Modality	Oxford Flower-17				Oxford Pet-37			
	AUC	APS	ACC	F1S	AUC	APS	ACC	F1S
Color	99.39	93.62	87.65	87.65	96.04	63.05	58.87	58.86
Gray	98.81	89.50	83.24	83.35	96.15	59.29	53.86	53.98

TABLE S4
Image quality of synthesized color images generated by six different GANs from grayscale images

Synthesis Model	Oxford Flower-17				Oxford Pet-37			
	MAE	MSE	SSIM	PSNR	MAE	MSE	SSIM	PSNR
GAN-d	74.69	92.42	2.38	6.73	60.25	73.18	9.99	8.77
GAN-f	21.69	25.59	63.83	18.59	20.68	23.89	61.51	19.34
GAN-p	10.01	16.09	82.20	21.89	11.47	15.30	74.85	23.08
PixGAN	11.48	17.89	77.18	20.97	11.10	15.02	75.75	23.27
FGAN	20.31	26.33	57.36	17.61	20.10	23.29	62.41	19.64
FPixGAN	12.95	18.48	77.34	20.69	11.71	15.35	74.86	23.05

TABLE S5
Classification results of DSNet on those synthesized color images generated by six different GANs from grayscale images

Synthesis Model	Oxford Flower-17				Oxford Pet-37			
	AUC	APS	ACC	F1S	AUC	APS	ACC	F1S
GAN-d	49.33	7.56	5.59	1.68	56.84	3.99	3.65	1.22
GAN-f	98.26	87.93	81.76	81.75	93.80	50.85	48.42	48.44
GAN-p	96.54	80.57	72.65	72.34	91.13	37.30	35.47	33.71
PixGAN	95.26	75.92	68.53	68.40	90.99	35.56	32.05	28.49
FGAN	98.14	89.06	82.94	82.97	94.05	52.14	48.48	48.38
FPixGAN	97.19	84.93	77.94	77.81	93.74	48.34	46.25	45.88

TABLE S6
Classification results while averaging the predicted scores of each synthesized color image and its corresponding grayscale image

Synthesis Model	Oxford Flower-17				Oxford Pet-37			
	AUC	APS	ACC	F1S	AUC	APS	ACC	F1S
GAN-d	97.72	87.23	64.41	69.31	93.96	54.86	42.00	46.80
GAN-f	99.24	92.29	87.94	87.88	96.25	61.75	55.82	55.91
GAN-p	98.91	90.83	83.82	83.68	95.60	56.83	50.92	50.84
FGAN	99.28	93.02	88.53	88.48	96.30	62.32	56.93	55.89
PixGAN	98.65	88.71	81.47	81.28	95.57	56.47	49.62	49.41
FPixGAN	98.93	90.96	86.18	86.02	96.17	61.04	56.10	56.06

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1125–1134.
[5] M. . Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, 2006, pp. 1447–1454.
[6] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 3498–3505.
[7] J. P. Cohen, M. Luck, and S. Honari, "Distribution matching losses can hallucinate features in medical image translation," in *Proc. Int. Conf. Med. Image Comput. Computer Assisted Intervention*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 529–536.