Multi-modal data generation with a deep metric variational autoencoder

Josefine Vilsbøll Sundgaard*[1], Morten Rieger Hannemose[1], Søren Laugesen[2], Peter Bray[3], James Harte[2], Yosuke Kamide[4], Chiemi Tanaka[5], Rasmus R. Paulsen[1], and Anders Nymark Christensen[1]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[2]Interacoustics Research Unit, c/o Technical University of Denmark, Denmark
[3]Interacoustics A/S, Middelfart, Denmark
[4]Kamide ENT clinic, Shizuoka, Japan
[5]Diatec Japan, Kanagawa, Japan

Abstract

We present a deep metric variational autoencoder for multi-modal data generation. The variational autoencoder employs triplet loss in the latent space, which allows for conditional data generation by sampling new embeddings in the latent space within each class cluster. The approach is evaluated on a multi-modal dataset consisting of otoscopy images of the tympanic membrane with corresponding wideband tympanometry measurements. The modalities in this dataset are correlated, as they represent different aspects of the state of the middle ear, but they do not present a direct pixel-to-pixel correlation. The approach shows promising results for the conditional generation of pairs of images and tympanograms, and will allow for efficient data augmentation of data from multi-modal sources.

1 Introduction

Deep generative models are able to generate new data within the distribution of the training dataset, and can be used for advanced data augmentation in cases where data are costly to annotate or difficult to acquire [19]. A widely used generative model is the variational autoencoder (VAE) [13]. The VAE is a probabilistic model consisting of an encoder that learns an approximation of the posterior distribution of the data and a decoder that learns to reconstruct the original input from a latent representation. An advantage of VAEs over generative adversarial networks (GANs) [4] is that the VAE learns a smooth latent representation of the input data [3]. The latent space can therefore be used for sampling new latent representations and thus generate new examples from the distribution of the training dataset using the VAE decoder.

Conditional data generation, e.g., the conditional VAE [14], allows us to specify which class in the dataset to generate new data from. Here, both the latent representations and the input data are conditioned by, e.g., class label. Instead of conditioning the model for class specific data generation, Karaletsos et al. [11] proposed the triplet-loss based VAE for generation of interpretable latent representations that separate the classes in the latent space with deep metric learning. Karaletsos et al. [11] put their main focus on learning the latent representations, whereas we are interested in using the triplet-loss-based VAE for data generation. We expand the approach to include estimation of the class distributions in the latent space, and generation of new conditional examples. Furthermore, we propose a multi-modal network structure with a common latent space, which allows for the generation of new paired examples from multiple modalities.

*Corresponding Author: josh@dtu.dk

https://doi.org/10.7557/18.6803

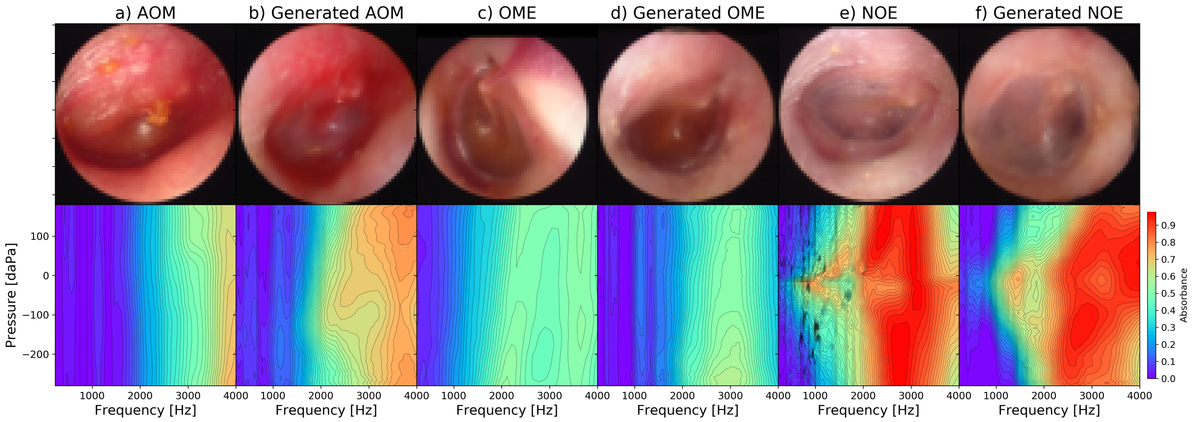

Figure 1: Examples from the dataset and generated examples: otoscopy images (top) and WBT measurements (bottom). Acute otitis media (left two images), otitis media with effusion (middle two images), no effusion (right two images).

The multi-modal dataset consists of pairs of otoscopy images of the tympanic membrane and wideband tympanometry (WBT) measurements, examples of which are presented in Figure 1. The two types of data are very different, as the first is an image acquired with a camera, and the other is the result of an acoustic measurement. Furthermore, they reflect different aspects of the state of the middle ear. The otoscopy image gives a visual impression of the tympanic membrane, which can show signs of e.g. infection or effusion, while the WBT measurement provides quantitative indications about the presence of fluid in the middle ear, the mobility of the tympanic-ossicular system, and the volume of the external auditory canal. The two types of data are therefore correlated but do not have a direct pixel-to-pixel relation.

Several studies have developed different approaches for otitis media classification based on either otoscopy images [18, 20, 16] or WBT measurements [5, 22, 21]. Binol et al. proposed a combined deep learning classification approach based on standard single-frequency tympanograms and otoscopy images [1]. Otitis media can be separated into two main diagnostic groups: acute otitis media (AOM) and otitis media with effusion (OME). Figure 1 shows the difference between these two diagnostic groups, where AOM is an acute infection with redness and a bulging eardrum, and OME is a build-up of fluid in the middle ear. An example of a normal eardrum with no effusion (NOE) is also shown. The WBT measurements in Figure 1 show how the absorbance across the pressure axis does not change in AOM or OME measurements, while the NOE measurements typically show a general increase in absorbance around 0 daPa, compared to negative or positive relative pressures. Furthermore, the general absorbance level at lower frequencies is lower for AOM and OME, than for NOE measurements. Both types of data can be used for the diagnosis of otitis media.

The aim of this paper is to generate new pairs of otoscopy images and WBT measurements from each of the three diagnostic groups: AOM, OME, and NOE, and for this task, we propose the multi-modal triplet VAE. The generated otoscopy image and WBT pairs can be used as advanced data augmentation for a multi-modal classification pipeline. Our multi-modal generative model can also be used in other domains such as pairs of cardiac images and electrocardiograms, or brain scans and electroencephalograms. These modalities have a correlation, while reflecting different aspects - visual and functional - of the condition of the examined organ. This work can also be used for the training of doctors and models while preserving patient privacy. Generated data ensures anonymity and allows for data to be shared without regulations such as EU's GDPR, and some studies have already shown the usability of variational autoencoders in this field [15, 19].

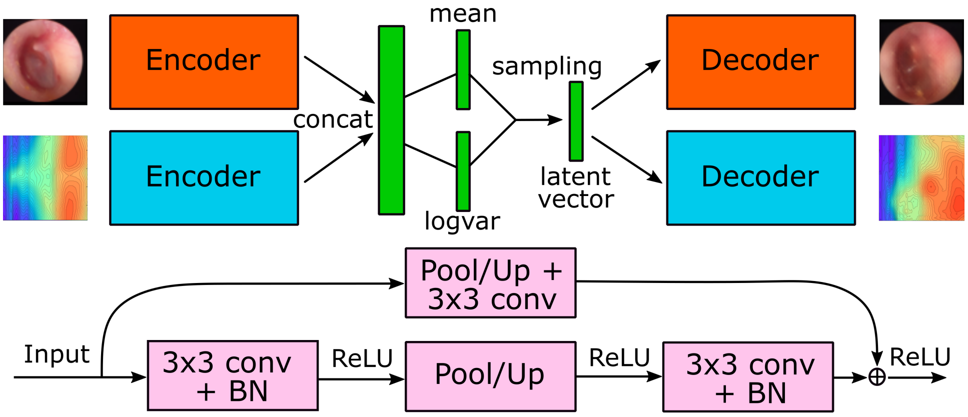

Figure 2: Structure of the multi-modal triplet VAE. Top figure shows the overall structure with two encoders, concatenation of the outputs, sampling, and two decoders. The bottom figure shows the residual blocks used in both encoders and decoders. BN refers to batch normalization.

2 Methods

The multi-modal triplet VAE consists of two encoders and two decoders - one for each modality, and the structure is shown in Figure 2 together with the structure of the upsampling and downsampling blocks used to construct the encoders and decoders. The architecture was inspired by Hou et al. [9] with several modifications, including residual connections [6] and changes in kernel size of the convolutional layers. The encoders define the approximate joint posterior distribution $q(z|x_1, x_2)$, where x_1 represents the otoscopy image input and x_2 represents the WBT input. Each encoder network consists of five downsampling blocks using 2D average pooling, and take the $64 \times 64 \times 3$ otoscopy images and the $64 \times 64 \times 1$ WBT measurements as input. The architecture starts with 64 feature maps in the first block, and the number of features is doubled in each consecutive block, while the size of the feature maps is halved. The output feature maps from each encoder ($2 \times 2 \times 512$) are concatenated, and two 2×2 convolutional layers are used to obtain the mean and variance in the 128-dimensional latent space, z. The size of the latent space was chosen based on the experiments presented in the papers by Schroff et al. [17] and Hermans et al. [8]. Because the encoder outputs are concatenated, we achieve a joint latent space, which allows for sampling in the latent space to generate new paired examples of both modalities. The decoders will thus receive information from both the image and WBT for the reconstruction of each modality. Using the reparameterization trick [14], a latent vector is sampled from the joint posterior distribution $q(z|x_1, x_2)$ and passed to both decoders. The decoders, $p(\bar{x}_1|z)$ and $p(\bar{x}_2|z)$, reconstruct the inputs ($\bar{x}_1$ representing the reconstructed otoscopy image and $\bar{x}_2$ the reconstructed WBT measurement) provided the latent vector. The decoder networks consist of six residual upsampling blocks using nearest-neighbor upsampling, where the number of features is halved for each block starting at 512 and the feature maps size is doubled. The final layer is a single 3×3 convolutional layer that goes from 32 feature maps to the desired number of output channels, one channel for WBT measurements and three for otoscopy images.

The training loss function consists of several parts. The difference between reconstructed WBT and input WBT is penalized using binary cross entropy (BCE) loss. The reconstruction of the image is evaluated using structured similarity index (SSIM) loss [24], which is a local measurement that compares the reconstruction and original image based on luminance, contrast, and structural information. In the latent space, both Kullback–Leibler (KL) divergence and triplet loss [17] are computed. The KL divergence forces the latent embeddings close to a standard normal distribution, while the triplet loss forces examples from the same class to cluster together and pushes examples from different classes further apart [17]. The loss function terms related to the embedding space are weighted lower than the rest of the terms, and the value 0.1 was experimentally chosen, leading to a loss function defined as:

$$Loss = L_{SSIM} + L_{BCE} + 0.1 \cdot (L_{KL} + L_{triplet}) \quad (1)$$

Balanced sampling is performed during training, with a batch size of 60 (20 pairs from each class) to ensure a balanced representation of every class in each training batch and to cope with the class imbalance in the dataset. The triplets are sampled in each batch using semi-hard mining [17] based on the encoder-generated mean vector from each input pair. The VAE is trained for 5000 epochs using the Adam optimizer [12] with a learning rate of 0.0004. Data augmentation is performed using random erasing [26] on both image and WBT measurement, while horizontal flipping and rotation with ±20 degrees are also performed on the images.

The individual distributions of each of the three classes in the latent space are not necessarily equal to the prior distribution $p(z) = N(0, 1)$, since the model is trained with both KL divergence loss and triplet loss to regulate the latent space. For generation of new image and WBT pairs, we thus need to estimate the posterior distribution of each of the three classes. Once the network is trained, the test set is passed through the encoders, obtaining latent representations of each image and WBT pair in the test set. The distribution of each class in the latent space is approximated using kernel density estimation for each class. Gaussian kernel density estimation estimates the probability density function in the latent space by placing a Gaussian kernel on each sample. The bandwidth of the kernel is fine-tuned using five-fold cross-validation. Kernel density estimation is performed only on the test set, estimating the joint posterior distribution, $q(z|x_1, x_2)$, of each class. New latent vectors, z can then be sampled within each of the estimated class distributions, which are run through the decoders, , $p(\bar{x}_1|z)$ and $p(\bar{x}_2|z)$, generating new pairs of images and WBT measurements.

2.1 Data

The dataset consists of 1420 pairs of images and WBT measurements collected at Kamide ENT clinic, Shizouka, Japan, from patients aged between 2 months and 12 years. Each pair was assigned one of the three classes: NOE (537 pairs), OME (419 pairs), and AOM (211 pairs) by an experienced ENT specialist based on signs, symptoms, patient history, otoscopy examination, and WBT measurements. The data was collected and handled under the ethical approval from the Non-Profit Organization MINS Institutional Review Board (reference number 190221), with either opt-out consent, or informed consent from their parent or guardian.

An otoscopy image is captured using an endoscope (dedicated video otoscope) inserted into the ear canal, allowing a visual inspection of the tympanic membrane. The original image size was 640×480 pixels, which was cropped and downsampled to 64×64 to fit the proposed architecture. A WBT measurement is performed by inserting and hermetically sealing an acoustic probe with an appropriately sized silicone ear tip into the patient's ear canal. The probe repeatedly presents a transient stimulus with a frequency range encompassing 226 Hz to 8 kHz, while modifying the pressure in the external acoustic canal relative to the ambient pressure from 200 to -300 daPa [7]. The measurements were performed using the Titan system (Interacoustics, Denmark). From the WBT measurement, it is possible to derive conclusions about both tympanic membrane mobility and the condition of the middle ear, and thus additional diagnostic power can be gained over visual inspection alone. WBT measurements were bilinearly resampled to a common grid from 180 daPa to -280 daPa in 64 steps on a linear scale for the pressure axis, and from 226 Hz to 4 kHz in 64 steps for the frequency axis. Examples of both images and WBT measurements are shown in Figure 1.

The dataset is split into a train (80%) and test (20%) set. It was ensured that data from one patient was only used for either training or testing, to prevent data leakage.

3 Results

The test embeddings are shown in Figure 3. The 128-dimensional latent representation of each image has been reduced to two dimensions using t-SNE dimensionality reduction [23] to visualize the latent space. The test embeddings clearly show

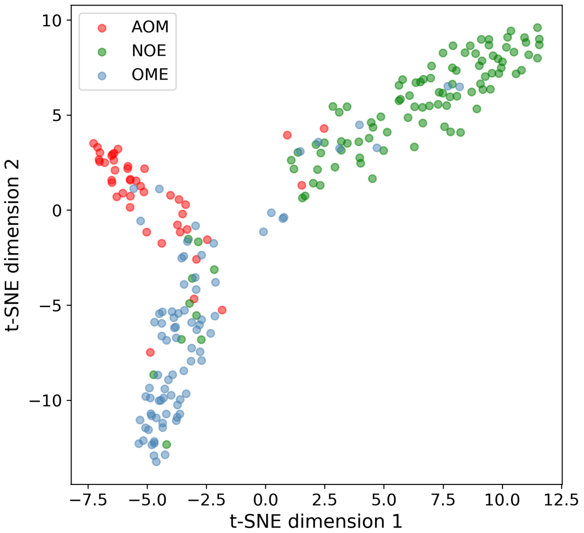

Figure 3: t-SNE visualization of test data latent embeddings.

three clusters, but they blend in the transition areas between the classes, as the images and WBT measurements can look quite similar across the diagnostic groups. Some of the overlap could also arise from the drastic dimensionality reduction from 128 to two dimensions. It is not possible to fully represent and visually inspect a 128-dimensional embedding space in a 2D plot, but t-SNE attempts to preserve the topology neighbourhood structure of the data. The clusters would therefore likely more separable in the high-dimensional space, as some information is lost in the dimensionality reduction [23].

New latent representations are sampled in the full 128-dimensional space within the three-class distributions estimated with kernel density estimation, and examples of generated otoscopy images and WBT measurements are plotted in Figures 4, 5, and 6. Figure 4 shows examples of generated images in the three diagnostic groups. The images look realistic, as they all contain a tympanic membrane, clear diagnostic markers, and the malleus bone is seen in several examples. The top row of AOM images shows signs of redness and bulging eardrum, and the OME cases clearly have effusion behind the eardrum. The NOE cases appear pale and translucent, as expected.

Other examples of generated pairs of otoscopy images and WBT measurements are shown side by side with original examples from the dataset in Figure 1. These are not reconstructions, but new generated images. In this figure, it is possible to compare the diagnostic markers of the conditions across modalities, while also comparing the generated examples with original examples. Figure 1 a) and b) show similar signs of AOM redness and infection and reduced absorbance in the WBT, which is relatively flat across the pressure axis. The two OME cases in Figure 1 c) and d) show very similar diagnostic signs on both the original and generated data with yellow effusion behind the tympanic membrane. Likewise, the absorbance is much lower with very little variation across pressures. The NOE cases in Figure 1 e) and f) show normal tympanic membranes and high absorbance in the WBT with a change across pressures.

The generation of WBT measurements is summarized in Figure 5, where generated examples are shown together with the average WBT of the generated samples as well as the original dataset for each of the three diagnostic groups. The average of the generated samples is computed from 500 samples in each diagnostic group. The two average WBT measurements look very similar. This shows that the generated WBT measurements within each diagnostic group follow the same pattern as the mean of the original dataset, thus the distribution of the classes has been captured quite well. The generated examples also indicate great variation within each class.

There is, of course, great variability in the appearance of both otoscopy images and WBT measurements, depending on the severity of symptoms. The generated data show the same range from mild to severe symptoms, and the generated pairs show that the two modalities are representing the same disease severity. This is shown in Figure 6. Figure 6 a) and c) show mild cases of AOM and OME, where the respective otoscopy image shows no severe signs of otitis media, and the absorbance in the WBT is also high. On the other hand, in Figure 6 b) and d) the otoscopy images show a severe infection in the AOM case and effusion in the OME case, accompanied by very low absorption values in both WBT measurements. Both NOE cases in Figure 6 e) and f) show pale eardrum with no sign of infection, and high absorbance with an increased absorbance at ambient pressure.

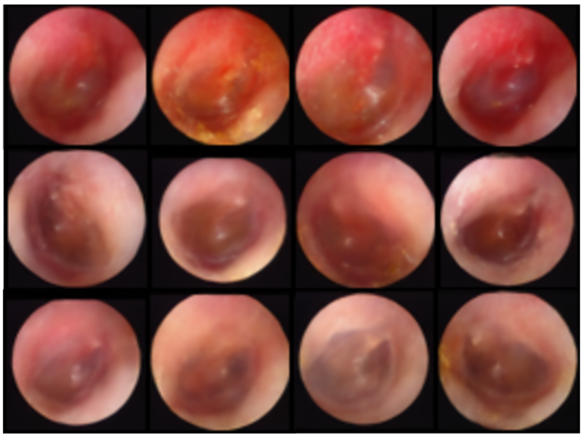

Figure 4: Examples of generated otoscopy images. Top row: AOM, middle row: OME, bottom row: NOE. Best viewed with zoom.

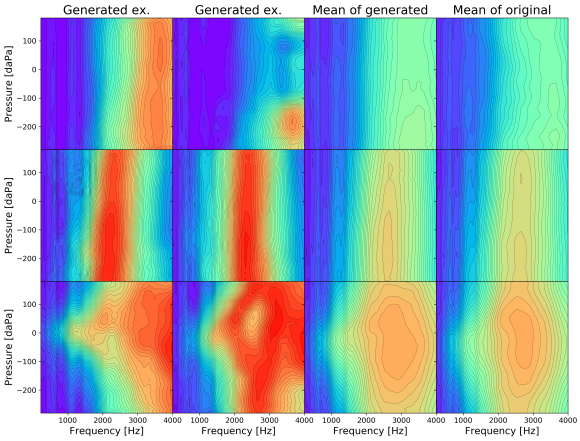

Figure 5: Overview of generated WBT measurements. Top row: AOM, middle row: OME, bottom row: NOE. Best viewed with zoom.

4 Discussion and Conclusion

The proposed multi-modal triplet-loss based VAE is able to generate highly realistic conditional pairs of otoscopy images and WBT measurements. The generated images in Figures 1, 4, and 6 show that the proposed triplet-loss based VAE generates images with a large variation in appearance, and with clear diagnostic markers. The generated images are slightly blurry, which is a common problem with VAEs [2]. The use of SSIM loss [24] has dramatically improved the quality of the generated images, compared to the use of BCE loss. Other studies have found ways to improve the quality even further and have thus synthesized high-resolution images using VAEs [10, 25], which can be incorporated into our approach in future work. As the WBT is a simpler type of data to generate as it does not contain the same level of detail as an image, BCE loss is sufficient for this modality, and the results in Figures 1, 5, and 6 show that the generated WBT measurements correspond very well to the appearance and structure of the original WBT measurements.

In this study, we propose a VAE structure for conditional multi-modal data generation, even when no direct pixel-to-pixel correlation exists between the different modalities. This multi-modal VAE structure is very flexible, as the encoder and decoder for each modality are completely decoupled from the other modality. This allows different architectures to be used for each modality depending on the specific needs of the individual modalities. The network architecture employed for the otoscopy images can be changed to allow the generation of larger and higher-quality images. Similarly, the architecture can be altered to fit temporal data, such as electrocardiograms or electroencephalograms, if our method is employed in other domains. Furthermore, the results show how conditional data generation can be accomplished by employing triplet loss in the latent space of the VAE. In this way, conditioning the input or latent space is not needed, as one can simply sample within a certain class cluster.

This work shows how we are able to generate new data pairs. We do not transform between modalities, such as generating an otoscopy image from a WBT measurement input. The level of information in these two modalities is very different, as the WBT contains much less information about the middle ear, compared to the otoscopy images. It is thus not feasible to synthesize an image by analysing only a WBT measurement. The main use of this model is thus for data augmentation. As seen in Figure 3, combining these two modalities in a single model allows for a good separation of the three classes in the embedding space, which we will further explore in future work on a multi-modal classification pipeline.

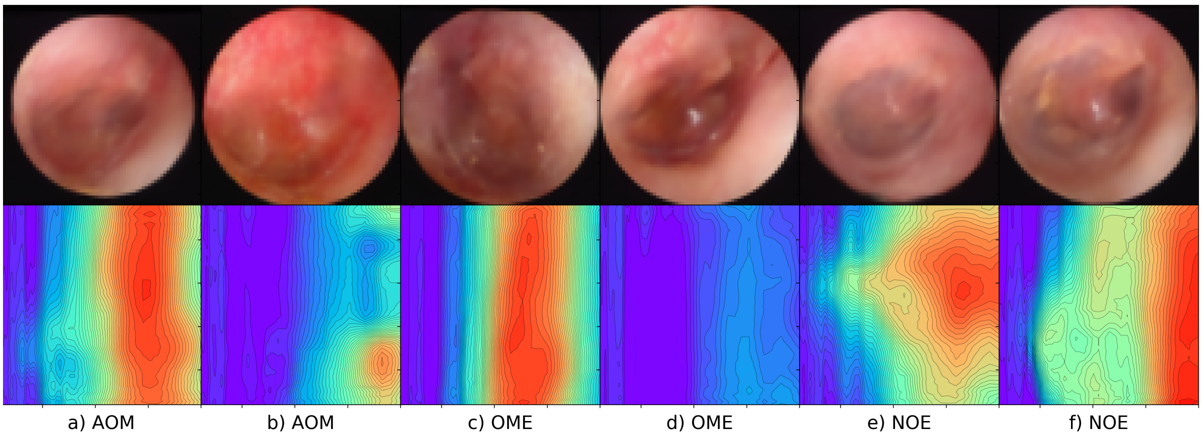

Figure 6: Generated pairs of otoscopy images and WBT measurements. a) mild AOM case, b) severe AOM case, c) mild OME case, d) severe OME case, e) NOE case, and f) NOE case.

5 Acknowledgements

This study was financially supported by William Demant Foundation. Josefine Vilsbøll Sundgaard now works for Novo Nordisk A/S and Technical University of Denmark.

References

[1] H. Binol, A. C. Moberly, M. K. K. Niazi, G. Essig, J. Shah, C. Elmaraghy, T. Teknos, N. Taj-Schaal, L. Yu, and M. N. Gurcan. Decision fusion on image analysis and tympanometry to detect eardrum abnormalities. *Proc. SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis*, 11314, 2020. ISSN 16057422. doi: 10.1117/12.2549394.

[2] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in Neural Information Processing Systems*, 2016. ISSN 10495258.

[3] P. Ghosh, M. S. M. Sajjadi, A. Vergari, M. Black, and B. Scholkopf. From variational to deterministic autoencoders. In *International Conference on Learning Representations*, 2020.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[5] E. M. Grais, X. Wang, J. Wang, F. Zhao, W. Jiang, and Y. Cai. Analysing wideband absorbance immittance in normal and ears with otitis media with effusion using machine learning. *Scientific Reports*, 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-89588-4.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[7] T. A. D. Hein, S. Hatzopoulos, P. H. Skarzynski, and M. F. Colella-Santos. Wideband Tympanometry. In *Advances in Clinical Audiology*. BoD – Books on Demand, 2017. doi: 10.5772/67155.

[8] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[9] X. Hou, L. Shen, K. Sun, and G. Qiu. Deep feature consistent variational autoencoder. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages

1133–1141, 2017. doi: 10.1109/WACV.2017.131.

[10] H. Huang, Z. Li, R. He, Z. Sun, and T. Tan. IntroVAE: Introspective variational autoencoders for photographic image synthesis. *Advances in Neural Information Processing Systems*, 2018. ISSN 10495258.

[11] T. Karaletsos, S. Belongie, and G. Rätsch. Bayesian representation learning with oracle constraints. *4th International Conference on Learning Representations, ICLR*, 2016.

[12] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. ISSN 09252312.

[13] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[14] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling. Semi-supervised learning with deep generative models. *Advances in Neural Information Processing Systems*, 4, 2014. ISSN 10495258.

[15] S. Li, B. Tai, and Y. Huang. Evaluating variational autoencoder as a private data release mechanism for tabular data. In *2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC)*. IEEE, 2019. doi: 10.1109/PRDC47002.2019.00050.

[16] H. C. Myburgh, S. Jose, D. Swanepoel, and C. Laurent. Towards low cost automated smartphone- and cloud-based otitis media diagnosis. *Biomedical Signal Processing and Control*, 39, 2018. ISSN 17468108. doi: 10.1016/j.bspc.2017.07.015.

[17] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7298682.

[18] C. Senaras, A. C. Moberly, T. Teknos, G. Essig, C. Elmaraghy, N. Taj-Schaal, L. Yua, and M. N. Gurcan. Detection of eardrum abnormalities using ensemble deep learning approaches. *Proceedings SPIE, Medical Imaging 2018: Computer-Aided Diagnosis*, 10575, 2018. doi: 10.1117/12.2293297.

[19] H. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International workshop on simulation and synthesis in medical imaging*. Springer, 2018. doi: 10.1007/978-3-030-00536-8_1.

[20] J. V. Sundgaard, J. Harte, P. Bray, S. Laugesen, Y. Kamide, C. Tanaka, R. R. Paulsen, and A. N. Christensen. Deep metric learning for otitis media classification. *Medical Image Analysis*, 71, 2021. ISSN 13618423. doi: 10.1016/j.media.2021.102034.

[21] J. V. Sundgaard, P. Bray, S. Laugesen, J. Harte, Y. Kamide, C. Tanaka, A. N. Christensen, and R. R. Paulsen. A deep learning approach for detecting otitis media from wideband tympanometry measurements. *IEEE Journal of Biomedical and Health Informatics*, 26(7):2974–2982, 2022. doi: 10.1109/JBHI.2022.3159263.

[22] S. Terzi, A. Özgür, Erdivanli, Z. Coşkun, M. Ogurlu, M. Demirci, and E. Dursun. Diagnostic value of the wideband acoustic absorbance test in middle-ear effusion. In *Journal of Laryngology and Otology*, 2015. doi: 10.1017/S0022215115002339.

[23] L. Van Der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008. ISSN 15324435.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13, 2004. doi: 10.1109/TIP.2003.819861.

[25] S. Zhao, J. Song, and S. Ermon. Towards Deeper Understanding of Variational Autoencoding Models. *arXiv preprint arXiv:1702.08658*, 2017.

[26] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. doi: 10.1609/aaai.v34i07.7000.