

Conditioned Generative Transformers for Histopathology Image Synthetic Augmentation

Meng Li, Chaoyi Li, Can Peng, and Brian Lovell

The University of Queensland, School of ITEE, QLD 4072, Australia

Abstract. Deep learning networks have demonstrated state-of-the-art performance on medical image analysis tasks. However, the majority of the works rely heavily on abundantly labeled data, which necessitates extensive involvement of domain experts. Vision transformer (ViT) based generative adversarial networks (GANs) recently demonstrated superior potential in general image synthesis, yet are less explored for histopathology images. In this paper, we address these challenges by proposing a pure ViT-based conditional GAN model for histopathology image synthetic augmentation. To alleviate training instability and improve generation robustness, we first introduce a conditioned class projection method to facilitate class separation. We then implement a multi-loss weighing function to dynamically balance the losses between classification tasks. We further propose a selective augmentation mechanism to actively choose the appropriate generated images and bring additional performance improvements. Extensive experiments on the histopathology datasets show that leveraging our synthetic augmentation framework results in significant and consistent improvements in classification performance.

Keywords: Conditional GAN · Transformer · Histopathology image · Image synthesis · Data augmentation · Multi-task learning.

1 Introduction

Vision transformer (ViT) [25] models have started to outperform convolutional neural networks (CNNs) in various image tasks [5,24,8,3]. Recently, a few pioneering studies have published convolution-free ViT-based GANs and demonstrated good performance on image generation tasks [15,17]. Different from the local receptive field in CNN, ViTs have exhibited promising superiority in modeling non-local contextual dependencies [17]. One known attribute of histopathology images is that they commonly comprise various non-local or long-range information [27]. As a result, at the infancy stage, these convolution-free ViT-based GAN models are considered to have great potential for histopathology image analysis tasks.

Deep learning has benefited from abundantly labeled data and achieves good performance on the histopathological image analysis tasks. However, most of the works come at the expense of acquiring much labeled data and require the extensive participation of domain experts [20]. The whole process is labor-intensive

and time-consuming, which is impractical for rare diseases, early clinical studies, or new imaging modalities [19]. A few works have been proposed to mitigate the aforementioned issues by employing GANs to generate synthetic images for data augmentation and achieved satisfactory results [20,9,7]. Nonetheless, these works split the image generation and data augmentation into two stages, which results in different models for each stage and requires more effort to train.

We address the aforementioned challenges by introducing a framework that adopts a convolution-free ViT-based GAN model for histopathology image synthetic augmentation. The framework merges models from two stages into one, which not only conditionally generates synthesized images but also performs classification prediction. This is a challenging task, as GANs are known to be prone to notorious training instability [15], combining multiple losses would lead to extra unstable performance. As a result, we first propose a conditioned class projection to assist in the separation of the conditional information during training. Inspired by the success of multi-task learning works [16,22], we then introduce a novel multi-loss weighing function to balance the losses when training our conditional GAN model. To improve synthetic augmentation performance even further, we introduce a selective augmentation mechanism that actively selects appropriate images for augmentation.

To summarize, the major contributions of this paper can be listed as follows:

- To the best of our knowledge, this work is among the first to explore conditional convolution-free ViT-based GANs on histopathology images, which is of great practical value and less studied by previous literature.
- We introduce a novel conditioned class projection method to facilitate class separation during training.
- We address the problem that two-stage image synthetic augmentation in previous works results in low efficiency. To this end, we propose a novel all-in-one framework by using a multi-loss weighing function. This helps stabilize training and improves performance on each task.
- We propose a selective augmentation mechanism to further increase the synthetic augmentation performance.
- Extensive experiments are conducted on lymph node histopathology datasets. Compared with baseline models, experimental results show that our proposed method significantly improves the augmented classification performance.

2 Method

The overall framework of our proposed transformer-based GAN is shown in Fig. 1. Unlike other transformer-integrated GANs [29,6] for medical image synthesis, our model is constructed entirely of transformer components and is CNN-free. We adopt a conditional training strategy by introducing multiple class projection modules in the generator, and an additional class token in the discriminator to perform the classification task. The outputs of class and source heads are adaptively combined through our proposed multi-loss weighing function.

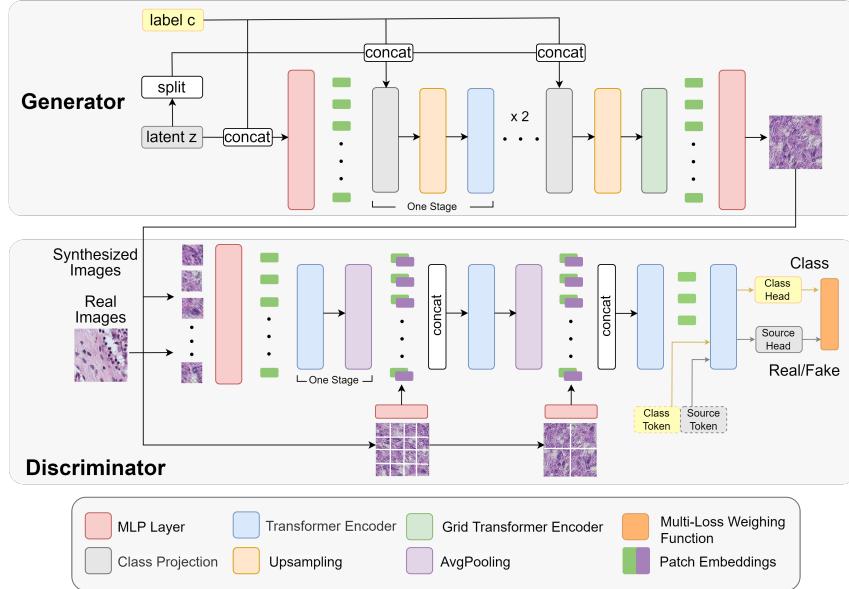


Fig. 1. Overview of our proposed network. In the generator, the latent vector z is split into multiple chunks and concatenated with conditional label c as the inputs to class projection layer. One generator block includes a class projection layer, an upsampling layer, and a transformer encoder. One discriminator block includes a transformer encoder and an average pooling layer. A multi-scale pyramid technique is employed to learn global and local information. The output of discriminator is the adaptive weighing function of the class and source losses.

2.1 Transformer-based GAN

Our proposed model is a pure transformer-based GAN following the design of TransGAN [15]. The primary component is the ViT encoder [8], which comprises multi-head self-attention modules, including query, key, and value representations to conduct a self-attention scaled dot-product computation. Next, the output goes through a feed-forward multi-layer perceptron (MLP) layer with GELU non-linearity [12]. Both parts employ residual connection and layer normalization techniques.

The generator of the proposed GAN comprises four-stage blocks that gradually learn and upsample the latent vector z . Initially, the latent vector noise z is concatenated with the one-hot label c and passed through a linear projection layer to generate embedding tokens. Each block consists of a class projection layer, an upsampling module, and four ViT encoders. By stages, the generator progressively increases the feature map resolution until it reaches the predefined dimension $H \times W$. In the first stage, the embedded feature map is upsampled from $X_i \in \mathbb{R}^{H_i \times W_i \times C}$ to $X'_i \in \mathbb{R}^{2H_i \times 2W_i \times C}$ by cubic interpolation without dimension reduction, which guarantees the early feature learning. The next three

stages adopt a pixel-shuffle module [23] that upsamples the feature map while decreasing the channels to a quarter. The feature map then converts from $X_i \in \mathbb{R}^{H_i \times W_i \times C}$ to $X'_i \in \mathbb{R}^{2H_i \times 2W_i \times \frac{C}{4}}$. This reduces memory requirements and makes the network more efficient. As the self-attention layer is prone to learn global correspondence while neglecting local information [15], for the third and fourth stages, we use grid transformer encoders to constrain the model to learn local details of the image. This is achieved by reshaping the feature map size to the predefined dimension that goes through the ViT encoders. In our model, we set the predefined window size to 32×32 for the third stage and 16×16 for the fourth stage.

The discriminator consists of four stages. Each of the first three includes a ViT encoder and an average pooling layer. Between the stages, a multi-scale technique is employed by combining the outputs from the last block with the varying sizes of patches from the same input image. After an MLP layer, the patch information is encoded in a sequence of embeddings for concatenation. This is to let the model learn both semantic structure and texture details. To enable conditional learning, at the end of the third block, a class token and a source token are appended at the beginning of the 1D sequence and go through the fourth stage with two ViT encoders. The tokens are taken by the classification head and the source head to output the class and real/fake predictions.

2.2 Conditioned Class Projection

To further improve the conditional image generation, we first adopt a direct skip connection (skip-z) [4] method that splits noise vector z into multiple chunks and adds connections to the four stages in our proposed generator. The intuition behind this design is to enable the generator to use the latent space that directly affects features generated at different resolutions. More specifically, we split z into one chunk per stage, and concatenate each chunk to the class vector c . The combination is then passed into the class projection layer.

Next, we implement the class projection onto the token embeddings. We call this approach *conditional layer normalization*. In contrast to the class-conditional [4] method that uses batch normalization, we find that employing layer normalization [2] for ViT-based conditional GAN obtains superior performance. The approach transforms a layer’s activations a into a normalized activation \bar{a} specific to class condition, which is shown as:

$$\bar{a} = \frac{a - \hat{\mu}}{\sqrt{\hat{\sigma} + \epsilon}} * \gamma + \beta. \quad (1)$$

Where $\hat{\mu}$ and $\hat{\sigma}$ stands for mean and variance of the input, ϵ is a small value added to the denominator for numerical stability. Specifically, we inject class-conditional information by parameterizing γ and β as linear transformations of the class embedding, c , where $\gamma := W_\gamma^\top [c, z]$ and $\beta := W_\beta^\top [c, z]$, $[c, z]$ denotes the concatenation of c and z in skip connections.

2.3 Selective Data Augmentation

We further improve synthetic augmentation performance by employing a selective data augmentation mechanism. To ensure a high fidelity image generation, we first adopt a truncation method [18] which resamples z from a truncated normal distribution by a truncation value τ . After generating the images conditionally, we use our trained classification head of the same model to assist in selecting qualified augmentation images. Only the image prediction confidence value greater than threshold λ will be used. We find $\tau = 0.7$ and $\lambda = 0.6$ achieves best performance in our experiments.

2.4 Multi-Loss Weighing Function

The learning strategy of our network follows AC-GAN [21] by incorporating an auxiliary class head in the discriminator. However, the auxiliary classifier in [21] primarily serves as a control for diversified image generation yet obtains downgraded classification performance. In our work, we aim to address this problem by incorporating a multi-loss weighing function that adapts WGAN-GP loss [1] and classification cross-entropy loss, so that the generation and classification both achieve satisfying results.

We apply Wasserstein loss with gradient penalty [1,10] to our GAN. The objective is given by:

$$\mathcal{L}_S = \min_G \max_{D \in \mathcal{D}} \mathbb{E}_{x \sim \mathbb{P}_r} [D(x)] - \mathbb{E}_{z \sim \mathbb{P}_z} [D(G(c, z))] + \lambda \mathbb{E}_{\hat{x} \sim \mathbb{P}_{\hat{x}}} \left[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2 \right], \quad (2)$$

where D is the critic head of discriminator that describes a group of 1-Lipschitz functions and G is the generator. \mathbb{P}_r , \mathbb{P}_z , and $\mathbb{P}_{\hat{x}}$ denote the distribution of real data x , the normal distribution of a random z , and the distribution of pairs of points from \mathbb{P}_r and \mathbb{P}_z . Every generated sample has a corresponding class label $c \sim \mathbb{P}_c$ besides z . G uses conditioned noise to generate images $X_{\text{fake}} = G(c, z)$. Note that the critic head does not involve conditioned information.

To enable conditional learning, we employ an auxiliary classification head $P(C | X)$ on discriminator, the log-likelihood objective function is given by:

$$\mathcal{L}_C = \mathbb{E} [\log P(C = c | X_{\text{real}})] + \mathbb{E} [\log P(C = c | X_{\text{fake}})], \quad (3)$$

both D and G are trained to maximize \mathcal{L}_C . We find that simply combining \mathcal{L}_C and \mathcal{L}_S is prone to result in training failure as WGAN loss value is relatively large at most times, which results in cross-entropy loss overwhelmed during training. In addition, the model tends to be good at one classification task on discriminator which results in an imbalance problem. To this end, we employ the concept from the multi-task learning realm [16] to weigh classification losses. Kendall *et al.* [16] proposed a strategy to balance multi-task losses by considering the uncertainty of each task, which is shown as:

$$\mathcal{L}_{\text{joint}} (\gamma_1, \gamma_2) = \frac{1}{2\gamma_1^2} \mathcal{L}_1 + \frac{1}{2\gamma_2^2} \mathcal{L}_2 + \log \gamma_1 + \log \gamma_2, \quad (4)$$

where $\mathcal{L}_{1,2}$ represent the losses of classification tasks, $\gamma_{1,2}$ are trainable parameters representing the positive scalar of each task.

During training, we observe that the range of scalar values affects model performance (details in Section 3) as they can only be positive in *log* form, we therefore propose a modified multi-loss weighting strategy which allows them to be in the range $[-\infty, +\infty]$. Specifically, we use $\exp(-\sigma)$ as the trainable scaling parameter to replace $\frac{1}{2\gamma_1^2}$, which allows σ to be an unlimited regularization value. Let $f^\theta(x)$ be the output of the critic head with weights θ , our adapted classification likelihood of the model output through softmax function can be written as:

$$p(\mathbf{y} | \mathbf{f}^\theta(\mathbf{x}), \sigma) = \text{Softmax}(\exp(-\sigma)\mathbf{f}^\theta(\mathbf{x})) \quad (5)$$

with an unlimited range trainable scalar σ . The scaling process can be regarded as a Maxwell–Boltzmann distribution, where $\exp(-\sigma)$ is commonly referred to as temperature of the input. The learnable parameter’s magnitude determines how “flat” the discrete distribution is. The output is then related to uncertainty by using the log likelihood, which can then be written as:

$$\begin{aligned} -\log p(\mathbf{y} = c | \mathbf{f}^\theta(\mathbf{x}), \sigma) &= -\log \text{Softmax}(\exp(-\sigma)\mathbf{f}_c^\theta(x)) \\ &= -\log \frac{\exp[\exp(-\sigma)\mathbf{f}_c^\theta(x)]}{\sum_{c'} \exp[\exp(-\sigma)\mathbf{f}_{c'}^\theta(x)]} \\ &= -\exp(-\sigma)\mathbf{f}_c^\theta(x) + \log \sum_{c'} \exp[\exp(-\sigma)\mathbf{f}_{c'}^\theta(x)] \\ &= -\exp(-\sigma) \log \left[\frac{\exp(\mathbf{f}_c^\theta(x))}{\sum_{c'} \exp(\mathbf{f}_{c'}^\theta(x))} \right] \\ &\quad + \log \frac{\sum_{c'} \exp[\exp(-\sigma)\mathbf{f}_{c'}^\theta(x)]}{(\sum_{c'} \exp(\mathbf{f}_{c'}^\theta(x)))^{\exp(-\sigma)}} \\ &\approx \exp(-\sigma)\mathcal{L} + \sigma, \end{aligned} \quad (6)$$

where $-\exp(-\sigma) \log \left[\frac{\exp(\mathbf{f}_c^\theta(x))}{\sum_{c'} \exp(\mathbf{f}_{c'}^\theta(x))} \right] = -\exp(-\sigma) \log \text{Softmax}(\mathbf{y}, \mathbf{f}^\theta(x))$ is our scaled loss function. We write $\mathcal{L} = -\log \text{Softmax}(\mathbf{y}, \mathbf{f}^\theta(x))$ for the cross entropy loss (not scaled). In the last transition, an explicit simplifying assumption is introduced:

$$\exp(-\sigma) \sum_{c'} \exp[\exp(-\sigma)f_{c'}^\theta(\mathbf{x})] \approx \left(\sum_{c'} \exp(f_{c'}^\theta(\mathbf{x})) \right)^{\exp(-\sigma)}, \quad (7)$$

When $\sigma \rightarrow 0$, Eq 7 becomes equality. This helps simplify the optimisation objective and empirically demonstrates promising results.

Given multiple classification outputs from the critic head, we often define the likelihood to factorise over the outputs. Our multi-task likelihood is shown as follows:

$$p(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3 | \mathbf{f}^\theta(\mathbf{x})) = p(\mathbf{y}_1 | \mathbf{f}^\theta(\mathbf{x})) p(\mathbf{y}_2 | \mathbf{f}^\theta(\mathbf{x})) p(\mathbf{y}_3 | \mathbf{f}^\theta(\mathbf{x})), \quad (8)$$

where $\mathbf{y}_{1,2,3}$ represent classification predictions of fake data from generator, and real and fake data from discriminator. Next, the joint loss of weighted log likelihood of multi classification tasks is given as:

$$\begin{aligned}\mathcal{L}_{weighting} &= -\log p(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3 = c_1, c_2, c_3 | \mathbf{f}^\theta(\mathbf{x}), \sigma_1, \sigma_2, \sigma_3) \\ &= -[\log p(\mathbf{y}_1 = c_1 | \mathbf{f}^\theta(\mathbf{x}), \sigma_1) + \log p(\mathbf{y}_2 = c_2 | \mathbf{f}^\theta(\mathbf{x}), \sigma_2) \\ &\quad + \log p(\mathbf{y}_3 = c_3 | \mathbf{f}^\theta(\mathbf{x}), \sigma_3)] \\ &= \exp(-\sigma_1)\mathcal{L}_1(\theta) + \exp(-\sigma_2)\mathcal{L}_2(\theta) + \exp(-\sigma_3)\mathcal{L}_3(\theta) + \sigma_1 + \sigma_2 + \sigma_3\end{aligned}\tag{9}$$

The final weighted objective can be regarded as learning each task weights of each output losses. Larger temperature value σ gives less contribution of loss function, while smaller σ increases its weight. The function is regulated by the last three terms when σ is too small.

3 Experiments

3.1 Datasets

Our experiments are conducted on the public PatchCamelyon (PCam) benchmark dataset [26]. The PCam dataset consists of 327,680 color images derived and digitized using a 40x objective (resultant pixel resolution of 0.243 microns) from lymph node sections. Each patch is of 96×96 resolution and split into two classes based on whether a metastatic tissue is present in the center region. The dataset is divided into 75%:12.5%:12.5% for training, validation, and testing utilizing a hard-negative mining regime. To mimic the situation where limited training data is available, we randomly choose 10% of the images from training data (32768). The whole test set is used to evaluate the augmentation performance.

3.2 Implementation Details

Our network is built based on a pure transformer architecture [15], details are shown in supplementary materials. In the experiment, we use two Tesla V100 GPUs of 16G of RAM each. An Adam optimizer with $\beta_1 = 0$, $\beta_2 = 0.99$, and the learning rate $1e-4$ is adopted to tune both generator and discriminator. The batch size is 12 for both generator and discriminator. During training, we run 500 epochs for all experiments and choose the best epoch based on the Frechet Inception Distance (FID) [13] score between generated images and validation data. We implement DiffAug. [30] as our augmentation strategy during the training process. The augmentation policy adopted in our experiment includes random color transforms, translation, cutout, scaling, and rotation.

3.3 Discussion and Comparisons

We report quantitative evaluation scores between all selected baseline models and our synthetic augmentation method. They are the accuracy, the area under the

Table 1. Evaluation of different models for our data synthesis augmentation method. Aug: augmentation. CP: conditional projection. SA: selective augmentation.

	Accuracy	AUC	Sensitivity	Specificity
Resnet34 [11]	0.881 ± 0.079	0.945 ± 0.022	0.846 ± 0.086	0.916 ± 0.020
Resnet34+Aug(MedViTGAN)	0.912 ± 0.154	0.968 ± 0.089	0.882 ± 0.092	0.942 ± 0.071
Resnet34+Aug(CP)	0.900 ± 0.060	0.949 ± 0.064	0.868 ± 0.031	0.929 ± 0.293
Resnet34+Aug(SA)	0.908 ± 0.083	0.958 ± 0.082	0.866 ± 0.122	0.929 ± 0.029
Resnet34+Aug(CP+SA)	0.916 ± 0.144	0.954 ± 0.042	0.891 ± 0.100	0.951 ± 0.109
Resnet50_Cbam [28]	0.899 ± 0.131	0.955 ± 0.045	0.863 ± 0.096	0.935 ± 0.040
Resnet50_Cbam+Aug(MedViTGAN)	0.918 ± 0.230	0.930 ± 0.021	0.821 ± 0.123	0.917 ± 0.066
Resnet50_Cbam+Aug(CP)	0.920 ± 0.180	0.959 ± 0.055	0.880 ± 0.040	0.937 ± 0.088
Resnet50_Cbam+Aug(SA)	0.916 ± 0.039	0.959 ± 0.062	0.876 ± 0.112	0.941 ± 0.137
Resnet50_Cbam+Aug(CP+SA)	0.922 ± 0.077	0.962 ± 0.046	0.879 ± 0.047	0.945 ± 0.110
Densenet169 [14]	0.894 ± 0.059	0.955 ± 0.036	0.881 ± 0.094	0.908 ± 0.032
Densenet169+Aug(MedViTGAN)	0.927 ± 0.025	0.974 ± 0.040	0.867 ± 0.049	0.966 ± 0.052
Densenet169+Aug(CP)	0.900 ± 0.177	0.946 ± 0.074	0.879 ± 0.099	0.904 ± 0.107
Densenet169+Aug(SA)	0.917 ± 0.021	0.975 ± 0.066	0.877 ± 0.131	0.956 ± 0.099
Densenet169+Aug(CP+SA)	0.928 ± 0.076	0.960 ± 0.049	0.898 ± 0.125	0.967 ± 0.092
MedViTGAN Cls Head	0.939 ± 0.059	0.980 ± 0.073	0.906 ± 0.079	0.974 ± 0.054
HViTGAN Cls Head (ours)	0.916 ± 0.070	0.971 ± 0.098	0.900 ± 0.033	0.923 ± 0.081
HViTGAN Cls Head (ours)+Aug(CP)	0.931 ± 0.029	0.966 ± 0.101	0.897 ± 0.277	0.957 ± 0.123
HViTGAN Cls Head (ours)+Aug(SA)	0.929 ± 0.110	0.977 ± 0.093	0.889 ± 0.042	0.961 ± 0.069
HViTGAN Cls Head (ours)+Aug(CP+SA)	0.945 ± 0.054	0.981 ± 0.045	0.910 ± 0.063	0.977 ± 0.102

ROC curve (AUC), sensitivity, and specificity for a comprehensive comparison. All models were executed for five repetitions with random initialization for a fair comparison. The mean and standard deviation were reported. As is shown in Table 1, we conducted our experiments on representative baseline models with same settings for comprehensive study, including Resnet34 [11], Resnet50 CBAM [28], Densenet169 [14], and our model’s classification head for both baseline and augmentation training. The results consistently showed that with the augmentation of our synthesized images, the models obtained superior performance compared to baselines. We also analyzed how different loss functions affect the generated images. As shown in Fig. 3, we compared seven scenarios for loss functions. Experiments showed that WGAN-GP loss was significantly scaled-down after applying weight to it which resulted in inferior performance. Hence, we omitted the weight applied to WGAN-GP loss and formed our multi-loss weighing function. Our method demonstrated the best image separation performance. More details can be found in the supplementary materials. As demonstrated in Fig. 2, the synthetic images have comparable quality in terms of both fidelity and diversity. We therefore believe that our proposed model is capable of handling other histopathology image-associated tasks with insufficient training data.

4 Conclusion

In this paper, we introduce a pure ViT-based conditional GAN framework for histopathology image synthetic augmentation that incorporates several novel techniques. To enable conditional learning, we employ an auxiliary discriminator head for classification. We further improve the training by proposing several novel techniques including conditioned class projection, multi-loss weighing function, and selective augmentation mechanism. Experimental results demonstrate

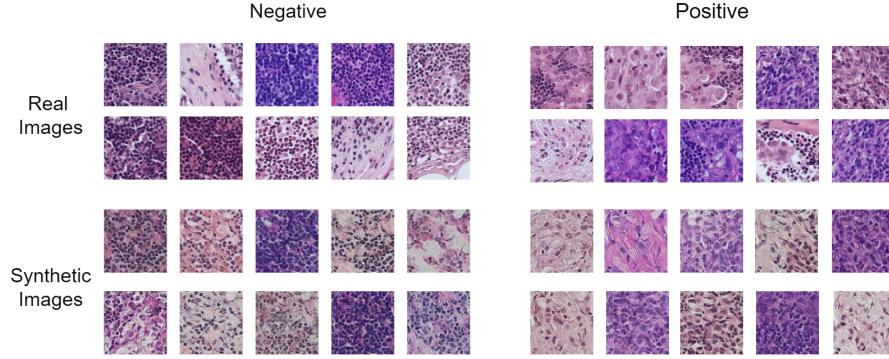


Fig. 2. Selected real and synthetic images generated by the proposed model trained on 10% of PCam dataset. SA: selective augmentation. CP: conditional projection

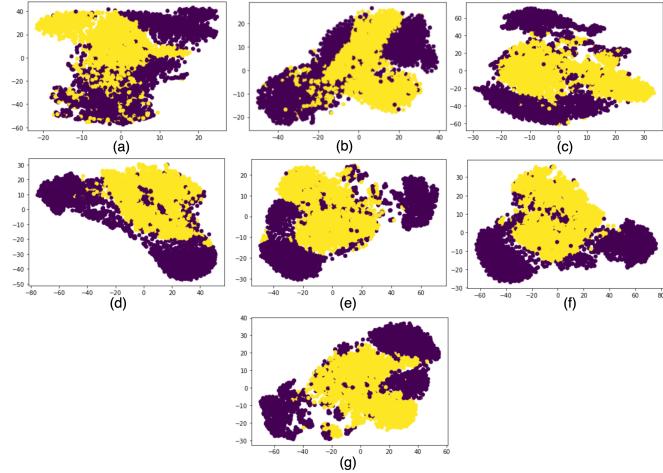


Fig. 3. T-SNE graphs for generated images with different loss scenarios. (a) No weighting function applied to individual loss. (b) Weighting function applied to all losses. (c) Kendall *et al.* [16] weighting function. (d) Our weighting function. (e) Our weighting function with CP. (f) Our weighting function with SA. (g) Our weighting function with CP+SA. CP: conditional projection. SA: selective augmentation.

that our approach makes a significant improvement over baseline methods. In future work, we plan to generalize our model to natural image tasks. Additionally, larger networks will be examined to generate high-resolution images.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017)

2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095 (2021)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
5. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International Conference on Machine Learning. pp. 1691–1703. PMLR (2020)
6. Dalmaç, O., Yurt, M., Çukur, T.: Resvit: Residual vision transformers for multi-modal medical image synthesis. arXiv preprint arXiv:2106.16031 (2021)
7. Dar, S.U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., Çukur, T.: Image synthesis in multi-contrast mri with conditional generative adversarial networks. IEEE transactions on medical imaging **38**(10), 2375–2388 (2019)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 289–293. IEEE (2018)
10. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028 (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
14. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
15. Jiang, Y., Chang, S., Wang, Z.: Transgan: Two transformers can make one strong gan. arXiv preprint arXiv:2102.07074 **1**(2), 7 (2021)
16. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482–7491 (2018)
17. Lee, K., Chang, H., Jiang, L., Zhang, H., Tu, Z., Liu, C.: Vitgan: Training gans with vision transformers. arXiv preprint arXiv:2107.04589 (2021)
18. Marchesi, M.: Megapixel size image creation using generative adversarial networks. arXiv preprint arXiv:1706.00082 (2017)
19. Medela, A., Picon, A., Saratxaga, C.L., Belar, O., Cabezón, V., Cicchi, R., Bilbao, R., Glover, B.: Few shot learning in histopathological images: reducing the need of labeled data on biological datasets. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 1860–1864. IEEE (2019)
20. Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., Shen, D.: Medical image synthesis with context-aware generative adversarial networks. In: Interna-

- tional conference on medical image computing and computer-assisted intervention. pp. 417–425. Springer (2017)
21. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: International conference on machine learning. pp. 2642–2651. PMLR (2017)
 22. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
 23. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
 24. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
 25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
 26. Veeling, B.S., Limmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: International Conference on Medical image computing and computer-assisted intervention. pp. 210–218. Springer (2018)
 27. Wang, J., Kong, J., Lu, Y., Qi, M., Zhang, B.: A modified fcm algorithm for mri brain image segmentation using both local and non-local spatial constraints. Computerized medical imaging and graphics **32**(8), 685–698 (2008)
 28. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
 29. Ye, J., Xue, Y., Liu, P., Zaino, R., Cheng, K.C., Huang, X.: A multi-attribute controllable generative model for histopathology image synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 613–623. Springer (2021)
 30. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient gan training. arXiv preprint arXiv:2006.10738 (2020)