Automated Annotation and Classification of Catheters in Chest X-Rays

Akash Karthikeyan(✉) and Saravana Perumaal Subramanian

Department of Mechanical Engineering, Thiagarajar College of Engineering, Madurai, India
akashk1@student.tce.edu, sspmech@tce.edu

Abstract. Catheters are usually used to deliver drugs and medications close to the heart and to monitor the vital organs around the chest region for patients who undertook critical surgery. Radiologists often check for the presence of catheters, puncture-needles, guiding sheaths, and various other tube-like structures in interventional radiology. The clinical analysis of X-ray requires a manual pixel-wise annotation which is an excruciating process. In order to address this issue, we attempt to auto-annotate the CXRs using a Self-Supervised Learning approach. Further, the classification task on the catheter is performed based on semantic and perceptual clues (object shapes, colors, and their interactions) of color and class distributions. A generative adversarial network is utilized to learn a mapping to annotate (colorize and identify end-tip points) and classify the given grayscale CXR. The additional number of classes, custom loss function, and attention heads introduced in the model is a unique attempt to ensure robust results in the radiological inferences. It is evident that the qualitative and quantitative results of annotation and classification are viable which resembles how humans perceive such problems. The results are consistent and outperform's the state-of-the-art supervised learning models in terms of metrics and inference durations. The model being end-to-end in nature, can be integrated along with the existing in-hospital pipeline and will be ready to use instantly.

Keywords: GAN · Chest X-rays · Self-supervised learning

1 Introduction

Chest X-rays are commonly performed radiologic examination of the human body for patients kept under critical care. Portable anterior-posterior (AP) CXRs are often used to detect malpositions if any and verify the placement of catheters. These catheters are inserted through the subclavian or jugular veins and are typically blindly operated upon [12]. After placement Chest X-Rays (CXRs) are obtained to analyze their presence, and identity to avoid any mispositioning or other complications. Traditionally it requires years of training, experience, and skill-set to accomplish such a task. With the recent developments in computer

© IFIP International Federation for Information Processing 2022
Published by Springer Nature Switzerland AG 2022
E. J. Neuhold et al. (Eds.): ICCCSP 2022, IFIP AICT 651, pp. 151–162, 2022.
https://doi.org/10.1007/978-3-031-11633-9_12

vision and the advent of Artificial Intelligence, autonomous report generation in radiology will considerably expedite clinical workload, Such a system would be able to remotely generate reports for CXRs in a matter of seconds.

Table 1. Distribution of the samples over various classes of catheters based on their position

Description		No. samples
Endo-tracheal tube	Abnormal	16
	Borderline abnormal	192
	Normal	1423
Naso-gastric tube	Abnormal	61
	Borderline abnormal	95
	Incompletely imaged	507
	Normal	887
Central venous catheter	Abnormal	640
	Borderline abnormal	1596
	Normal	4220
Swan-Ganz	Presence of Swan-Ganz catheter	139
Total classes: 11		Total: 30083

Previous attempts at producing such reports, under in-hospital conditions - lack the diagnostic interpretation and accuracy. The need for consistent and structured reports are imminent and plays a pivotal role in clinical care. Failure to report (multilabel classifications as shown in Table 1, and corresponding problems based on their positioning and insertions if any) in a clear and concise manner reflects in sub-optimal care.

The popular approach involves, the use of Deep Convolutional models. To be able to learn unique discriminative spatio-temporal features for CXRs is a difficult task. Hence recognition and classification of catheters from direct whole images yield poor results, as these needle-like structures account for less than 1% of the footprint in the whole image [12]. It is evident from the Class Activation Map as shown in Fig. 1 of traditional convolutional models struggles with classification tasks due to overlap in receptive fields. Different classes of catheters arise, owing to the difference in tip locations, functions, and the target organ as shown in Table 1.

While conventional CNNs could be used for this task, they usually require a large amount of paired data which is to be manually annotated, yet will still suffer from class imbalances. Moreover, almost all the CNNs are used to 3-channel input rather than single-channel input in the case of CXRs limiting the use of ImageNet weights of state-of-the-art models.

This paper presents an approach to address the automatic classification of unlabeled samples and the classification of peripheral and central catheter positions through a GAN model that is based on semantic and perceptual (end-tip points, pins, object shapes, colors, and their interactions) nature. We also

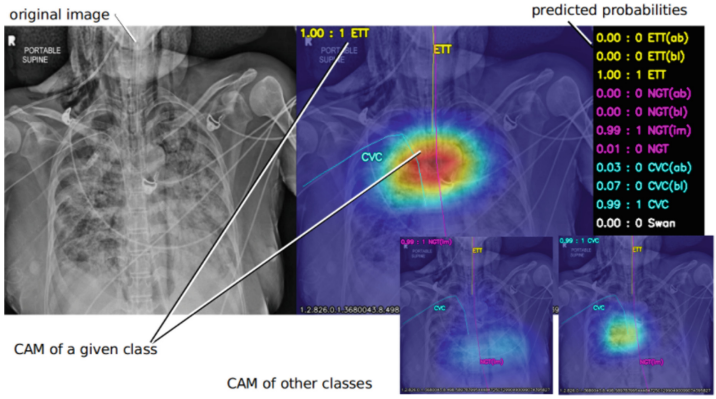

Fig. 1. Class activation maps of a multi-labelled sample, with logits

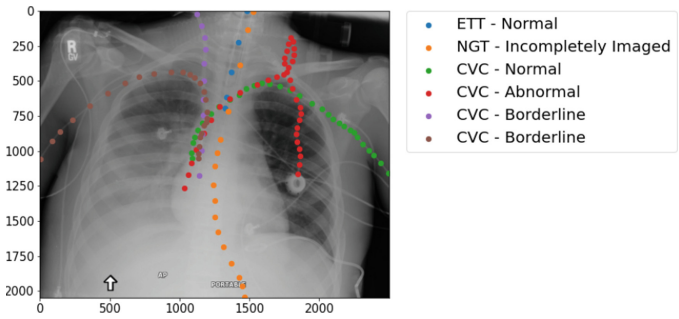

Fig. 2. Example of annotated sample, multi-labeled and annotated sample

explore how colorization affects the class distribution in CXRs. The model is designed in an end-to-end fashion and trained in a weakly supervised manner. Initial tests involving perpetual cues provided good initializations. In particular, adding color to the classification task leads to better initializations and forces the model to learn proper feature representations as seen in warm-starting of ImageNet models for various transfer learning applications.

1.1 Related Work

The absence of colors being a major reason for the failure of ImageNet models in CXRs. Colors also seemed to add class-specific semantic clues which are known to boost performance. Zhang *et al.* proposed an automatic colorization algorithm where colorization is treated as a multinomial classification problem [16] proposed a colorization model that learns to map a monochromatic image along with user-provided hints and cues that later is used to fuse along with high-level semantic information to provide realistic outputs. These processes again make

the task multistage or involve humans in a loop making it difficult to build a standalone system.

"Generative Adversarial Networks (GANs) have been a popular choice to learn a mapping between gray-scale photographs. In recent times, Wasserstein GANs have been used with the gradient penalty to generate paired image-mask samples from CXRs" [10]. Authors of [2] proposed to use such generated masks to be paired along with real-time images to train in a supervised manner. A few methods were employed for non-paired data generation of medical and clinical images, entailing these data from an auxiliary imaging modality in a domain adaptation setting [8]. Performance evaluation of these techniques has not been evaluated. The author [6] proposed the "use of conditioned GANs to map the images using U-Net-based architecture. They adopted Dice Loss and Wasserstein loss to generalize the same to high-resolution images, stabilize and converge faster" [9].

Notice that GANs were essentially utilized to generate synthetic data which is used to provide better domain adaptation in various transfer learning tasks utilizing deep convolutional networks. Upon using a shared module trained in a joint manner (Semantic annotation and Classification) in an adversarial method yields better results and reduced inference period, as these learned representations aid in the classification as well acts as a form of additional supervision.

2 Method

2.1 Annotation-GAN

When the gray-scale image L which resembles the CXR, the model aims to learn a mapping $G\colon L \rightarrow c$ such that $I = (L, c)$ which represent a probable color distribution, and $\pi(z)$ denotes real image distribution. A bijective mapping such that $c = G^{-1}(z; L)$ is near the ground truth values is learned. Thus, the generated annotated sample is expected to encompass semantic, perceptual, and geometric characters to that of the ground truth image.

As in [14] a generator module learns a function G which is also invertible in nature and in an end-to-end fashion. While parallelly a discriminator module validates the similarity between the colorized output to actual ground truth. $I = (L, c)$ of L. The GAN model tries to tune the Ψ and w, parameters of the model. This ensures a conditioned GAN trained in a weakly-supervised nature using the annotated samples, with model initialized with weights from ImageNet [4], Especially when using an annotated sample $I_z = (L, c)$, the model learns the color information in a split fashion: The mono-chromatic channel L and the chrominance channels (c) similar to the method in [14].

The generator G_Ψ also tries to learn a $[nx1]$ classification matrix. The advantage of using such an approach enables us to learn the perpetual and semantic image distribution contained in L that makes the model similar to how humans perceive color. Let us subnets of generator be defined as follows, $G_{\Psi_1^1}\colon L \rightarrow (c)$, and $G_{\Psi_2^2}\colon L \rightarrow y$.

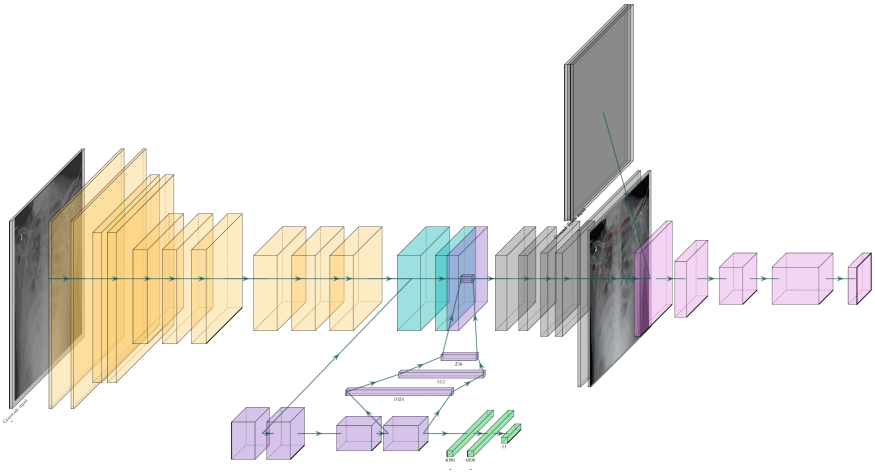

Fig. 3. Auto-Annotation GAN can colorize the tubular structures of catheters in CXR's based on different classes, It is a combination of Discriminator Network, D_Λ in pink, and the Generator Network, G_Ψ which consist of two-subnets: $G^1_{\Psi_1}$ (yellow, aqua-green, purple and gray layers) and $G^2_{\Psi_2}$ (yellow, purple and gray layers). (Color figure online)

2.2 Objective Formulation

The error function is given by:

$$L(G_\Psi, D_\Lambda) = L_\epsilon(G^1_{\Psi_1}) + \lambda_g L_g(G^1_{\Psi_1}, D_\Lambda) + \lambda_s L_s(G^2_{\Psi_2}) \tag{1}$$

The initial term represents ***reconstruction loss*** which is defined as follows:

$$L_\epsilon(G^1_{\Psi_1}) = E_{(L,c)\sim P_c}[-\log(q)(L, c')||G^1_{\Psi_1}(L) - c||^2_2] \tag{2}$$

where, P_c represents the distribution of ground truth images, the initial part represents negative log-likelihood loss added along with $||\cdot||_2$ for the euclidean norm to retain the structural similarity in the image and prevent from losing data. To improve the sensitivity to perpetual color, we use ***class distribution loss***

$$L_s(G^2_{\Psi_2}) = E_{L\sim P_z}[KL(y_v||G^2_{\Psi_2}(L)))] \tag{3}$$

Herein, P_z represents the distribution of CXR input, and $y_v \in labels$ as in Table 1 the softmax output is obtained from VGG-16 classification branch [11]. $KL(\cdot||\cdot)$ represents the Kullback-Leibler divergence. The L_g denotes the ***WGAN loss*** which represents Wasserstein GAN [1]. WGAN provides a few niceties like avoiding vanishing gradient issues, preventing mode collapse, and faster convergence.

$$L_g(G^1_{\Psi_1}, D_\Lambda) = E_{L' \sim P_c}[D_\Lambda(L')] - E_{(c) \sim P_{G^1_{\Psi_1}}}[D_\Lambda(c, L)] \\ -E_{L' \sim P_{\hat{I}}}[(||\nabla_{\hat{I}} D_\Lambda(\hat{I})||_2 - 1)^2] \tag{4}$$

with $P_{G^1_{\Psi_1}}$ representing the probability density of the model $G^1_{\Psi_1}(L)$ denoting the generator distribution, with $L \sim P_z$. $P_{L'}$ sampled uniformly along straight lines between pairs of points from data distribution P_c and $P_{G^1_{\Psi_1}}$ The negative sign in (Eq. 4) ensures that the model minimizes the loss with respect to discriminator parameters.

This now reduces to a min-max problem that tries to converge in an alternate fashion by adjusting the weights of generator and discriminator modules. Later in the ablation experiments, it was also observed that the reconstruction loss enables quicker convergence.

$$\min_{G_\Psi} \max_{D_\Lambda \in D} L(G_\Psi, D_\Lambda) \tag{5}$$

Conditional GAN Loss and Adversarial Strategy L_g. The general adversarial min-max game between a generator and discriminator deals with learning suitable parameters so that the generator mimics the probability distribution of real data.

The WGAN is a better alternative to the non-overlapping supports approach. In contrast, KL divergence causes a vanishing gradient problem. The JS divergence may be non-continuous with parameters.

The results for objective formulation were in agreement with [7]. This approach outperforms the colorization produced by only L_2 or L_1 color loss terms and prevents model collapse.

Learning per-pixel probability distribution allows the use of a variety of classification losses. **Classification Loss.** The $L_s(G^1_{\Psi_1})$ Eq. 3 tries to minimize the difference between generated data with respect to actual distribution, to accomplish that we adapt convolutional layers from VGG-16 model pre-trained on ImageNet dataset. So to make the input competent with the model we create copies of the grayscale channel and reshape the input as (L, L, L) such that it could be used to generate the density distribution as stated earlier.

2.3 Model Architecture

The model architecture is as shown in Fig. 3. It is comprised of 3 sub-parts. Two of those focusing on chrominance information and classification are the same. The last one belongs to the discriminator network which to distinguishes ground truth from synthetic data.

Generator Architecture G_{Ψ}. The generator subnetworks $G^1_{\Psi_1}$ and $G^2_{\Psi_2}$ outputs chrominance related data, $(c) = G^1_{\Psi_1}(L)$ and classification vectors, $y = G^2_{\Psi_2}(L)$. Both the subnetworks are trained jointly through a single step backprop as in proposed in [14].

The basic blocks responsible for global feature extraction are shared by both subnets. Which is initialized with pre-trained ImageNet weights.

The first subnetwork (displayed aqua-green in Fig. 3) proceeds with a form of Convolutional(3×3)-BatchNorm-AReLu to learn $G^1_{\Psi_1}$ and similarly the second subnetwork (displayed purple in Fig. 3) learns the $G^2_{\Psi_2}$ using four modules of form Convolutional(1×1)-BatchNorm-AReLu, sufficed by fully connected layers (in purple) providing us the classification vector.

The first stage results (displayed yellow in Fig. 3) are shared to both the subnetworks and has the same architecture of VGG-16 [17] and are initialized with ImageNet weights [17]. Once $G^1_{\Psi_1}$ is learned it is used to generate useful information to help the colorization process. Later these two subnetworks are fused and are used to predict (c) by up-sampling with the help of Convolution-AReLu layers. The class distribution loss Eq. 3 only subnet $G^2_{\Psi_2}$ is affected. Whereas the color error loss Eq. 1 affects the whole network. We use AReLU [3] as activation for all layers and use a Softmax in the final layer to obtain the logits distribution. The AReLU enables a learnable activation function and formulates an element wise attention mechanism. The attention map forwards scaled positive elements which enable us to capture discriminative features amongst different classes. This amplifies positive elements and suppresses the negative ones.

Discriminator Architecture D_{Λ}. The discriminator is adapted from PatchGAN [7] a markovian discriminator. It operates in a sliding window fashion instead of considering the CXR as a whole to focus on local patches. Instead of classifying a whole image the discriminator convolutionally runs across the image and provides the average of those responses.

3 Experimentation and Results

Quantitatively studying the effect of colorization is a difficult task so in order to analyze how each term in the loss function affects the model, we can perform an ablation study and evaluate the different variants. *GAN* model using adversarial learning and classification approach and *GAN w/o class* [$\lambda_s = 0$]. We also compare the results with EfficientNet backbones and Knowledge Distilled models for endpoint detection and classification with custom loss function as shown in Fig. 4.

3.1 Dataset

We took the RANZCR CLiP dataset which is in-turn built upon Chest X Ray14 dataset [15], one of the largest publicly available datasets, where 30000 labeled (but non- annotated) CXRs spread across 11 different classes as in Table 1 were taken into consideration. A subset of these samples, around 950 were manually

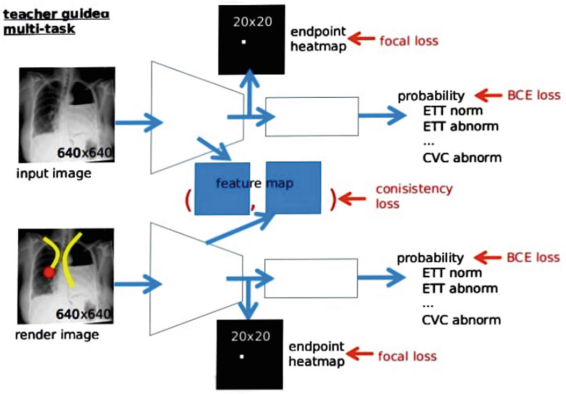

Fig. 4. Teacher - Student model with catheter enpoint heatmap, Source: kaggle

annotated with catheters and tube positions as shown in Fig. 2 to indicate pixels that belong to catheters or tubes. The class wise distribution of the labeled samples are as shown in Table 1. The CXRs are of varied resolutions ($>2048 \times 2048$). We resize these samples to 1024×1024 by means of learned image resizing technique [13]. Replacing the typical linear resizers with learned resizers can substantially improve the results as they produce machine friendly visual manipulations. This resizer model is trained jointly with the classifier subnet model.

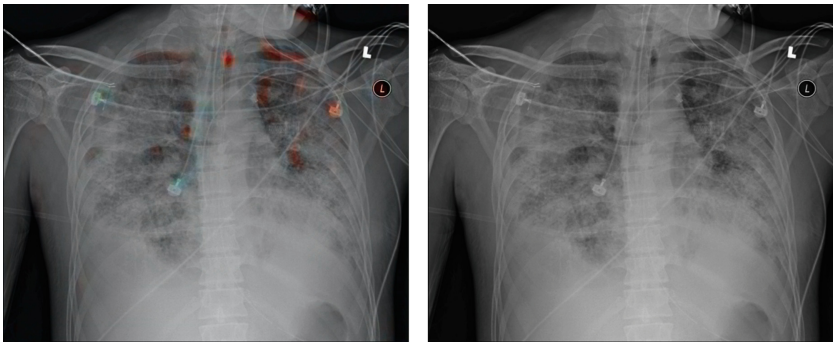

Fig. 5. Segemeted results show the presence of a multi-labeled sample (i.e. ETT - Abnormal, ETT - Borderline positive sample), Image to the left represents unannotated sample and the one on right represents auto-annotation

3.2 Implementation Details

Table 2. Optimal Training Parameters and Hyper-parameters, Note that first and last layers consist of 7×7 kernels. With a larger kernel size allows more receptive fields, it is also verified from [13] and provides about 1–2% increase in accuracy. We also use batch normalization layers followed by AReLU activations [3].

Optimizer	
GAN: AdamW	$1e-5$
Hyperparameter	
$\lambda_g = 0.1$	$\lambda_s = 0.003$
Activation	
GAN: ReLU	$G_{\Psi_2^2}$: AReLU
Kernel	
$n = 16$	3×3
Device: NVIDIA P100	
Batch-Size: 10	

We trained the Annotation - GAN Model with about 950 annotated CXRs resized to 2048×2048 with parameters as in Table 2. A single epoch took about 5 h to train. We had to retain the original image size to make the annotations less ambiguous. The inference phase takes around 2 s Fig. 5. We minimize the objective loss using AdamW optimizer with the parameters as given in Table 2. The intensity regularization makes the model develop a mean color response over the regions of the catheter and tubes. For reliable evaluation of the data, we took a 5-fold cross-validation using F1 scores, pixel-wise precision, and recall metrics Table 3, owing to the thin structures of the catheters and tubes we enlarged the manually annotated samples to a 5-pixel dilation radius which resulted in 70% cases with over 50% of overlap.

Metrics such as Structural Similarity Index (SSIM) check for variation in contrast, luminance in high frequency region. Similarly, the Hausdorff distance measures the closeness to the ground truth thus representing the resemblance amongst the images.

Table 3. Evaluation of Auto-Annotated Samples with respect to ground truth annotation. Note these are the results of auto-annotated data averaged over a batchsize of 10.

Data	SSIM	PSNR	Hausdorf
Auto - Annotated CXR	0.98184	28.5766	4 ± 0.3
GANw/oclass	0.99411	33.9304	4 ± 0.5

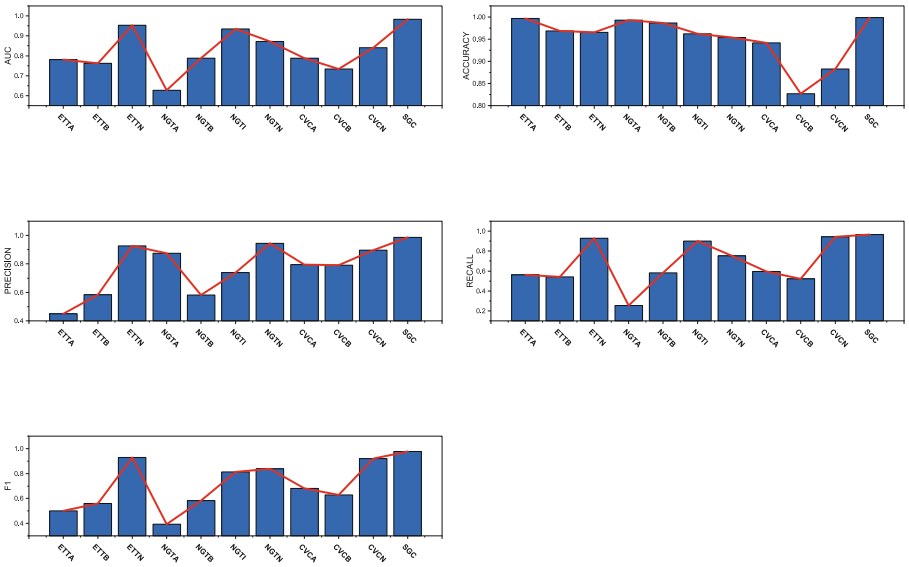

Fig. 6. Results of the Annotation GAN for each class, Mean values of the same are as shown in Table 4

3.3 Limitations

While the model offers quick inference and has a higher potential to improve workflow efficiency, there are a few drawbacks. The system confuses the catheters with tips with bone edges, often around the ribs. This can be avoided by adding additional background annotation classes (i.e., ribs, collar bones) and retraining the network so that the model will learn a differential feature to separate out the various classes.

The class imbalance can be mitigated via class frequency weights as an additional loss term. Further data augmentation techniques can be studied to provide better initializations to the model and prevent the mixing of background with labeled data.

3.4 Future Scope and Direction

Chest radiography analysis is an important stage in post-surgical care it involves detection and identification of tip locations, endpoints of catheters. The proposed model is capable of detecting tip locations of various different catheters and the inference takes about 2 s this could accelerate the mundane but essential tasks required in the health care system.

Our initial target was to detect catheter endpoints and reduce the false negatives, such as to make the solution viable in real-time applications. However, as a part of distinguishing the catheters from one another, it will work to our

Table 4. Results of the catheter classification (Mean ± standard deviation) AUC: Area under ROC, Acc: Accuracy, P: Precision, R: Recall). MTSS: Multi-Teacher Single Student trained and distilled network model with EfficientNet backbone.

Model	AUC	Accuracy	P	R	F1
Resnet200D*	0.884 ± 0.055	0.822	0.33	0.27	0.61
EfficientNet* - B0*	0.883 ± 0.055	0.732	0.4	0.39	0.62
EfficientNet - MTSS	0.917 ± 0.005	0.903	0.50	0.55	0.65
MoCo [5]	0.815 ± 0.5	0.711	0.56	0.58	0.51
AnnotationGAN	0.969 ± 0.005	0.952	0.77	0.68	0.73

benefit if we could capture all the tubular structures present in the given sample. This also aids the clinical setting without having to redesign the system to incorporate the model. The current approach involves analyzing the CXR from a single viewpoint. It could benefit us to analyze both frontal and lateral views for detecting any abnormalities.

4 Conclusion

We propose an autonomous report generation system that can autonomously classify and identify unlabeled samples and provide reports for central and peripheral catheter positions (Table 4, Fig. 6). Experiments show that our end-to-end model performs equivalent to the state-of-the-art models even though it was trained 0.033% (950 annotated samples) of labeled data. Auto-annotation GAN can further be explored with the help of shape constraints and incorporating spatial priors to improvise the results. Future prospects include autonomous report generation and validation by various trials, which results in robust reporting with minimal costs which is well under regulatory requirements. However, the proposed approach could be expanded scope for similar tasks in ophthalmology, 3D volume colorization, motion forecasting or satellite image analysis, and beyond.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. arXiv abs/1701.07875 (2017)
2. Bailo, O., Ham, D., Shin, Y.M.: Red blood cell image generation for data augmentation using conditional generative adversarial networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1039–1048 (2019). https://doi.org/10.1109/CVPRW.2019.00136
3. Chen, D., Xu, K.: Arelu: attention-based rectified linear unit. arXiv abs/2006.13858 (2020)

4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
5. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning (2020)
6. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976 (2017). https://doi.org/10.1109/CVPR.2017.632
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
8. Lei, Y., et al.: Male pelvic CT multi-organ segmentation using synthetic MRI-aided dual pyramid networks. Phys. Med. Biol. **66**(8), 085007 (2021). https://doi.org/10.1088/1361-6560/abf2f9
9. Nazeri, K., Ng, E., Ebrahimi, M.: Image colorization using generative adversarial networks. In: Perales, F.J., Kittler, J. (eds.) AMDO 2018. LNCS, vol. 10945, pp. 85–94. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-94544-6_9
10. Neff, T.: Data augmentation in deep learning using generative adversarial networks. Ph.D. thesis, Master's thesis, Graz University of Technology (2018). https://www.tugraz.at
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2015)
12. Subramanian, V., Wang, H., Wu, J.T., Wong, K.C.L., Sharma, A., Syeda-Mahmood, T.: Automated detection and type classification of central venous catheters in chest X-rays. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 522–530. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_58
13. Talebi, H., Milanfar, P.: Learning to resize images for computer vision tasks (2021)
14. Vitoria, P., Raad, L., Ballester, C.: Chromagan: adversarial picture colorization with semantic class distribution. In: The IEEE Winter Conference on Applications of Computer Vision, pp. 2445–2454 (2020)
15. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.: Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3462–3471 (2017)
16. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM Trans. Graph. (TOG) **9**(4) (2017)
17. Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5219–5227 (2017). https://doi.org/10.1109/ICCV.2017.557