2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428076

SAMPLE EFFICIENT LUNG SEGMENTATION USING GROUP STRUCTURED CONDITIONAL VARIATIONAL DATA IMPUTATION

Yan Li[1], *Guitao Cao*[1,*], *Wenming Cao*[2,†]

[1]MOE Research Center for Software/Hardware Co-Design Engineering, East China Normal University
[2]College of Information Engineering, Shenzhen University
52194501020@stu.ecnu.edu.cn, gtcao@sei.ecnu.edu.cn, wmcao@szu.edu.cn

ABSTRACT

Patients infected with COVID-19 can lead to their Chest X-rays (CXRs) with opacifications rendered regions, which may produce incomplete lung segmentation in automated image analysis models. To tackle this issue, we propose a *Group structured Conditional Variational data Imputation* model to capture the missing data accurately with conditional distribution, where the high-dimensional probability distribution is narrow down to a small latent space to account for unobserved features. This work particularly arises in the fight against COVID-19 that effectively modeling a segmentation of plausible can be presented to a subsequent automated risk scoring and treatment. We train this model with limited CXRs data to demonstrate the abilities on the task of data imputation and proved to be effective though with relatively small datasets.

Index Terms— COVID-19, Chest X-rays, Lung Segmentation, Data Imputation, Group-CNNs, cVAEs

1. INTRODUCTION

The coronavirus COVID-19 continues to spread around the world and poses a huge challenge to the strained medical facilities [1], due to high contagiousness and severe respiratory complications. So far, nucleic acid test is the most important tool for the diagnosis of patients at present, but there are many problems, such as too few medical institutions in the epidemic area meeting the requirements of nucleic acid test, an insufficient supply of nucleic acid test boxes, too long nucleic acid test approval process, too low detection rate, leading to a large number of false negatives. Under these conditions, Chest X-rays (CXRs) and computed tomography (CT) provide a non-invasive tool to monitor the evolution of the disease, and play an essential role in triage of COVID-19 patients and allocation of hospital resources [2]. In particular, the CXRs are relatively easier, cheaper and faster obtained than CT, even in emergency settings. As such, CXRs systems have become a part of standard procedure for COVID-19 early findings [2]. Recently, many researchers attempted to exploit automatically CXRs-based AI diagnosis models, for example, Tartaglione *et al*. [3] provide a methodological guide on what is reasonable to expect by applying deep learning to COVID-19 classification. Further, Cohen *et al*. [4] build a model that predicts the severity of COVID-19 pneumonia, especially important in escalation or de-escalation of care as well as monitoring treatment efficacy. Signoroni *et al*. [5] design an end-to-end semi-quantitative scoring system based on multi-network deep learning architecture, and show significant value in one of the hospitals that experienced one of the highest pandemic peaks in Italy.

Building these models usually involves four stages: data pre-processing, lung segmentation, feature extraction, and final prediction. Among these stages, lung segmentation is one of the most important steps to obtain accurate AI diagnosis models. Several approaches have been proposed for lung segmentation tasks, such as encoder-decoder architectures U-Net [6] and fully convolutional FCNs [7]. These approaches are well suited to capture fine-grained details of lung tissue. However, most segmentation methods only provide pixel-wise probabilities that ignore all co-variance between pixels, which makes extreme levels of opacification obfuscate lung segmentation much more difficult or even impossible.

A body of work with different approaches towards segmenting lungs from high opacity CXRs. Souza *et al.* [8] built two stages cascaded CNNs models to tackle opacity: initial lung segmentation model and reconstruction model. The training of these two models, however, requires a very complex and heterogeneous dataset that contains exams with a large variety of lung abnormalities, which can be challenging to collect abnormal in realistic scenarios. To address the lack of abnormal cases, Tang *et al.* [9] propose a data augmentation strategy using an image-to-image translation to construct a large number of abnormal cases. This method does offer a useful synthetic scan over the pixel-space, however the diversity of opacifications are limited by the training samples. Selvan *et al.* [10] address the above shortcomings by using

*Guitao Cao is corresponding author, †Wenming Cao is co-corresponding author. This work was supported in part by the National Natural Science Foundation of China under Grant 61871186 and 61771322, and in part by the Shanghai Natural Science Foundation under Grant 18ZR1411400.

deep latent variable generative models to learn a distribution, which is then jointly decoder to obtain the segmentation, however this method also relies on specialized data augmentation similar to [9].

In this work, we tackle these challenges by introducing a group-structured generative U-Net for lung segmentation that utilizes the group convolutional network to facilitate the effective representation learning of the data distribution, and allows the extraction of reliable unobserved information. Our key contributions are as follows:

- We propose an expressive generative model that can be conditioned on observed features to restore missing data in segmentation maps.
- We build the generative models via group convolution neural networks to improve the data efficiency when only limited training data is available.
- The proposed model can combine with downstream tasks such as disease classification and detection, to provide a comprehensive and accurate computer-aided model on CXRs images.

2. BACKGROUND

For most image segmentation tasks, convolutional neural networks (CNNs) have become default backbones, however, CNNs typically require a large of labeled data to train on, which often difficult to obtain in the medical domain. So, how to train our model on a limited dataset is the first problem, especially the COVID-19 datasets [11], up to date, are still limited. Recent work by Bekkers [12] propose a Group Convolutional Neural Networks (G-CNNs) that can be used to improve data efficiency of vanilla CNNs by equipping them with the geometric structure of groups. These networks have shown considerable gains in terms of performance and speed of convergence compared to regular CNNs, but have not been extended to the segmentation domain. In this paper, we leverage group convolutions to replace the default backbones of segmentation architecture with G-CNNs, so that it can learn more representative features under the limited data condition.

The second question is how to segment lungs from high opacity CXRs, since CXRs exams in some COVID-19 patients present regions of high opacification. CXRs with such rendered regions, making it difficult to perform automated segmentation tasks on them. In this work, we try to infer the missing data in high opacity regions by building a conditional generative model to learn a distribution over a latent space, and then sampling from this distribution and joint a decoder we can obtain a more complete lung segmentation.

3. PRELIMINARIES

3.1. Group Theory

Definition 3.1 (Group). A group G is a set, equipped with an associative binary operator $\cdot : g \times g \to g$, having the following properties: $\cdot$ is associative, has an identity element e, for each $g \in G$ there exists an inverse element $g^{-1} \in G$.

Definition 3.2 (Group action). Let G be a group, $\mathcal{X}$ be a set, we say a map $\rho : G \to Aut(\mathcal{X})$ is a group action, which maps each $g \in G$ to a corresponding transformation on $\mathcal{X}$, where ρ is a homomorphism, satisfying $\rho(gv) = \rho(g)\rho(v)$ for all $g, v \in G$ and $Aut(\mathcal{X})$ is a bijective endomorphism. We use $gx := \rho(g)x$ for any $x \in X$.

Definition 3.3 (Group representation). Let G is a group, there exists a set of invertible linear maps $\mathbb{L}_2(V)$ on some vector spacce V such that for each $g \in G$ we has $\mathcal{L}_g^{G \to \mathbb{L}_2(V)} : \mathbb{L}_2(V) \to \mathbb{L}_2(V)$, this map is a group action on vector space, here we call $\mathcal{L}_g^{G \to \mathbb{L}_2(V)}$ a regular group representation.

Definition 3.4 (Group equivariance). Let us consider G be a group and V_1, V_2 be a vector spaces, there exist some mapping $\Psi : \mathbb{L}_2(V_1) \to \mathbb{L}_2(V_2)$ to commute with the group actions on the domain $\mathbb{L}_2(V_1)$ and codomain $\mathbb{L}_2(V_2)$. It is group equivariant under the actions of g if and only if there exists a $\mathcal{L}_g^{G \to GL(\cdot)}$ such that

$$\forall_{g \in G} : \mathcal{L}_g^{G \to \mathbb{L}_2(V_2)} \circ \Psi = \Psi \circ \mathcal{L}_g^{G \to \mathbb{L}_2(V_1)} \tag{2}$$

Similar to CNNs, which keep translation equivariant in each layer, we also want to preserve equivariance in each layer but not just translation equivariant. To achieve this, one can define a group convolution between two functions over the group, which is described in the following theorem.

Theorem 3.1. *Let $\mathcal{K}$ be a bounded equivariant operator, $\kappa : G \times G \to Hom(X, Y)$ be a two-argument linear operator-valued kernel, and a Radon measure $d\mu$ on X, then*

1. *An equivariant operator can always be written as a convolution-like integral $\mathcal{K}f(y) = \int_x \kappa(y, x) f(x) d\mu_x$;*
2. *One can define a one-argument kernel using the above equivariance constraints of Equ.* (2)

$$\kappa(y, x) = \frac{d\mu_{g_y^{-1}x}}{d\mu_x} k(g_y^{-1}x) \tag{3}$$

Corollary 3.1. *If $X = G$ and $d\mu_x$ is a Haar measure on G then $\frac{d\mu_{g_y^{-1}x}}{d\mu_x} = 1$.*

Proofs can be found in [12].

3.2. Group Convolutional Network Layers

Under the Thm 1 on groups, we define three types of convolution operators.

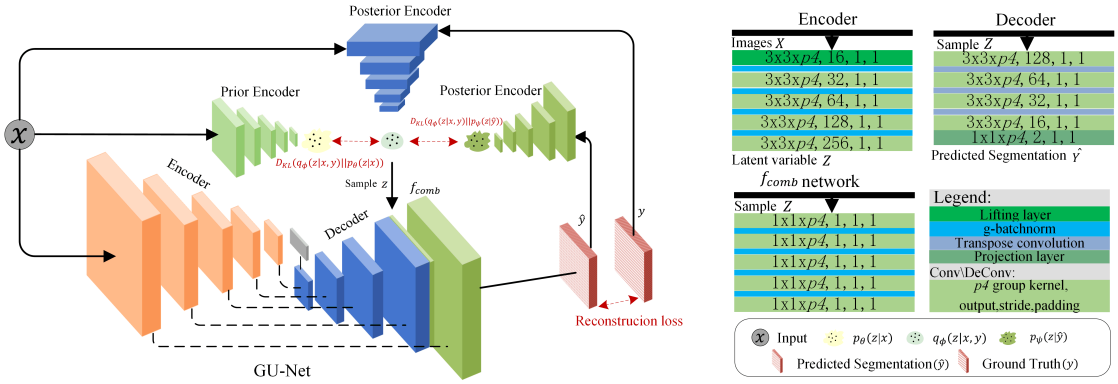

Fig. 1. Our framework consists of group structured prior encoder, GU-Net and two group structured posterior encoders.

- **The lifting layer** ($\mathbb{R}^d \to G$): In this layer, we lift $X = \mathbb{R}^d$ image data to $G = \mathbb{R}^c \rtimes H$ space. Thus, we define the lifting convolution $\mathcal{K}$ as follows:

$$(\mathcal{K}f)(g) = (k\tilde{\star}f)(g) := \frac{1}{|\det h|}(\mathcal{L}_y^{G\to\mathbb{L}_2(\mathbb{R}^c)}k, f)_{\mathbb{L}_2(\mathbb{R}^c,dx)} = \int_G k(g^{-1}(x'-x))f(x')dx' \tag{4}$$

- **Group convolution layer** ($G \to G$): At this layer, the in- and out- feature fields all in G space, we arrive at the following form of our group convolution:

$$(\mathcal{K}F)(g) = (K\tilde{\star}F)(g) := (\mathcal{L}_g^{G\to\mathbb{L}_2(\mathbb{R}^c)}[K], F)_{\mathbb{L}_2(\mathbb{R}^c,d\mu)} = \int_G K(g^{-1}\tilde{g})F(\tilde{g})d\mu_{\tilde{g}} \tag{5}$$

- **Projection layer** ($G \to \mathbb{R}^d$): In this layer, we project the data on $G = \mathbb{R}^c \rtimes H$ back to $X = \mathbb{R}^d$ via

$$(\mathcal{K}F)(g) = (k_h\tilde{\star}F)(x) = \int_H F(x,\tilde{h})d_{\mu_{\tilde{h}}} \tag{6}$$

These layers can be embedded in any excellent architecture by simply replacing standard convolutions with group convolutions. In this paper, we leverage *p4*-group (SE(2,N) with N=4) group to build our *p4*-group convolutional networks.

3.3. VAEs

Recent efforts (e.g., [13], [10]) have shown that generative model, variational auto-encoders(VAEs) [14] in particular, can achieve superior performance on performing data imputation. In general, VAEs learn a generative distribution by maximizing a variational lower bound on the log-likelihood $\log p(x)$:

$$L_{\text{VAE}}(x;\theta,\phi) := \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\text{KL}}(q_\phi(z|x)||p(z)) \leq \log p(x) \tag{7}$$

where the approximate posterior $q_\phi(z|x)$ is modeled by a CNN encoder, $p_\theta(x|z)$ is modeled by a decoder CNNs, the Kullback-Leibler divergence (also called Relative Entropy) term D_{KL} essentially measures how different the distribution of the true and the approximate posterior [14]. Looking closely at this model, we could see that the encoder models the latent variable z based on x, which doesn't take any account of some additional conditioning input information. This may fail to model complex structured representation that effectively infers high opacity regions of CXRs, especially in the extremely low data regimes.

To address this issue, we introduce a *Group structured Conditional Variational Data Imputation* (G-CVDI) model, which deals with incomplete data by modeling a group structured generative model.

4. GROUP STRUCTURED CONDITIONAL VARIATIONAL DATA IMPUTATION

To learn a complete segmentation automatically, we aim to model a distribution of possible segmentation given the ground-truth segmentation masks. This observed features ensures that the encoder generates an unobserved segmentation map according to the true underlying data distribution. The generative process of our model is similar to the generative process of cVAE [15] by conditioning on an observation. However, the naive cVAE failed to learn structured latent variables without constraining the learned representation, which may limit the performance of the encoder and decoder.

In this work, we leverage group convolution operations as a new way to give cVAEs the structures of *p4*-group as mentioned above and propose a new architecture to learn

an equivariant conditional density model over segmentations given on the opacity image. Specifically, the central component of our architecture is illustrated in Figure 1. There are three types of networks in our deep conditional generative model: a group structured prior net, group structured U-Net (GU-Net) and group structured posterior net. The group structured prior net generates low-dimensional embedding sample $z \sim q_\phi(z|x)$ for given a input images x, and then sample is concatenated to last activation map of a group structured U-Net by a combination function f_{comb}, this is, the output (eg. segmentation map) corresponding to the given sample is $\widehat{y} = f_{\text{comb}}(f_{\text{GU-Net}}(x; w), z)$. To find a useful embedding space for data imputation in the training stage, we first introduce a posterior net that learns a posterior distribution $q_\phi(z|x, y)$ by conditioning the input x and the ground-truth segmentation masks y as a reference axis to calibrate the prior distribution. In order to reflect the difference between sampling z from prior and posterior distribution, a Kullback-Leibler divergence $D_{\text{KL}}(q_\phi(z|x, y)||p_\theta(z|x))$ is adopted to estimate the similarity between the two distributions. From an information-theoretic perspective, this D_{KL} term encourages the latent bottleneck z to efficiently transmit information from x and y, this is,

$$
\begin{aligned}
&\mathbb{E}_{p_D(x,y)}[D_{\text{KL}}(q_\phi(z|x,y)||p_\theta(z|x))] \\
&\triangleq \mathbb{E}_{p_D(x,y)}\mathbb{E}_{q_\phi(z|x,y)}[\log q_\phi(z|x,y) - \log p_\theta(z|x)] \\
&= \sum_{x,y} q(x,y,z) \log \frac{q(x,y,z)}{p(x,y,z)} \\
&= \mathbb{E}_{x,y,z} \log \frac{q(x,y|z)q(z)}{p(x,y)p(z)} \\
&= \mathbb{E}_{q_\phi(z)}\mathbb{E}_{q_\phi(x,y|z)} \log \frac{q(x,y|z)q(z)}{p(x,y)p(z)} + \mathbb{E}_{q_\phi(x,y,z)} \frac{q(z)}{p(z)} \\
&= \underbrace{H(x,y) - H(x,y|z)}_{\triangleq I_{q_\phi(z;x,y)}} + \underbrace{\mathbb{E}_{q_\phi(z)} \frac{q_\phi(z)}{p(z)}}_{\triangleq D_{\text{KL}}(q_\phi(z)||p(z))}
\end{aligned}
$$

As the $\triangleq I_{q_\phi(z;x,y)}$ is non-negative in the last line, we could see that minimizing all the relative entropy $D_{\text{KL}}(q_\phi(z|x, y)||p_\theta(z|x))$ is equivalent to match the aggregated posterior distribution of the latent variable $q_\phi(z)$ to the aggregated prior distribution $p(z)$. By doing so, it is more likely to contain the unobserved features of the segmentation maps when sampling a z from the given posterior $q_\phi(z|x, y)$. And then, combining the sample z with the last activation map of the group structured U-Net by f_{comb} function can predict a high-quality segmentation $\widehat{y}$. We use negative expected log-likelihood,

$$
\mathcal{L}_{\text{rec}}(y, \widehat{y}; \theta, \phi) := \mathbb{E}_{z\sim q_\phi(z|x,y)} [\log p_\theta(y|f_{\text{comb}}(x, z))] \quad (9)
$$

as a reconstruction loss to penalize difference between the prediction $\widehat{y}$ and y. To combine the $D_{\text{KL}}(q_\phi(z|x, y)||p(z|x))$ with $\mathcal{L}_{rec}(y, \widehat{y})$, we get the following objective function:

$$
\begin{aligned}
\mathcal{L}(x, y; \theta, \phi) := &\mathbb{E}_{z\sim q_\phi(z|x,y)} [\log p_\theta(y|f_{\text{comb}}(x, z))] \\
&- D_{\text{KL}}(q_\phi(z|x, y)||p_\theta(z|x))
\end{aligned} \quad (10)
$$

In order to encourage embedding space of the predicted segmentation $\widehat{y}$ on a set of representational axes (e.g, close in data space of $q_\phi(z|x, y)$), we introduce another posterior net that model variational posterior $p_\psi(z|\widehat{y})$ to get close to a shared coding space (a set of representational axes). Following the above, we use relative entropy $D_{\text{KL}}(q_\phi(z|x, y)||p_\psi(z|\widehat{y}))$ to characterize the degree of overlap between the posterior distribution across the shared coding space, which will tend to result in the prediction towards ground-truth. Adding this term to the above objective function we get a weighted mixture loss function,

$$
\begin{aligned}
\mathcal{L}_{\text{G-CVDI}}(x, y; \theta, \psi, \phi) := &\mathbb{E}_{z\sim q_\phi(z|x,y)} [\log p_\theta(y|f_{\text{comb}}(x, z))] \\
&- \alpha \cdot D_{\text{KL}}(q_\phi(z|x, y)||p_\psi(z|\widehat{y}) \\
&- \beta \cdot D_{\text{KL}}(q_\phi(z|x, y)||p_\theta(z|x)))
\end{aligned} \quad (11)
$$

where α and β balance the objections, note that when $\beta = 1$ and $\alpha = 0$, we recover the cVAEs objective, when $\alpha = 0$, the objective is similar to the conditional β-TCVAE [16].

For all the network as mentioned above, the equivariant are guaranteed by repace the 2D-convolution operations with *p4*-group convolutions (e.g., the lifting layer as the first layer of networks to produce *p4*-feature maps, the group convolutions layer act on the group structured feature map, the last layer of the networks is usually a projection layer to output a 2D image). But we haven't discussed how to ensure the equivariance of sampling a sample from a distribution. In other words, we want the posterior to keep *invariant* under the deterministic action of the group.

Specifically, let G be a finite group which acts on the image $\mathbb{R}^2$ via a left-regular representation $\mathcal{L}_g^{G\to\mathbb{L}_2(\mathbb{R}^2)} : \mathbb{L}_2(\mathbb{R}^2) \to \mathbb{L}_2(\mathbb{R}^2)$, we consider the densities ρ of a distribution, which is assumed to be invariant w.r.t. to some symmetry transformation e.g., $C_4, SE(2, N)$. This allows us to construct equivariance very naturally:

Theorem 4.1. *Let ρ be a G-invariant density on $\mathbb{R}^2$, if $\forall_{x\in\mathbb{R}^2, g\in G} : f(\mathcal{L}_g^{G\to\mathbb{L}_2(\mathbb{R}^2)}x) = \mathcal{L}_g^{G\to\mathbb{L}_2(\mathbb{R}^2)}f(x)$, then f is equivariant to transformations sampled from ρ:*

Proof. *Supposing that G acts linearly on the image $x = \mathbb{R}^2$, and let $\mathcal{L}_g^{G\to\mathbb{L}_2(\mathbb{R}^2)} : \mathbb{L}_2(\mathbb{R}^2) \to \mathbb{L}_2(\mathbb{R}^2)$ be a representation of G over x, f be a function that is equivariant to some symmetry transformation g, then the model f is equivariant to transformations sampled from ρ:*

$$
\begin{aligned}
f(hx) = \mathbb{E}_{g\sim\rho} g^{-1} f(ghx) &= \mathbb{E}_{g\sim\rho} h(gh)^{-1} f(ghx) \\
&= h\mathbb{E}_{u\sim\mu} u^{-1} f(ux) \\
&= hf(x)
\end{aligned}
$$

where $u = gh$ and for any measurable set S, $\forall_{g\in G} : \mu(S) = \mu(gS)$ according to the corollary 3.1, this is, ρ is an G-invaraint under some transformation $\mathcal{L}_g^{G\to\mathbb{L}_2(\mathbb{R}^2)}$, e.g., we can assume that the transformation is uniform then the ρ is invariant the actions of group. So the desired posterior is being *invariant* under a *left-regular representation* $\mathcal{L}_g^{G\to\mathbb{L}_2(\mathbb{R}^2)}$:

$$q_\phi(z|\mathcal{L}_g^{G\to\mathbb{L}_2(\mathbb{R}^2)}[x], \mathcal{L}_g^{G\to\mathbb{L}_2(\mathbb{R}^2)}[y]) = q_\phi(z|x, y); \forall_{g\in G}$$

$$q_\phi(z|\mathcal{L}_g^{G\to\mathbb{L}_2(\mathbb{R}^2)}[x]) = q_\phi(z|x); \forall_{g\in G}$$

therefore the objective of our model stays in its original form.

5. EXPERIMENTS

In this section, we conduct experiments to verify the effectiveness of our proposed method. Specifically, we want to answer the above two questions: (i) how good is the data efficiency or the rate of convergence of our model ? and (ii) how good is the segmentation in lung with opacity regions.

5.1. Dataset and Evaluation Metrics

To train and compare the models investigated in this study, we use publicly available COVID-19 datasets, currently totaling 679 frontal chest X-ray images from 412 people from 26 countries. But only 517 images contain lung masks, we filtered 425 CXRs as our training/validation set, and following the same experimental setup in [10], each input images is rescaled to 640x512px and mitigate the grayscale variability in the images by applying a histogram equalization, and another publicly available CXRs datasets from Shenzhen and Montgomery hospitals [10], this dataset contains 704 CXRs images, we use 493 for training, 211 for testing. We use dice similarity coefficient and binary accuracy as the evaluation metric, and make all experiments in triplicate and report the results as mean and standard deviation.

5.2. Results

5.2.1. *Data efficiency and Rate of convergence*

To verify the data efficiency and rate of convergence, we compared our model (G-CVID) with a conventional $\mathbb{Z}^2$-CVID (replacing the group convolution with a standard 2D convolution). Table 1 shows the Dice Overlap for standard translation $\mathbb{Z}^2$-CVID baseline and our G-CVID with different training set sizes N. Overall, we observe that our method is capable of obtaining a better dice accuracy over multiple training set sizes when compared to the standard CNNs baseline. Furthermore, we plot the training loss per epoch. For training runs with dataset 50,100,150 and 200 in Figure 2, we can find that our proposed model show a faster decline than regular $\mathbb{Z}^2$-CVID. These characteristics of data efficiency and faster convergence can be attributed to the fact that each gradient signal comes from multiple *p4* feature maps.

Table 1. Overall score for all training set sizes N.

N	$\mathbb{Z}^2$-CVDI	G-CVDI
50	0.8659±0.07	0.8995±0.09
100	0.8971±0.10	0.9307±0.13
150	0.9137±0.12	0.9358±0.08
200	0.9259±0.09	0.9470±0.11

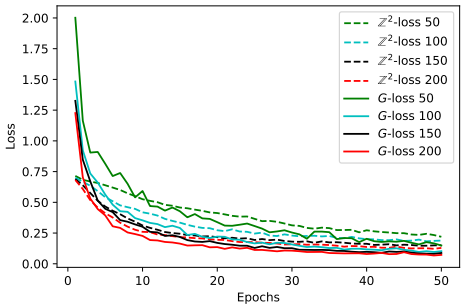

Fig. 2. Learning curves for the all networks trained on different training set sizes N.

5.2.2. *Segmentation in lung with opacity regions*

We demonstrate the effectiveness of our approach in lung segmentation with opacity regions. First, we train all models on the CXRs dataset to verify the segmentation ability when the data is low opacification. Second, we chose the COVID-19 dataset as the second dataset to verify the performance when data with extreme levels of opacification obfuscate regions in the lungs. Table 2 shows the result on CXRs and COVID-19 datasets. The proposed model outperforms all the baseline models, this is due to the fact that the information contained in the observed data is aligned with the label in the conditional latent space, which provides sufficient information to predict more complete segmentation.

Figure 3 shows a qualitative example, it can be seen that our model captures well the ground truth compared with other models. Intuitively, this further demonstrates the effectiveness of our model.

6. CONCLUSIONS

In this study, we propose a group structured generative model that conditions on a set of observed features, allowing for latent representations that contain more mutual information across group structural latent variables. The learned latent representations are used to generate the missing data. We show that our model unambiguously outperformed the baseline CNN on the lung segmentation, especially on small datasets, without any further tuning, which is particularly im-

Table 2. Performances on two datasets(*Aug($+D$, $+D+B$) denote diffusion and diffusion+block augementation respectively).

CXRs			COVID-19		
Models	**Dice Overlap**	**Accuracy**	**Models**	**Dice Overlap**	**Accuracy**
U-Net [6]	0.9578±0.04	0.9733±0.08	U-Net [6]	0.9481±0.07	0.9564±0.02
XLSor$_R$ [9]	0.9551±0.07	0.9724±0.09	XLSor$_R$ [9]	0.9302±0.06	0.9405±0.05
XLSor$_{RA}$ [9]	0.9579±0.04	0.9740±0.03	XLSor$_{RA}$ [9]	0.9472±0.07	0.9515±0.06
VID [10]	0.9573±0.11	0.9730±0.04	VID [10]	0.9432±0.05	0.9531±0.04
VID (Aug+D) [10]	0.9555±0.10	0.9722±0.08	VID (Aug+D) [10]	0.9488±0.04	0.9571±0.07
VID (Aug+$D+B$) [10]	0.9434±0.03	0.9663±0.08	VID (Aug+$D+B$) [10]	0.9343±0.04	0.9473±0.09
$\mathbb{Z}^2$-CVDI(ours)	0.9585±0.05	0.9734±0.09	$\mathbb{Z}^2$-CVDI(ours)	0.9399±0.03	0.9519±0.03
G-CVDI(ours)	**0.9611±0.07**	**0.9747±0.02**	G-CVDI(ours)	**0.9496±0.05**	**0.9578±0.06**

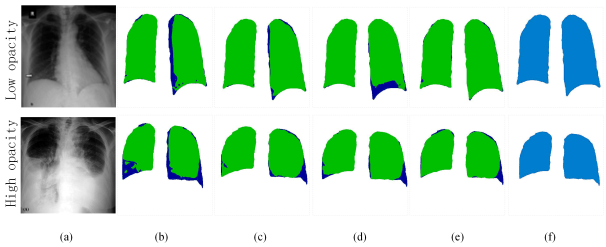

Fig. 3. (a) Two test samples with low and high opacity. (b) U-Net prediction. (c) XLSor$_{RA}$ prediction. (d) VID prediction. (e) our model prediction. (f) ground truth; Green:True postive, Blue: False Negative.

portant in reducing the collection of medical data. We hope that our model can be used as a basic segmentation component to accelerate the development of a highly accurate yet high quality solution for detecting COVID-19 cases from CXRs images.

7. REFERENCES

[1] UNDP, "Coronavirus disease covid-19 pandemic," `http://www.undp.org/content/undp/en/home/coronavirus.html`, 2020.

[2] Adam Jacobi, Michael Chung, et al., "Portable chest x-ray in coronavirus disease-19 (covid-19): A pictorial review," *Clinical Imaging*, 2020.

[3] Enzo Tartaglione, Carlo Alberto Barbano, et al., "Unveiling covid-19 from chest x-ray with deep learning: a hurdles race with small data," *arXiv preprint arXiv:2004.05405*, 2020.

[4] Joseph Paul Cohen, Lan Dao, et al., "Predicting covid-19 pneumonia severity on chest x-ray with deep learning," *arXiv preprint arXiv:2005.11856*, 2020.

[5] Alberto Signoroni, Mattia Savardi, et al., "End-to-end learning for semiquantitative rating of covid-19 severity on chest x-rays," *arXiv preprint arXiv:2006.04603*, 2020.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[7] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[8] Johnatan Carvalho Souza, João Otávio Bandeira Diniz, Jonnison Lima Ferreira, et al., "An automatic method for lung segmentation and reconstruction in chest x-ray using deep neural networks," *Computer methods and programs in biomedicine*, vol. 177, pp. 285–296, 2019.

[9] Youbao Tang, Yuxing Tang, Jing Xiao, et al., "Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation," *arXiv preprint arXiv:1904.09229*, 2019.

[10] Raghavendra Selvan, Erik B Dam, et al., "Lung segmentation from chest x-rays using variational data imputation," *arXiv preprint arXiv:2005.10052*, 2020.

[11] Joseph Paul Cohen, Paul Morrison, et al., "Covid-19 image data collection," *arXiv 2003.11597*, 2020.

[12] Erik J Bekkers, "B-spline cnns on lie groups," *arXiv preprint arXiv:1909.12057*, 2019.

[13] Alfredo Nazabal, Pablo M Olmos, et al., "Handling incomplete heterogeneous data using vaes," *Pattern Recognition*, p. 107501, 2020.

[14] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[15] Kihyuk Sohn and et al. Lee, Honglak, "Learning structured output representation using deep conditional generative models," in *NeurIPS*, 2015, pp. 3483–3491.

[16] Ricky TQ Chen, Xuechen Li, et al., "Isolating sources of disentanglement in variational autoencoders," in *NeurIPS*, 2018, pp. 2610–2620.