Strengths and Limitations of Normalizing Flows in Geospatial Data Modeling

Mathias Niemann Tygesen [1] Daniele Gammelli [1] Sergio Garrido [1] Filipe Rodrigues [1]

Abstract

Geospatial density estimation problems often exhibit complex and highly multi-modal structures. In this work, we explore the use of Normalizing Flows (NFs) in such geospatial settings. NFs offer a principled approach to represent rich and flexible distributions and in this work we experiment with conditional affine coupling layers and planar transformations layers, showing their strengths and highlighting their limitations. We also experiment with a proposed mixed flow showing how using both kind of layers can increase the flexibility of the flow.

1. Introduction

Much of the research in Normalizing Flows (NF) focuses on creating new and expressive flows for a high-dimensional setting. However many real-world problems such as geospatial density estimation are not high-dimensional. As such we are interested in analyzing how the NFs developed for high-dimensional problems work in lower dimensional settings.

We experiment with NFs based on conditional affine couplings and conditional planar transformations, highlight some of their limitations in the low dimensional setting and show how combining affine coupling layers and planar transformations can remedy this.

2. Background

2.1. Normalizing Flows

Normalizing flows model a distribution by transforming a simple, known distribution through a series of transformations. As long as the transformations are differentiable and have differentiable inverses the change in probability can be calculated via a change of variable. Let T be a transformation with inverse T^{-1} such that T and T^{-1} are both differentiable and that $\boldsymbol{u} = T^{-1}(\boldsymbol{x})$, where $\boldsymbol{u}$ comes from some known distribution $p_u(\boldsymbol{u})$. Then the distribution of $\boldsymbol{x}$ can be found by $p_x(\boldsymbol{x}) = p_u(\boldsymbol{u})|\det \boldsymbol{J}_T(\boldsymbol{u})|^{-1}$ If the flow consists of multiple transformations this can be extended using the property that the transformations are composable. For the case of K transformations the modelled distribution becomes

$$p_x(\boldsymbol{x}) = p_u(\boldsymbol{u}) \prod_{k=1}^{K} |\det \boldsymbol{J}_{T_k}(\boldsymbol{z}_{k-1})|^{-1}, \tag{1}$$

where $\boldsymbol{z}_k$ is the value after the k'th transformation with $\boldsymbol{z}_0 = \boldsymbol{u}$ and $\boldsymbol{z}_K = \boldsymbol{x}$. Often these transformations are defined as a function $f_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$.

Two common transformations are the affine coupling layers (Dinh et al., 2016) and the planar transformations (Rezende & Mohamed, 2015). The affine coupling layer uses an autoregressive structure where half of the dimensions in the input are used to condition the other half. Let $\boldsymbol{x}$ denote the input to the transformation and $\boldsymbol{x'}$ denote the output of the transformation. The transformation is then

$$\begin{aligned} \boldsymbol{x'}_{<d} &= \boldsymbol{x}_{<d} \\ (\boldsymbol{h}_{d+1}, \ldots, \boldsymbol{h}_D) &= c(\boldsymbol{x}_{<d}) \\ x'_i &= \tau(x_i; \boldsymbol{h}_i) \;\; i > d, \end{aligned} \tag{2}$$

where τ is an affine transformation

$$\tau(x_i; \boldsymbol{h}_i) = \alpha_i x_i + \beta_i \;\; \boldsymbol{h}_i = \{\alpha_i, \beta_i\} \tag{3}$$

and c is a feed forward neural network with parameters $\boldsymbol{\theta}$. The planar transformation uses a transformation of the form

$$\boldsymbol{x'} = \boldsymbol{x} + \boldsymbol{v}\sigma(\boldsymbol{w}^\top \boldsymbol{x} + b), \tag{4}$$

where the parameters are $\boldsymbol{\theta} = \{\boldsymbol{w}, \boldsymbol{v}, b\}$ and σ is the hyperbolic tangent function. A diagram of the affine coupling layer and the planar transformation can be seen in Appendix A. For an in-depth presentation of NFs see (Papamakarios et al., 2019) and (Kobyzev et al., 2019).

2.2. Conditional Normalizing Flows

In order to extend normalizing flows to conditional density estimation that condition the dependent variables $\boldsymbol{y}$ on the

[1]Technical University of Denmark, Kgs. Lyngby, Denmark, 2800. Correspondence to: Mathias Niemann Tygesen <s153583@student.dtu.dk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

independent variables $\boldsymbol{x}$, each of the transformations in the flow is conditioined on $\boldsymbol{x}$. As such the log conditional probability of an observation is:

$$\log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \log p_u(\boldsymbol{u}) - \sum_{i=1}^{K} \log \left|\det \boldsymbol{J}_{f_{\boldsymbol{x},\boldsymbol{\theta}}}(\boldsymbol{z}_{k-1})\right|.$$

The affine coupling layer can be extended to a Conditional Affine Coupling Layer as in (Lu & Huang, 2019). The independent variables $\boldsymbol{y}$ are first sent through a neural network cn for feature extraction. The extracted features are then sent through c along with half of the dimensions

$$\begin{aligned} \boldsymbol{x}_r &= cn(\boldsymbol{x}) \\ \boldsymbol{y'}_{<d} &= \boldsymbol{y}_{<d} \\ (\boldsymbol{h}_{d+1}, \ldots, \boldsymbol{h}_D) &= c(\boldsymbol{y}_{<d}, \boldsymbol{x}_r) \\ y'_i &= \tau(y_i; \boldsymbol{h}_i), \quad i > d, \end{aligned} \tag{5}$$

where $\boldsymbol{y}$ are the dependent variables before the transformation and $\boldsymbol{y'}$ are the dependent variables after the transformation.

The planar transformation can be extended to a conditional planar transformation as in the implementation in the probabilistic programming language *Pyro* (Bingham et al., 2018). The independent variables $\boldsymbol{x}$ are sent through a feature extraction network cn and then the features are sent through a neural network c in order to generate the parameters for the transformation $\boldsymbol{\theta} = \{\boldsymbol{w}, \boldsymbol{v}, b\}$. The full transformation is then

$$\begin{aligned} \boldsymbol{x}_r &= cn(\boldsymbol{x}) \\ (\boldsymbol{w}, \boldsymbol{v}, b) &= c(\boldsymbol{x}_r) \\ \boldsymbol{y'} &= \boldsymbol{y} + \boldsymbol{v}\sigma(\boldsymbol{w}^\top \boldsymbol{y} + b). \end{aligned} \tag{6}$$

A diagram of the conditional affine coupling layer and the conditional planar transformation can be seen in Appendix A

2.3. Noise Regularization

As conditional generative models are prone to overfitting, we regularize the normalizing flows using *noise regularization* as presented in (Rothfuss et al., 2019). The regularization works by pertubing the training data with a small bit of noise, often Gaussian, before each epoch of training. It can be shown that this is equivalent to adding a cost for the second derivative of the modelled distribution in the loss function. Let l be a loss function and $\boldsymbol{z}$ some observation. The expected loss with noise regularization is then

$$\mathbb{E}_{\boldsymbol{\xi} \sim K(\boldsymbol{\xi})} \left[l\left(\boldsymbol{z} + \boldsymbol{\xi}\right)\right] \approx l\left(\boldsymbol{z}\right) + \frac{h^2}{2} \operatorname{tr}\left(\mathbf{H}\right) \tag{7}$$

where $K(\boldsymbol{\xi})$ is a noise distribution and $\boldsymbol{H} = \nabla^2_{\boldsymbol{z}} l(\boldsymbol{z})$.

In maximum likelihood estimation of conditional density estimation $l(\boldsymbol{x}, \boldsymbol{y}) = -\log p(\boldsymbol{y}|\boldsymbol{x})$ and the noise regularized loss function becomes

$$\begin{aligned} l(\boldsymbol{x}, \boldsymbol{y}) \approx & -\log p\left(\boldsymbol{y}|\boldsymbol{x}\right) - \frac{h^2}{2} \sum_{j=1}^{d_y} \frac{\partial^2 \log p(y|x)}{\partial y^{(j)} \partial y^{(j)}} \\ & - \frac{h^2}{2} \sum_{j=1}^{d_x} \frac{\partial^2 \log p(y|x)}{\partial x^{(j)} \partial x^{(j)}} \end{aligned} \tag{8}$$

where d_x and d_y are the dimension of $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively. The second term is a regularization of the modelled distribution while the third term is a regularization of the conditioning of the density estimation.

3. Experiments

3.1. Three kinds of normalizing flows

In order to see how conditional affine coupling flows perform in a low dimensional settings, we construct a full flow along with two comparison flows. The conditional affine coupling flow consists of conditional affine coupling layers along with batchnorm layers and permutation layers. The batchnorm layers are used to better propagate the training signal as presented in (Dinh et al., 2016) and the permutation is just a simple switch of the dimensions. The first comparison flow is a conditional planar flow with batchnorms and the second comparison flow is a mixed flow with alternating conditional affine coupling layers and planar transformation layers starting with a conditional affine coupling layer. Before each layer we have a batchnorm layer and after each conditional planar transformation layer we have a permutation layer[1]. The layout of all of the flows can be seen in Appendix A.

3.2. Synthetic data

In order to visually compare the expressiveness of the different flows we first train them on a synthetic data set. The data set is the two moons data set as used in (Atanov et al., 2019) but it has been extended to rotate around the center with uniform probability. We train three flows of depth 4, 24 and 48. Each flow is conditioned on a rotation of 1.5 radians. In the first row of Figure 1 we can see that the conditional affine coupling flow becomes better at defining the edges of the distribution the deeper the flow is. However, for all of the depths the conditional affine coupling flow has clearly visible paths of non-zero probability density between the two half moons. We will call such paths "probability paths". In the second row we can see that the conditional planar flow has problems with defining the edges of the distribution for shallower flows but get better with depth. For all of the conditional planar flows it is clearly visible that the

[1] Implementation can be found at https://github.com/MathiasNT/Normalizing_Flows_on_Geospatial_Data.

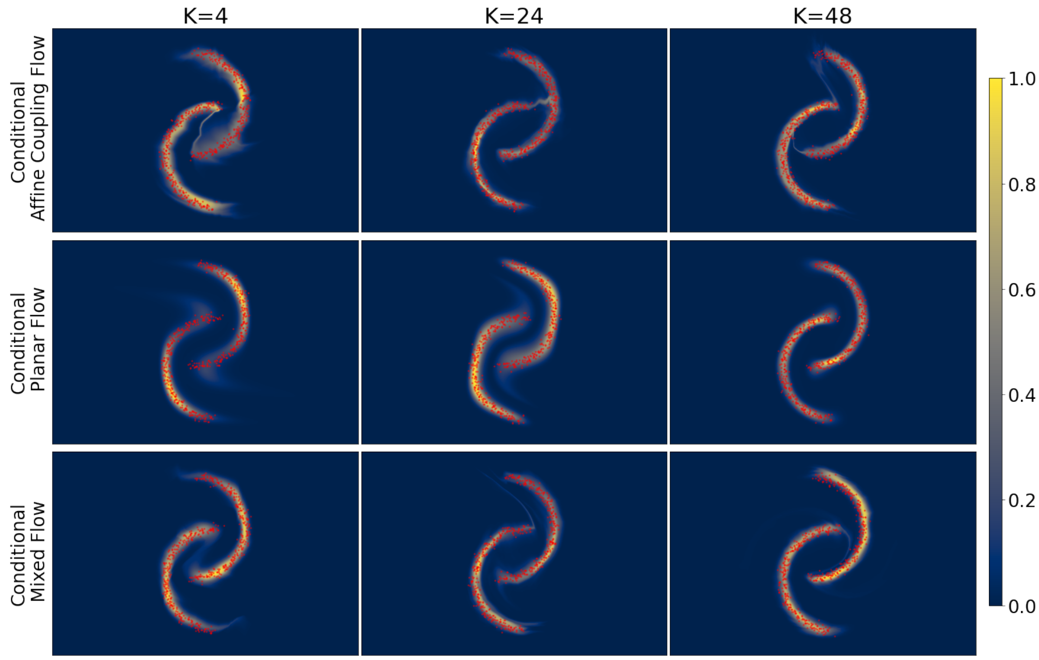

Figure 1. Here we see three different kinds of conditional normalizing flows fitted to the rotating two moons data with the a depth K of 4, 24 and 48. The first row are conditional affine coupling flows, the second row is conditional planar flows and the third row are mixed flows with alternating affine coupling layers and planar transformation layers. All of the flows are conditoned on a rotation of 1.5 radians. The red dots are samples from the corresponding two moons.

planar transformation has no problem with modelling the disjoint moons. In the third row, we can see how a mixed flow with both conditional affine coupling layers and planar transformations have the ability to define clear edges due to the conditional affine coupling layers and is better than the conditional affine coupling flow at creating disjoint distributions due to the planar transformations.

3.3. Geospatial transport data

One real world geospatial setting where conditional normalizing flows could be used is modelling the distribution of the demand for transport. One example of this, prevalent in literature, is the problem of modelling taxi dropoff locations on Manhattan. The data set is a processed NYC Yellow Taxi data set originally presented in (Dutordoir et al., 2018) and also used in (Rothfuss et al., 2019). The data set has the longitude and latitude of the taxi drop off location as dependent variables while the coordinates of the taxi pickup, hour of the day and day of the week are independent variables. We trained flows of each kind with a depth of 48 to the data. All of the flows were conditioned on a pickup at 18:00 on a Tuesday on Manhattan. In Figure 2 we have visualized the conditional affine coupling flows modelled distribution in the area between Manhattan and JFK airport. In the figure there are clear probability paths between Manhattan and JFK airport. In Figure 3 we have visualized the same area for the conditional planar flow. For the conditional planar flow there is no visible probability path and the area around JFK is modelled as its own disjoint probability. However the distribution on Manhattan extends out over the East River. In Figure 4 the same area is visualized for the conditional mixed flow. Again the area around JFK is modelled as a disjoint probability, this time extending out to cover the airport train stations. Furthermore the distribution around the Brooklyn area has complex enough shape to allow it to avoid the East River. As such we can see that mixing together both affine coupling flows and planar flow in the same flow gives both the ability to model disjoint probabilities and complex distributions at lower flow depth than a single type of layer can do on its own.

4. Discussion

As shown by the experiments, the conditional affine coupling flow is capable of modelling complex distributions but have problems with creating disjoint areas of distribution. The affine transformation of a dimension is done given the same dimension, albeit with parameters conditioned upon

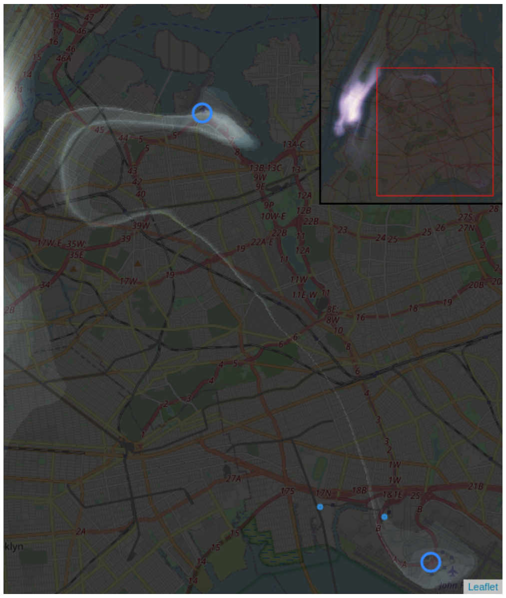

Figure 2. A conditional affine coupling flow conditioned on 18:00 Tuesday with pickup on Manhattan. The blue circles show the JFK and LaGuardia Airport and the blue dots show the JFK airport train stations. Upper right corner shows overview of NYC.

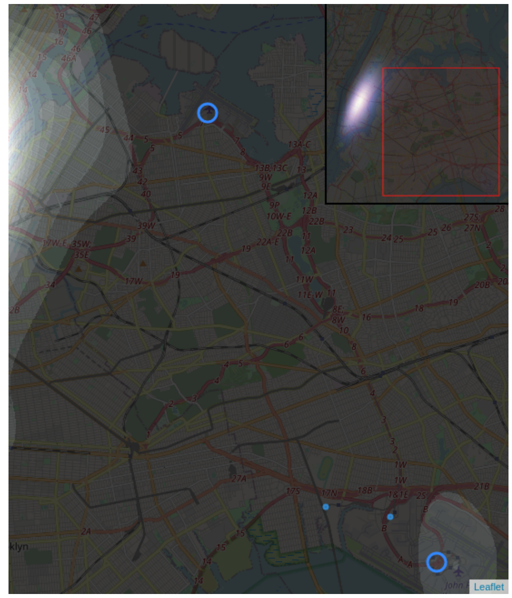

Figure 3. A conditional planar flow conditioned on 18:00 Tuesday with pickup on Manhattan. The blue circles show the JFK and LaGuardia Airport and the blue dots show the JFK airport train stations. Upper right corner shows overview of NYC

the other dimension. As such the transformation of a dimension can only be a scale and translation, corresponding to spreading and pushing the distribution around. Therefore the affine coupling layer is unable to create disjoint probabilities, however for deeper flows these probability paths can be stretched thin. As deeper NFs need more regularization to avoid overfitting as they become more expressive, perhaps shallower mixed NFs with the right kind of expressiveness are easier to regularize. While this problem of probability paths is clearly visible in low dimensions it could theoretically persist in higher dimensions.

The planar transformation can be seen as expanding or contracting the distribution perpendicular to a hyperplane. As such it has no problem creating disjoint areas of distribution but as the conditioning between the dimensions is much simpler this leads to distributions that have a much simpler shape. This can be remedied by resorting to deeper flows.

However as we have demonstrated by combining the two types of layers we can have both. The mixed flow showed the ability to create disjoint areas of probability while maintaining flexible enough shape to model complex distributions.

This raises multiple questions: Are the probability paths a problem in higher dimensions in practice? If they are, how do they impact the modelling of e.g. images? Could more complex transformers in coupling layers solve this problem? What is the best way to combine different kinds of layers together in a single flow model?

Figure 4. A conditional mixed flow conditioned on 18:00 Tuesday with pickup on Manhattan. The blue circles show the JFK and LaGuardia Airport and the blue dots show the JFK airport train stations. Upper right corner shows overview of NYC

References

Atanov, A., Volokhova, A., Ashukha, A., Sosnovik, I., and Vetrov, D. Semi-Conditional Normalizing Flows for Semi-Supervised Learning. may 2019. URL http://arxiv.org/abs/1905.00505.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 2018.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using Real NVP. may 2016. URL http://arxiv.org/abs/1605.08803.

Dutordoir, V., Salimbeni, H., Deisenroth, M., and Hensman, J. Gaussian Process Conditional Density Estimation. oct 2018. URL http://arxiv.org/abs/1810.12750.

Kobyzev, I., Prince, S., and Brubaker, M. A. Normalizing Flows: An Introduction and Review of Current Methods. aug 2019. URL http://arxiv.org/abs/1908.09257.

Lu, Y. and Huang, B. Structured Output Learning with Conditional Generative Flows. may 2019. URL http://arxiv.org/abs/1905.13288.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing Flows for Probabilistic Modeling and Inference. dec 2019. URL https://arxiv.org/abs/1912.02762.

Rezende, D. J. and Mohamed, S. Variational Inference with Normalizing Flows. may 2015. URL http://arxiv.org/abs/1505.05770.

Rothfuss, J., Ferreira, F., Boehm, S., Walther, S., Ulrich, M., Asfour, T., and Krause, A. Noise Regularization for Conditional Density Estimation. jul 2019. URL http://arxiv.org/abs/1907.08982.

A. Diagrams of different transforms

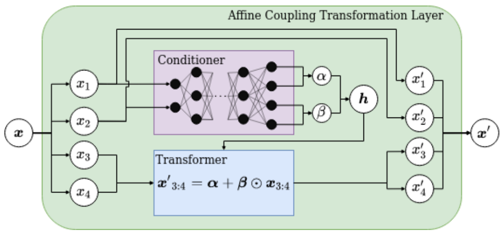

Figure 5. The Affine Coupling Layer visualized for four dimensions.

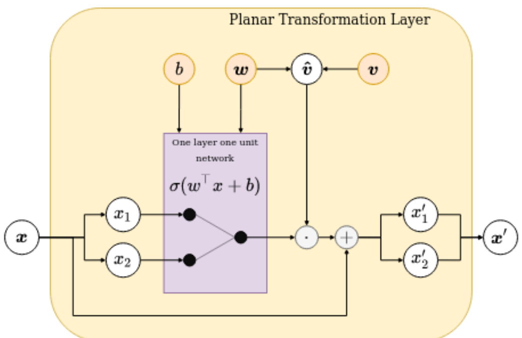

Figure 6. The Planar Transformation visualized for two dimensions.

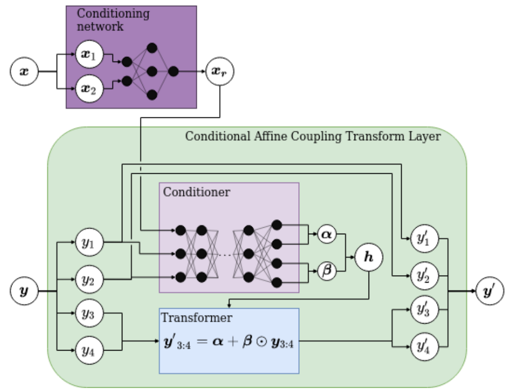

Figure 7. The Conditional Affine Coupling Layer visualized for a flow with four dimensional dependent variables and two dimensional independent variables.

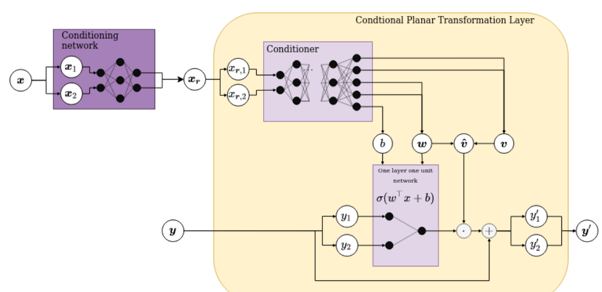

Figure 8. The Conditional Planar Transformation visualized for a flow with two dimensional dependent variables and two dimensional independent variables.

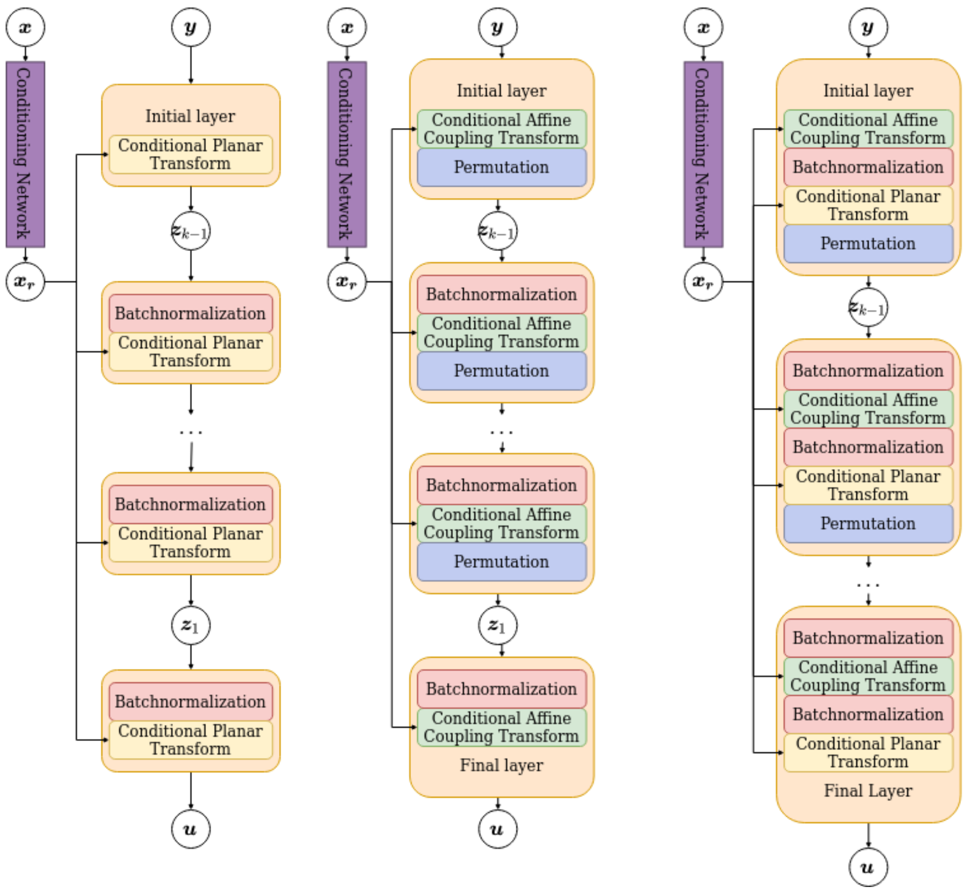

Figure 9. To the left we see a conditional planar flow with batchnorm layers. In the middle we see a conditional affine coupling flow with batchnorm layers and permutation layers. To the right we a mixed flow with both conditional affine coupling layers and planar transformation layers along with batchnorm and permuation layers. Note that all of the layers have no initial batchnorm and the layers with permutation layers have the last permutation removed.