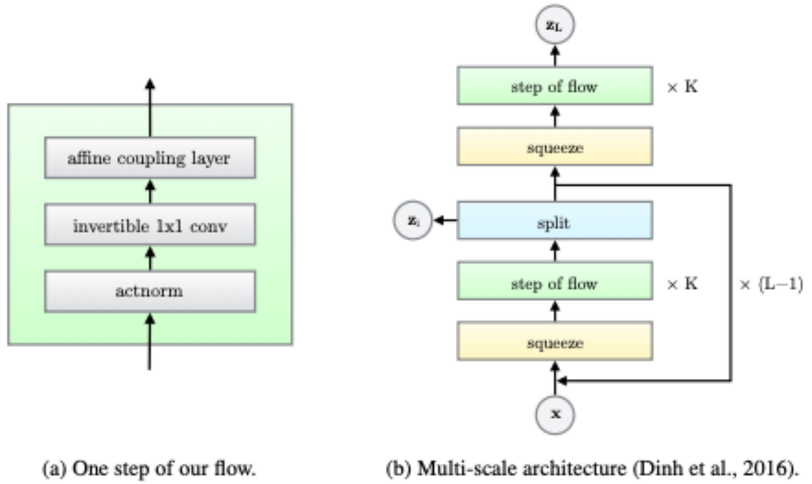
# Normalizing Flow history

$$
\begin{aligned}
\text{Conv} &= \text{Input} \to \text{Nonlinearity} \\
&\to \text{Conv}_{3\times3} \to \text{Nonlinearity} \to \text{Gate} \\
\text{Attn} &= \text{Input} \to \text{Conv}_{1\times1} \\
&\to \text{MultiHeadSelfAttention} \to \text{Gate}
\end{aligned}
$$

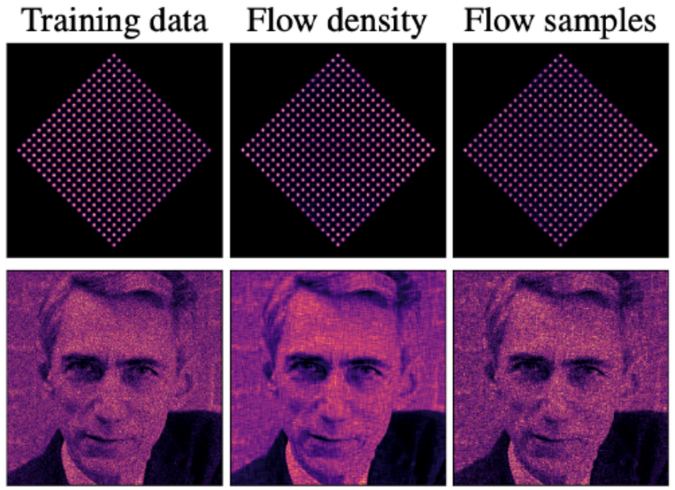
where Gate refers to a $1 \times 1$ convolution that doubles the number of channels, followed by a gated linear unit (Dauphin et al., 2016). The convolutional layer is identical to the one used by PixelCNN++ (Salimans et al., 2017), and the multi-head self attention mechanism we use is identical to the one in the Transformer (Vaswani et al., 2017). (We always use 4 heads in our experiments, since we found it to be effective early on in our experimentation process.)




(a) One step of our flow. (b) Multi-scale architecture (Dinh et al., 2016).

**2018**
Kingma, Durk P., and Prafulla Dhariwal.
"Glow: Generative flow with invertible 1x1 convolutions."

**2019**
Ho, Jonathan, et al.
“Flow++: Improving flow-based generative models with variational dequantization and architecture design."




**2019**
Durkan, Conor, et al. "Neural spline flows."

# Why do we care about generative models?



“What I can not create I do not understand”
-Richard Feynman

“…generative capabilities are the key to intelligence…”
-Max Welling



https://youtu.be/mmDw5glry9w?t=2985