Deep Generative Learning via Variational Gradient Flow
Supplementary Material

We here include proofs, hyperparameter settings and network architectures.

1. Proofs

In this section we give detailed proofs for the main theory in the paper.

Lemma 1.1. Let $\frac{\delta \mathcal{F}}{\delta q}(q) : \mathbb{R}^d \to \mathbb{R}$ denote the first variation of $\mathcal{F}[\cdot]$ at q, then $\left(\frac{\delta \mathcal{F}}{\delta q}(q)\right)(\mathbf{x}) = f'(r(\mathbf{x}))$ where $r(\mathbf{x}) = \frac{q(\mathbf{x})}{p(\mathbf{x})}$.

Proof. For any $w(\mathbf{x})$, define the function $\eta(s) = \mathcal{F}[q + sw] : \mathbb{R} \to \mathbb{R}$. The chain rule and direct calculation shows $\eta'(s)\big|_{s=0} = \langle \frac{\delta \mathcal{F}}{\delta q}(q), w \rangle = \int f'(r(\mathbf{x}))w(\mathbf{x})d\mathbf{x}$. □

Lemma 1.2.

$$\frac{d}{dt}\mathcal{F}[q_t] = -\mathbb{E}_{\mathbf{X}\sim q_t}[\|\mathbf{v}_t(\mathbf{X})\|^2]$$

Proof. Follow the expression 10.1.16 in (Ambrosio et al., 2008) (section E of chapter 10.1.2, page 233). □

Theorem 1.1. For any $\mathbf{g} \in \mathcal{H}(q_t)$, if the vanishing condition $\lim_{\|\mathbf{x}\|\to\infty} \|f'(r_t(\mathbf{x}))q_t(\mathbf{x})\mathbf{g}(\mathbf{x})\| = 0$ is satisfied, then

$$\left\langle \frac{\delta \mathcal{L}}{\delta \mathbf{h}}[\mathbf{0}], \mathbf{g} \right\rangle_{\mathcal{H}(q_t)} = \langle f''(r_t)\nabla r_t, \mathbf{g} \rangle_{\mathcal{H}(q_t)}.$$

Proof. For any $\mathbf{g} \in \mathcal{H}(q_t)$, define $\eta(s) = \mathbb{D}_f(\mathbb{T}_{s,\mathbf{g}\#}q_t\|p)$ as a function of $s \in \mathbb{R}^+$. Let $\theta_{\mathbf{g}}(s) = \mathbb{T}_{s,\mathbf{g}\#}q_t/p$. By definition, $\mathcal{L}[\mathbf{g}] = \eta(s) = D_f(\mathbb{T}_{s,\mathbf{g}\#}q_t\|p) = \int p(\mathbf{x})f(\theta_{\mathbf{g}}(s))\mathrm{d}\mathbf{x}$. Since $\eta'(s)\big|_{s=0} = \langle \frac{\delta \mathcal{L}}{\delta \mathbf{h}}[\mathbf{0}], \mathbf{g} \rangle_{\mathcal{H}(q_t)}$, we need calculate the derivative of $\eta(s)$ at $s = 0$. Since

$$(\mathbb{T}_{s,\mathbf{g}\#}q_t)(\mathbf{x}) = q_t(\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))|\det(\nabla_{\mathbf{x}}\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))|,$$

by the chain rule, we get

$$\eta'(s)\big|_{s=0} = \int p(\mathbf{x})[f'(\theta_{\mathbf{g}}(s))\theta'_{\mathbf{g}}(s)]\big|_{s=0}\mathrm{d}\mathbf{x},$$

where

$$\theta'_{\mathbf{g}}(s)\big|_{s=0} = \frac{1}{p(\mathbf{x})}\Big\{[q_t(\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))]'\big|_{s=0}|\det(\nabla_{\mathbf{x}}\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))|\big|_{s=0} + q_t(\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))\big|_{s=0}[|\det(\nabla_{\mathbf{x}}\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))|]'\big|_{s=0}\Big\}.$$

By definition, $\theta_{\mathbf{g}}(s)\big|_{s=0} = \frac{q_t(\mathbf{x})}{p(\mathbf{x})} = r_t(\mathbf{x})$. We claim that

$$\begin{aligned}\theta'_{\mathbf{g}}(s)\big|_{s=0} &= \frac{1}{p(\mathbf{x})}\{-\mathbf{g}(\mathbf{x})^T\nabla q_t(\mathbf{x}) - q_t(\mathbf{x})\nabla\cdot\mathbf{g}(\mathbf{x})\}\\ &= -\frac{1}{p(\mathbf{x})}\nabla\cdot[q_t(\mathbf{x})\mathbf{g}(\mathbf{x})].\end{aligned}$$

Indeed, recall that

$$\mathbb{T}_{s,\mathbf{g}}(\mathbf{X}) = \mathbf{X} + s\mathbf{g}(\mathbf{X}).$$

We get

$$\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{X}) = \mathbf{X} - s\mathbf{g}(\mathbf{X}) + o(s),$$

and

$$\mathbb{T}^{-1}_{s,\mathbf{g}}\big|_{s=0}(\mathbf{X}) = \mathbf{X}.$$

Then it follows that

$$\begin{aligned}[q_t(\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))]'\big|_{s=0} &= \lim_{s\to 0}\frac{q_t(\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x})) - q_t(\mathbf{x})}{s}\\ &= -\mathbf{g}(\mathbf{x})^T\nabla q_t(\mathbf{x}),\end{aligned}$$

and

$$|\det(\nabla_{\mathbf{x}}\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))|\big|_{s=0} = 1, q_t(\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))\big|_{s=0} = q_t(\mathbf{x}).$$

We finish our claim by calculating

$$\begin{aligned}&[|\det(\nabla_{\mathbf{x}}\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))|]'\big|_{s=0}\\ =&[\exp^{\log(|\det(\nabla_{\mathbf{x}}\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))|)}]'\big|_{s=0}\\ =&|\det(\nabla_{\mathbf{x}}\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))|\big|_{s=0}[\log|\det(\nabla_{\mathbf{x}}\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))|]'\big|\big|_{s=0}\\ =&\lim_{s\to 0}\frac{\log|\det(\nabla_{\mathbf{x}}\mathbb{T}^{-1}_{s,\mathbf{g}}(\mathbf{x}))| - \log|\det(\mathbf{I})|}{s}\\ =&\lim_{s\to 0}\frac{\log|\det(\mathbf{I} - s\nabla_{\mathbf{x}}\mathbf{g}(\mathbf{x}))| - \log|\det(\mathbf{I})| + o(s)}{s}\\ =&-\operatorname{tr}(\nabla_{\mathbf{x}}\mathbf{g}(\mathbf{x})) = -\nabla\cdot\mathbf{g}(\mathbf{x}).\end{aligned}$$

Thus

$$\begin{aligned}\eta'_{\mathbf{g}}(s)\big|_{s=0} &= \int p(\mathbf{x})\cdot[f'(\theta_{\mathbf{g}}(s))\cdot\theta'_{\mathbf{g}}(s)]\big|_{s=0}\mathrm{d}\mathbf{x}\\ &= -\int f'(r_t(\mathbf{x}))\nabla\cdot[q_t(\mathbf{x})\mathbf{g}(\mathbf{x})]\mathrm{d}\mathbf{x}\\ &= \int q_t(\mathbf{x})\mathbf{g}(\mathbf{x})^T\nabla f'(r_t(\mathbf{x})) - \nabla\cdot[f'(r_t(\mathbf{x}))\mathbf{g}(\mathbf{x})]\mathrm{d}\mathbf{x}\\ &= \int q_t(\mathbf{x})f''(r_t(\mathbf{x}))[\nabla r_t(\mathbf{x})]^T\mathbf{g}(\mathbf{x})\mathrm{d}\mathbf{x}\\ &= \langle f''(r_t(\mathbf{x}))\nabla r_t(\mathbf{x}), \mathbf{g}(\mathbf{x})\rangle_{\mathcal{H}(q_t)},\end{aligned}$$

where the fourth equality follows from integral by part and the vanishing assumption. □

Theorem 1.2. The evolving distribution q_t under the infinitesimal pushforward map $\mathbb{T}_{s,\mathbf{v}_t}$ satisfies the Vlasov-Fokker-Planck equation.

Proof. Similar to the proof of equation (13) in (Liu, 2017). We present the detail here for completeness. The proof of Theorem 1.1 shows that,

$$q_t(\mathbb{T}^{-1}_{s,\mathbf{v}_t}(\mathbf{x})) = q_t(\mathbf{x}) - s\mathbf{v}_t(\mathbf{x})^T\nabla q_t(\mathbf{x}) + o(s),$$

and

$$|\det(\nabla_{\mathbf{x}}\mathbb{T}^{-1}_{s,\mathbf{v}_t}(\mathbf{x}))| = -s\nabla\cdot\mathbf{v}_t(\mathbf{x}) + o(s).$$

Then by the Taylor expansion,

$$\begin{aligned}
&\log(\mathbb{T}_{s,\mathbf{v}_t\#}q_t)(\mathbf{x})\\
=&\log q_t(\mathbb{T}^{-1}_{s,\mathbf{v}_t}(\mathbf{x})) + \log|\det(\nabla_{\mathbf{x}}\mathbb{T}^{-1}_{s,\mathbf{v}_t}(\mathbf{x}))|\\
=&\log q_t(\mathbf{x}) - s\frac{\mathbf{v}_t(\mathbf{x})^T\nabla q_t(\mathbf{x})}{q_t(\mathbf{x})} - s\nabla\cdot\mathbf{v}_t(\mathbf{x}) + o(s)\\
=&\log q_t(\mathbf{x}) - \frac{s}{q_t(\mathbf{x})}(\mathbf{v}_t(\mathbf{x})^T\nabla q_t(\mathbf{x})\\
&+ q_t(\mathbf{x})\nabla\cdot\mathbf{v}_t(\mathbf{x})) + o(s).
\end{aligned}$$

Let $\tilde{q}(\mathbf{x})$ denote the density of $\mathbb{T}_{s,\mathbf{v}_t\#}q_t$, then

$$\begin{aligned}
\frac{\tilde{q}(\mathbf{x}) - q_t(\mathbf{x})}{s} &= \frac{q_t(\log\tilde{q} - \log q_t)}{s}\\
&= -\nabla\cdot(q_t(\mathbf{x})\mathbf{v}_t(\mathbf{x})) + o(s).
\end{aligned}$$

Let $s\to 0$, we get the desired result. □

Lemma 1.3. Let $(\mathbf{X}, Y)$ be a random variable pair admitting $p(\mathbf{x}, y)$ with the binary random variable $Y\sim p(y)$ taking the value in $\{-1,+1\}$. Denote $q(\mathbf{x}) = p(\mathbf{x}|Y=-1)$, $p(\mathbf{x}) = p(\mathbf{x}|Y=1)$ and $r(\mathbf{x}) = \frac{q(\mathbf{x})}{p(\mathbf{x})}$. Let $d^*(\mathbf{x}) = \arg\min_{d(\mathbf{x})}\mathbb{E}_{(\mathbf{X},Y)\sim p(\mathbf{x},y)}\log(1+\exp(-d(\mathbf{X})Y))$. If $p(Y=1) = p(Y=-1)$, then $r(\mathbf{x}) = \exp(-d^*(\mathbf{x}))$.

Proof. $d^*(\mathbf{x})$ is the minimizer of

$$\begin{aligned}
&\min_{d(\mathbf{x})}\mathbb{E}_{(\mathbf{X},Y)\sim p(\mathbf{x},y)}\log(1+\exp(-d(\mathbf{X})Y))\\
=&\min_{d(\mathbf{x})}\int p(\mathbf{x},y)\log(1+\exp(-d(\mathbf{x})y))\mathrm{d}\mathbf{x}\mathrm{d}y\\
=&\min_{d(\mathbf{x})}\{\int p(y=1)p(\mathbf{x}|y=1)\log(1+\exp(-d(\mathbf{x})))\mathrm{d}\mathbf{x}\\
&+\int p(y=-1)p(\mathbf{x}|y=-1)\log(1+\exp(d(\mathbf{x})))\mathrm{d}\mathbf{x}\}.
\end{aligned}$$

The above criterion is a functional of $d(\cdot)$. By setting the first variation to zero yields

$$\exp(-d^*(\mathbf{x})) = \frac{p(y=1)p(\mathbf{x}|y=1)}{p(y=-1)p(\mathbf{x}|y=-1)},$$

i.e. $r(\mathbf{x}) = \exp(-d^*(\mathbf{x}))$. □

Theorem 1.3. At the population level, the logD-trick GAN (Goodfellow et al., 2014) minimizes the "logD" divergence $\mathbb{D}_f(q(\mathbf{x})\|p(\mathbf{x}))$, with $f(u) = (u+1)\log(u+1) - 2\log 2$, where $q(\mathbf{x})$ is the distribution of generated data.

Proof. By definition,

$$\begin{aligned}
&\mathbb{D}_f(q(\mathbf{x})\|p(\mathbf{x}))\\
=&\int p(\mathbf{x})f\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right)d\mathbf{x}\\
=&\int (p(\mathbf{x})+q(\mathbf{x}))\log\left(\frac{p(\mathbf{x})+q(\mathbf{x})}{p(\mathbf{x})}\right)d\mathbf{x} - 2\log 2\\
=&2\mathrm{KL}(\tfrac{p(\mathbf{x})+q(\mathbf{x})}{2}\|p(\mathbf{x}))
\end{aligned}$$

At the population level, the objective function of the logD-trick GAN (Goodfellow et al., 2014) is

$$\begin{aligned}
&\max_D \mathbb{E}_{\mathbf{X}\sim p(\mathbf{x})}[\log D(\mathbf{X})] + \mathbb{E}_{\mathbf{Z}\sim p_{\mathbf{Z}}}[\log(1-D(G(\mathbf{Z})))],\\
&\min_G -\mathbb{E}_{\mathbf{X}\sim p(\mathbf{x})}[\log D(\mathbf{X})] - \mathbb{E}_{\mathbf{Z}\sim p_{\mathbf{Z}}}[\log D(G(\mathbf{Z}))],
\end{aligned}$$

where $p_{\mathbf{Z}}$ is the simple low-dimensional reference distribution. Denote $q(\cdot)$ as the distribution of $G(\mathbf{Z})$. Then the losses of D and G are equivalent to

$$\begin{aligned}
&\max_D \mathbb{E}_{\mathbf{X}\sim p(\mathbf{x})}[\log D(\mathbf{X})] + \mathbb{E}_{\mathbf{X}\sim q(\mathbf{x})}[\log(1-D(\mathbf{X}))],\\
&\min_G -\mathbb{E}_{\mathbf{X}\sim p(\mathbf{x})}[\log D(\mathbf{X})] - \mathbb{E}_{\mathbf{X}\sim q(\mathbf{x})}[\log D(\mathbf{X})].
\end{aligned}$$

The optimal discriminator is $D^*(\mathbf{x}) = \frac{p(\mathbf{x})}{p(\mathbf{x})+q(\mathbf{x})}$. Substituting this D^* into the G criterion, we get

$$\begin{aligned}
&-\mathbb{E}_{\mathbf{X}\sim p(\mathbf{x})}[\log D^*(\mathbf{X})] - \mathbb{E}_{\mathbf{X}\sim q(\mathbf{x})}[\log D^*(\mathbf{X})]\\
=&\mathbb{E}_{\mathbf{X}\sim p(\mathbf{x})}\left[\log\tfrac{p(\mathbf{X})+q(\mathbf{X})}{p(\mathbf{X})}\right] + \mathbb{E}_{\mathbf{X}\sim q(\mathbf{x})}\left[\log\tfrac{p(\mathbf{X})+q(\mathbf{X})}{p(\mathbf{X})}\right]\\
=&\mathbb{D}_f(q(\mathbf{x})\|p(\mathbf{x})) + 2\log 2.
\end{aligned}$$

□

Proof of the relation between VGrow and SVGD

Proof. Let $f(u) = u \log u$. Let $\mathbf{g}$ in a Stein class associated with q_t. By the proof of Theorem 1.1, we know,

$$
\begin{aligned}
&\left\langle \frac{\delta \mathcal{L}}{\delta \mathbf{h}}[\mathbf{0}], \mathbf{g} \right\rangle_{\mathcal{H}(q_t)} \\
=&\langle f''(r_t)\nabla r_t, \mathbf{g}\rangle_{\mathcal{H}(q_t)} \\
=&\int \mathbf{g}(\mathbf{x})^T \frac{\nabla r_t(\mathbf{x})}{r_t(\mathbf{x})} q_t(\mathbf{x}) \mathrm{d}\mathbf{x} \\
=&\int \mathbf{g}(\mathbf{x})^T \nabla \log r_t(\mathbf{x}) q_t(\mathbf{x}) \mathrm{d}\mathbf{x} \\
=&\mathbb{E}_{\mathbf{X}\sim q_t(\mathbf{x})}[\mathbf{g}(\mathbf{x})^T \nabla \log q_t(\mathbf{X}) - \mathbf{g}(\mathbf{x})^T \nabla \log p(\mathbf{X})] \\
=&\mathbb{E}_{\mathbf{X}\sim q_t(\mathbf{x})}[\mathbf{g}(\mathbf{x})^T \nabla \log q_t(\mathbf{X}) + \nabla \cdot \mathbf{g}(\mathbf{x})] \\
&- \mathbb{E}_{\mathbf{X}\sim q_t(\mathbf{x})}[\mathbf{g}(\mathbf{x})^T \nabla \log p(\mathbf{X}) + \nabla \cdot \mathbf{g}(\mathbf{x})] \\
=&\mathbb{E}_{\mathbf{X}\sim q_t(\mathbf{x})}[\mathcal{T}_{q_t}\mathbf{g}] - \mathbb{E}_{\mathbf{X}\sim q_t(\mathbf{x})}[\mathcal{T}_p \mathbf{g}] \\
=&-\mathbb{E}_{\mathbf{X}\sim q_t(\mathbf{x})}[\mathcal{T}_p \mathbf{g}],
\end{aligned}
$$

where the last equality is obtained by restricting $\mathbf{g}$ in a Stein class associated with q_t, i.e., $\mathbb{E}_{\mathbf{X}\sim q_t(\mathbf{x})}\mathcal{T}_{q_t}\mathbf{g} = 0$. □

2. Hyperparameter Settings

For the real data, we set the batch size to be 64 and use RMSProp as the SGD optimizer to train neural networks. The learning rate is 0.0001 for both the deep sampler and the deep classifier except for 0.0002 on MNIST for VGrow-JF. Inputs to samplers are vectors generated from a 128 dimensional standard normal distribution on all the datasets. Meta-parameters of VGrow are listed in Table 1 where IL denotes the number of inner loops in each outer loop.

Table 1. Meta-parameter values in VGrow

Parameter	Value
s	0.5
ℓ	128
N	1280
IL	20

3. Network Architectures

Deep samplers and classifiers are parameterized with residual networks. Each ResNet block has a skip-connection. The skip-connection takes upsampling / downsampling of its input if necessary or 1×1 convolution if not. The upsampling is nearest-neighbor upsampling and the downsampling is achieved with mean pooling. Details concerning the networks are listed in Table 2, 3, 4, 5. We use c to denote the number of image channels, i.e. $c = 1$ or $c = 3$.

Table 2. ResNet sampler with $32 \times 32 \times c$ resolution.

Layer	Details	Output size
Latent noise	$\boldsymbol{z} \sim \mathcal{N}(0, I)$	128
Fully connected	Linear	2048
	Reshape	$4 \times 4 \times 128$
ResNet block	ReLU	$4 \times 4 \times 128$
	Upsampling	$8 \times 8 \times 128$
	Conv3 $\times$ 3, BN, ReLU	$8 \times 8 \times 128$
	Conv3 $\times$ 3	$8 \times 8 \times 128$
ResNet block	ReLU	$8 \times 8 \times 128$
	Upsampling	$16 \times 16 \times 128$
	Conv3 $\times$ 3, BN, ReLU	$16 \times 16 \times 128$
	Conv3 $\times$ 3	$16 \times 16 \times 128$
ResNet block	ReLU	$16 \times 16 \times 128$
	Upsampling	$32 \times 32 \times 128$
	Conv3 $\times$ 3, BN, ReLU	$32 \times 32 \times 128$
	Conv3 $\times$ 3	$32 \times 32 \times 128$
Conv	ReLU, Conv3 $\times$ 3, Tanh	$32 \times 32 \times c$

Table 3. ResNet classifier with $32 \times 32 \times c$ resolution.

Layer	Details	Output size
ResNet block	Conv3 $\times$ 3	$32 \times 32 \times 128$
	ReLU, Conv3 $\times$ 3	$32 \times 32 \times 128$
	Downsampling	$16 \times 16 \times 128$
ResNet block	ReLU, Conv3 $\times$ 3	$16 \times 16 \times 128$
	ReLU, Conv3 $\times$ 3	$16 \times 16 \times 128$
	Downsampling	$8 \times 8 \times 128$
ResNet block	ReLU, Conv3 $\times$ 3	$8 \times 8 \times 128$
	ReLU, Conv3 $\times$ 3	$8 \times 8 \times 128$
ResNet block	ReLU, Conv3 $\times$ 3	$8 \times 8 \times 128$
	ReLU, Conv3 $\times$ 3	$8 \times 8 \times 128$
Fully connected	ReLU, GlobalSum pooling	128
	Linear	1

References

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NIPS*, 2014.

Liu, Q. Stein variational gradient descent as gradient flow. In *NIPS*, 2017.

Table 4. ResNet sampler with $64 \times 64 \times c$ resolution.

Layer	Details	Output size
Latent noise	$\boldsymbol{z} \sim \mathcal{N}(0, I)$	128
Fully connected	Linear	2048
	Reshape	$4 \times 4 \times 128$
ResNet block	ReLU	$4 \times 4 \times 128$
	Upsampling	$8 \times 8 \times 128$
	Conv3 $\times$ 3, BN, ReLU	$8 \times 8 \times 128$
	Conv3 $\times$ 3	$8 \times 8 \times 128$
ResNet block	ReLU	$8 \times 8 \times 128$
	Upsampling	$16 \times 16 \times 128$
	Conv3 $\times$ 3, BN, ReLU	$16 \times 16 \times 128$
	Conv3 $\times$ 3	$16 \times 16 \times 128$
ResNet block	ReLU	$16 \times 16 \times 128$
	Upsampling	$32 \times 32 \times 128$
	Conv3 $\times$ 3, BN, ReLU	$32 \times 32 \times 128$
	Conv3 $\times$ 3	$32 \times 32 \times 128$
ResNet block	ReLU	$32 \times 32 \times 128$
	Upsampling	$32 \times 32 \times 128$
	Conv3 $\times$ 3, BN, ReLU	$32 \times 32 \times 128$
	Conv3 $\times$ 3	$64 \times 64 \times 128$
Conv	ReLU, Conv3 $\times$ 3, Tanh	$64 \times 64 \times c$

Table 5. ResNet classifier with $64 \times 64 \times c$ resolution.

Layer	Details	Output size
ResNet block	Conv3 $\times$ 3	$64 \times 64 \times 128$
	ReLU, Conv3 $\times$ 3	$64 \times 64 \times 128$
	Downsampling	$32 \times 32 \times 128$
ResNet block	ReLU, Conv3 $\times$ 3	$32 \times 32 \times 128$
	ReLU, Conv3 $\times$ 3	$32 \times 32 \times 128$
	Downsampling	$16 \times 16 \times 128$
ResNet block	ReLU, Conv3 $\times$ 3	$16 \times 16 \times 128$
	ReLU, Conv3 $\times$ 3	$16 \times 16 \times 128$
	Downsampling	$8 \times 8 \times 128$
ResNet block	ReLU, Conv3 $\times$ 3	$8 \times 8 \times 128$
	ReLU, Conv3 $\times$ 3	$8 \times 8 \times 128$
Fully connected	ReLU, GlobalSum pooling	128
	Linear	1